UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

DATA RESCUE: DEFINING A COMPREHENSIVE WORKFLOW THAT INCLUDES THE ROLES AND RESPONSIBILITIES OF THE RESEARCH LIBRARY

by

LOUISE HILDA PATTERTON

Submitted in fulfilment of the requirements for the degree of

DOCTOR PHILOSOPHIAE (RESEARCH)

in the

DEPARTMENT OF INFORMATION SCIENCE,
FACULTY OF ENGINEERING,
THE BUILT ENVIRONMENT AND
INFORMATION TECHNOLOGY

at the

UNIVERSITY OF PRETORIA

Supervisors:
Prof. T.J.D. Bothma
Dr M.J. van Deventer

January 2023

DECLARATION REGARDING PLAGIARISM

Full names	Louise Hilda Patterton
Student number	04670133

Declaration

1. I understand what plagiarism is and am aware of the University’s policy in this regard.
2. I declare that this thesis is my own original work. Where other people’s work has been used (either from a printed source, internet or any other source), this has been properly acknowledged and referenced in accordance with the requirements as stated in the University's plagiarism prevention policy.
3. I have not used another student’s past written work to hand in as my own.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Signature LPatterton ________________________________

ACKNOWLEDGEMENTS

My most profound appreciation goes to my study supervisors, Prof. Theo Bothma and Dr Martie van Deventer, for their excellent guidance, scholarly knowledge, insightful comments, enthusiasm and encouragement to complete this thesis.

I also want to thank the external examiners for their thoughtful feedback and efforts toward improving this manuscript.

The completion of this endeavour would not have been possible without the funding provided by my employer, the Council for Scientific and Industrial Research (CSIR). The generous doctoral study leave offered by the CSIR is also highly appreciated.

I would also like to convey my appreciation to all the research participants. Without their valuable feedback and opinions the research would not have been accomplished.

As a recipient of postgraduate merit bursaries I wish to express my sincere thanks to the University of Pretoria for this gesture.

I would like to extend my utmost gratitude to David Kaplan for his outstanding work as proofreader of the thesis.

Many thanks to everyone who has been there for me emotionally and intellectually during the past four years.

Lastly, I would like to recognise the invaluable support provided throughout this journey by my constant and faithful companion Mia, also known as Shiny. You might only be a cat, but you were human enough to be a source of great comfort, solace and joy in times of frustration, despair, anxiety and writer’s block.

DATA RESCUE: DEFINING A COMPREHENSIVE WORKFLOW THAT INCLUDES THE ROLES AND RESPONSIBILITIES OF THE RESEARCH LIBRARY

ABSTRACT

This study, comprising a case study at a selected South African research institute, focused on the creation of a workflow model for data rescue indicating the roles and responsibilities of the research library. Additional outcomes of the study include a series of recommendations addressing the troublesome findings that revealed data at risk to be a prevalent reality at the selected institute, showing the presence of a multitude of factors putting data at risk, disclosing the profusion of data rescue obstacles faced by researchers, and uncovering that data rescue at the institute is rarely implemented.

The study consists of four main parts: (i) a literature review, (ii) content analysis of literature resulting in the creation of a data rescue workflow model, (iii) empirical data collection methods , and (iv) the adaptation and revision of the initial data rescue model to present a recommended version of the model.

A literature review was conducted and addressed data at risk and data rescue terminology, factors putting data at risk, the nature, diversity and prevalence of data rescue projects, and the rationale for data rescue.

The second part of the study entailed the application of content analysis to selected documented data rescue workflows, guidelines and models. Findings of the analysis led to the identification of crucial components of data rescue and brought about the creation of an initial Data Rescue Workflow Model. As a first draft of the model, it was crucial that the model be reviewed by institutional research experts during the next main stage of the study.

The section containing the study methodology culminates in the implementation of four different empirical data collection methods. Data collected via a web-based questionnaire distributed to a sample of research group leaders (RGLs), one-on-one virtual interviews with a sample of the aforementioned RGLs, feedback supplied by RGLs after reviewing the initial Data Rescue Workflow Model, and a focus group session held with institutional research library experts resulted in findings producing insight into the institute’s data at risk and the state of data rescue.

Feedback supplied by RGLs after examining the initial Data Rescue Workflow Model produced a list of concerns linked to the model and contained suggestions for changes to the model. RGL feedback was at times unrelated to the model or to data and necessitated the implementation of a mini focus group session involving institutional research library experts. The mini focus group session comprised discussions around requirements for a data rescue workflow model.

The consolidation of RGL feedback and feedback supplied by research library experts enabled the creation of a recommended Data Rescue Workflow Model, with the model also indicating the various roles and responsibilities of the research library.

The contribution of this research lies primarily in the increase in theoretical knowledge regarding data at risk and data rescue, and culminates in the presentation of a recommended Data Rescue Workflow Model. The model not only portrays crucial data rescue activities and outputs, but also indicates the roles and responsibilities of a sector that can enhance and influence the prevalence and execution of data rescue projects. In addition, participation in data rescue and an understanding of the activities and steps portrayed via the model can contribute towards an increase in the skills base of the library and information services sector and enhance collaboration projects with relevant research sectors. It is also anticipated that the study recommendations and exposure to the model may influence the viewing and handling of data by researchers and accompanying research procedures.

Keywords: Data rescue, data at risk, data conservation, data curation, digital curation, data management, research data management, research library roles and responsibilities, library and information services roles and responsibilities.

Contents

List of Tables

List of Figures

Table i: Clarification of key terms

TERM	CLARIFICATION
Closed repository	A repository where the content is only accessible to certain members
Data assessment	Evaluating data to determine whether the data are worthy of rescue
Data documentation	Documentation that ensures that the data will be understood and similarly interpreted by any user
Data management plan	A formal document that outlines how data are to be handled both during a research project, and after the project is completed
Data rescue	A collection of processes, including photography, scanning and converting, that stores historical and modern data in a usable format
Digital Object Identifier	A persistent identifier or handle used to identify objects uniquely. Also commonly referred to as a DOI
Digitising/digitisation	The process of converting information into a digital format
Early digital data	In this study, the term refers to data in older digital formats including data on floppies, stiffies, older audio and videotapes, CDs, DVDs, and older magnetic tapes
Full rescue	A data rescue process that involves all rescue stages included in the study's recommended Data Rescue Workflow Model
Handle	A globally unique identifier assigned to an object
Imaging	Replicating paper data into a digital format
Institutional repository	An archive for collecting, preserving and disseminating digital copies of the intellectual output of an institution
Inventory	A complete list of items in a collection
Keying of data	Entering data by means of a computer keyboard
Metadata	Metadata summarise basic information about data, making finding and working with particular instances of data easier
Metadata standard	A requirement which is intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and users
Open data	Data that anyone can access, use and share
Open repository	A repository providing free access to research data for users outside the institutional community
Partial rescue	A data rescue process involving only selected stages included in the study's recommended Data Rescue Workflow Model
Preservation	A combination of policies, strategies and actions to ensure longer-term access to reformatted and born digital content regardless of the challenges of media failure and technological change

TERM	CLARIFICATION
Preservation format	Recommended format used for storage in a data archive
Repository	An information system that ingests, stores, manages, preserves and provides access to digital content
Scanning of data	The process of converting non-digital materials to a digital format, mainly by use of a scanner or digital camera
SHEQ	An acronym for Safety, Health, Environment and Quality
Stream A	Data undergoing full rescue. Also see: ‘full rescue’
Stream B	Data undergoing partial rescue. Also see: ‘partial rescue’
Uncrawlables	Information that cannot be found via Google and other search engines

Table ii: Clarification of acronyms

ACRONYM	CLARIFICATION
ACRE	Atmospheric Circulation Reconstructions over the Earth
C3S	Copernicus Climate Change Service
CODATA	Committee on Data of the International Science Council
COVID-19	Coronavirus Disease of 2019
CSIR	Council for Scientific and Industrial Research, South Africa
CSIRIS	Council for Scientific and Industrial Research Information Services
CSV file	Comma-separated values file
DIRISA	Data Intensive Research Initiative of South Africa
DMP	Data Management Plan
DRPP	Data Rescue Project Plan
EPA	Environmental Protection Agency, USA
ICT	Information and Communications Technology
I-DARE	Data Rescue Portal
IEDRO	International Environmental Data Rescue Organization
IG	Interest Group
LIS	Library and Information Science (as opposed to Library and Information Services)
MB	A megabyte: a unit of digital data that is equal to about one million bytes
MLA	Music Library Association, USA
PDF	Portable Document Format
RDA	Research Data Alliance
RGL	Research Group Leader
SA	South Africa
SET	Science, Engineering and Technology

ACRONYM	CLARIFICATION
SHEQ	Safety, Health, Environment and Quality
TB	A terabyte: a unit of digital data that is equal to about one trillion bytes
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States (of America)
USA	United States of America
WMO	World Meteorological Organization
XLS	A Microsoft Excel 97-2003 Worksheet file that stores spreadsheet data

CHAPTER 1: STUDY BACKGROUND AND RESEARCH QUESTIONS

1.1 Introduction

This chapter introduces the study and provides insight into the research problems answered, and the contribution that the research makes to the field of Information Science. The introductory chapter also provides details on the study's methodology, its scope and limitations, and clarification on the chapter exposition.

1.2 Context of the research problem

Many research institutes, research groups and researchers invariably end up with data that can be termed 'at risk'. Data at risk are often defined as sets of scientific data which are not in modern electronic formats and the information of which is therefore not accessible to the research that needs it (CODATA, 2013). Other definitions of research data at risk describe it as the state of data when economic models and infrastructure are not in place to ensure access and preservation (Thompson, Davenport Robertson & Greenberg, 2014: 843), or when there is not a dedicated plan to ensure that the data not be at risk (Mayernik *et al.*, 2020: 1). The implications of this range of definitions are that data at risk often entail the default state of data unless active management measures are taken, and that all types of data from a range of disciplines can be affected by risk factors.

A range of risk factors can hinder, constrain, or limit the current or future use of data (Mayernik *et al.*, 2020: 1). These factors include catastrophic loss (World Meteorological Organization (WMO), 2016: 2), deterioration of the medium (WMO, 2014: 2), lack of use (Mayernik *et al.*, 2020: 5), loss of knowledge around context or access (Mayernik *et al.*, 2020: 5), cybersecurity breach (Mayernik *et al.*, 2020: 5), poor management (bwestra@uoregon.edu, 2017), direct attempts at reducing access (bwestra@uoregon.edu, 2017), and dispersion of data assets over time (WMO, 2014: 2).

It is crucial that data at risk be rescued, as such data can often not be regenerated (e.g., historic data linked to a specific period, location or event), and recollection of data can be costly and impractical. In addition, rescue of data at risk ideally should occur while scientists or others familiar with the data are still available to provide information about the data, its origin, collection and management. The benefits of data rescue, as a precursor to the enhancement of knowledge in a disciplinary field, are vast and numerous. Examples of such outcomes in a range of subject areas include:

- the rescue of 70-year-old, handwritten stream data by South African hydrologists leading to insight into the effect of alien trees on water distribution (Griffin, 2017: 267),

- the transcriptions of thousands of logs recorded during ship voyages in past centuries viewed as a 'bonanza' for studying weather patterns today (Griffin, 2017: 267),
- decoded medical records on abandoned 1950 punch cards aiding the understanding of the effect of varying levels of cholesterol on later disease (Griffin, 2017: 267),
- historic paper records found in a decaying Congo research station providing evidence of the response of vegetation to climate change (Grossman, 2017), and
- reanalysis of stratospheric data from 1816 enabling researchers to show a link between the eruption of Mount Tambora in 1815, and subsequent weather impacts (Brohan *et al.*, 2016).

The contributions of rescued data as illustrated in the examples above are typical of the benefits and outcomes of data rescue projects. The environmental sciences in particular benefit from rescued data, with the International Environmental Data Rescue Organization (IEDRO), an entity funding and assisting with environmental data rescue, stating that access to such data can minimise the unnecessary loss of infrastructure, agriculture, the economy and human life (IEDRO, 2014).

Data rescue is described as any effort to preserve data at risk. It includes but is not limited to any of the following activities: digitisation, format migration, treating damaged materials, or adding metadata. It is any action to make data accessible in the long term (Hsu *et al.*, 2015). Best practice guidelines of the WMO on climate data rescue highlight a number of rescue activities, including searching for and locating paper media, preserving and storing the located data, and creating inventories of the data. Imaging, digitising and archiving are also steps regarded as vital to data rescue (WMO, 2016). In a similar vein, a Canadian manual on the rescue of marine species data underlines activities such as inventories creation, digitising, data description, archiving, adding metadata, sharing of data, and data harvesting of the rescued datasets (Kennedy, 2017).

Descriptions of the rescue process range from brief descriptions to lengthy and intricate manuals, with limited details on the roles and responsibilities of various participatory parties. In addition, the potential role of the library and information services sector, crowdsourcing, citizen scientists, or parties other than the research community is still lacking in many published data rescue models.

A scrutiny of the published data rescue manuals/models/guidelines reveals that many of the rescue tasks form part of, or are linked to data management. With data management already included in the curriculum of many LIS courses, and data management increasingly forming part of the services provided by the library and information services sector, it begs the question: 'How can the library and information services sector participate in data rescue?'

The Council for Scientific and Industrial Research[1] (CSIR) is a leading African scientific and technology research organisation that 'researches, develops, localizes, and diffuses technologies' to accelerate socioeconomic prosperity in South Africa (SA) (CSIR, 2022). Multidisciplinary research is conducted, resulting in the generation of vast quantities of data in varied formats encompassing a range of subject areas. In addition, with the institute formed in 1945 (CSIR, 2022), it is likely that more than seven decades of data generation would invariably lead to some data being at risk. The author of the proposed study is a qualified librarian and interested in delving into the data rescue status quo of the research institute and in establishing the role that the library could play in a workflow that will assist in ensuring that valuable data are rescued. Aspects of importance to the study include the following characteristics linked to the institute:

- data at risk (formats, volumes, location, priorities),
- current data rescue expertise,
- current data rescue workflow and activities,
- current data rescue tools, and
- current data rescue role-players.

Research libraries and information specialists worldwide are already actively managing research data in digital formats, and the CSIR is also moving in that direction. Whilst a data librarian was recently appointed at the institute, and a data management procedure still under institutional review at the start of this study's data collection stages, data management at the selected research institute can be regarded to be in its infancy. A logical future step would be the involvement of the CSIR's research library in the data rescue process. This study intends to define the role and responsibilities of all the stakeholders in the data rescue workflow of the institute's data at risk. An investigation into the various required data rescue activities and the potential roles and responsibilities will all form part of the framework that will guide the broader rescue workflow. The intention is to document how a research library can contribute to data rescue projects other than in trendy 'data refuge' or 'guerrilla archiving' projects (Mayernik *et al.,* 2017; Wright, 2017), and how a study such as this can enhance the current exploratory LIS perspective on data at risk (Thompson, Davenport Robertson & Greenberg, 2014).

1.3 Study objectives

The main research objective of this study is to contribute towards the rescue of data at risk by establishing ways in which parties, other than researchers and scientists, may be involved.

[1] www.csir.co.za

Sub-objectives include the following:

- establish how data rescue is currently done, both locally and internationally,
- adding to the above: determine who is currently involved in data rescue, and their roles,
- establish current data rescue perceptions, needs and challenges,
- create a data rescue workflow based on information gained via the three previous points,
- within the model: stipulate how library and information services professionals can be involved in data rescue,
- adapt the data rescue model based on project outcomes, and
- contribute towards future data rescue activities by ensuring the model is freely available to all interested parties.

These objectives were used to formulate a number of research questions, as outlined in the section to follow.

1.4 Research problem/questions

The previous sections have revealed the following:

- The occurrence of data at risk is a prevalent global phenomenon, with a range of formats in a variety of research disciplines affected by risk factors.
- In many disciplines, the rescue of data at risk is crucial to the knowledge base of the subject area.
- Data at risk can often not be regenerated as their collection stems from a specific period in the past.
- Published data rescue guidance ranges from rudimentary to detailed, with minimal indication of roles, responsibilities and involvement of parties outside the research sector.
- The library and information services sector is already involved with data management; data rescue involves several data management activities.
- The library and information services sector displays untapped and vast potential for involvement in data rescue projects and activities.
- The LIS discipline will benefit from the inclusion of a data rescue module in its curriculum.

As a result of these aspects and findings, a main research question and several sub-questions were formulated, and are described in the sections below.

1.4.1 Research problem/main question

The study attempts to find an answer to the following question:

What are the roles and responsibilities of the research library within a comprehensive workflow for data rescue?

The intention is to answer this question by studying the practice and phenomenon of data rescue in detail, reviewing the various data rescue models and workflows, gaining insight into data rescue best practices, and determining the areas and tasks that would benefit from the participation of the library and information services sector. The answering of the main research question is accompanied by a series of applicable recommendations.

1.4.2 Research sub-questions

To be able to answer the study's research question, it was necessary to also address the following eight research sub-questions:

- *What do the current data rescue frameworks/workflows look like?*
- *How do current South African workflows in data rescue compare with best practices/guidelines internationally?*
- *What is the current documented state of library and information services involvement with data rescue, seen in the global context?*
- *What is the current documented state of data rescue awareness within the South African library and information services community?*
- *What is the current documented state of data rescue involvement/participation within the South African library and information services community?*
- *What is the current documented state of data rescue globally and in SA?*
- *To what extent can the theory and practice be formalised in a model for a data rescue workflow?*
- *What suggestions could be made to include data rescue topics in the LIS curricula?*

A brief discussion on each of these questions, detailing the reason for it to be included in the study, as well as the way each question's answer was obtained, follows below.

1.4.2.1 Research sub-question 1: The current data rescue frameworks/workflows

To answer this research question, the published data rescue models, frameworks, guidance and workflows were reviewed and analysed. This was achieved via the study's literature review in general, and through a process of content analysis in particular. Published data rescue workflows and frameworks were evaluated to gather information on the stages forming part of data rescue workflows, the tasks and activities incorporated within data rescue stages, and the roles and responsibilities of participatory parties.

Information obtained by the answering of this research question would provide a global perspective on data rescue workflows, indicate best practices, and give an indication of various roles and responsibilities within the workflows. It was anticipated that such details would assist in the drafting of a data rescue workflow model that would have wider use than in the research community only, as recommendations regarding the various ways the library and information services community can participate in data rescue projects would also be provided.

1.4.2.2 Research sub-question 2: The current SA workflows in data rescue vs international best practices

Answering this question enabled the gathering of information about the data rescue workflows used in SA, including details about the differences and similarities between local data rescue workflows and workflows used elsewhere. It was anticipated that investigating South African data rescue workflows would also provide a better idea of the number of South African data rescue workflows being used, as well as the range of rescue models, and typical stages, typical features, and common activities and tasks. Insight into local data rescue workflows would also indicate if and how local data rescue is lacking, whether there is adherence to best practices, and whether specified data rescue roles and responsibilities can be identified via the models.

To answer this sub-question, published sources detailing South African data rescue projects and the accompanying workflows were inspected and reviewed. In two instances, email communication was used to supplement information traced via conventional literature searches. After receiving all requested information and reviewing supplied details together with information found via literature searches, comparisons with international workflows were made.

1.4.2.3 Research sub-question 3: The current documented state of library and information services involvement with data rescue in the global community

Through answering this research question an understanding was formed of the extent of library and information services involvement in data rescue. It was anticipated that scrutinising global data rescue models would produce details about the data rescue activities typically performed by research library professionals, the indirect participation of the library and information services sector (e.g., providing a storage location, or providing data management guidance), disciplines more or less likely to involve the research library sector in rescue ventures, and data formats that are more or less likely to involve the research library sector in their rescue activities. It was also anticipated that the review of data rescue publications would deliver an understanding of areas that could potentially benefit from library and information services input, such as repository-related tasks, and metadata activities. In addition, gaining insight into the potential library and information services involvement within data rescue would pave the way for the indication of research library roles and responsibilities during data rescue.

This question was answered via literature searches into data rescue and determining whether and how the library and information services sector was involved in such ventures. Published sources were scrutinised for overt mentions of library involvement, such as the project being managed by a librarian, or specific mentions of library involvement, library positions, or a library providing a suitable data rescue location. Additionally, literature was also reviewed for covert LIS involvement, such as a data rescue workflow describing a data rescue activity that could also be executed by a library and information services professional.

1.4.2.4 Research sub-question 4: The current documented state of data rescue awareness in the South African library and information services community

Through answering this research question an idea would be formed of the awareness of, and knowledge about data at risk and data rescue within the South African library and information services community. Such details would indicate whether awareness training is required among South African LIS sector professionals and whether there is a need for inclusion of a data rescue module at tertiary LIS level in SA.

This question was answered via literature searches into applicable published sources detailing South African data rescue efforts.

1.4.2.5 Research sub-question 5: The current documented state of data rescue involvement in the South African library and information services community

Insight into current South African LIS sector involvement would be an extension of the answer gained via the previous research question, where the data rescue awareness of the same sector was discussed. Through answering the research question in the current section, insight into the current involvement of the LIS sector in data rescue would be obtained, with findings indicating data rescue roles and responsibilities. In addition, findings would also reveal the extent of involvement, the current research library positions suited towards data rescue, and whether the physical library space would also play a role during data rescue projects. Conversely, findings would also indicate whether there is a lack of LIS sector involvement in South African data rescue projects.

As with the previous research sub-question, this question was answered via literature searches into applicable published sources detailing South African data rescue efforts.

1.4.2.6 Research sub-question 6: The current documented state of data rescue globally and in South Africa

It was anticipated that through answering this research question insight would be gained into various aspects related to data rescue, such as data rescue organisations and funders, discipline-specific data-related issues, data formats being rescued, the nature of data rescue projects, data rescue outcomes/deliverables, and data rescue events and training opportunities. An investigation into the state of data rescue would also bring forth details about data at risk, and factors leading to data being at risk.

To answer this research question a wide range of published sources was consulted, including scholarly articles, conference papers, popular magazine and newspaper articles, blog posts, Twitter and other social media posts, blogs, university websites, and university library websites. In addition, the search for South African data rescue information entailed email correspondence between this researcher and the study leaders of local data rescue projects.

Apart from the study's literature review activities, additional details about data rescue in SA were obtained via the study's empirical stage. Information about data at risk and data rescue as experienced by experts based at a South African research institute was gathered via the various data collection methods briefly described in Section 1.8.2.

1.4.2.7 Research sub-question 7: Formalising the theory and practice in a workflow model for data rescue

Through answering this research question, it was possible to create a data rescue workflow model indicating the generic stages and activities for rescuing data. The model also describes outputs forming part of each data rescue stage, and is supplemented by guidelines and templates accompanying certain rescue activities or stages. In addition, the model provides an alternative route for handling data at risk should there be inadequate data rescue infrastructure and resources available at the time of data identification and assessment. The model is discipline-agnostic, and caters for the rescue of a range of data formats.

To answer this research sub-question, a range of different research activities was performed which entailed the following:

- literature review into published data rescue projects,

- content analysis of a representative sample of data rescue models,
- creating an initial data rescue model based on the steps and data collected,
- a process of review of the initial data rescue model (involving academics and peers),
- incorporating online questionnaire data and interview data into the model feedback data,
- amending the initial Data Rescue Workflow Model based on review feedback and empirical data collected,
- discussion (involving research library experts) of an amended Data Rescue Workflow Model created by this researcher, and
- presenting a recommended Data Rescue Workflow Model based on all feedback, data and reviews obtained from the different expert samples.

The continual model reviews and feedback (i.e., empirical activities), supplemented by insight gained through literature review and its content analysis enabled the creation of a Data Rescue Workflow Model showcasing vital rescue activities and stages, best practices, and accompanying guidelines, templates and outcomes.

1.4.2.8 Research sub-question 8: Inclusion of data rescue topics in the LIS curricula

The most important outcome of the inclusion of data rescue in the LIS curricula includes an increased awareness of data at risk and data rescue by a sector showing good potential for data rescue involvement. Apart from heightened awareness regarding data at risk and data rescue, the suggested data rescue module should also cover the recommended Data Rescue Workflow Model, thereby providing students with details of data rescue stages, activities, outputs, roles and responsibilities. Exposure to data rescue as a vital practice, and the recommended data rescue model as a segment of that practice, would therefore not only enlarge the knowledge base of LIS students, but also promote the practice of data rescue within the LIS community. In addition, a new generation of LIS practitioners would be entering the workforce already armed with an understanding of this vital component of data management and digital curation.

The research question was answered via cumulative LIS-related information gained through the study's literature review, analysis of published data rescue workflows, and data gathered by means of the empirical stage of the study.

1.5 Field of research

This study, investigating various aspects of data at risk and data rescue, and proposing a Data Rescue Workflow Model also indicating areas of LIS sector involvement, intends to make an essential contribution to the field of Information Science in SA. It is especially focused on increasing the prevailing knowledge, information and scientific studies available in the subfields of data management and digital curation. Apart from contributing to the field of Information Science, the findings of the study will also assist SET researchers in understanding what needs to be done to rescue their longitudinal data.

1.6 Relevance of study for the subject field

It is anticipated that the study will contribute to the subject field in the following manner:

- increased awareness of data at risk,
- identification of factors putting data at risk,
- promotion of correct handling of data at risk (prior to a rescue project),
- increased awareness of data rescue as an activity,
- better understanding of the current state of local and global data rescue activities,
- increased awareness of data rescue workflows and models (local and abroad),
- better understanding of best practices regarding data rescue,
- identification of data rescue challenges,
- increased awareness of data rescue requirements as stated by researchers,
- increased awareness of data rescue requirements as stated by library and information service experts,
- development of a data rescue workflow model,
- identification of the potential ways in which the LIS sector can be involved in data rescue activities and projects,
- the recommended data rescue model and its described LIS sector roles and responsibilities to provide impetus to the involvement of the research library sector in data rescue projects, and
- identification of the ways in which the topic of data at risk and data rescue can be included in the LIS syllabus.

In addition, the outcome(s) of the study may influence the processes of linked science, engineering and technology (SET) research disciplines.

1.7 Overview of the literature

This section provides a brief overview of the process of the literature review implemented during this study. More detailed overviews are provided in Chapter 2 and Chapter 3.

The anticipated outcome of the literature review was to help define the key concepts, identify best practices regarding data rescue, and to lay a theoretical framework for the empirical study. The literature review comprises two chapters, and its contents and objectives are summarised below.

1.7.1 Chapter 2: Literature review

This chapter contains an overview of the literature related to data at risk and data rescue. It includes sections on terminology related to data at risk and data rescue, touches on the factors putting data at risk, describes how to assess data at risk and the feasibility of data rescue, and details the outcomes of data rescue. The literature review chapter also includes topics such as the disciplines commonly involved in data rescue, geographic prevalence of data rescue projects, data formats featuring in data rescue projects, data rescue project participants, and meetings focused on data at risk and data rescue.

The objective of this chapter is to acquaint this researcher (and the reader) with the available body of knowledge around data at risk and data rescue. The chapter also serves to improve the research methodology and to contextualise the findings of the study through demonstrating what this researcher proposes to examine and what has already been examined.

1.7.2 Chapter 3: Literature and the creation of a data rescue workflow model

This chapter contains a scrutiny of 15 selected published data rescue workflow models, and includes the description and results of the process of content analysis applied to the selected models.

The objective of the chapter is to determine commonalities and differences between data rescue workflows, determine data rescue best practices, and use these findings to assist with the creation of an initial Data Rescue Workflow Model. The review of the initial model by an expert sample formed part of the empirical data collection stage of this study.

1.7.3 Summary

The literature review activities of this study comprise two chapters, with Chapter 2 providing a conventional review of literature related to data at risk and data rescue, and Chapter 3 comprising a review and content analysis of 15 selected published data rescue models.

1.8 Research methodology

This section provides a brief overview of the research process used during this study. A more detailed overview is provided in Chapter 4.

1.8.1 Research philosophy, research approach, and research design

This study implements the interpretivist research philosophy, emphasises qualitative analysis over quantitative analysis, focuses on meaning, and employs multiple methods to reflect different aspects of data at risk and data rescue.

The research approach used in this study entails qualitative research. As stated by Teherani *et al*. (2015: 669), qualitative research involves the systematic inquiry into phenomena in natural settings. Such phenomena may include aspects such as investigating how individuals and/or groups behave, or how organisations function, with the researcher being the main data collection instrument (Teherani *et al*., 2015: 669). This researcher will therefore examine why events linked to data at risk and data rescue occur, what happens during data rescue, and aspects related to the data rescue experience of experts based at a selected South African research institute. Although there are elements of quantification to learn more about the phenomenon at a specific stage of the research, the study is predominantly qualitative.

The research design used consisted of both empirical and non-empirical study. The non-empirical part of this study includes a literature review, and the empirical part involves several data collection stages and methods.

This study required an in-depth investigation into data at risk, data rescue practices and experiences, and data rescue services and infrastructure required by experts at a selected South African research institute. Based on these research components, case study research was used as a research design for this study. Case study research involves the detailed and intensive analysis of a particular event, situation, organisation, or social unit (Schoch, 2019: 245). These traits, together with associated case study advantages such as diverse kinds of data collected, not being bound by specific methods, and the accompanying principles and lessons learnt enabling transferability (Schoch, 2019: 245-246) made case study research an ideal design for this study.

1.8.2 Data collection methods and tools

The empirical phase of this study involved multiple data collection methods. Using these methods allowed the researcher to collect data that would generate details about the following:

- data at risk at the selected research institute,
- data rescue activities at the selected research institute,
- data rescue challenges experienced at the selected research institute,
- requirements for, and perspectives on a data rescue workflow model created by this researcher, and

- perspectives on data rescue roles and responsibilities.

The data collection tools employed in this study initially entailed the following:

- a web-based questionnaire involving research group leaders based at the selected research institute (Sample A),
- in-depth one-on-one interviews with a subgroup of Sample A (i.e., Sample B), and
- feedback supplied by Sample B after reviewing a data rescue workflow model created by this researcher.

It was anticipated that should the three data collection stages produce an insufficient amount and depth of feedback after the data rescue model had been reviewed, an additional data collection stage involving a different sample of experts would be required. As it turned out, this additional step was needed.

The intention of the process of systematic review of the created data rescue workflow model is to eventually arrive at a recommended model that would incorporate not only established theory, but also be based on model feedback received and recommendations put forward by two expert research samples.

1.8.3 Target population

The target population of this study is the entire population or group that this researcher was interested in researching and analysing. In this study, researchers employed at the selected research institute, and library and information services professionals based at the selected institute’s research library comprised the target population.

1.8.4 Sampling

Sampling is the process of selecting a representative group from the target population under study. Purposive sampling is a non-probability sampling method, and according to Laerd Dissertation (2012) it involves a sampling technique where the units being investigated are based on the judgement of the researcher. In this study, purposive sampling was the sampling technique used in that research experts, as a subgroup of the target population, were selected to form the respective samples completing the study’s web-based questionnaire, being interviewed by this researcher, and requested to supply feedback on a data rescue workflow model.

Similarly, the study’s mini focus group sample also entailed purposive sampling in that three research library experts, based at the institute’s research library, formed the selected sample invited to participate in the mini focus group session.

Purposive sampling was used, as it was judged to result in a group of experts who would best enable this researcher to answer the study's research questions. Three purposive samples were used, each differing in terms of size, nature and characteristics.

- Sample A comprised all 49 of the institute's research group leaders (RGLs), who were invited to complete a short web-based questionnaire featuring questions on data at risk and data rescue.
- Sample B comprised 18 RGLs, all of whom were respondents who had stated via the web-based questionnaire to either have data at risk in their research group or to have performed data rescue. Sample B members were invited to participate in one-on-one interviews where data at risk and data rescue would be discussed. Sample B were also requested to provide feedback on an initial Data Rescue Workflow Model.
- Sample C comprised three library and information services professionals based at the institute's research library. The three library professionals were invited to a mini focus group session where two data rescue workflow models would be discussed. Invited research library professionals were regarded as experts in the fields of (i) institutional workflows and research–library collaborations, (ii) records management, archival procedures and data management, or (iii) SET-based research and having been employed as a researcher prior to joining the research library.

1.8.5 Ethical considerations

Despite no safety and health implications being anticipated during the execution of the study, several potential ethical issues accompanied the study and their adherence had to be ensured before ethical clearance could be granted.

An important ethical consideration pertained to the treatment of confidential data collected during the study. It was crucial that all personal, institutional and identifying details be removed from data collected via the web-based questionnaire, from transcribed interview data, from feedback data, and from data gathered during a possible mini focus group session. Data would only be uploaded to a public repository once the data had been anonymised or de-identified.

Other ethical considerations featuring in this study involved obtaining institutional managerial consent, study participants being informed of their rights, and study participants informed how data would be treated, stored, shared and used. Informed consent had to be obtained prior to any data collection stage.

Both the University of Pretoria and the research institute, where this researcher is based (and where the study was conducted), reviewed the study's ethical application and granted ethical clearance approval for the study. Documented proof of ethical approval, granted with reference number **EBIT/114/2020**, was communicated to this researcher on 23 June 2020 and is attached as Appendix 20.

1.8.6 Data collection

This study predominantly makes use of qualitative data, collected via the four different data collection methods described in Section 1.8.2.

The web-questionnaire data comprise both quantitative and qualitative information, as several questions comprise a 'yes/no/unsure' response. The web-based questionnaire contains several quantitative questions, however, this was done to gather more information about data at risk and data rescue at this preliminary stage of research. Web-questionnaire data will be exported to an Excel spreadsheet.

The interview phase of the study contains collected qualitative data from eight RGLs through the use of a semi-structured interview schedule. This qualitative data comprised responses regarding the nature of the group's data at risk, factors placing data at risk, previous data rescue activities and data rescue obstacles. The virtual one-on-one interviews were conducted by this researcher and recorded via the online platform's recording feature and a mobile phone as recording backup. Audio data were transcribed by this author to text documents. To limit researcher bias and ensure consistency, the same basic set of semi-structured interview questions was used with all participants. Deviation from the schedule, prompts and additional questions were implemented when the nature of feedback necessitated such steps.

The third data collection phase of this study dealt with qualitative information in the form of textual feedback received from four RGLs after they had reviewed the initial Data Rescue Workflow Model created by this researcher. This qualitative data comprised suggestions, opinions, recommendations, criticism and commendations linked to the reviewed model. Three of the respondents supplied feedback via email while a fourth RGL used email and Skype as the feedback method.

The final data collection method entailed the collection of qualitative data supplied during a mini focus group session. This researcher acted as facilitator and note-taker during the session where participants comprised three library and information services experts based at the institute's research library. As with the feedback stage, qualitative data collected comprised opinions, recommendations, queries and clarifications after viewing the data rescue model. To limit researcher bias all participants were encouraged to contribute and were made comfortable with voicing different opinions.

The different data collection methods and samples enabled the application of triangulation during this study, and are briefly discussed in Section 1.8.7: Data analysis, and in more detail in Section 4.11.2.3: Reliability and trustworthiness in this study.

1.8.7 Data analysis

With the data collection tools gathering qualitative data, the method of analysis used in this study entails content analysis and thematic analysis. Content analysis was the method chosen for analysis of the various published data rescue models or guidelines, as this method is a powerful tool when used in combination with other research methods such as archival records and interviews, and is also useful when analysing historical documents (Columbia University, Mailman School of Public Health, 2019).

Thematic analysis was implemented for analysis of the empirical data, as this method is deemed as ideal when exploring patterns or themes across qualitative data (Maguire & Delahunt, 2017: 3352).

Triangulation was implemented during data analysis. This strategy enhanced the validity and credibility of the findings and played a role in mitigating any research biases in the work. A diagrammatical summary of triangulation techniques used in this study is presented in Figure 1.1.

The two triangulation strategies entailed methods triangulation and data triangulation. The former strategy involved making use of different methods (i.e., information obtained via the web-based questionnaire and the information gathered through in-person interviews) to obtain an understanding of the selected institute’s data at risk and data rescue activities. The latter method formed part of data triangulation, and involved collecting data from different samples to obtain feedback regarding the study’s data rescue workflow models.

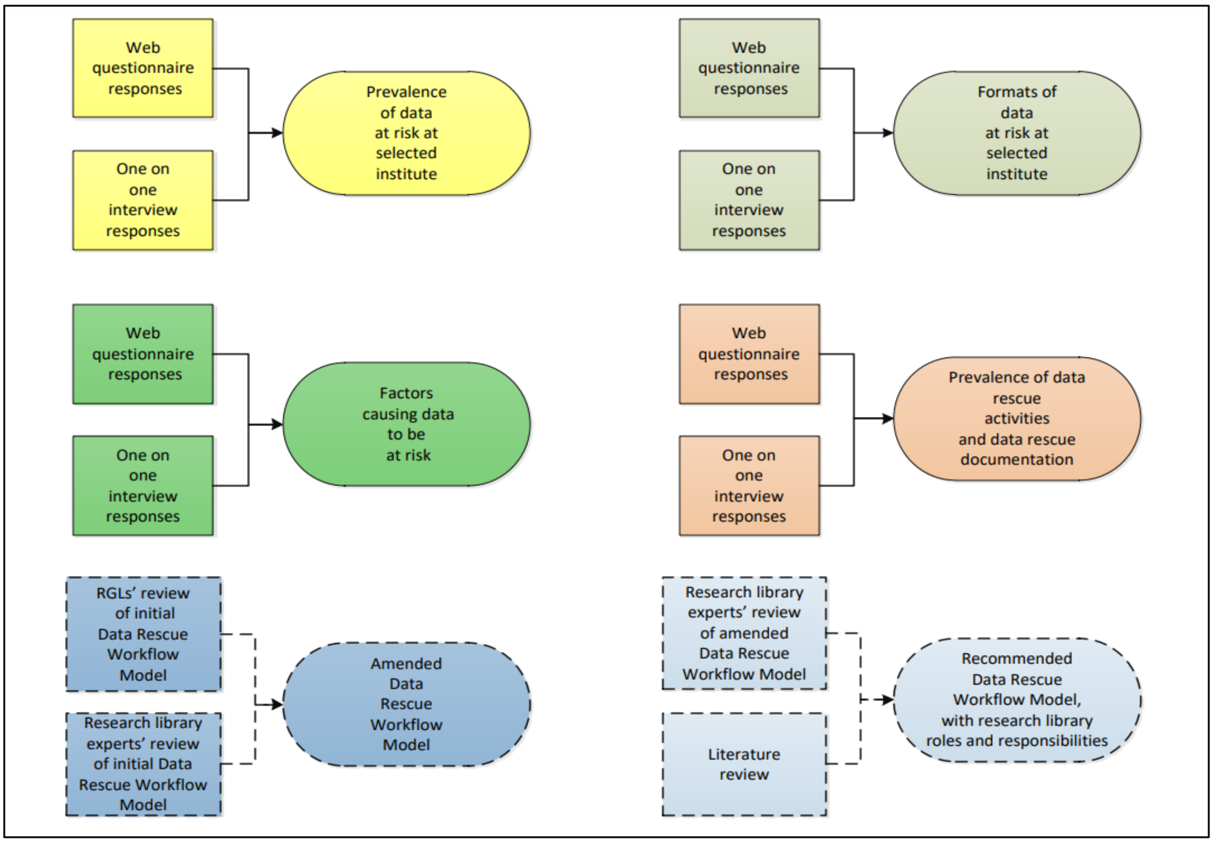

Web questionnaire responses
One on one interview responses
Prevalence of data at risk at selected institute
Web questionnaire responses
One on one interview responses
Formats of data at risk at selected institute
Web questionnaire responses
One on one interview responses
Factors causing data to be at risk
Web questionnaire responses
One on one interview responses
Prevalence of data rescue activities and data rescue documentation
RGLs' review of initial Data Rescue Workflow Model
Research library experts' review of initial Data Rescue Workflow Model
Amended Data Rescue Workflow Model
Research library experts' review of amended Data Rescue Workflow Model
Literature review
Recommended Data Rescue Workflow Model, with research library roles and responsibilities

Figure 1.1: Use of triangulation in this study

1.8.8 Data management plan

It was considered important and a value-adding step to include the study's data management plan (DMP) as part of the study methodology. As an indicator of the ways in which this researcher will be managing the data collected during this study, including storage of data, sharing of data, and long-term preservation of data, the DMP (see Appendix 21) adds to the reader's understanding of the study methodology used.

1.8.9 Methodology challenges

The researcher anticipated challenges linked to the methodology of the study. One of the major hurdles included possible survey fatigue at the selected institute resulting in lower response rates, with similar data collection tools frequently distributed to institutional employees. Another potential hurdle related to the virtual interview tool to be used, and its associated complications such as connectivity problems, quality of audio, and electricity load-shedding issues.

The implementation of an additional data collection stage, specifically dealing with the review and critique of a data rescue workflow model by institutional experts, can be viewed as a methodological challenge. The mini focus group discussion involving research library experts was not anticipated at the start of the study, but was considered vital after the planned review activities by research experts had failed to produce sufficient in-depth feedback pertaining to the initial Data Rescue Workflow Model.

Arranging a mini focus group session during a period when even small gatherings were subject to a range of stringent COVID-19 restrictions is regarded as a methodology challenge. Strict adherence to required protocols was crucial, while it was also vital to ensure the safety of participants, accompanied by the free and unencumbered flow of ideas, critique, suggestions and recommendations pertinent to the proposed Data Rescue Workflow Model.

1.8.10 Research process

The chronological steps forming part of the research process are described in detail in the chapter dealing with the study methodology. The summarised version of the research steps is portrayed in the following graphic:

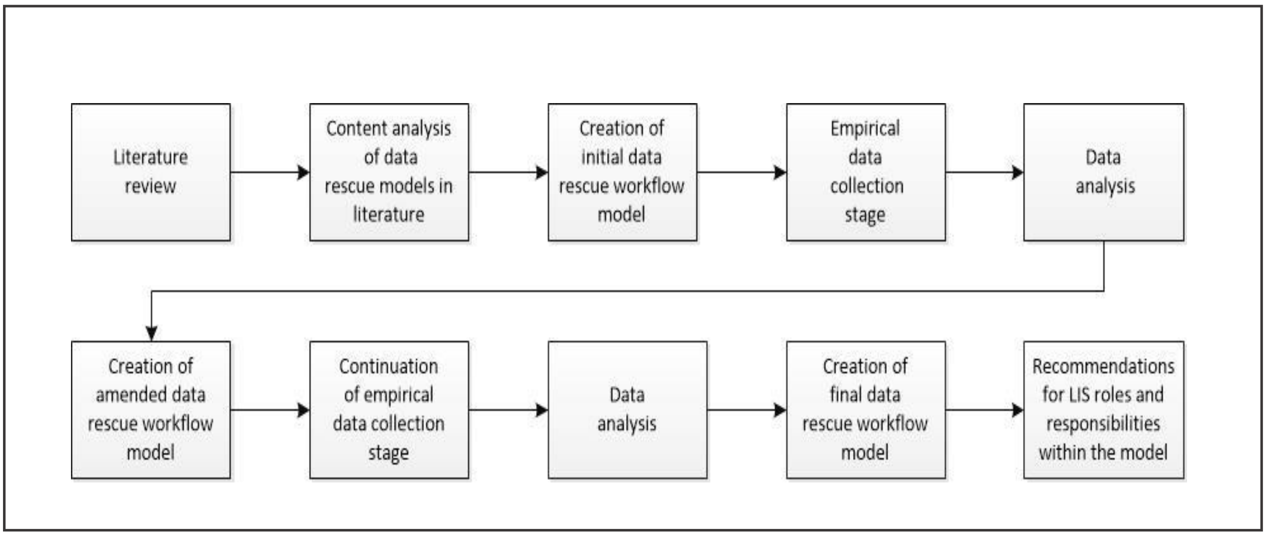

Literature review
Content analysis of data rescue models in literature
Creation of initial data rescue workflow model
Empirical data collection stage
Data analysis
Creation of amended data rescue workflow model
Continuation of empirical data collection stage
Data analysis
Creation of final data rescue workflow model
Recommendations for LIS roles and responsibilities within the model

Figure 1.2: Summary of research process

1.9 Scope and limitations of the study

This study focuses on the rescue of data at risk, with the data collected during the empirical research stage of this case study emanating from a sample of experts based at the selected South African research institute. As a result of this interplay of factors, several issues should be highlighted as forming part of the scope and limitations of this study.

1.9.1 Selected institute

The empirical data collection stage of the study involved participants from the selected research institute only. Although content analysis was applied to global and South African data rescue workflows and models (see Section 1.10.3 below), the data collection phases were limited to respondents based at the selected research institute. As such, the study's resultant data and findings should be seen as influenced by aspects such as the institute's mandate and available funding, resources, infrastructure and systems. Despite this limitation, it is anticipated that many of the study findings and learnings will be applicable to other research institutes, especially those involving SET-based research.

As the study's recommended Data Rescue Workflow Model is discipline-agnostic, its relevance is not limited to SET-based research institutes. The model will certainly be of use to any organisation dealing with research data, and the LIS-related recommendations will be of use to an institute with library and information services professionals familiar with data management. Such institutions include university departments, university libraries and special libraries.

In addition to institutional use, the model will also be of value to individuals who wish to embark on data rescue without having the support of a linked library.

1.9.2 Science, engineering and technology

An aspect linked to the point discussed in Section 1.9.1 is the fact that the empirical phase of the study involves experts from the SET sphere only. As such, participation by and input from experts in non-SET fields such as the humanities, arts, literature, and management studies are not included in the study.

Despite the empirical data collection stage entailing input from employees based at an SET research institute, the study findings and resultant discipline-agnostic data rescue workflow model are anticipated to be relevant to all research institutes, tertiary establishments, and individuals concerned with the correct treatment of data at risk.

1.9.3 Research areas

An aspect linked to the two points mentioned above concerns the research areas forming part of the selected institute, and the research areas of actual respondents. A list of the selected institute's research areas is shown in Table 4.4: Full population (all research groups) and Sample A (indicated in bold).

While the study data emanated from experts performing research in the aforementioned research disciplines, it is anticipated that the recommended model and study findings will be applicable to a wider group than the selected institute, and its relevant research areas.

1.9.4 Geographic location

Information gathered about data at risk and data rescue during this study emanates from South African and international published sources, however, the focus of this study and the organisational entity featuring in the case study is a South African research institute. This choice of location is based on reasons of practicality. Although it is not anticipated that the geographical area will limit the universal applicability of the findings and usability of the study's data rescue model and recommended library and information services participatory activities, the aspects mentioned below should be kept in mind.

- Terminology used in the study is based on South African usage. Terms and phrases such as 'data librarian' and 'honours degree' are words that may have a different meaning than the South African usage in other parts of the globe. These South African terms might also be known by another word elsewhere. The reference to South African currency (e.g., R25 000) when describing the cost of data storage should also be kept in mind; in this instance the amount indicates 25 000 South African Rands.
- The status of data management and the resultant data at risk and data rescue practices are anticipated to be at a different maturation stage when compared with research institutes in the United States of America (USA), United Kingdom (UK) or Europe.
- Research areas deemed to be of vital importance to South African development and innovation may have a lesser status in certain parts of the world.

1.9.5 Sample

Three different purposive samples were used during this study's empirical stage, and respectively comprised research experts based at the selected research institute (Sample A), a sub-group of Sample A (Sample B), and research library experts based at the selected research institute (Sample C). While the input and feedback supplied by sample members was linked to their own research groups/discipline, and to their own experiences, challenges and requirements, the findings of the study and its resultant data rescue workflow model are anticipated to be applicable to a wider audience.

1.9.6 Data origin/use

This study involves an investigation into mainly research data, even though respondents from the selected institute also work with innovation and development data. The empirical stage of the study does not involve input of non-research sectors of the institute handling non-RDI[2] data, such as financial data, institutional human resources data and institutional maintenance data. Identified data at risk and the described data rescue practices within this study emanate from research activities.

1.9.7 Data at risk

The study is concerned with the identification, handling and rescue of data at risk. For the purposes of this study, such data are defined as scientific data which are not in modern electronic formats and the information of which is therefore not accessible to the research that needs it.

Even though the data collected during the empirical stage of the study focused on data at risk, the study's resultant data rescue workflow model can be used to include more than data at risk. Several of the stages form part of best practice with regard to data management and can therefore be viewed as a model portraying the conservation of any type of data generated during research, irrespective of the risk status of the data.

1.9.8 The concept of 'research library'

A primary outcome of this study is an indication of potential roles and responsibilities of the research library within the recommended Data Rescue Workflow Model. A subsection of the study's collected data emanates from responses by experts based at the selected institute's research library. In contrast, the study's literature review and content analysis sections also involve academic libraries, public libraries and special libraries.

It is anticipated that the resultant Data Rescue Workflow Model will be of value to any library dealing with data at risk, and with such libraries ideally having a library and information services professional familiar with data management, particularly the role of metadata, repository selection and use, and long-term data preservation.

1.9.9 Resource availability

The topic of resource availability plays a role when it comes to identifying factors putting data at risk, data rescue challenges experienced, the creation of a data rescue workflow model, and the indication of research library roles and responsibilities within a data rescue workflow model. These resources

[2] Research, Development, Innovation

can be specific to an institute, a specific research group, and the library involved in a prospective data rescue project.

Resources discussed in this study may refer to time, skills, expertise, infrastructure and available funding. Workload and institutional priorities may also influence available resources. These are potential limitations that should be borne in mind by the reader when studying the findings of the empirical part of the study, the recommended Data Rescue Workflow Model and its indicated research library roles and responsibilities.

1.10 Exposition of chapters

The division of chapters is described below.

1.10.1 Chapter 1: Introduction

The study's first chapter is an introduction to the study and describes the research problem and research questions, elaborates on the scope and limitations, discusses the significance of the study, and provides an outline of the methodology to be followed.

1.10.2 Chapter 2: Literature review

Chapter 2 contains the literature review of this study and includes an overview of the published theoretical knowledge base regarding data at risk and data rescue. The chapter contains sections on terminology associated with data at risk and data rescue, and details the factors leading to data being at risk. In addition, the chapter investigates features linked to data rescue projects such as disciplines involved, data rescue participants, geographical areas linked to data rescue projects, and data formats forming part of rescue activities.

1.10.3 Chapter 3: Literature and the creation of a data rescue workflow model

Chapter 3 contains a discussion of 15 different data rescue workflows/models found in published literature, and the resultant process of content analysis applied to these models. The chapter also describes the steps used by this author when applying the results of content analysis to create an initial Data Rescue Workflow Model. These steps are followed by a description of the model and each of the data rescue stages forming part of the model.

1.10.4 Chapter 4: Methodology

The methodology chapter presents the description of the research process and the justification of included methods, samples and techniques. The chapter commences by providing information concerning the research paradigm, research approach and research design used in this study. The chapter also describes the study participants, including the sampling method/s used and the justification of these methods. Descriptions of the various data collection tools forming part of the study's empirical stage form a major part of the chapter. The data analysis methods implemented during this study are described in this chapter. The methodology chapter also discusses the ethical clearance process, and the ways in which this study addresses the topics of validity, reliability and transparency. The study's data management plan, attached as Appendix 21, is also mentioned during this chapter, as the way the data is handled is regarded as part of the study's methodology.

1.10.5 Chapter 5: Results and findings

This chapter contains the results of the study and includes a discussion of the study findings. The chapter consists of four main sections, with each section making use of a different data collection tool to collect data. Collected data and the information gleaned from the data were essential in meeting the study's research objectives and answering the study's research questions. The four main sections of this results chapter, with the respective data collection tool/method used, therefore discuss the results of the web-based questionnaire involving Sample A, the results of one-on-one interviews with Sample B, results of feedback supplied by Sample B after they had reviewed the initial Data Rescue Workflow Model, and results of a mini focus group session held with Sample C after discussing the initial Data Rescue Workflow Model, and the revised model.

Detailed information on the findings emanating from each of the data collection stages, and a discussion of findings are provided in the remainder of the chapter.

1.10.6 Chapter 6: Recommendations

This closing chapter presents the study's main research question and research sub-questions and discusses how these questions have been addressed. The section containing research questions is followed by a summary of the study's main findings, after which a conclusion regarding the findings is presented.

Based on the study findings and considering the main research questions of the study, several recommendations have been put forward. Ideas for future research emanating from this study are also stipulated.

This chapter ends with a study conclusion.

1.10.7 Appendices

Twenty-three appendices are attached to the manuscript. These appendices contain clarifying details regarding the study methodology, or guiding information linked to the data management tasks forming part of the various data rescue workflow models.

1.11 Conceptual framework

A conceptual framework for the study has been created and is presented in Figure 1.3.

Patil and Aditya describe a conceptual framework as the interrelated system of the study's research problem, the context, variables and phenomena linked to the research problem, presumed relationships that exist between these variables and the research methods the researcher will be implementing to investigate these surmised relationships (2020). Adding to this is the statement by Varpio *et al*. that the conceptual framework is the justification for why a given study should be conducted, as it describes the state of known knowledge (usually through a literature review), identifies gaps in our understanding of a phenomenon or problem and outlines the methodological underpinnings of the research project (2020).

With the conceptual framework bringing together different aspects of scientific research in a single framework (Patil & Aditya, 2020), this researcher has opted to portray the framework in graphical format, with the framework displaying key concepts, variables, research methods and anticipated outcomes. This visual framework illustrates the overall structure that will shape this research project and can be viewed as a roadmap for the study.

Charlesworth Author Services mentioned the importance of developing a conceptual framework in the early stages of a study as this guides the researcher's thinking and enables the visualisation of linkages between various concepts (2022). According to Cueva (2022) the conceptual framework also enables readers to have a general understanding of the researcher's planned work. Afribary (2020) supports this viewpoint by stating that while a conceptual framework presents a researcher's perception about the research problem, it is also an arranged and self-explanatory method drafted for the readers.

The benefits of a shared conceptual framework are numerous and in this study it not only serves to clarify what the researcher intends to investigate, but also assists readers in understanding the purpose, context and logical justification behind the investigation of a given phenomenon or presumed relationships among sets of phenomena.

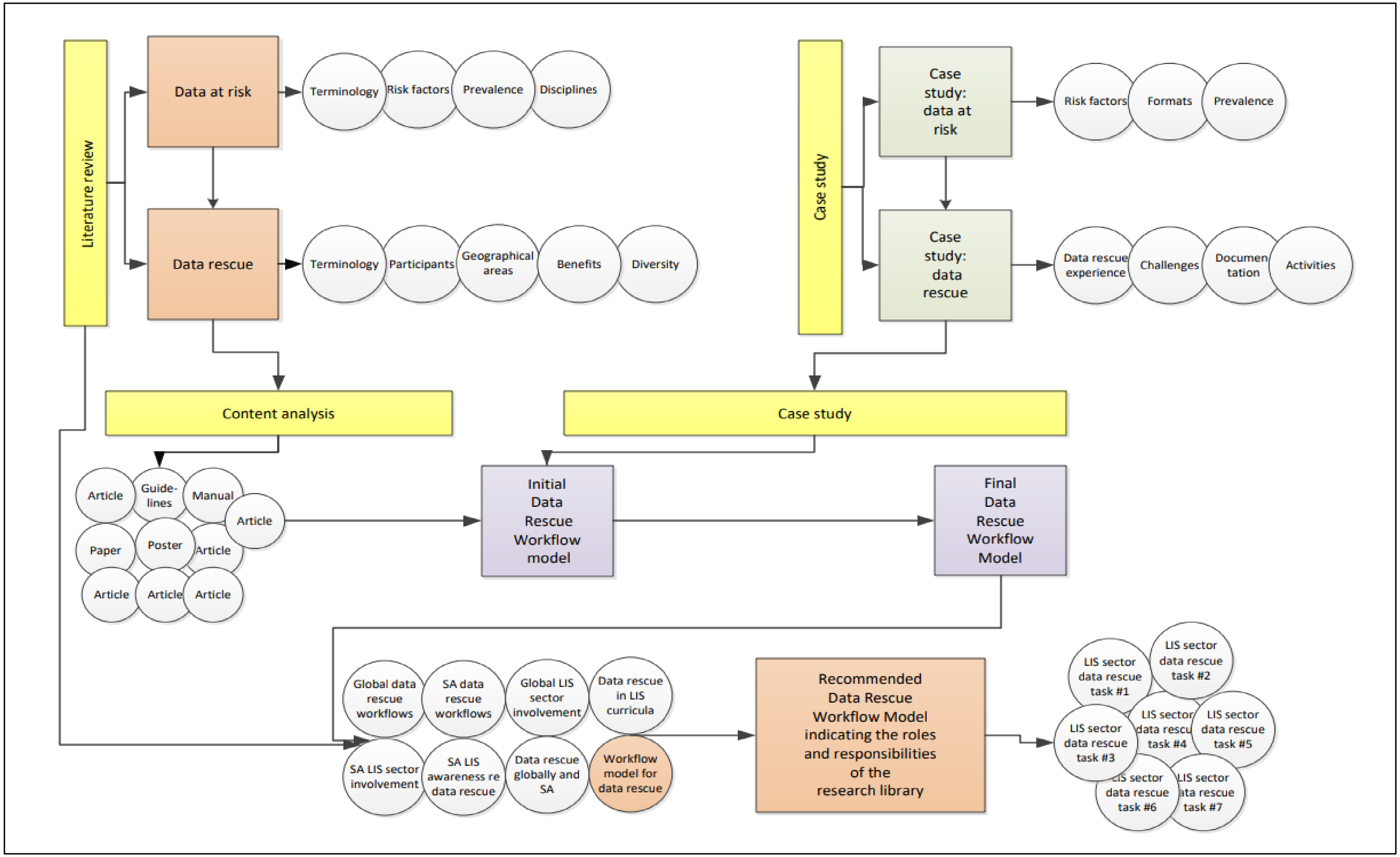

Literature review
Data at risk
Terminology
Risk factors
Prevalence
Disciplines
Data rescue
Terminology
Participants
Geographical areas
Benefits
Diversity
Case study
Case study: data at risk
Risk factors
Formats
Prevalence
Case study: data rescue
Data rescue experience
Challenges
Documen tation
Activities
Content analysis
Case study
Article
Guide- lines
Manual
Article
Paper
Poster
Article
Article
Article
Article
Initial Data Rescue Workflow model
Final Data Rescue Workflow Model
Global data rescue workflows
SA data rescue workflows
Global LIS sector involvement
Data rescue in LIS curricula
SA LIS sector involvement
SA LIS awareness re data rescue
Data rescue globally and SA
Workflow model for data rescue
Recommended Data Rescue Workflow Model indicating the roles and responsibilities of the research library
LIS sector data rescue task #1
LIS sector data rescue task #2
LIS sector data rescue task #3
LIS sector data rescue task #4
LIS sector data rescue task #5
LIS sector data rescue task #6
LIS sector data rescue task #7

Figure 1.3: Conceptual framework

The legend to the framework is as follows:

🟧	Key concepts	🟨	Methodology stages
○	Variables/components	╲	Links (not causal)

The four key concepts of this study have been identified as data at risk, data rescue, a data rescue workflow model, and the involvement of the LIS sector in data rescue. The key concepts were selected based on their particular importance to the study context, and the fact that the four concepts form the essence of the study. With this study investigating the nature and prevalence of data at risk, the nature of data rescue, the components of a data rescue workflow model, and the involvement of the LIS sector in data rescue the identification of the four concepts as key to the study is evident.

Variables and concepts linked to the key concepts (not causally) comprise :

- the different data collection activities used in this study,
- factors leading to data being at risk,
- diverse factors forming part of data at risk,
- challenges linked to data rescue,
- activities forming part of data rescue,
- publications providing guidance on data rescue,
- the study's research questions, and
- rescue activities to be performed by library and information services professionals.

Patil and Aditya (2020) and Maxwell (2005) have reported on the tentative nature of a conceptual framework, and this characteristic of the portrayed structure (see Figure 1.3) is anticipated to lead to an updated and revised version during a later stage of the study.

1.12 Summary

This chapter described the context of the research problem and listed the research problem/question as well as its linked sub-questions. Next, the relevance of the study for the subject field was highlighted. Following this, a brief overview of the literature review steps, and the research methodology followed in this study was supplied. This chapter also provided details on the scope and limitations of the study. Lastly, an exposition of the structure of the study's six chapters was given.

In the next chapter, topics related to data at risk and data rescue, as traced via published sources, will be discussed and reviewed.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter contains a discussion of data rescue as portrayed in literature and other documented sources. Preference was given to data rescue-related research articles from scholarly journals, and papers presented at conferences. Additional sources consulted and scrutinised include data rescue documentation found in:

- non-conference presentations and posters,
- online reports,
- library webpages,
- books and book chapters,
- websites of organisations and institutes, and
- social media pages such as relevant blogs and Twitter.

The reasoning behind the extension of the literature review beyond scholarly sources is that a substantial proportion of current data rescue reporting involved popular and general online publications. Examples of such publishing include universities publishing the details of a planned data rescue event on their website, data rescue events and calls for volunteers published in general online newspapers, data rescue interest groups reporting on activities and achievements using a blog, and presenters uploading data rescue workshop presentations to a platform such as SlideShare[3] or YouTube[4]. These reports, papers and blog entries are vital to this study, as they contain valuable information pertaining to the data being rescued, parties involved, activities planned, and outcomes of the efforts.

The intention of this chapter is twofold. The focus is firstly to provide the reader with an overview of the nature and prevalence of data at risk as found in documented outputs. The second objective is to provide an overview of the nature, prevalence and features of data rescue projects as found via published literature on the mentioned topics. Therefore, the chapter contains sections devoted to factors causing data to be at risk, terminology applicable to data rescue, reasons for data rescue, and benefits and outcomes of data rescue. The different vocations found to be involved with data rescue activities are also discussed, with the involvement of the LIS sector during data rescue highlighted.

[3] https://www.slideshare.net/
[4] https://www.youtube.com/

The chapter demonstrates the interesting mix between the universality of data at risk, and the diversity and range of features found when studying data rescue activities, data rescue outcomes, data rescue participants, and data rescue projects.

This chapter does not contain a description of data rescue workflows, a discussion of which forms the backbone of Chapter 3.

2.2 'Data rescue' and related terminology

This study is focused on the concept of 'data rescue': what it entails, the benefits involved, and the roles and responsibilities of parties taking part. A review of relevant literature has revealed that a clarification of terms related to 'data rescue' is required, and this is done to address the issues briefly listed below.

- Several related words and terms are often mentioned in the same breath as the term 'data rescue'.
- The terms referred to in the point above are often used as synonyms of 'data rescue', but not necessarily so. Alternative terms, when compared with 'data rescue', often contain a slight difference in nuance. There is a need to clarify the intent of the use of the alternative terms.
- The term 'data rescue' was also found to be used when referring to the process of conventional rescue (digitisation of data at risk and sharing the converted data with the public), or to the period-specific (2016–2019) data rescue activities in the USA when data deemed to be in danger of being wiped from federal websites were copied, and saved elsewhere by concerned citizens.

To address the aspects listed above, this section of the chapter elaborates on the various terms closely related to 'data rescue'.

2.2.1 Data rescue

Different definitions or explanations of the term were found in the various literature sources consulted; however, there are definite similarities present in the longer, more complete definitions or explanations of the term. These similar concepts centre around the **subjects/items** being rescued, the state of the **pre-rescued** item, the state of the item **post-rescue**, and the rescue **activity** itself.

An example of such a definition was that of Diwakar, Kulkarni and Talwai (2008: 139), who stated that data rescue is the 'ongoing process preserving all data at risk of being lost due to deterioration of the medium and digitising current and past data into computer compatible form for easy access'. Here it is seen that the rescue subjects are mentioned ('all data', 'current and past data'), the pre-rescue condition is stated ('deterioration of the medium'), the post-rescue ideal is mentioned ('computer compatible form for easy access'), and the activities are included ('preserving' and 'digitising'). In a similar vein is the definition of the Atmospheric Circulation Reconstructions over the Earth (ACRE) describing data rescue as the process of salvaging paper-based climate histories, and that it involves the discovery, identification and cataloguing of worldwide records. In addition, ACRE stated that records are imaged, and then transcribed into an electronic format. Joining these definitions is Levitus (2012: 48), who describes data rescue as an effort to save data at risk by digitising manuscript data, transferring it to electronic media, and archiving the data into an internationally available electronic database. Also fitting the mentioned detailed description is Diviacco *et al.* (2015: 45), who describe data rescue as the extraction of data from a storage system that cannot guarantee their preservation, to a more reliable and accessible system.

Shiue, Clarke and Fenlon (2020) placed greater emphasis on the reasons for data being at risk and stated that they define data rescue as a framing overarching concept of data curation of data that are particularly vulnerable to disappearance, corruption, or obsolescence.

Data rescue may entail various curation processes, including digitisation (the conversion of materials from physical to digital format, e.g., scanning), data recovery (the retrieval of deleted or damaged digital data, e.g., as from a corrupted hard drive), metadata creation, digital preservation activities, and other processes to ensure the continued management, accessibility and usability of data. Whilst physical formats are a major focus of data rescue, data rescue projects increasingly target aging digital data that are remastered (Wyborn *et al*., 2015) into more sustainable and robust digital formats.

Janz (2018: 5) is not alone in her idea that data rescue means a great many things to many people. While data rescue pre-2016 was mostly used to describe the efforts to ensure the preservation of deteriorating data, or preserving data at risk of loss, the year 2016 saw the arrival of data rescue activities directed at rescuing federal data in the USA from expected deletion. While these 2016–2019 efforts are mostly referred to as forming part of data refuge (see separate heading further on), the term 'data rescue' is also used to describe these efforts. A fitting example is Monahan (2017:2), who describes data rescue as 'methods and techniques to identify, store and preserve datasets, predominantly associated with government entities'. Monahan expanded on this description by stating that data rescue is especially critical during election transitions.

Further examples of the use of 'data rescue' pertaining to government or federal data include Gaudin's (2017) description, which states that it brings together software programmers, librarians and other volunteers who are trying to safely archive scientific data from government websites. Similarly, 'datarescue' (used as a single word) is described by Datarescuedenton (2017) as a one-day hack-a-thon type event where trustworthy copies of federal climate change and environmental data will be created.

While pre-2016 data rescue efforts were mostly, to quote Brandsma (2007: 1), concerned with the digitisation of data, and making it available to the public, data rescue efforts linked to the 2016 US elections entailed the copying and storing of datasets, already in digital format, on a public open access repository. This meant that the targeted data pre-2016 was most often historical data, while more recent efforts could also entail the rescue of US federal data. Despite these differences, the two types of efforts are fairly similar, in that parties are interested in providing access to valuable scientific data that are at risk of being lost. The WMO (2014: i) summarises this resemblance by stating that most countries have data requiring a form of data rescue, and that this is usually for one of two reasons: record deterioration, and/or catastrophic loss. Hsu *et al.* capture the essence of data rescue concisely by stating that it entails any effort to preserve data at risk (2013: 3).

Mayernik (2017: 1) mentioned data rescue, data refuge and guerrilla archiving in the same sentence, and added that these initiatives involved the creation of new copies of earth science data generated and held by US federal agencies.

2.2.2 Data conservation

The RDA's Data Rescue Interest Group, after using the term 'data rescue' within its name for years, was renamed as the RDA Data Conservation Interest Group during the latter part of 2019 (RDA Data Rescue IG, 2019). The group's former name and outputs made it particularly valuable for researchers who require data from the past, and the group's website contained the sharing of best practices, and information on hardware and software for archivists and others with responsibilities to oversee the preservation of historical information in the humanities and social sciences.

The newly selected group name was selected to better demonstrate the new scope of the group – the group would carry forward the work of the Data Rescue Interest Group with an expanded scope, as all data, digital and analogue, would be included in its range of focus. The group would also incorporate aspects such as conservation processes, data rescue prioritisation and decision-making, workflows, and data rescue funding in its expanded scope.

As stated on the group's webpage, use of the term 'data conservation' signifies that the group is concerned with data rescue of all types of data (RDA Data Conservation IG, 2021), and would also focus on decision-making topics including the securing of data rescue funding, defining risk factors, assessing risk factors, and sharing best practices related to the prioritisation of data rescue or conservation.

2.2.3 Data Refuge

As stated in a previous section, data refuge is a term used to describe the US data rescue initiative directed at rescuing digital federal/government data deemed to be at risk of inaccessibility in the near future. Guest Blogger's definition, describing it as a distributed, grassroots effort in the USA, where various parties (scientists, researchers, programmers, librarians, and other volunteers) endeavour to preserve government data, is typical of the definition supplied on various data refuge websites.

Wiggin, as quoted by Guest Blogger (2017), expands on this definition by stating that data refuge events have several goals: making research-quality copies of federal environmental data, showcasing the role environmental data plays with regard to health and safety, advocating for robust archiving of, and access to, born-digital materials, and building a consortium of research libraries to scale data refuge tools and practices. This definition reveals data refuge to be more than merely the rescue of federal environmental data; it is a wider-reaching effort working towards environmental literacy as well as laying the foundation for the rescue of other types of federal data.

Falling under the banner of a data refuge event, but labelled differently, is the use of the term 'emergency data-archiving event' to describe an event in New York during 2017 (Phiffer, 2017).

2.2.4 Data archaeology

From a chronological point of view, it can be said that data archaeology precedes data rescue. It is an activity leading up to, or even demonstrating, the need for data rescue. Caldwell (2012: 3) described data archaeology as the process of seeking out, restoring, evaluating, correcting and interpreting historical datasets. An example of the outcome of data archaeology activities, as stated by Bradshaw, Rickards and Aarup (2015: 9), is the filling of gaps in a global environmental dataset.

2.2.5 Data at risk

Simplistically seen, data rescue entails saving data at risk. The term 'endangered data' is often used interchangeably with 'data at risk'. While many factors could lead to a dataset being classified as at risk (see Section 2.3: Data risk factors), Murillo (2014: 207) stated it best, by describing it as essentially being endangered scientific data that are at risk of being lost.

An investigation of literature pertaining to data at risk reveals that, as with the term 'data rescue', it refers to data in a non-digital format, or to data in a digital format but at risk of being deleted, or both formats.

Examples of the first viewpoints include the definition of DARTG[5], which states that data at risk are scientific data not in a format that permits full electronic access to the information they contain. Likewise, Earls *et al.* (2013: 1) described it as data that are fragile or deteriorating, or data lacking sufficient metadata, or data in a format not allowing electronic access. Similar to this is Griffin's stance, whereby data at risk is a blanket term for non-electronic data which are subject to various hazards (2015: 93). Griffin further stated that this is the type of data that cannot be reproduced completely or reliably. The succinct definition of Hsu *et al.* (2013: 2), namely that data at risk are scientific data not in formats that permit full electronic access, captures this stance. These definitions and descriptions exemplify the viewpoint that data at risk are non-digital data that cannot be accessed electronically.

With the section on data rescue showing that digital data can also require rescue, it follows that not all data at risk will necessarily be in a non-digital format. This assumption is supported by the stance of Mayernik *et al*., who state that the term 'at risk' is used to imply that data may be deleted or become inaccessible to the public now or in the future; either being visible as a broken link, or a portal that has been removed (2017: 2). This definition is especially geared towards the data targeted in data refuge activities. Murillo (2014: 207) includes digital data in the 'at risk' description by stating that some born-digital data can be at risk if it cannot be ingested into managed databases, due to lack of adequate formatting or adequate metadata.

2.2.6 Endangered data

As mentioned previously, the term 'endangered data' and 'data at risk' are often used interchangeably. A survey by Murillo (2014: 207) revealed that users ascribe several attributes to endangered data, such as data unavailability, lack of context, or potential endangerment. These risk factors are discussed in more detail in a section focusing on data risk factors (Section 2.3: Data risk factors). An interesting panel discussion around endangered data and data refuge revealed that different approaches to this type of data were taken (Schell, 2018). In addition, one of the panellists took the concept of endangered data a step further by including in this definition data that might endanger people in its collection and use, i.e., data endangering others (Sheble, 2018).

2.2.7 Endangered Data Week

Following in the footsteps of the Data Refuge initiative of 2016 and beyond, the concept of Endangered Data Week was established in the USA in 2018. Repeated annually, the effort is described as a collaborative effort coordinated across campuses, nonprofits, libraries, citizen science initiatives, community activist groups, and cultural heritage institutions, to foster an environment of data

[5] CODATA Data at Risk Task Group: http://www.codata.org/task-groups/data-at-risk

consciousness (Endangered Data Week, 2020). The event anticipated to shed light on public datasets perceived to be in danger of being deleted, repressed, mishandled, or lost, according to definitions found on the websites of many US universities taking part in the event (Illinois Tech Library Guides, 2017; University of Virginia Library, 2018; Yale University, 2019). A similar definition was located on the website of a university in Utrecht, located in the Netherlands (Utrecht Universiteit, 2019).

The event, as described on a webpage of the University of Maryland (2018), pursues goals such as promoting the care of endangered collections, increases critical engagement with these datasets, encourages activism for open data policies, and fosters data skills through workshops.

As can be gathered from the online platforms announcing and promoting the event, the concept of a week centred on endangered data targets more than the rescue of endangered data. It is an initiative involving several stakeholders and strives at contributing towards data refuge activities, improving scientific and environmental literacy, and laying a foundation for future events.

2.2.8 Heritage data

This term is often used in data rescue literature to describe pre-electronic historic environmental data. More specifically, the Research Data Alliance (RDA) sees heritage data as data that are either in analogue or primitive digital formats, and essential to many fields of science due to their long timestamp (2015: 1).

Heritage data at risk are therefore pre-electronic data that cannot be used optimally in research until it can be accessed electronically, or as mentioned by Griffin (2015: 93), data that have come down to modern scientists more by circumstance than by design, thereby revealing a lack of any planning for future use.

2.2.9 Legacy data

Legacy data is a term not often found in data rescue literature, and when mentioned, not formally defined. An investigation of references reveals legacy data to be data that have been collected in the past (Smith, 2015: 74), and are often the best, or sometimes only source of information revealing data differing from the norm (Griffin, 2015: 207).

‘Legacy product’ (USGS, 2019a) is a related term used by the United States Geological Survey (USGS) and falls under the banner of legacy data. This label refers to unreleased data, software, physical samples, and publications developed during USGS projects before 2011. These records are stored in an old or obsolete format, computer system or facility, and because of this are unavailable to scientists and the public.

2.2.10 Vulnerable data

Although not a term commonly used in data rescue literature, the term ‘vulnerable data’ was used by the University of Minnesota when referring to federally produced data hosted on government websites. The data were regarded as vulnerable as their availability was subject to government funding, administrative policy, and the government’s perceived usefulness for the data (2017). The fear was expressed that data, only available on government websites, will be unavailable during government shutdowns. As seen during the US shutdown activities of 2018 and 2019, government shutdowns are a reality, not an anomaly, and a shutdown took place in the USA for 34 days, dating from 22 December 2018 up to 25 January 2019 (Zaveri, Gates & Zraick, 2019). In addition to the shutdown data dilemma, historic data can be lost if governmental archival practices change over time.

2.2.11 Magnetic media crisis

The XRF Collective (Barnard Library & Academic Information Services, 2019) used the term ‘magnetic media crisis’ when postulating about all analogue video and audio formats being unplayable in the year 2030. The Collective also stated that this situation will result in the loss of a great deal of valuable historical documentation and historical data.

The next section elaborates on documented factors leading to data being at risk.

2.3 Data risk factors

The concept of data at risk was defined in the previous section. This section examines more closely the most common reasons why data are categorised as being at risk.

2.3.1 Deterioration of record

The danger of non-digital data deterioration, or data fragility, is a risk factor commonly stated in data rescue reports and articles. Ross and Gow (1999: 2), Nordling (2010), Brunet (2011: 30), Levitus (2012: 46), IEDRO (2014), Downs (2015), Hachileka (2015) and Arrouays *et al.* (2017) all mentioned that paper-based data are subject to deterioration over time. Specific degradation issues include paper data damage due to humidity or tropical conditions (Ross & Gow, 1999; WMO, 2014; Arrouays *et al.*, 2017), paper damage due to rodents and pests (Fry, 2010a; Hachileka, 2015), and fading handwritten records (Hachileka, 2015).

As most datasets were paper based during the pre-digital era, this makes it a risk factor very pertinent in research centres harbouring older data.

2.3.2 Catastrophic loss of records

Environmental catastrophes such as fire (Levitus, 2012; Arrouays *et al.*, 2017), floods (Levitus, 2012; WMO, 2014; Hachileka, 2015), lightning strikes (Ross & Gow, 1999) and storms (Arrouays *et al.*, 2017) are examples of calamities that can contribute towards total and instantaneous data loss. Civil conflict, and war-related damage or destruction (Nordling, 2010; WMO, 2014; Hachileka, 2015; Arrouays *et al.*, 2017) are additional examples of events leading to catastrophic loss of data. Hachileka (2015) specifically mentioned that thousands of historical climate datasets were lost during the civil war in Sierra Leone.

2.3.3 Loss of human knowledge/skills

Data are seen to be at risk when the knowledge or technical knowledge pertaining to the use of the data is no longer available. Examples of the factors included within this risk category include the resignation and retirement of researchers and custodians (Levitus, 2012; WMO, 2014; Wyborn *et al.*, 2015), manpower changes (WMO, 2014), and researcher death (Wyborn *et al.*, 2015). Mavraki *et al.* mentioned that problems with data integration and understanding often remain unresolved when the original authors can no longer be consulted (2016: 28).

Muller (2015a, 2015b) has published a number of documents stating the importance of data rescue capabilities and has expressed the need for an organisation that supports the preservation of these skills. According to Muller, the abilities and expertise within such a group would be of support to scientists, historians, librarians and archivists. The collected details of skills and capabilities spanning government, academic and commercial projects would ensure that the tools do not disappear when projects end, or when staff retire.

While many studies detail the data rescue of paper-based data, the efforts around capabilities of data rescue related to older digital assets are less prevalent in literature. This study’s researcher agrees with Muller regarding the need for preservation and documenting of data rescue skills and capabilities, and is intending to use this study to provide a contribution to the skills set required when implementing a data rescue initiative.

2.3.4 Outdated format/media

The risk of outdated format/media pertaining to magnetic or digital records is a factor often mentioned in data rescue discussions. It would seem to be the second-most data rescue reason/trigger, after the risk of data in paper format. The essence of this risk factor can be summarised as follows: digitised data in a primitive digital format, as well as data on magnetic tapes, are often at risk due to the dangers of format obsolescence, media degradation and media incompatibility.

The studies and reports of the WMO (2014), Wippich when discussing the USGS (2012), Levitus (2012), Murillo (2014), Wyborn *et al*. (2015), Muller (2015a) and the Research Data Alliance (2019) all mentioned the problem of data being inaccessible due to being in an outdated format, a difficult-to-read format, or in a format not interoperable with current formats. In addition, Mayernik (2017: 4) mentioned the issue of data readers, in that the instrument required to read the data might no longer be usable, repairable or available.

Apart from risk factors relating to accessibility and interoperability, microfilm and microfiche formats may be subject to damage due to heat and humidity (Fry, 2010a). This aspect was supported by Muller (2015a: 4), who stated that the combination of time and fragile storage media can result in decay.

The rescue of Nimbus satellite data, older than 40 years and in an outdated film and tape format, is a good example of this risk factor being recognised and addressed (Gallaher *et al*., 2015). Despite the rescue project being described as difficult and time consuming (2015: 124), the fact that the Nimbus data were collected long before any other satellite observations underlined the importance of the rescue venture. The recovered data would be of use for studies of tropical and mid-latitude weather and cloud variations, and lead to an improved understanding of the US hurricane and typhoon season.

2.3.5 Substandard quality

As stated by Brunet (2011: 30), even if digitised data are available and accessible, it might not reach required standards of quality and homogeneity. The thought was expressed that it is doubtful whether data, described as sub-standard, can be used when, for instance, undertaking climate analysis, applications or services. The viewpoint by Mavraki *et al*. supports this idea, and mention was made of inconsistent data, or data showcasing obvious errors, as contributing towards data use being especially challenging (2016: 28).

2.3.6 Missing/displaced data

Data can be at risk when the data are hidden away in a researcher's desk (Wyborn *et al*., 2015: 106), or when researchers retire or pass away without leaving accessible documentation describing the dataset, or fail to leave instructions on locating and using the data (Levitus, 2012: 48; Wyborn *et al*., 2015: 106).

2.3.7 Perceived data value

Data thought to be of little or no value are at risk of being discarded, deleted or destroyed. This risk factor is applicable to all data formats and is mentioned in several data rescue studies. One such example is Nordling (2010), who refers to researcher oblivion regarding data value, and how this can lead to the destruction of data. Another fitting instance of the outcome of data not being cherished, or seen as valuable, is Hachileka's account of a chief Zambian meteorologist learning that a colleague had set fire to several paper-based weather data collections to create more office space (2015).

Griffin also stated that data, both digital and analogue, that are not accessed often tend to be seen as unwanted, and risk being thrown away (2015: 93).

2.3.8 Lack of awareness

Data are at risk when data owners or custodians fail to realise that data could be at risk, or that their data habits are sub-standard. Schumacher and VandeCreek found a strong correlation between faculty members' digital data loss, and their data management practices (2015: 96), and stated that university researchers producing digital objects can help themselves by becoming aware that these materials are indeed subject to loss.

In a similar vein was the statement by Griffin (2015: 93) that lack of planning, and low or non-existing levels of maintenance can be classified as data risk factors.

2.3.9 Changing priorities

It is a reality that research institutes and scientific centres are often subject to budget cuts, restructuring, and downsizing. As a result of these changes there might be changes in priorities, and the possibility of assigning a less vital role to data rescue and data preservation activities. This risk factor was mentioned by Muller (2015a: 4), Arrouays *et al.* (2017), Gaudin (2017) and McGovern (2017), to name a few studies.

In addition, an archive that is regarded as no longer being sustainable, leading to data eventually being inaccessible or destroyed, belongs in this risk category (Downs & Chen, 2017: 263).

2.3.10 Metadata/documentation issues

Data having inadequate or missing accompanying documentation, missing metadata, or sub-standard metadata is a risk factor often mentioned in data rescue studies. Studies and documentation by Levitus (2012: 64), Murillo (2014: 212), Thompson, Davenport Robertson and Greenberg (2014: 843), Wyborn *et al*. (2015: 106) and Mavraki *et al.* (2016: 28) underlined the importance of metadata being vital to data comprehension and use. Wyborn *et al*. (2015: 106) were of the opinion that long-tail data especially are subject to this risk, as these datasets often have less established documentation standards and formats. Additionally, supporting information of data collected in laboratories have most likely been recorded in analogue laboratory notebooks.

Furthermore, Thompson's account of Canadian health data rescue illustrates the importance of documentation, and how lack of supplementary information can render data useless (2017: 34). Health data files, despite being in digital format and not subject to concerns about outdated media format, were released without any accompanying documentation, and therefore difficult or impossible to understand. Thompson also stated that the risk is amplified when data without documentation is separated from the data creator, as it makes the tracking down of contextual information impossible (2017: 34).

Although inadequate/missing documentation is a risk factor in itself, Murillo stated that this factor could have as a ramification the inability of being ingested and managed by databases (2014: 207).

The term 'metadata archaeology' was encountered during the perusal of data rescue studies. Hills (2016: 20) described it as an activity often performed during data rescue, when the metadata are neither sufficient nor reliable. The activity of figuring out the missing metadata is likened to an archaeological venture.

2.3.11 Government funding/administrative policy

'Administrative policy' is somewhat similar to 'Changing priorities' (see Section 2.3.9) as a risk factor and is especially pertinent to the Data Refuge initiative. This factor is rooted in the perceived danger of US federal/government data deletion, due to administrative changes associated with the 2016 election of Donald Trump as the 45th president of the USA. As stated on a webpage of the University of Nevada's libraries' event page, federal data availability is particularly vulnerable to 'changes in government funding and administrative policy' (University of Nevada, 2017). Adding to this is the view of Gaudin (2017), who stated that even though there is data loss with any administration, there was more concern since Donald Trump took office. Reasons for the concern, according to Gaudin (2017) include the fact that his administration stated that it doubted the reality of climate change and proposed cuts to the budgets of the EPA[6] as well as the NOAA[7]. Gaudin stated that due to the aforementioned factors, there were fears that federal data would be intentionally lost or altered. Gaudin further quoted Margaret Janz, founder of the Data Refuge initiative, who feared that data might have been taken offline and only been available to the public as Freedom of Information Act requests (2017).

These stated fears appeared not to have been unfounded; Endangered Data Week (2019) reported that in the first few weeks of the Trump administration, the EPA was ordered to remove climate

[6] Environmental Protection Agency, US (https://www.epa.gov/)
[7] National Oceanic and Atmospheric Administration, US (https://www.noaa.gov/)

change information from their website, animal welfare data were removed from the USDA[8] website, and a bill was passed which would exclude changes to the Affordable Care Act from mandatory long-term analysis.

2.3.12 Non-digital formats

Studies have stated that data not being in a digital format is a risk factor. An example of this viewpoint includes Levitus (2012: 46), who stated that many historical oceanographic vertical profile and plankton datasets are not available to the international community as they exist only in manuscript form. Furthermore, Thompson, Davenport Robertson and Greenberg (2014: 843) mentioned that the inaccessibility and resulting non-use of non-digital photographic plates and plant specimens can result in them being ignored, and eventually destroyed.

The deduction can be made that there is a high probability that scientists will be oblivious to the existence of many older and valuable non-digitised datasets, and that this lack of knowledge puts the data at risk.

2.3.13 Archiving/preservation policy

As stated by Downs and Chen (2017: 263), an existing archive that is no longer sustainable will result in the archived data being at risk. Data Refuge founder Margaret Janz, as quoted by Gaudin (2017), mentioned the importance of having trustworthy data, and of the inherent risk of an archiving policy not requiring multiple copies. In addition, low levels of data and archive maintenance should be seen as a risk factor (Griffin, 2015: 93).

2.3.14 Data in one location only

Janz, a co-founder of the Data Refuge initiative, is quoted as stating that the movement's goal is to make trustworthy copies of data so it will be available to the public and suitable for research (Gaudin, 2017). Gaudin (2017) also reported that Janz was of the opinion that (federal) data should never have been in just one place. Murillo's viewpoint, namely that data can be viewed as at risk when the data are not backed up, supports this stance (2014: 207).

Moreover, the danger of storing a single set of unique records is mentioned by the WMO (2014), while Guest Blogger's post on the Sunlight Foundation[9] website claims that data are less vulnerable when stored in multiple locations (2017). The post also mentioned that it is easy to limit or block access to

[8] United States Department of Agriculture (https://www.usda.gov/)

[9] The Sunlight Foundation is a US-based, nonpartisan non-profit organisation that advocates for open government (https://sunlightfoundation.com/)

public federal data when data are in one location, and that a true library adage supports the LOCKSS idea: 'Lots of Copies Keeps Stuff Safe'.

2.3.15 Sub-standard data management practices

Data management practices that do not adhere to a high standard will lead to data being at risk. The WMO mentioned the reality of archive deterioration due to unmanaged use (2014), while Levitus stated the danger of data neglect (2012). In addition, data are at risk when not kept properly (Murillo, 2014), not properly inventoried or archived (Thompson, Davenport Robertson & Greenberg, 2014), or when there is lack of attention to the data (Muller, 2015a: 4).

2.3.16 Non-trustworthy copy

It is clear that a non-trustworthy copy of a dataset is indicative of risky data archiving practices. The extent of such actions can be far-reaching, and a post on the Sunlight Foundation's website warns against such a potential disaster when stating that one can imagine the consequences when a faulty copy is created, either by accident, technical error, or deliberate action, and there is proliferation of the faulty copy (Guest Blogger, 2017). Such instances can contribute to the epidemic of fake data, and the issue can be compounded as the data often resemble trustworthy records. In such situations, the opposite of the LOCKSS principle might well be achieved: instead of 'stuff being kept 'safe', a large quantity of bad copies is spread instead.

2.3.17 Additional factors contributing to data at risk

Downs (2018) has compiled a list of additional factors that were observed for using earth science data and related research data products and services. These factors include:

- legality (whether permission for data use has been obtained),
- integrity (whether the data are complete and correct),
- interpretability (whether the intended audience can understand the data),
- accessibility (whether the intended users can view the data),
- discoverability (whether the intended users can discover the data),
- security (whether the data are protected from loss, theft, tampering),
- confidentiality (whether confidential parts of the data have been identified, and access prevented),
- recoverability (whether data will be available in future or in case of a disaster), and
- sustainability (whether data risks in future can be addressed).

Downs also described ways in which many of these factors could be mitigated. Examples of mitigating steps are described below.

- Encouraging and using open licenses and permissions statements that are attached to data and data products will reduce confusion about intellectual property rights. The aspect therefore addresses the risk pertaining to legality.
- A template that contains documentation procedures for data producers and distributors to follow for the description of data products can mitigate interpretability risks.
- Discoverability risks can be mitigated by establishing a unique persistent identifier to provide an access point for each dataset that has been disseminated.
- Establishing designated safe locations (offsite and onsite) for long-term storage of any media containing data can address security risks.

2.4 Data assessment and appraisal

The previous section discussed factors leading to data at risk and were listed in no particular order. Probing the literature related to data at risk has shown that several methods are employed by entities when assessing whether data are at risk, and whether proceeding with data rescue activities is in the best interest of the particular institute. These assessment steps are especially pertinent when there are budgetary constraints or a lack of resources, including lack of manpower, skills, and the required data rescue tools and equipment. Several of the data at risk assessment steps are described below.

Ritchey (2017: 2–3) mentioned that it is not possible to save all data, resulting in the need to determine how to prioritise data rescue efforts. A risk matrix can be of value during risk assessment and is used to define the level of risk by considering probability, or likelihood, against consequence severity. Ritchey (2017: 6) stated that the following aspects are typically included in a risk matrix:

- the value of a particular dataset,
- the impact of losing the dataset,
- the effort/cost to reproduce the dataset, and
- the effort to rescue the dataset.

The Digital Preservation Coalition’s Bit List of Digitally Endangered Species (2018), while an advocacy tool, can be regarded as a tool to classify risk to digital objects. The coalition placed items in one of the following classes:

- minimal risk,
- vulnerable,

- endangered,
- critically endangered,
- practically extinct, and
- of concern.

The USGS makes use of a scoring system when determining which legacy products[10] are at greatest risk of loss/destruction. In addition, this system is used to establish the products having the biggest potential impact on the USGS and the science community (USGS, 2019b). A first step entails each legacy product being scored on the following aspects:

1. geospatial extent,
2. temporal extent,
3. species extent,
4. significance to USGS efforts/themes, and
5. risks related to the loss or destruction of the legacy product.

A next step in the USGS scoring system requires using the five scores to generate three calculated scores that are used to create greater differentiation between submissions. An example of one such score is the risk weighted average, which is the sum of the geospatial, temporal and taxonomic extents, and the significance score, multiplied by the risk score, weighting the score towards records that have more significant risk factors. Thirdly, the previous calculated scores are added together, resulting in what is termed a sum calculation score. It is interesting to note that all USGS reports forming part of the online Legacy Data and Information Reporting System allow sorting by all calculated metrics, thereby paving the way towards flexibility when determining priorities.

Shiue, Clarke and Fenlon (2020) and also Shiue *et al*. (2021) have identified a list of 18 assessment factors after conducting three case studies to investigate potential issues of curating collections with the purpose of data recovery and reuse. These factors are listed below.

- Extent: According to Hoffman *et al.* (2020: 21), this aspect refers to the size of the collection, and the extent to which the data are already documented, organised, digitised and curated.
- Reuse value: According to Hoffman *et al.* (2020: 21), this aspect refers to the intended, demonstrated, anticipated or plausible reuse opportunities for the collection. The novel uses of the data is the crucial issue to be considered under this aspect.

[10] According to the USGS: Unreleased data, software, physical samples, and publications developed by USGS projects *completed* prior to 2011 that are stored in an old/obsolete format, computer system or facility and are, therefore, unavailable to the scientific community and greater public

- Historical value: According to Hoffman *et al.* (2020: 22), this aspect refers to the important or noteworthy scientific approaches, results, or advances linked to the data.
- User communities: According to Hoffman *et al.* (2020: 21), this aspect refers to the groups of potential users who should be able to understand and use the data.
- Stakeholders: According to Hoffman *et al.* (2020: 21), this aspect refers to groups, institutions or communities (as opposed to direct users) who have or could have an ongoing interest in the data.
- Reuse objects: Hoffman *et al.* (2020: 22) described this aspect as any specific components of the collection that carry reuse opportunities.
- Historical objects: According to Hoffman *et al.* (2020: 22), this aspect refers to components of the collection that carry historical value or can be used as potential evidence for the history of science.
- Fit for purpose: Hoffman *et al.* (2020: 23) described this aspect as the extent to which the data are ready or suitable for actual or potential uses identified in reuse value, historical value and reproducibility. It involves determining how much additional documentation, interpretation and processing are required to prepare data either for reuse, or to serve as historical evidence.
- Data objects: According to Hoffman *et al.* (2020: 21), this aspect refers to the kind of data in the collection, the file formats used, and whether the data stand alone or are embedded in other documents.
- Obstacles for recovery: Hoffman *et al*. (2020: 23–24) described this aspect as the anticipated or observed obstacles to recovering data from the collection.
- Associated publications: According to Hoffman *et al.* (2020: 23), this aspect refers to the existence of identifiable publications associated with the data, such as scientific journal articles that report, rely on, or cite the data.
- Relevant collections: Hoffman *et al.* (2020: 23) described this aspect as referring to the existence of other collections of research materials that are relevant to the data, and which demonstrate a wider network of interest or investment in the research documented.
- Completeness: According to Hoffman *et al.* (2020: 22), this aspect refers to the completeness of the data, and establishing whether there are gaps in the collection that would limit either use or historical value.
- Sensitivity: Hoffman *et al.* (2020: 22) described this factor as aspects of the data that may be considered sensitive to unintended or undesirable access, use or interpretations, whether from the standpoint of privacy, ethics, security, or scientific accuracy.

- Access and use restrictions: Hoffman *et al.* (2020: 22) described this aspect as the constraints that will be placed on access to and use of the data, such as intellectual property constraints or factors in sensitivity that affect how the data should be made accessible.
- Priorities: According to Hoffman *et al.* (2020: 24), this aspect refers to the most immediate priorities for data recovery, as opposed to the optimal or long-term objectives of recovery.
- Reproducibility: Hoffman *et al.* (2020: 23) stated that this aspect refers to the ways, if any, of the data being reproducible.
- Rarity and uniqueness: Hoffman *et al.* (2020: 23) described this factor as referring to any part of the data duplicated elsewhere, or actively stewarded, curated or maintained by another other group or institution. This factor may also be used to address other, distinctive strands of rarity: whether the data are fundamentally irreplaceable, or whether aspects thereof could be recreated.

The assessment aspects described in this section provide a glimpse of the various methods used by different entities to establish the value of data, and whether a data rescue project is in the best interest of the organisation. As stated by Shiue *et al*. (2021), there is no one-size-fits-all solution to process and appraise data-rich collections. It would be prudent of institutes contemplating data rescue projects to establish an institute-specific assessment framework to assist in determining the feasibility and need of the rescue of specific data deemed to be at risk.

2.5 Rationale for data rescue

While the obvious benefit of data rescue is ensuring that data are digitised, safe and accessible, having a closer look at the advantages gained by data rescue activities, as well as the knowledge gained, will underline the importance of rescue efforts.

As stated on the charter of the RDA's Data Rescue Interest Group (2015: 1), the relevance and importance of heritage scientific data are most aptly seen in terms of evidence of change in the natural world, and supported by the prevalence of environmental data rescue within this section of the chapter. While other disciplines are also included – an example being Curry's article (2011) on the outcomes of physics data rescue – the benefits and outcomes of data rescue are mostly discussed and reported on in the environmental sciences, especially in the areas related to climatology.

2.5.1 Data rescue provides access to data

This outcome is an obvious yet important benefit of data rescue initiatives. Data are often rescued for the simple reason that their current format makes the data inaccessible to researchers. For example, formats such as the paper-based data hidden in a researcher's office, or the magnetic data held in an archive that no longer has the data reader, or a ship's logs and diaries housed in a museum are not conducive to the data being traced, accessed or shared. Rescuing these datasets, by way of digitising, adding metadata, and depositing it in a free open access data repository, will ensure that the data are available for use by researchers and other parties.

An excellent example of a rescue project providing access to data are the deliverables produced by GODAR, the international Global Oceanographic Data Archaeology and Rescue Project. It was initiated in 1993, and described by the Intergovernmental Oceanographic Commission (IOC) as striving towards an increase in the volume of historical oceanographic data available to researchers, by locating data not yet in digital form, digitising these datasets, and submitting them to national data centres (IOC, 2003; International Oceanographic Data and Information Exchange, 2013). As stated by Griffin (2015: 94), the campaign focused on locating historical oceanic measurements in any storage medium, whether paper manuals, microfiche, early electronic media of non-contemporary readability, or computer-ready files. Griffin further stated that the data were transferred into science-ready electronic files through various steps of digitisation and quality control (2015: 94). Caldwell (2003: 1) stated that the GODAR Project, initiated by the National Oceanographic Data Centre (NODC) and World Data Centre, had by 2003 already increased records by over three million historical ocean temperature profiles, 140 000 chlorophyll profiles, and 1.4 million plankton observations. Griffin adds on to this by stating that the GODAR Project was able to triple the number of ocean profile datatypes recorded in pre-1992 years (2015: 94).

The importance of this data rescue benefit is supported by the likes of Wippich (2012: 1), who stated that data rescue activities provide access to datasets that are both valuable and critical. Mention was also made of the rescued geological data then being available in the USA as well as the international science community. Adding to this is the European Space Agency (ESA) reporting that data rescue activities have resulted in centuries-old religious texts now being freely available online through the Vatican Digital Library (2018). These two vastly different disciplines – geology and religious history, both engaging in data rescue initiatives – emphasise this primary benefit mentioned by the RDA's Data Rescue Interest Group, namely that data rescue is required, as data can often not be used in research until the data are digitised and can be accessed electronically (2019).

2.5.2 Historical data are essential in many fields of science

In many disciplines the access to historical data is essential to research and discovery. A prime example of this requirement entails climate research, where the unique timestamp of historical data is critical for the quantification of changes and trends. Bradshaw, Rickards and Aarup (2015: 9) stated that at least 60 years of data is the recommended timeline when long-term trends are being analysed. The

RDA Data Rescue Interest Group stated that older data are needed when studying environmental changes, and that this historical data are critical when differentiating between natural changes or changes due to anthropogenic reasons (2019).

Levitus (2012: 46) agreed with these viewpoints, with mention being made of the international scientific community, in this instance oceanographic and climate scientists, requiring access to the most complete electronic databases when advising national and international bodies on issues such as climate change. The need for environmental data to be available in perpetuity is also vital for studies investigating concepts such as global change, global warming, and fisheries research.

2.5.3 A better understanding of the past

This factor is especially pertinent to historical environmental data, and a reason environmental data from the pre-digital era feature commonly in data rescue studies.

Caldwell (2003: 1) and ACRE (2019) stated that historical environmental data hold clues to past environmental conditions and are crucial to our understanding of past weather phenomena. Historic sea-level data can, for instance, aid in the understanding of decadal variations and climate change (Caldwell, 2003: 6). These tide-gauge datasets are of value to the oceanographic community in that the datasets assist in our understanding of the time scales of sea-level change, and in particular sea-level rise associated with climate change (Caldwell, 2012: 1). Mavraki *et al*. (2016: 29) support this viewpoint pertaining to environmental understanding by mentioning that historical data provide valuable insights into ecosystems of the past.

Agreeing with the opinions above was the Data Rescue division of the Copernicus Climate Change Service (C3S), which asserted that an improved estimation of climate variability, and better detection of climate trends are obtained when historic data are consulted (2018).

2.5.4 Indicators to future conditions and climate change

Rescued historical environmental data provide a fascinating glimpse of the living world of the past, but can also be used to provide clues to future conditions. This benefit is especially relevant in climate studies; an apt summary of this data rescue benefit is the statement by Caldwell that historical environmental data help in understanding climate change and related variations (2012: 2).

This viewpoint is present in many studies, reports and publications discussing historical data and future conditions. Examples include ACRE's statement that historic data can aid in the prediction of future climate change (ACRE, 2019), and C3S (2018) stating that a data time series of 30 years or more is needed for developing seasonal forecasts, and much longer for decadal forecasts. According to

Brunet and Jones (2011:30), high-quality, high-resolution and long-term climate data are of paramount importance for more confidently undertaking a wide range of climate assessments, services and applications. It is further mentioned that the data can be used for generating climate change scenarios.

The benefit of older environmental data is mentioned by Antico, Aguiar and Amsler (2018: 1368), who stated that preserving historical Parana hydrometric data is required for understanding South American and global hydroclimate changes. In addition, the WMO considered the need for long hydrological records for purposes of water planning, as well as flood estimation (2014: 1). The WMO expanded on this by describing the recent need for long-term datasets to understand the response of hydrological regimes to climatic variations and anthropogenic influences. Extended data are regarded by the WMO to help in identifying national as well as regional hydrological trends and changes, due to being a long-term record that captured the natural variability in hydrological systems.

Agreeing with the opinions above is the Data Rescue division of the Copernicus Climate Change Service, which mentioned the benefit of a better estimation of climate variability, and better detection of climate trends being possible when historic data are consulted (CS3, 2015). Maximising the availability of digitised historical data and metadata, according to Page *et al*. (2004: 1483), is important for long-term climate monitoring, and is especially vital when aspiring to analyse trends accompanying the occurrence of extreme events.

2.5.5 Assists in understanding current conditions

Previous benefit categories have discussed the use of historical data to gain a better understanding of the past environment, as well as future environmental conditions. Adding to this is the use of historical environmental data to assist in comprehending current environmental conditions. ACRE's view that historic data can help us understand today's weather and climate (2019), as well as Cooper's statement on 19th century weather data allowing a greater understanding of climate (2015), support this benefit of data rescue.

2.5.6 Extends the knowledge of the subject

Data rescue and access to the rescued data will result in an increase in scientists' subject knowledge. The benefit was mentioned by Gallaher *et al.* (2015: 124), who stated that access to early climate data extends the climate record and provides important context regarding the climate change record. Wood *et al.* (2012) referred to a similar benefit with rescued habitat data and explained how the rescued data will provide novel information on the impact of land use as a driver of ecological variation in plant species composition and soils. The authors also stated that habitat data dating from three decades ago make it possible to characterise the mosaic of habitats as they existed before.

Adding to the above was the statement by Knockaert *et al.* (2019: 1) pertaining to the recovery of data in theses and reports, with these outputs containing data resulting from marine and coastal research activities in the Eastern African region conducted between 1984 and 1999. The authors stated that the recovered dataset provides a better insight into the different types of research conducted between 1985 and 1996 involving the Kenya–Belgium cooperation in marine sciences (KBP) project.

2.5.7 Increases scientific accuracy

Using historical data to make science more accurate is a factor especially pertinent to climate science. This benefit was mentioned by several authors (Brohan, 2009; Jones, 2008; Hawkins *et al*., 2013), underlining the fact that the precision and validity of climate models, forecasts and analyses are improved when there is access to historical data. According to Eveleth (2014), the International Environmental Data Rescue Organization (IEDRO) estimated that there are 100 million paper strip charts – i.e. records that list weather conditions – sitting in meteorological storage facilities throughout the world. This unused data amount to approximately 200 million research observations, access to which could lead to a substantial improvement in climate models. Eveleth further mentioned that climate researchers ideally require global records reaching back hundreds of years to produce reliable models (2014).

Brunet and Jones (2011: 29) agree with the above and stated that lack of access to historical data hampers environmental researchers' ability to carry out more robust assessments of the climate. These assessments are in turn required to better understand, detect and respond to global climate variability and change.

The benefit of access to historical data leading to improved understanding of the environment is supported by Hachileka (2015), who mentioned that historical data allow the establishment of trends, aiding in a better understanding of climate, and leading to the development of climate models, satellite-based instruments, and seasonal forecasts. The fact that the effect of historical data even extends to providing foundational data for adaptation studies at local, national and regional scales is a further outcome touched on by Hachileka (2015). He also stated that the implementation of historic climate data can lead to the building of societal resilience against the effects of climate change (Ibid.).

As stated by Hachileka (2015) and Mukhondia, Swaswa and Bojang (2017: 1), a knock-on effect of a better understanding of the discipline of climate science is an enhancement of climate information services. Such an improvement would lead to better forecasting and earlier warnings of natural disasters or certain catastrophic events. In addition, IEDRO's presentation on their data rescue activities (2010: 8–13) stated that a better understanding of weather and climate patterns applied pro-actively can result in a more climatically secure society. The rescued data may be applied to contribute towards:

- starvation prevention,
- disease prevention,
- safer construction,
- better flood forecasting,
- better weather forecasts, and
- improved climate change prediction.

Besides the climatic benefits discussed above, rescued data and documentation in the field of wildlife diseases not only provide an early warning system for human or domestic animal disease outbreaks, but also help identify the local, regional and national level of environmental health (Wippich, 2012: 2).

Another example of the reuse of endangered data was stated by Rudin *et al.* (2014) to be the use of historic records for predicting events in electric grid systems. They further mentioned that these predictions would not have been possible without access to long-term data. Krotz (2011) using historic punch cards for the longest analysis of cholesterol and heart disease, Griffin discussing the use of historic stellar spectra for determining telluric O3 column densities (2005), and Axelsson, Östlund and Hellberg (2002) describing the use of historic data in a retroactive gap analysis to analyse changes in forest conditions in Sweden, support the idea that historic data increase a discipline's accuracy.

2.5.8 Extends the coverage in data repositories

Data rescue efforts and initiatives provide as tangible benefit the expansion of a discipline's data holdings, and topics covered. An example of such an outcome is the extension of temporal and spatial coverage of national and international sea-level data repository holdings after tide gauge data had been rescued (Caldwell, 2003: 6).

The University of Hawaii's Sea Level Center (2018) reported that their rescue programme resulted in the extension of 91 hourly tide gauge series backwards in time for 1 411 years. Subsequently, the quality-controlled versions of rescued data have been ingested into GLOSS data centres, thereby furthering the inventory of historic, non-digital tide gauge data.

2.5.9 Cost-effectiveness

Data rescue activities, while requiring resources in terms of time, manpower and costs, are invariably less expensive resource-wise than regenerating the data. This reality is supported by Diviacco *et al.* (2015: 44), who claimed that similar measurements would be difficult, and often impossible to acquire. They explained further that the recollecting of data often involve logistical difficulties such as encountering problems when acquiring an exploration permit, or not being given permission to use dynamite. For these types of data collecting activities, access to the historical data is the cheaper – and often only – approach to follow. Even with a permit, the expense of recollecting such data would be immense.

Schmidt (2017) supported this notion of data rescue being time-efficient by stating that while researchers can gather data about the planet at the rate of one year per year, data rescue activities can add data much faster. The WMO contributed towards this topic by declaring that data rescue provides an avenue whereby hydrological records can be extended at a fraction of the cost of repeating the historical experiments (2014: 3).

2.5.10 Often the only record of a particular phenomenon

The uniqueness and irreplaceability of historical datasets is a major contributing factor or trigger for many data rescue initiatives. Due to their unique timestamp, historical environmental datasets cannot be recollected. When only one copy of an irreplaceable dataset is available, it heightens the risk factor associated with the data. More details on this are listed in Section 2.3: Data risk factors.

Studies supporting this concept include the USGS fact sheet by Wippich (2012: 1), which states that historical environmental data represent observations and events that can never be repeated. Examples of irreplaceable environmental data include the USGS data on volcanoes, which contain a unique timeline of Alaskan volcanoes as well as views of the changing and dynamic landscape captured over decades. Wippich mentioned the invaluableness of a USGS bird phenology dataset, and noted the uniqueness of the dataset’s coverage, temporal breadth, comprehensiveness, and comparative value to modern measures of climate change (2012: 4). Wippich further stated that the data’s rescue, digitisation, and distribution provide unprecedented insight into the migratory behaviours of birds, including the response of the species to climate change.

Besides the often-mentioned uniqueness of historical environmental data in literature, there have also been instances of data in other disciplines described as irreplaceable. An example is Curry (2011: 694) discussing the usefulness of the Jade project’s data, the fact that many past physics experiments are unique in that they will not be repeated under the same low-level energy conditions, and how modern energy-level physics experiments cannot replace the historical measurements.

Diviacco *et al.* mentioned that it would be near impossible to recollect vintage seismic data, as many of the seismic areas are currently subject to restrictions in obtaining exploration permits (2015: 44). Adding to this is that many of the historic seismic datasets cover large areas and would be expensive to generate today. A factor making these datasets even more valuable is that the acquisition of data was done via dynamite blasting; hence, permission for use of such a high energy source would currently be difficult. As a result of these limitations, the recovery of such datasets is stated by Diviacco *et al.* to be vital to the scientific as well as commercial communities (2015: 44).

In agreement with the above opinions regarding the value of irreplaceable historic datasets is the description by Levitus (2012: 46) of the goals of the GODAR Project, which is to ensure that all oceanographic data available for international exchange are archived at two or more international data centres in electronic form.

2.5.11 Creation of new theories, publications, projects and products

With data rescue efforts making previously inaccessible and unused data available to scientists, it follows that the research efforts making use of this data will result in new publications, new projects and new products. Examples of such outcomes are listed below.

- Data rescue can lead to scholarly research articles discussing the data rescue initiative, such as the publications by Brunet and Jones (2011), Diviacco *et al.* (2015), and Gallaher *et al.* (2015).
- Researchers can make use of the resurrected data and produce high-impact discipline-specific scientific articles which would not have been possible pre-data rescue (Curry, 2011: 694).
- The creation of new maps is made possible through data rescue; an example of this is Arrouays *et al.* mentioning the creation of new soil property maps according to GlobalSoilMap specifications (2017: 1), and the formation of regional atlases (Levitus, 2012: 46).
- Data rescue can lead to new and fascinating projects (Muller, 2015b: 23).
- Data rescue can result in the construction of traditional climate data products (Thorne *et al.*, 2011).
- Data rescue can lead to the production of long-term climatic reanalysis products (Thorne *et al.*, 2011).
- The construction of normals and extremes climate indices for use by science, industry and society are possible following data rescue projects (Thorne *et al.*, 2011).
- Data rescue can result in the creation of derived products such as heat stress/thermal comfort [sic] (Thorne *et al.*, 2011).

- Curry (2011: 694) mentioned that studying rescued data can confirm theories; for example, the use of rescued physics data confirming quantum chromodynamics played a vital role in the awarding of the 2004 Nobel Prize in physics to Gross, Politzer and Wilczek.
- Wippich (2012: 8) stated that the use of rescued data can lead to new insights and interpretations that never formed part of the intentions of the original data collectors.
- Curry (2011: 694) mentioned that by reanalysing rescued particle physics data, a research team was able to publish more than a dozen high-impact scientific publications. Griffin (2015: 95) stated that the results of data rescue efforts should be published in both domain-specific journals and popular media. Use of the latter will ensure that information about data rescue reaches the wider community and is likely to encourage and inspire other data rescue endeavours. Griffin further stated that sharing rescue stories can capture the imaginations of eager rescue volunteers, while the sharing of experiences can lead to the exchange of information on methodologies, hardware and software (2015: 95). Moreover, Griffin mentioned that such sharing will result in greater output at a smaller cost (2015: 95).
- IEDRO projects are not regarded as complete until the rescued data are used to inform something such as a weather model, disaster recommendations, or improved city planning (Eveleth, 2014).
- The ACRE team, as mentioned by Eveleth (2014), plugs recovered climate data into current weather models to create pictures of what the global climate was like a century ago.
- Rescued data can contribute towards new tools/programmes; an example of this was Rountree *et al.* (2002: 242) mentioning the use of fish sounds as a tool to study temporal and spatial distribution patterns, habitat use, and spawning, feeding and predator avoidance behaviours of fishes. The rescued fish audio data can also be implemented to develop programmes using fish sounds as a tool to map essential fish habitats of soniferous species.
- Levitus (2012: 46) stated that rescued data leading to a better understanding of fisheries variability can lead to the enhanced management of fisheries and other marine resources.
- Griffin (2015: 95) stated that the sharing of rescued climate data can inspire a whole portfolio of innovative ideas and new concepts.
- Other data rescue initiatives can transpire: Mattson and Schell (2018) reported on efforts that have transpired since data rescue had taken place, and mention Data Together[11], an ongoing collaborative open-source data stewardship group. Another initiative is the Data Refuge Storybank[12], which is a collection of the ways in which data exist in the world, and how it connects people, places and non-human species.

[11] A new model for distributed community-driven stewardship of data (https://Datatogether.org)

[12] Stories.datarefuge.org

- Thorne *et al.* (2011) referred to the myriad of scientific contexts in which historic data can be applied once converted to digital format and made available through one or more recognised repositories.
- The rescued Montevergine Observatory meteorological data collected in the Southern Apennines were anticipated to be used by the scientific community for a range of meteorological and climatological purposes (Capozzi *et al.,* 2020: 1482–1483). The data had peculiar features uncommon to long and old climatological time series; its high time resolution of the weather observations, the variety of recorded meteorological parameters, and the uniqueness of the geographical context are three interesting attributes. The reason for the profusion of applications lies in the data encompassing many fields and studies, and in it having links to the socioeconomic impacts of climate change, hydrology, and agricultural planning.

2.5.12 Infrastructure

Enhanced infrastructure, emanating from rescue projects funded by IEDRO, is a potential benefit of funded rescue projects. As reported by Eveleth, IEDRO sets up weather stations with their own scanners in the countries where data rescue is carried out (2014). Once the project has been completed, the scanners and other equipment are donated to the local weather station.

2.5.13 Job creation

Data rescue initiatives, such as the IEDRO projects where local people are hired to do digitising, can assist with job creation (Eveleth, 2014).

2.5.14 Community involvement

Both IEDRO (2014) and ACRE (2016) make use of crowdsourcing to find volunteers who would participate in data input activities. A website such as Old Weather[13] exemplifies the way in which citizen science is used during data rescue, as it enlists members of the public to assist with the digitisation of US ship logbooks dating from the mid-19th century onwards. As stated by the RDA Data Rescue IG (2019), citizen science projects such as Old Weather, or projects involving the recovery of lost tapes from early space missions, also appeal strongly to the public.

2.5.15 Environmental literacy

An entry on the Sunlight Foundation's website stated that one of the aims of the Data Refuge movement is to advocate for environmental literacy with storytelling projects (Guest Blogger, 2017).

[13] Old Weather is an online weather data project that currently invites members of the public to assist in digitising weather observations (https://www.oldweather.org/)

The storytelling projects showcase how federal environmental data support health and safety in local communities, and the point that data funded by taxpayers should be public and visible is emphasised.

2.5.16 Provides impetus for other data recovery projects

In the field of space exploration, the spectacular success of the Lunar Orbiter and Nimbus data rescue efforts is spurring other satellite data recovery projects such as at ESA/ESRIN (Griffin, 2015: 95). Similarly, the extensive data recovery project involving data in theses and reports resulting from marine and coastal research activities in the Eastern African region conducted between 1984 and 1999 was expected to facilitate further coastal biodiversity research in Kenya (Knockaert *et al*., 2019).

2.5.17 Sharing of data knowledge and data rescue expertise

One of the benefits of the International Data Rescue Award in the Geosciences is the sharing of the ways in which research data are processed, stored and used (Elsevier, 2016). Contest participants are required to submit written descriptions of projects involving the recovery of data not initially accessible. Moreover, this award wishes to draw attention to the high importance of legacy data, and to stimulate the sharing of data rescue tools, standards and knowledge required for making data accessible and reusable.

The two open access guidelines published by the WMO on the rescue of hydrological data (WMO, 2014) and climate data (WMO, 2016) are good examples of knowledge, tools and expertise being shared with parties eager to undertake data rescue activities.

Lastly, data rescue efforts can result in the identification of improvements in data rescue workflows and establish tools and services to be of use with future projects (Hsu *et al*., 2013: 18).

2.5.18 Data provision to relevant organisations, initiatives or repositories

A beneficial outcome of many data rescue efforts entails the provision of rescued data to initiatives that will in turn reanalyse the data. Kaspar *et al.* (2015: 60) reported that the International Surface Temperature Initiative, the Atmospheric Circulation Reconstructions over the Earth (ACRE), the International Comprehensive Ocean-Atmosphere Data Set, and other European projects have all been recipients of digitised data. Additional examples of similar enterprises mentioned by Kaspar *et al.* include the provision of data of 13 land stations that have been gifted to the China Meteorological Administration, signal station data collected at the Polish coast transferred to the Institute of Meteorology and Water Management, and documents from former colonial stations in Cameroon handed to the Data Rescue @ Home project (2015). Brunet *et al.* (2020b: 6) mentioned the provision of rescued data to global data centres and repositories; the International Surface Pressure Databank, the Global Precipitation Climatology Centre, and the ECMWF's Mars Archive are examples of recipient repositories.

2.5.19 Data archive establishment

According to the WMO (2014), a comprehensive inventory of data holdings can be useful for demonstrating the value of a data archive and justifying the ongoing expense of running such an archive.

2.5.20 Breaking down silos

The Sunlight Foundation website, via an entry by Guest Blogger (2017), reported that one of the endeavours of the Data Refuge movement is the building of a consortium of research libraries. By breaking down silos, the consortium will scale Data Refuge's tools and practices and make it easier for other interested parties to make copies of other kinds of federal data beyond environmental data. The consortium, supported by the Association of Research Libraries, will be supplementing the existing system of federal depository libraries, and will actively copy public materials from federal agencies.

2.6 Data rescue projects: diversity and range

Investigating the literature around data rescue initiatives and activities exposes the reader to the diversity in the range of rescue tasks performed, scientific disciplines covered, data formats rescued, geographic areas targeted, and professions and vocations involved. This section will briefly discuss the miscellany within documented data rescue projects, and how it is impossible to describe a quintessential data rescue project, or state that certain characteristics always form part of data rescue.

2.6.1 Disciplines

This section consists of two parts: a section providing details regarding the range of research disciplines involved with data rescue, and a segment describing disciplinary differences during data rescue activities.

2.6.1.1 Range of research disciplines

Due to the unique timestamp associated with environmental data, most data rescue projects described in literature and other publications involve the discipline of the natural environment. Climate and weather data in particular are targeted for data rescue; the research of Page *et al.* (2004), Brunet and Jones (2011), Eveleth (2014), Cooper (2015), Ashcroft (2016), Allan and Willett (2017), Boehm *et al*. (2017), Raughley (2017), Brönnimann *et al*. (2018) and Park *et al.* (2018) are only a few of the publications detailing efforts to rescue climate and weather data.

In addition to historical data from the environmental sciences being frequently discussed in literature, the presence of environment-related data rescue organisations is also common. Several groups, consortia or organisations exist, and one of their objectives is to support environmental data rescue. IEDRO, involved in the rescue of at-risk climate data from around the globe, is one such group (IEDRO, 2014). Another group heavily invested in environmental data rescue is the ACRE initiative (ACRE, 2019). The ACRE initiative works to retrieve global climate history by harnessing the efforts of professionals and volunteers to recover historical terrestrial and marine weather observations (ACRE, 2020). The WMO is another global entity that assists with environmental data rescue via comprehensive data rescue guidelines for climate data (WMO, 2016) and hydrological data (WMO, 2014), respectively.

Examples of more groups or organisations involved with environmental data rescue are listed below.

- The data rescue portal of the Copernicus Climate Change Services/C3S[14] reveals a consortium providing a climate data rescue service, building upon current WMO data rescue activities.
- The IGAD Climate Prediction and Applications Centre (ICPAC)[15] provides data rescue support and backup services.
- CLIMARC is a jointly funded collaborative project involved with the digitisation of Australian Climate Archives[16].
- Deutscher Wetterdienst[17] is involved with the digitisation of historical weather data from overseas stations, ship logbooks, and signal stations.

Climate and weather data are by no means the only type of environmental data included in data rescue projects. A sample of other disciplines involved with data rescue, listed in alphabetical order, is as follows:

- astronomy (Griffin, 2006: 21),
- atmospheric sciences (Allan *et al.*, 2011),
- biodiversity (Griffin, 2006: 21; Güntsch *et al.*, 2012),
- biogeographic sciences (Mavraki *et al.*, 2016),
- earth science (Mayernik *et al.*, 2017: 1),
- geology, geosciences, and geomorphology (Wippich, 2012; Smith, 2015; Hills, 2016),

[14] https://data-rescue.copernicus-climate.eu/
[15] https://www.idare-portal.org/sites/default/files/COUNTRY%20PRESENTATIONS%20FROM%20THE%20DMs%20IGAD%20TRAINING.pdf
[16] http://www.bom.gov.au/climate/how/climarc.shtml
[17] https://www.dwd.de/EN/climate_environment/climatemonitoring/climatedatamanagement/datarescue/data_rescue_node.html

- global sea-level sciences (Caldwell, 2003),
- habitat science (Wood *et al.*, 2012),
- hydrological sciences (WMO, 2014),
- hydrometric sciences (PONS *et al.*, 2016; Antico, Aguiar & Amsler, 2018),
- marine biodiversity (Rountree *et al*., 2002; Mavaraki *et al*., 2016),
- oceanography (Levitus, 2012),
- satellite data (Knapp, Bates & Barkstrom, 2007; Gallaher *et al*., 2015; ESA, 2018),
- soil sciences (Arrouays *et al*., 2017), and
- tide gauge research (Caldwell, 2012; University of Hawaii, Sea Level Center, 2018).

The above list should not be viewed as a complete list of environmental data rescue topics, and serves simply to illustrate the range of environmental disciplines present in data rescue.

While the foregoing paragraphs have indicated the variety of environmental disciplines involved with data rescue, it should be noted that non-environmental data rescue initiatives have also been described. Examples of non-environmental data rescue projects include the following research fields:

- arts and music (Kijas, 2018; Cascone, 2022),
- cancer research (Murillo, 2014),
- health (Thompson, 2017),
- historical religious/Vatican research (ESA, 2018),
- particle physics (Curry, 2011),
- physical sample metadata rescue (Hills, 2015),
- physics education (Murillo, 2014),
- social sciences (Woolfrey, 2016; Khwela, 2018), and
- socioeconomic research (Downs & Chen, 2017).

These lists show that while the rescue of environmental data is the discipline featuring most prominently during data rescue efforts, it is not the only discipline with data at risk, or with historic data worth rescuing. A survey by Thompson, Davenport Robertson and Greenberg (2014: 847) revealed how researchers ranked the risk of data in different subject areas, with geology, climate and biology most mentioned as areas having data at risk. On the other side of the spectrum, thought to be least at risk, were data from the disciplines of psychology, demography and political science. In other words, while certain areas are certainly perceived to have more data at risk, studies show that disciplines other than the environmental sciences might also have data at risk.

2.6.1.2 Discipline-specific data rescue factors and activities

Disciplinary differences linked to data rescue are numerous; five data rescue categories and the manner in which different research disciplines are involved with each are briefly described in this section. Although differences are bound to exist in the themes of data formats, participants, sources of data and age of data, the study looks at contrasts found with regard to funding, repository selection, tools and platforms, metadata and other standards, and significance of the rescue.

Funding and support: An examination of published literature has shown that certain disciplines have access to a larger number of funding opportunities and resource assistance during rescue projects. The natural sciences, and climate and meteorological sciences in particular, are able to benefit from funding, support and assistance from external sources. Examples of ventures that were able to make use of funding include the rescue of pre-1850 instrumental meteorological data in the Netherlands, which was partly funded by the European Union (Brandsma, 2007), and the rescue and digitisation of climate records in the Mediterranean Basin, which benefited from funding advice by the WMO Resource Mobilization Office (Jones, 2008). IEDRO is a foremost provider of data rescue assistance and state on their website that during data rescue, an IEDRO volunteer assists the country's national meteorological service in readying the data rescue facility, then purchases and installs computer and camera equipment and trains the data rescue personnel (2014).

Examples of funder support in other disciplines have also been traced. The data rescue project involving threatened snow and ice data benefited from funding by the Council on Library and Information Resources and won the 2016 International Data Rescue Award (Maness *et al*., 2017). Wippich (2012) stated that geological data rescue projects in the United States are able to benefit from financial assistance provided by the Data Rescue Program of the US Geological Survey, receiving between 3 000 dollars and 65 000 dollars per project.

Page's statement made nearly two decades ago that lack of resources is a considerable challenge to digitising historical observations and maintaining observational networks and related infrastructure is still applicable today. In many cases, digitisation of records occurs on an ad hoc basis or is paid for by an external client.

Repository selection: Disciplinary differences regarding selection and use of repositories are evident in document literature, with most rescued data deposited in a disciplinary repository, and in certain instances a national disciplinary repository. Examples of different repositories used based on research discipline include the deposit of marine species data in OBIS Canada IPT and published to OBIS (Kennedy, 2017), agricultural data in Ag Data Commons (Clarke & Shiue, 2020), biodiversity data in the GBIF network (Global Biodiversity Information Facility, 2022) and seismic data uploaded to the Seismic data Network Access Point (SNAP).

Data rescue tools and platforms: A perusal of data rescue literature has shown that different disciplines have access to different data rescue tools, systems and platforms. An excellent example of a discipline-specific data rescue tool is the reBiND workflow for biodiversity data consisting of phases for data transformation into contemporary standards, data validation, storage in a native XML database, and data publishing in international biodiversity networks (Diviacco, 2015). NUNIEAU software was developed to assist with the rescue of hydrometry data (Pons, 2006), while the Integrated Publishing Toolkit (IPT) is a free, open-source software tool to publish and share biodiversity datasets through the GBIF network (Global Biodiversity Information Facility, 2022).

Conversely, the DRAW software used when rescuing historic weather data is open-source and also designed for easy repurposability to a variety of data rescue goals including other weather registers, medical records and student records (Slonosky, 2019).

Metadata and standardisation: Data at risk undergoing rescue need to be accompanied by metadata unique to the discipline, and must adhere to specific standardisation requirements. Rescued historic snow and ice data, for example, not only require the addition of glaciological and geospatial metadata, but also descriptions that may facilitate humanistic or social science research for instance (Maness *et al*., 2017). Some of the rescued photographic images have people in them and were taken using a variety of photographic techniques (e.g., silver gelatin or albumen). Photographic paper-type metadata are being added, and this opens the collection to photographic historians. Describing these aspects of the images may yield additional discoveries related to the history of the area or the research project.

Similar discipline-specific metadata requirements exist for legacy soil data; a rescue project documented by Brönnimann *et al*. (2018) reported that data had to be compiled and harmonised according to GlobalSoilMap specifications in a world-level database referred to as WoSIS. The rescue of Canadian marine species data needs to produce standardised datasets, and it is recommended that species or higher-level taxonomic names be mapped to the World Register of Marine Species (Kennedy, 2017). Additionally, the location where the taxa were collected or observed must be associated with latitude and longitude coordinates, and georeferencing procedures, although not compulsory, should be described in the dataset metadata.

Significance: The benefits and outcomes of rescued data can also be regarded to be discipline specific. While the rescue of historical climate data allow climatologists to establish long-term trends, which in turn helps them understand and better plan for future changes in climate (Hachileka, 2015), the rescue of the earliest satellite maps of Arctic and Antarctic sea ice shows how ice volumes have changed with time (Cowing, 2014). The rescue and publishing of historical South African apartheid era datasets by DataFirst enables the use of such information as an important policy research resource (2022), warrants the tracking of household livelihoods over time, provides insight into the social effects of removals on individuals and households, as well as the economic and political dimensions of these community upheavals.

The significance of rescued data in other disciplines includes the implementation of marine management and policy through access to rescued marine data (Hawkins, 2013) and an effort to save the Ukrainian cultural heritage through the efforts of the SUCHO group, creating digital archives for more than 1 500 websites, digital exhibits, open access publications, and other online resources from Ukrainian cultural organisations (Cascone, 2022).

While disciplinary differences regarding data rescue are prevalent, no culture-related differences were traced via an examination of published literature.

The next section provides details on the geographical extent of documented data rescue projects.

2.6.2 Geographical area

Data rescue is a truly global activity, with rescue initiatives recorded on all continents. Country-wise data rescue is just as encompassing, with rescue activities taking place in developed countries with advanced technology and funding, and also in many developing world areas.

2.6.2.1 Global data rescue activities

This geographical diversity is evident when glancing at the following list, containing examples of countries where data rescue has taken place:

- Argentina (ACRE, 2019),
- Australia (Ashcroft, 2016),
- Bolivia (IEDRO, 2018),
- Burkina Faso (WMO, 2019a),
- Canada (Fry, 2010b; Kennedy, 2017; Thompson, 2017),
- Chile (ACRE, 2010; IEDRO, 2018; Pizarro-Tapia *et al.,* 2020),
- China (ACRE, 2016),
- Cuba (Antuña *et al.*, 2008),
- El Salvador (IEDRO, 2018),

- Germany (Kaspar *et al.,* 2015),
- India (Diwakar, Kulkarni & Talwai, 2008),
- Ireland (Ryan *et al*., 2018),
- Italy (Diviacco *et al*., 2015; Stanley *et al*., 2020),
- Japan (ACRE, 2017),
- Malawi (IEDRO, 2018),
- Mali (Kimani, 2008; WMO, 2019a),
- Netherlands (Brandsma, 2007),
- Niger (IEDRO, 2018; WMO, 2019a),
- Norway (Hygen, Elo & Gjelten, 2021),
- Paraguay (IEDRO, 2018),
- Slovakia (Fasko *et al*., 2016; Bochnicek *et al*., 2017),
- South Africa (Woolfrey, 2016; ACRE, 2019; DataFirst, 2022),
- Sweden (Wern, 2017),
- Tanzania (ICPAC, 2016),
- Ukraine (Cascone, 2022),
- United Kingdom (Wood *et al*., 2012),
- United States of America (Libraries Plus Network, 2017; Ryan *et al*., 2018; University of Hawaii, 2018), and
- Uzbekistan (IEDRO, 2018).

Not all data rescue research involve data from one country only; certain initiatives target data from a wider geographical area, such as countries combined, or even an entire continent. Examples of combined areas where data rescue efforts have taken place include:

- Africa (Hachileka, 2015),
- Antarctica (ACRE, 2017),
- Europe (Brunet *et al*., 2020b),
- International territories (Knapp, Bates & Barkstrom, 2007),
- Meso-America (ACRE, 2017),
- South America (Antico, Aguiar & Amsler, 2018),
- Southeast Asia (Williamson *et al*., 2015),
- Southern Africa (Kaspar *et al*., 2015),
- South Pacific region (Page *et al*., 2004),
- The Arctic (ACRE, 2017),
- The Caribbean (Trotman, 2012),

- The entire globe (Arrouays *et al*., 2017),
- The Indian Ocean (ACRE, 2017),
- The Mediterranean Basin, the Mediterranean Sea (Jones, 2008; Mavraki *et al*., 2016),
- The Pacific (ACRE, 2017),
- Western Africa (WMO, 2012), and
- Western North Atlantic (Rountree *et al.,* 2002).

In addition to countries and areas, it is obvious that data can also be collected from a large area not easily delineated, such as data collected in space, or satellite data covering large expanses. Examples of such a rescue project would be the Nimbus project (Gallaher *et al.,* 2015) or the data rescue project involving astronomy images from the past 150 years (Gibney, 2017).

As stated by Hsu *et al.* (2013: 1), data rescue can involve data from the ocean bottom to the moon.

2.6.2.2 Data rescue in South Africa

As one of the study's research questions involves data rescue efforts in SA, two of the most prominent South African data rescue projects are worthy of further discussion. The rescue project entailing historic apartheid era datasets has been documented in several sources, including Woolfrey's presentation (2016), an article by Price (2018) in a popular LIS-related blog, an article by Khwela (2018) on a South African university news page, an article by Bernardo and Khwela (2018) on the news pages of a South African university library, and an article on the rescue organiser's website (DataFirst, 2022). The rescue project involved a partnership between the University of Cape Town's (UCT) DataFirst research data service, the Neil Aggett Labour Studies Unit (NALSU) and Rhodes Library at Rhodes University (RU), UCT Libraries, and the University of KwaZulu-Natal (UKZN). The rescue of the two paper-based datasets – the Keiskammahoek Rural Survey (KRS) and the Surplus People Project Survey (SPP) – as well as the collaboration and the resultant data publishing, underscores the importance of the original dataset. Additional details about the workflow of this data rescue project are provided in Section 3.3.11: Data Rescue Workflow: DataFirst (sociological data).

The second South African data rescue project deserving of elaboration is not a single project, but concerns the ongoing and extensive climatic data rescue activities managed by Stefan Grab of the University of the Witwatersrand, under direction of ACRE and funded by the C3S. The presentations by Grab (Grab, 2018a; Grab, 2018b), and Picas and Grab (2017) illustrate the scale of rescue efforts and the progress made. Rescued data include the full record of the Royal Astronomical Observatory in Cape Town, starting from 1834 and up to the present data (Picas & Grab, 2017). In addition, rescue efforts comprised the digitisation of the Table Bay Harbour Port Captain records from 1829 onwards (Picas & Grab, 2017). ACRE reported that the scope of rescue of historic South African terrestrial and marine weather observations amount to 25 000 digitised records, and 184 680 scanned items (2020). Brief details of the workflow of this ongoing data rescue project have been provided by the project leader via email communications (Grab, 2021a). Described by the project leader as a very informal workflow, the major rescue steps consist of:

- locating the paper-based data or media at risk,
- digitisation of located data/media by students involving adherence to a template,
- quality control of digitised data/media by project leader,
- digitised data/media sent to the ACRE International office in England, and
- sharing of data/media on platforms (sharing activities are beyond project leader jurisdiction).

Evidence of a third South African data rescue project was found via the study's empirical data collection stages (see Table 5.4; also Section 5.4.6.6); however, details supplied by the interviewee were brief in nature. This data rescue project could not be traced via the study's literature review activities.

2.6.3 Data sources and formats rescued

Data formats and types rescued, although mostly paper-based, on magnetic tapes, or in early digital format, display a range of interesting sources and formats. Examples of the range and diversity of documented rescued sources and formats include the following:

- 19th century KNMI yearbooks (Brandsma, 2007),
- 35mm volcano observatory slides (Wippich, 2012),
- diaries of explorers, missionaries, scientists (ACRE, 2019),
- farmers' diaries (Ashcroft, 2016),
- floppy disks (Thompson, Davenport Robertson & Greenberg, 2014),
- glass plates (Nordling, 2010),
- handwritten 18th and 19th century ship logs (Brandsma, 2007),
- handwritten and typewritten wildlife disease data pages (Wippich, 2012),
- handwritten observations of lighthouse keepers (I-DARE, 2019),
- handwritten observations of telegraphists (ACRE, 2019),
- historic punch cards containing cholesterol and heart disease data (Krotz, 2011),
- historic seismograms (Wyborn *et al.*, 2015),
- historic weather observations compiled by colonial surgeons and doctors (Williamson *et al.*, 2015),
- inexpensive recordable CDs (Schroeder, 2018),

- jaz drives[18] (Schroeder, 2018),
- lab notebooks (Murillo, 2014),
- metadata from physical samples (Hills, 2015),
- meteorologists' historic paper records dating from as early as the 17th century (ACRE, 2019; Brunet *et al.,* 2020b),
- microfilm (Hygen, Elo & Gjelten, 2021),
- military reports and records (Williamson *et al*., 2015),
- minerals exploration-assistance dockets: paper, carbon paper copies, copies of maps, graph copies, blueline copies of maps, linen or mylar maps, and other fragile and sometimes poor-quality media (Wippich, 2012),
- modern digitised data without proper documentation (Thompson, 2017),
- Nimbus satellite film data (Gallaher *et al*., 2015),
- Nimbus satellite tape data (Gallaher *et al*., 2015),
- nuclear explosion seismograms (An *et al*., 2015),
- obsolete analogue video (Barnard Library & Academic Information Services, 2019),
- ocean sediment samples (Stanley *et al.,* 2020),
- photo negatives (Thompson, Davenport Robertson & Greenberg, 2014),
- physics data on cartridges (Curry, 2011),
- physics data on magnetic tapes (Curry, 2011),
- plant specimens (Thompson, Davenport Robertson & Greenberg, 2014),
- pluviograph strip charts (Pizarro-Tapia *et al*., 2020),
- pro-formas (RDA DR IG, 2015),
- radiographs and photographs detailing wildlife diseases (Wippich, 2012),
- seismic shot hole drillers' lithostratigraphic logs (Wyborn *et al*., 2015),
- semaphore watchmen data (I-DARE, 2019),
- sound clips, sonograms, oscillograms pertaining to fish sounds (Rountree *et al.,* 2002),
- tide-gauge charts and tabulations (Caldwell, 2012),
- tissue samples (Murillo, 2014),
- video games (Janz, 2018: 5),
- water resources photographs, field measurements, reports, charts and maps (Wippich, 2012),
- weather strip charts (Brönnimann *et al*., 2018), and
- US federal websites, or webpages from US federal websites (Varinsky, 2017).

[18] A removable hard disk storage system sold by the Iomega company from 1995 to 2002 and capable of storing up to 2 GB

2.6.4 Data rescue participants and contributors

A scrutiny of data rescue literature reveals the diversity of disciplines, formats and geographic areas involved with the practice. Moreover, it reveals the range of participants involved. To the uninitiated, data rescue might seem to be the domain of scientists, aided by information custodians experienced in dealing with historic media. While this viewpoint is not entirely incorrect, these are not the only professions or vocations able to contribute towards data rescue. Data rescue efforts often entail collaborative efforts in which the input, experience and knowledge of different entities prove to be vital. This section of the chapter provides a brief glimpse into the various parties who might form part of a data rescue team.

While documented sources show that domain experts and information specialists often form the backbone of data rescue projects, data rescue participants can also include archivists, academics, repository professionals, indexers, data scientists, students, IT professionals, programmers, coders, and volunteering citizen scientists. Factors such as type of data rescued, the manpower available, and the training able to be provided can influence the range of participants. In addition to the current vocation of involved parties, there are also different types of involvement.

Examples of the involvement of mentioned parties are provided below.

2.6.4.1 Researchers/scientists

A study of data rescue literature showed that data rescue cannot be done without the input of parties familiar with the discipline of the data rescued. Subject specialists could be scientists or researchers, or even subject librarians. Examples of data rescue research where discipline specialists played a vital role are listed below.

- Brandsma mentioned that a professional should check a data source before it is digitised (2007: 2).
- The GLOSS project made use of many parties including a ‘group of experts’ (Bradshaw, Rickards & Aarup, 2015: 9).
- Muller (2015a: 18) referred to scientists and researchers who rescue, analyse, publish and repurpose data.
- The WMO (2016: 19) stated that climatic data rescue participants would include working or retired climatologists.
- ACRE’s data rescue initiatives involve several parties including climate researchers, meteorologists and social scientists (ACRE, 2019).

- The University Libraries website of Washington University in St. Louis invited scientists, among others, to a data rescue event (DataRescue WUSTL, 2017).

An interesting training initiative, developed at the Maynooth University after final-year undergraduate geography students successfully transcribed more than 1 300 station years of daily precipitation data and associated metadata for the period 1860–1939, should be included under this heading (Ryan *et al*., 2018). Furthermore, this innovative participatory learning rescue tool has been recognised by the WMO as an example of best data rescue practice (Maynooth University, 2017).

One can also differentiate between subject specialists who perform data rescue as part of their daily professional duties, and those who are participating as expert volunteers in a data rescue project. The US Data Refuge movement, concerned with the rescue of federal environmental data during the Trump administration period, featured the input of thousands of such volunteers. Examples of volunteering environmental experts are listed below.

- A popular online business magazine mentioned the involvement of scientists in data rescue events (Varinsky, 2017).
- Chassanoff (2017) referred to the data rescue contributions of concerned scientists.
- A graduate student declared that he is involved in data rescue as an IEDRO volunteer (Eveleth, 2014).
- McGovern (2017) mentioned the involvement of domain scientists during federal data rescue efforts.
- Researchers, environmental studies professors, scientists, historians, ecologists, laboratory managers and oceanographers were listed as data rescue participants (Beeler, 2017 and Schlanger, 2017b).

The next section discusses the involvement of the LIS sector during documented data rescue activities.

2.6.4.2 LIS sector (including archivists)

Librarians, data curators, information professionals as well as archivists form part of the LIS sector professionals mentioned in documented data rescue outputs. Descriptions of their respective roles and contributions vary between brief mentions and more elaborate detailing of rescue tasks performed.

The instances listed below are examples of cursory reporting on LIS sector involvement in data rescue.

- ACRE's data rescue initiatives involve a range of parties including information science experts and archivists (2016). Similarly, the University Libraries website of Washington University in St. Louis invited archivists (among others) to a data rescue event (DataRescue WUSTL, 2017).
- The WMO mentioned that parties interested in assisting with data rescue activities include working and retired librarians (2016: 8), while Gaudin (2017), Harmon (2017) and Morris (2017) referred to the involvement of librarians.
- Chassanoff (2017) discussed the data rescue contributions of research librarians.
- McGovern mentioned the involvement of information professionals in federal data rescue events (2017).
- The involvement of archivists during data rescue events was mentioned (BBC News, 2016 and Schlanger, 2017a).
- The A2 Data Rescue webpage mentioned the involvement of archivists, librarians and documentarians (among others) in a report on a data rescue event held at Ann Arbor (2017).
- The online newspaper of Georgetown University reported that librarians (among others) were involved in the DataRescueDC event (Carey, 2017).
- Slonosky *et al.* mention an interdisciplinary effort with experts from archival practices, information studies, data management, public participation, historical climatology, and software design (2019: 60).

The cursory mentions above refer to the conventional data rescue projects, and data rescue days involving the rescue of US federal data during 2016–2017. Not all documented literature referring to the involvement of the LIS sector in data rescue is brief in nature, as instances of more detailed LIS-sector involvement in data rescue can also be traced. Examples of mentioned LIS sector involvement in data rescue and data refuge provided below entail actual rescue involvement, as well as theoretical data involvement in activities related to data rescue or data at risk.

Levels of data rescue involvement: Eke's presentation on the ways in which libraries can get involved with data rescue mentioned four levels of LIS sector involvement in data rescue activities (2017: 2). The levels are dependent on resources available and are described below.

- Level 1 entails librarians surveying researchers to nominate data, and raising awareness in their research community. According to Eke, awareness is raised by attending and writing about a data rescue event, highlighting the ways in which one's repository can preserve data, and holding panels and workshops.

- Level 2 entails librarians making use of the Archive-It platform[19], performing deep web archiving, and documenting uncrawlables.
- Level 3 entails rescuing data needed by the research community, and regarded as 'high value, high priority' by the community.
- Level 4 entails the harvesting of uncrawlables through a data rescue event or a dedicated team. Level 4 is Eke's highest level of library commitment and involves adherence to established library preservation workflow protocols to maintain a trusted chain of custody. The workflow steps pertaining to this level are described in the referenced presentation.

Additional information on the links, spreadsheets, forms and GitHub workflows relevant to the four listed steps form part of the presentation.

Local LIS sector involvement: In the South African context, evidence of a single instance of library and information services being involved with a data rescue project was found (DataFirst, 2018; Khwela, 2018; DataFirst, 2022). DataFirst, a prominent South African data services entity had successfully rescued apartheid era datasets and had involved the library and information services of several South African university libraries in the rescue project. The exact nature of LIS sector roles and responsibilities had at the time of writing not yet been documented in the available publications, namely university webpages and the data organisation's own website. While a more detailed scholarly publication was still being finalised, the available online information confirmed that the LIS sector had been involved in this South African data rescue venture.

Data storage, curation, and long-term preservation: As stated by Thompson, Davenport Robertson and Greenberg (2014: 845), libraries and archives have historically organised and managed large collections of materials in both physical and digital formats, including the storage of documents, records and data. These are practices that are centuries old; the WMO aptly captures the involvement of library and information services with data storage and preservation:

- Libraries are appropriate places to search for and locate data (WMO, 2016: 2).
- Libraries are places where paper, microfilm/microfiche and digital copies of data are held (WMO, 2016: 3).
- National libraries often contain valuable datasets (WMO, 2016: 3).

[19] A web archiving service for collecting and accessing cultural heritage on the web, built at the Internet Archive (https://archive-it.org/)

- It is wise to approach national archives, agency/service libraries, or university libraries when requiring storage of data, and not having sufficient long-term storage resources (WMO, 2016: 4).
- Should national archives, agency/service libraries, or university libraries not be able to meet one's storage needs, international libraries or institutions for storage should be approached (WMO, 2016: 4).

The statement by Palmer, Weber and Cragin (2011: 8) that digital scientific data will not survive or become accessible without repositories dedicated to preserving these raw materials of research, underlines the crucial involvement of data curators with long-term data preservation.

Locating data: Thompson's article on the rescue of Canadian health files at risk concluded with the author stating that an important lesson learnt during the venture was that librarians without any technical or statistical background can still make valuable contributions to data rescue projects (2017: 35). The author further mentioned how the searching skills of library and information services staff, akin to detective work, and successfully locating items after searching through archives of neglected government documents, cross-checking details to track changes in content over time, or trawling departmental contact lists in anticipation of reaching that one person who knows where a file originated, formed a vital part of the documented data rescue project (2017: 35).

Data inventories: The WMO mentioned that international databases and libraries can be used when looking to download the known inventories of data, and that these entities are commonly involved when trying to assess a country's known meteorological assets (2016: 9).

Data rescue groups: Examples of library and information science professionals forming or being part of data rescue groups were traced. Two instances are described below.

Thompson (2017: 34) mentioned the Ontario Data Rescue Group, which was formed in 2015 and consisted of a small group of volunteers from the Ontario Data Community. The Data Rescue Group is comprised of subject librarians and technical support staff interested in improving access to data that has been deemed at-risk and of value to the academic community.

The NAL/iSchool Data Rescue Project is a strand of the Digital Curation Fellows Program, which is a multifaceted collaboration between the University of Maryland and the US Department of Agriculture (Hoffman *et al.*, 2020: 9). The project can be viewed as an example of a data rescue group, as the group is concerned with conducting research on curation at the National Agricultural Library and in the broader agricultural community (2020: 9).

Participation in data rescue projects: While many published outputs reported on the involvement of libraries or the LIS sector as distinct, or separate data rescue activities, evidence of libraries and library and information services experts being involved in the entire data rescue process has also been found. Examples of these descriptions are given below.

- A blogpost on the Libraries Plus Networks website described the data rescue activities performed by representatives from research libraries, and how these activities involved environmental scanning, virtual and in-person meetings, strategising and identifying data rescue efforts, performing rescue outreach to organisations and people, preparing for a rescue event, drafting a minimal data rescue workflow, rescuing data, creating metadata, and making the rescued data public (Kijas, 2018).
- Thompson's article described the rescue of Canadian Health Survey Files through the efforts of the Ontario Data Rescue Group, consisting of librarians. Data rescue activities forming part of this project included the creation of an inventory, assessment of data files, searching for and locating missing documentation, data sharing on an open portal, the capture of data rescue steps, and the transfer of rescue steps to a usable data rescue guide (2017: 33–35).
- Thompson further referred to a tradition of Canadian university data rescue work, including efforts at Carleton University and the University of Alberta (2017: 34). According to Fry (2010b: 7), activities forming part of the data rescue project involving the Carleton University libraries included the creation of data inventories after receiving boxes filled with data at risk, matching up files, deletion of confidential information, anonymisation of data, cleaning up the variable and value labels in the datasets, ensuring that metadata are available, and enabling of data download from the university data centre. The main rescue activities forming part of the University of Alberta project included retrieval of the digital data and all supporting (secondary) data sources, transfer of digital data from magnetic tape to compact disk, and the collection and preparation of relevant documentation describing the data (Kochtubajda, Humphrey & Johnson, 1995).
- Hoffman *et al.* (2020: 11) described the involvement of the US National Agricultural Library (NAL) in data rescue efforts. While not focused on scientific data, the described rescue overlapped with data rescue due to its concentration on rapid appraisal and rescuing potentially valuable materials at risk of loss (2020: 11). The rescue project entailed recovering an ample collection of personal files generated by a senior administrator at the NAL and resulted in securing an important segment of the entity's institutional history.

Data at risk survey participation: Thompson, Davenport Robertson and Greenberg's study on data at risk (2014) reported on the results of library and information services experts who had participated in a survey investigating types of data that are at risk as well as data practices and factors that endanger these data. The survey was intended for librarians, archivists and information custodians who were involved in any aspect of data curation.

Results of the survey showed that data preservation is important, and that participants viewed their lack of time as the greatest barrier to sharing these data.

The study has implications for data rescue and for training data custodians. As stated by the authors (2014), an understanding of the data at risk predicament can assist librarians, archivists and scientists in designing and funding successful data rescue efforts. While several universities in North America have developed graduate programmes to prepare information professionals for data curation, an understanding of the data at risk predicament will enhance educators' ability to prepare and mentor students who want to pursue careers in data curation.

Data rescue publications: Several examples of data rescue publications drafted by library and information services professionals are found. The Data Rescue Project of the US-based Digital Curation Fellows Program (see Section 2.6.4.2: Data rescue groups, for more details) published a report during 2020 (Hoffman *et al.,* 2020: 1). The white paper offers preliminary guidance on assessing the benefits and challenges of processing collections of scientific records for the purpose of data rescue (2020: 4). While the guide is oriented towards professional curators and curation institutions (including libraries, special collections and archives, and data repositories), the essential roles of data producers, potential data reusers, and domain experts in the data rescue process are acknowledged.

Thompson reported on a publication titled 'Data Rescue and Curation Guide for Data Rescuers', which is described as a how-to manual developed by the Ontario Data Rescue Group (2017: 35). More details on the group are provided in Section 2.6.4.2: Data rescue groups. The objective of the publication was stated to provide an accessible and hands-on approach to handling data rescue and digital curation of at-risk data for use in secondary research (2017: 35). The guide serves to improve librarians' and data curators' skills in providing access to high-quality, well-documented, and reusable research data (2017: 35). Aspects such as documenting data using standard metadata, file and data organisation; using open and software-agnostic formats; and curating research data for reuse are included in the manual.

Unique data at risk perspective: Thompson, Davenport Robertson and Greenberg (2014) mentioned that information professionals working in science, research or other special libraries offer a unique viewpoint on data at risk. The unique perspective offered by the library and information services sector stem from their position in the organisation and their LIS disciplinary training. The authors stated that understanding the data at risk predicament from the information custodian perspective is vital for preservation planning and efforts, and that there is value in examining this perspective. However, as stated by the authors, research to date had not investigated this perspective. The results of the survey described in section (h): Data at risk survey participation, provide more details on the library and information services perspective. In short, participants rated all formats as very or somewhat likely to become lost (2014: 846), only 25% of respondents reported storing data in external repositories (2014: 848), 46% of respondents indicated that there was no metadata-enabled catalogue or index for the endangered data (2014: 848), the majority (58%) of respondents felt that at-risk data were very important (2014: 849), and only 35% of respondents stated that their organisation complied with any standards or policies regarding data management, sharing or archiving. In addition, results showed that 72% of respondents viewed time involved in making data usable as a sharing limitation, while other obstacles were stated to be difficulty in accessing data files from storage media or repositories (59%), maintaining human subject confidentiality (35%), and protecting intellectual property rights (35%).

As stated by the authors, the unique perspective provided by information curators has implications for future data rescue projects, and the data rescue training of LIS students.

Data rescue collaboration: Documented sources refer to the importance of collaboration during data rescue, and that the library and information services sector should form part of such collaborations. McGovern's article on the data rescue observations of an archivist (2017: 25) shed light on the cumulative strengths of cross-domain expertise during rescue activities. She identified seven domains that usefully collaborate on data rescue: libraries, archives, records management, digital preservation, museums, software development and data science. Hoffman *et al.* (2020: 14) provided a summary of McGovern's article and stated that a practical example emanating from McGovern's stance pertains to the cross-sector contributions of libraries that can provide discovery services, archivists who are experts in provenance, and records management teams who are familiar with retention schedules of rescued data.

Documented sources also reveal the importance of the involvement of external sectors, together with the library and information services sector, during data rescue activities and projects. Hoffman *et al.* (2020: 13) stated that a data rescue project may involve librarians, archivists, domain experts and data scientists, and the involvement of the different parties offers a balance of roles. The authors also provide more details on this balance by describing how curators and data scientists advise on the structure and organisation of data and information, while subject matter experts provide insight into the past and future applications of the data. In addition, the authors are of the opinion that the combination of expertise will make it far more likely that rescued data will be put to future use (2020: 13), and that in general, successful data rescue initiatives in any domain will depend on collaboration between domain experts (disciplinary researchers) and data-curation professionals with a meta-disciplinary perspective and expertise in library and information science (2020: 32).

Data assessment: Palmer, Weber and Cragin referred to the role played by repository staff in establishing the preservation readiness of data, which included the conduct of integrity assessments for the data collection (2011: 7). The authors also mentioned the involvement of information professionals regarding the 'fit for purpose' assessment of data (2011: 4), and that the meta-science expertise of this sector of the LIS community can assist in identifying the range of potential communities that might find a dataset useful over time (2011: 4).

Adding to the above were statements by Hoffman *et al.* (2020: 18) regarding skills and knowledge required to be able to assess the reuse potential of data at risk. According to the authors, an appraiser should possess an understanding of several applicable subject areas. The appraiser should understand not just the primary research discipline community and their potential uses of the data, but also the likelihood that other potential user communities exist. Appraisers should also be able to determine whether the data suit the purposes of those communities (2020: 18). The authors further stated that the collaboration between data specialists (such as archivists, data scientists, or librarians) and subject matter experts (with knowledge of multiple disciplines) can enhance the data assessment stage of a data curation project. According to the authors, experts with knowledge of multiple disciplines, such as information specialists responsible for service provision to different disciplines, are better able to identify potential future uses of the data outside of the immediate research community.

Data sharing: An article by Thompson on the rescue of data documentation mentioned the eventual sharing of rescued data on an academic data portal (2017: 34). Kijas (2018) has also reported on a data rescue event organised by the MLA, whereby rescued data were deposited into a shared CKAN repository.

LIS training: Thompson, Davenport Robertson and Greenberg reported on the LIS perspective regarding data at risk and provided insight into the unique viewpoint of data custodians after participating in a survey on data at risk (2014: 842–861). Apart from the study findings assisting librarians, archivists and scientists in designing and funding successful data rescue efforts, results also had implications for data rescue and for training information custodians. According to the authors, certain North American universities had at the time of the study already developed graduate programmes to prepare information professionals for data curation (2014: 850). Insight into the data curator's perspective on data at risk was stated by the authors to enhance educators' ability to prepare and mentor students who want to pursue careers in data curation.

In SA, this researcher contributed towards enhancing the knowledge of LIS postgraduate students by presenting a guest lecture on data at risk and data rescue at a nearby university. More details about the lecture, and recommendations regarding future training interventions of LIS students, are presented in Section 6.4.12, under the heading 'Amend and adapt the LIS curriculum to include data rescue'.

Environmental scan of data rescue initiatives: Evidence of librarians performing an environmental scan was found in documented data rescue outputs. The US Music Library Association's environmental scan of existing data rescue initiatives, relevant to their discipline, is one such example (Kijas, 2018). The environmental survey was informal in nature and served to identify organised efforts to preserve government data. The scan also investigated the scope of rescue efforts, criteria and selection, methodology, and participation type. Identified efforts were grouped into either data archiving and distribution, or website crawling.

Host data rescue event: The involvement of the library and information services sector in hosting and organising data rescue events was mentioned by numerous sources. Kijas (2018) reported on a data rescue event organised by the MLA; the event involved the identification and rescue of data from the National Endowment for the Humanities, and the Institute of Museum and Library Services. Rescued data would be deposited into a shared CKAN repository[20]. The event also served as an outreach exercise and desired to encourage staff to deposit funded data into the shared repository.

Adding to this was the opinion of Eke (2017) that the LIS sector can be involved in four distinct levels of data rescue, and that one such level pertains to the holding of workshops and the hosting of relevant panels.

Other examples of the involvement of the library and information services sector regarding the organisation of data rescue events include Raughley reporting on the role of librarians during a federal data rescue event (2017), and the Libraries Plus Networks stating that representatives from research libraries were involved in the planning events regarding federal data access and preservation (2017).

Summary: Table 2.1 provides an overview of the documented data rescue participation activities of the library and information services sector.

[20] CKAN is an open-source data management system for powering data hubs and data portals (https://ckan.org/)

Table 2.1: Summary of documented LIS sector data rescue participation

DATA RESCUE DESCRIPTION OR ACTIVITY	REFERENCE
Cursory reporting on LIS sector involvement in data rescue: includes conventional data rescue and data refuge	ACRE, 2016; WMO, 2016; Chassanoff, 2017
Four levels of data rescue involvement: • Raise awareness • Use Archive-It platform for deep web archiving • Rescue 'high value, high priority' data • Harvest uncrawlables through a data rescue event or a dedicated team	Eke, 2017
SA LIS sector involvement	Khwela, 2018; DataFirst, 2022
Store and curate data; long term preservation of data	Thompson, Davenport Robertson & Greenberg, 2014; WMO, 2016
Locate data at risk	Thompson, 2017
Create data at risk inventories	WMO, 2016
Belong to data rescue groups	Thompson, 2017; Hoffman *et al.*, 2020
Participate fully in data rescue projects	Kochtubajda, Humphrey & Johnson, 1995; Kijas, 2018
Participate in data at risk surveys	Thompson, Davenport Robertson & Greenberg, 2014
Draft data rescue publications	Hoffman *et al.*, 2020; Thompson, 2017
Possess a unique data at risk perspective	Thompson, Davenport Robertson & Greenberg, 2014
Collaborate in terms of data rescue	McGovern, 2017; Hoffman *et al.*, 2020
Assess data at risk	Palmer, Weber & Cragin, 2011; Hoffman *et al.*, 2020
Share rescued data	Thompson, 2017; Kijas, 2018
Provide data rescue training to the LIS sector	Thompson, Davenport Robertson & Greenberg, 2014
Perform environmental scan of data rescue initiatives	Kijas, 2018
Host data rescue event	Eke, 2017; Libraries Plus Networks, 2017; Kijas, 2018

To recap, documented sources have mentioned or described the involvement of libraries and the library and information services sector workforce, including librarians, information custodians, data curators and archivists, in a diverse number of data rescue activities and concepts. Documented involvement ranged from vague, brief mentions, to descriptions of participation in a single rescue activity, to involvement throughout an entire rescue project.

Documented sources reporting on the LIS sector and data rescue have also reported on data rescue aspects indirectly related to data rescue, such as the data at risk perspectives of data curators, data rescue publications emanating from the LIS sector, and the data rescue training of LIS students.

The inference can be made that the library and information services workforce has the potential to play a role in most activities forming part of a data rescue project. In addition, library and information services professionals as well as semi-professional staff can be involved. As stated by Thompson (2017: 35), a lesson learnt from the experiences of the described Ontario Data Rescue Group is that librarians without any technical or statistical background can also make a valuable contribution to data rescue projects.

2.6.4.3 IT professionals, programmers, and coders

Examples of the involvement of IT-related experts in conventional data rescue (as opposed to the Data Refuge movement) were found to include programmers (Data Rescue: Archival and Weather (DRAW), 2019; Phiffer, 2017), IT professionals (Hsu *et al.*, 2015), and coders (Gaudin, 2017). The importance of infrastructure specialists is also mentioned by Mayernik *et al*. (2017: 3).

Expert volunteers in the US data refuge scenario were made up of a range of IT-related vocations and skills sets. The contributions of expert volunteers during US data refuge events are documented widely, and the following list contains some of the most common tasks performed:

- 'computer-savvy' archivists (BBC News, 2016; Schlanger, 2017a),
- expert volunteers backed up datasets and documents from the EPA and other government agencies dealing with environment and climate data (Beeler, 2017; Williams, 2017),
- experts from the coding community downloaded and scraped datasets (Allen, Stewart and Wright, 2017; Beeler, 2017),
- harvesters wrote code in three different languages to scrape data on topics ranging from water quality and snow cover to grain phenotype and genotype (Allen, Stewart and Wright, 2017; Beeler, 2017),
- seeders collected 1 100 URLs from the U.S. Fish and Wildlife Service pages and nominated them to the Internet Archive (Garcia, 2017; Lindsay, 2017),
- software engineers (Gaudin, 2017), and
- the University Libraries website of Washington University in St. Louis invited programmers and archivists (among others) to a data rescue event (DataRescue WUSTL, 2017).

2.6.4.4 Historians

The role of historians during data rescue activities has been mentioned (Muller, 2015b; WMO, 2016). These experts can also be regarded as subject specialists and placed within the earlier listed group.

2.6.4.5 Other expert volunteers

Many data rescue efforts rely on volunteers for the successful completion of the initiative. For this literature review, volunteering participants are divided into expert volunteers (discussed under this heading) and citizen scientists (discussed under the next heading). Expert volunteers are defined by this author as volunteers who have certain skills, knowledge or experience required in the rescue project.

The involvement of expert volunteers from environmental sciences, the LIS sector and the IT world has already been discussed. Expert volunteers from other areas were made up of a range of vocations and skills sets. The contributions of expert volunteers during US data refuge events are documented widely, and the following list contains some of the most common tasks performed:

- artists (Allen, Stewart and Wright, 2017; Phiffer, 2017),
- librarians (Gaudin, 2017),
- storytellers made a visualisation of #MyEPA tweets over time (Lindsay, 2017; Social Science Environmental Health Research Institute, 2017),
- storytellers worked on redesigning the EPA website (Lindsay, 2017), and
- storytellers created signs for the March for Science, (Lindsay, 2017; Social Science Environmental Health Research Institute, 2017).

2.6.4.6 Citizen scientists

This section of the chapter looks at the involvement and participation of members of the public who are deemed not to have previously acquired data rescue skills, knowledge or experience. An investigation of documented data rescue sources has revealed the vital role played by volunteers and citizen scientists in a range of data rescue projects. Well-known projects making use of citizen scientists in data transcription include Old Weather or the DRAW project. Additional examples of data rescue ventures employing similar steps are described below.

- 16 000 citizen scientists transcribed 350 years of UK archival rain records during the 2019–2020 COVID lockdown period (Currin, 2021).
- IEDRO is run by volunteers (IEDRO, 2014); the founder of IEDRO discussed the importance of making use of citizen scientists during environmental data rescue projects (Allan & Crouthamel, 2013).

- A well-documented ACRE citizen science initiative, namely Operation Weather Rescue, makes use of volunteers (ACRE, 2016); the founder of ACRE emphasised the role of volunteering citizens during data rescue projects (Allan & Crouthamel, 2013).
- The Global Surface Air Temperature (GLoSAT) project stated that a citizen science approach will be used to recover weather observations recorded more than a century ago (Global Surface Air Temperature, 2021).
- GLOSS stated that they are promoting a citizen science approach to discovering long-term records (Bradshaw, Rickards and Aarup, 2015: 11).
- Mayernik *et al*. (2017: 7) stated that the rescue of legacy data can benefit from volunteers.
- Kaspar *et al*. mentioned projects where the contributions of volunteer digitisers would be explored (2015: 60).
- The data rescue guidelines of the WMO provides a section on crowdsourcing, and mentions how many successful projects make use of volunteers and citizen scientists (2016: 20).
- Giusti (2022) mentioned an online forum for data rescue volunteers, and stated that the forum supports online discussion on historical climatology, especially data rescue, by anyone who is interested in this.
- The Weather Wizards application uses web-based graphical tools to enable volunteer communities to rapidly digitise climate charts (IEDRO, 2015).
- A climate data rescue study described making use of secondary school students through service–learning partnerships (Mateus, Potito and Curley, 2021).
- Ryan *et al*. described a data rescue project making use of undergraduate students (2018).
- Students can be used for data typing (Brandsma, 2007: 6), and can also form part of data rescue (WMO, 2016).

Benefits emanating from the involvement of citizen scientists in data rescue projects are mentioned in literature. Examples of advantageous outcomes of citizen science data rescue involvement are listed below.

- Both the ACRE founder and the IEDRO founder have described citizen science participation as beneficial, as it makes volunteers feel part of the data rescue enterprise and makes them feel part of scientific solutions (Allan & Crouthamel, 2013: 11).
- According to the IEDRO founder, citizen science participation can lead to said volunteers being more comfortable with climate concepts (Allan & Crouthamel, 2013: 10).
- According to the IEDRO founder, participation by citizen science volunteers results in said participants being more comfortable with the uses and limits of knowledge of climatic concepts (Allan & Crouthamel, 2013: 10).

Data Refuge activities, striving at saving US federal data deemed to be in danger of deletion/disappearance, were highly dependent on volunteers to achieve success. Prior to a Data Refuge project, an appeal was made (usually on the website of the university library where the event would be taking place), requesting volunteers. While the input of professionals, volunteering their time and skill, was included, the need for citizen scientists was present in most of these advertisements. Examples of such website requests are listed below.

- University of Minnesota's Facebook page requested data rescue volunteers, including researchers, librarians, archivists and passionate community members for a rescue event (2017).
- Fay (2017) stated that the MIT rescue event required volunteers with all levels of technical background.
- The Historical Canadian Climate Data Rescue Project was described as an entirely volunteer data rescue project (Jones, 2015).
- The DataRescueDenton 2017 event stated that the effort was an opportunity for volunteers of all kinds to identify, back-up, and help preserve publicly accessible federal data (DataRescueDenton, 2017).
- Slonosky referred to a citizen science, volunteer typing data project implemented with the assistance of the Canadian Meteorological and Oceanographic Society (2011).

Published outputs after the fact, reporting on the respective data rescue events, include:

- a 2021 article in a popular online publication giving details on the success of data rescue events several years prior (Calma, 2021),
- an article reporting on the valuable work performed by hackers to save environmental data at risk (Schlanger, 2017b; Wylie, 2021),
- an article describing how the US 2016–2017 data rescue activities had become a movement (Mandelbaum, 2017),
- the webpage of the Hesburgh Libraries website describing 'Data rescue' as a movement among concerned citizens to preserve primarily government-hosted scientific data (University of Notre Dame, Hesburgh Libraries, 2019), and
- an article in a popular computer magazine providing details of past data rescue events (Gaudin, 2017).

Published literature on citizen science involvement show that the most common rescue task performed by these participants entails the transcription of idiosyncratic handwriting, for example the

19th century handwritten observations recorded at the McGill Observatory (Slonosky, 2021), or the 19th century Ireland rainfall data (Ryan *et al.,* 2018).

2.6.4.7 Summary

The above list is by no means a complete index of all professions and parties able to contribute to data rescue. The list serves as an illustration of the collaborative nature of data rescue, and how the contributions of various backgrounds, skills and perspectives are required during rescue projects.

The role and input of volunteers should not be underestimated. The advent of data rescue activities, and in particular the Data Refuge initiative in 2017 has heralded an era where the input of volunteers is vital. To state it bluntly, the US federal data rescue and data refuge would not be possible without people from different backgrounds volunteering their time. The US federal data rescue efforts are not funded, and do not form part of any professionals' stipulated daily tasks or responsibilities. This is not to say that rescue events not forming part of the Data Refuge movement take place without the participation of volunteers. The input of the volunteer, as discussed earlier, is a regular feature in non-federal rescue projects, and examples of their indispensability are plentiful.

This section on data rescue contributors has shown that rescue initiatives are never the responsibility or domain of a single rescue specialist. At the very least, the input and skills of a domain specialist, as well as an information specialist, are required. An increasing number of projects require additional input from IT specialists, coders, programmers, and even enthusiastic citizens. While some data rescue efforts form part of a professional's tasks and remunerated duties, unpaid volunteers, experts as well as eager initiates form the backbone of many rescue undertakings.

The issue of collaboration between various skills sets and domain expertise is one that is mentioned in several rescue documents. According to Hsu *et al.*, the issue of domain knowledge (i.e., librarians versus scientists) is a frequent debate, and the mix of skills from researchers, librarians and IT professionals is necessary for proper data rescue (2015: 113).

Moreover, the reality of different data rescue skills sets required is apparent when looking at the volunteering tasks to be performed during federal data rescue events. A good example of a diverse skills set is the DataRescue Philly event (Appel, 2017) mentioning that the effort will include seeders, baggers, metadata creators, tool builders, storytellers on social media, and participants who will strategise Data Refuge into the future. Mayernik *et al.* (2017: 1) aptly claimed that cross-sector collaboration when preserving at-risk data results in a stronger partnership, and lauded efforts such as Data Refuge, where an energetic, diverse community of enthusiastic citizens and professionals with valuable skills and expertise had been brought together.

Another instance of the multifaceted character of data rescue skills is seen when analysing the professions of four panellists discussing endangered data. According to Schell (2018), the presenting group comprised a university archival coordinator, a biomedical research data specialist, an assistant professor in information science, and a director of the Shapiro Design Lab[21].

Published data rescue outputs have shown that relevant training before or during a data rescue event is vital, and not only when making use of citizen scientists, who are not experts in any data rescue skills. The WMO's climate data rescue guideline states that there will be new, unfamiliar equipment and software for the individuals who undertake data rescue work. Due to people transitioning, it is necessary that multiple persons be trained to ensure that there is no single point of failure (2016: 23). The WMO's guide on hydrological data rescue (2014: 9) emphasised the importance of assessing the skills required to undertake a rescue project and undertaking training where necessary. Another leading data rescue initiative stressing the aspect of training includes IEDRO (2014), who often coordinate with a country's meteorological services team to train staff in conducting data rescue. The data rescue project involving students at Maynooth University (Ryan *et al.*, 2017) is another good example of participant training forming part of the entire venture, as is the C3S webpage mentioning that training is supplied during their data rescue capacity building workshops (2018).

Many of the data refuge events involved the training of volunteers from academia, non-profit organisations, and the 'tech industry'. US federal data rescue events frequently mentioned training in their announcement, and an article in a popular online publication which stated that data rescue volunteers would receive training (Gaudin, 2017) is a good example.

In short: data rescue participants and contributors are linked to a range of professions and skills sets. Cross-collaboration is a feature of many successful rescue efforts.

2.6.5 Data rescue entities and interest groups

A number of groups are involved with data rescue, and their involvement ranges from the provision of training, right through to project implementation and the donation of the necessary basic data rescue equipment. Examples of data rescue entities are:

- ACRE: The international Atmospheric Circulation Reconstructions over the Earth initiative undertakes and facilitates the recovery of historical surface, terrestrial and marine global weather observations to underpin 3D weather reconstructions spanning the last 200–250

[21] The Shapiro Design Lab is described as an 'ever-evolving experimental space', forms part of the University of Michigan Library, and is focused on creating engaged learning opportunities and experiences across teaching, research, and artistic projects (https://ocs.umich.edu/certified-workplaces/shapiro-design-lab/)

years. This is achieved by linking international meteorological organisations and data rescue infrastructures, making these observations available to all, and ensuring that 3D reanalysis products can be tailored or downscaled to flow into various climate applications and models (ACRE, 2020).

- IEDRO: The International Environmental Data Rescue Organization is involved in locating, imaging, digitising and sharing of historic climate data, thereby enabling developing countries to adapt and mitigate the effects of climate change (IEDRO, 2014).
- I-DARE: The International Data Rescue Portal provides a single point of entry for information on the status of past and present worldwide data to be rescued and data rescue projects. The portal also provides guidance on best methods and technologies involved in data rescue, and on metadata for data that need to be rescued. The portal is available in English, Spanish and French (I-DARE, 2021).
- C3S: The Copernicus Climate Change Service is a service to facilitate climate data rescue that builds upon existing WMO International Data Rescue activities. The service also works with ACRE and with the I-DARE Portal. The C3S portal collects and shares information on past, current and planned data rescue projects. It also feeds into international repositories and promotes data rescue tools, best practice, and standards for all aspects of the data rescue process (C3S, 2018).

Another highly regarded international body is the World Meteorological Organization (WMO). The WMO, whose mandate covers weather, climate and water resources, has published two data rescue manuals, is involved in global data rescue projects, workshops and meetings, and has an active data rescue web presence.

Evidence of the ACRE-supported ‘ACRE South Africa’ group, concerned with the rescue of South African environmental data, was traced in several online sources (Picas & Grab, 2017; Grab, 2018a; Grab, 2018b; Grab, 2021b).

Interest or task groups, being a group of people drawn or acting together in support of a common interest, exist within the data rescue community. An example of a smaller national data rescue interest group is the interest group of the US Music Library Association (Kijas, 2018). Another example is the Ontario Data Rescue Group, formed in December 2015 by a small group of volunteers from the Ontario Council of University Libraries (Thompson, 2017: 34). The regional Canadian data rescue interest group contributed to the wider data rescue community via the publication of the Data Rescue & Curation Best Practices Guide, which is a how-to manual developed by the group (OCUL Data Community Data Rescue Group, 2020).

While smaller national and even regional data rescue interest groups exist and are mostly focused on regional support and membership, global data rescue interest groups support common global data rescue issues and concerns. The data rescue groups most prominent during the past decade were found to be the CODATA Data at Risk Task Group (CODATA, 2013), the RDA Data Rescue Interest Group, and the RDA Data Conservation Group (Research Data Alliance, 2019).

CODATA's Data at Risk Task Group, now historic and no longer active, was concerned with the predicament of many scientific datasets not in modern electronic formats and thus not accessible to the researchers who needed it (CODATA, 2013). The contributions of the task group and task group members included a two-hour panel discussion at the 2013 Digital Heritage Symposium held in France, and a 2014 panel discussion at a digital preservation conference held in India (CODATA, 2015).

During 2015, the CODATA Data at Risk Task Group became affiliated to the Research Data Alliance through the formation of an RDA Interest Group for Data Rescue. This expansion and collaboration increased the range of activity of the Data at Risk Task Group, while the combined groups shared the benefits of the two supporting organisations, i.e., CODATA and RDA.

Participatory events of the RDA interest group included a workshop on data rescue of data at risk held in Colorado during 2016 (UNIDATA, 2016), presentations delivered at the 2017 NAGARA Conference in Idaho, USA (RDA, 2017), and a free workshop on data rescue held at Harwell, UK during 2018 (Marine Environmental Data and Information Network, 2018).

The RDA Interest Group has since undergone a shift in focus and a name change and has been known as the RDA Data Conservation Interest Group since late 2019, with the RDA Data Rescue Interest Group now historic in status. The newly named group, concerned with data conservation, is building on the work of the previous interest group, and has expanded its scope to all types of data. The shift in focus also meant that the group would be discussing issues such as datasets prioritisation for rescue, defining, assessing and categorising data risk factors, and prioritising and securing data rescue funding (Research Data Alliance, 2019).

In SA, there are currently no groups focusing exclusively on the rescue of data. Collaborations that have taken place during the few documented rescue projects are temporary in nature.

2.6.6 Data rescue manuals and guidance

The WMO's data rescue guidelines – one pertaining to the rescue of climate data (WMO, 2016), the other to the rescue of hydrological data (WMO, 2014) – are recommended reading for any data rescuer embarking on the rescue of any discipline's paper-based data. Published manuals or guidelines by other parties, such as the Canadian Marine Species Data Rescue Cookbook (Kennedy, 2017), or the poster by Ryan *et al*. illustrating classroom data rescue settings (2017), are also highly endorsed by this researcher. Guidelines and manuals, and the data rescue workflow contained within them, are discussed in more detail in Chapter 3: Literature and the creation of a data rescue workflow model.

While the manuals mentioned above are regarded highly, they are not the only examples of recommended published data rescue manuals. Additional outputs include the following:

- a document titled 'Best Practice Guidelines for Climate Data Rescue' (Wilkinson *et al.*, 2019),
- a publication titled 'Best Practice Guidelines for Climate Data and Metadata Formatting, Quality Control and Submission' (Brunet *et al*., 2020a), and
- a guide titled 'Final Report and Recommendations of the Data Rescue Project at the National Agricultural Library' (Clarke & Shiue, 2020).

No evidence of manuals for the rescue of South African data at risk was found. It is presumed that the ACRE-affiliated data rescue activities taking place in SA would have received guidance from ACRE.

No guidance or data rescue manuals are currently in place at the selected institute. The single instance of a complete data rescue project had enlisted the assistance and guidance of an external data digitisation entity, and other discipline-specific experts.

2.6.7 Other data rescue outputs

Searching for documented data rescue outputs has revealed that data rescue reporting, publishing and sharing involve a wide range of publications. The reporting or published relaying of data rescue ventures has been found in the following publication types:

- articles on the websites of recognised data rescue entities (e.g., the 2020 article on data rescue in Uzbekistan on the IEDRO website),
- blog posts (e.g., the blog post by Schmidt, 2017),
- chapters (e.g., the chapter by Downs & Chen, 2017),
- dedicated data rescue project websites (e.g., the website of DRAW, 2019),
- popular journals or online newspapers (e.g., the article by Eveleth, 2014),
- posters and presentations at a workshop (e.g., the 2017 presentation by Mukhondia, and the poster by Ryan *et al.*, 2017),
- scholarly articles in data-oriented journals (e.g., the *Data Science Journal* article by Griffin, 2006),
- scholarly articles in LIS-oriented journals (e.g., the 2014 *College & Research Libraries* article by Thompson, Davenport Robertson & Greenberg),
- SlideShare presentations (e.g., the presentation by Eke, 2017),

- social media platforms (e.g., the Twitter post by the Marine Environmental Data and Information Network, 2018),
- technical, progress or annual reports (e.g., the report by Wood *et al.*, 2012),
- entries on a university library website (e.g., the 2018 article published on the University of Virginia Library Services page),
- university news page articles (e.g., the article by Khwela, 2018), and
- YouTube videos (e.g., the published video by Schell, 2018).

Documented outputs of South African environmental data rescue projects were traced in annual reports or progress reports of ACRE, supplemented by mentions on the ACRE website. Basic details of the South African sociological data rescue project could be found in online university news articles and webpages (Khwela, 2018), on the website of the data centre managing the project (DataFirst, 2018; DataFirst, 2022), and in a popular online LIS-related journal (Price, 2018). A presentation on the data rescue venture was forwarded upon request (Woolfrey, 2016), while the rescue project's manager indicated via email correspondence that a scholarly article pertaining to the latter project was being finalised (Woolfrey, 2021).

2.6.8 Data rescue meetings and related ventures

A number of data rescue workshops and training events have been held during the last decade. These events were organised by a range of involved data rescue entities. The list below contains examples of data rescue events that have taken place around the globe during the last decade.

- An event titled the 'International Workshop on Data Rescue and Digitization of Climate Records for Countries in West Africa' was held in 2012 and organised by the World Meteorological Organization (WMO, 2012).
- An event titled 'Sub-regional Training Workshop on Climate Data Rescue and Digitization', held in 2013 in Amman, Jordan was organised by the Economic and Social Commission for Western Asia (United Nations Economic and Social Commission for Western Asia, 2013).
- An event title 'Indian Ocean Data Rescue (INDARE) Capacity Building Workshop' was held in 2015 in Arusha, Tanzania and organised by the WMO with the scientific leadership of the Centre of Climate Change (WMO, 2015).
- The Regional Centre for Mapping of Resources for Development (RCMRD) reported on a 2016 event titled 'Tanzania Meteorological Agency Climate Data Rescue Assessment Workshop', held in Tanzania, and organised by the PREPARED Project, an NGO based in Kenya (RCMRD, 2016).

- An event titled '1st C3S Data Rescue Service Capacity Building Workshop and 10th ACRE Workshop' was held in Auckland, New Zealand in 2017 and organised by the C3S (C3S, 2017).
- An event titled 'Joint C3S Data Rescue Service Capacity Building Workshop and 12th ACRE Workshop' was held in Argentina during 2019 and organised by C3S Data Rescue Service and ACRE (WMO, 2019b).
- An event titled 'Mauritius Data Rescue Workshop' organised by the Centre of Climate Change and ACRE was due to be held during 2020 but postponed due to COVID-19 travel restrictions (WMO, 2020).
- An event titled 'Workshop on Sea Level Data Archaeology', organised by the IOC, was held in Paris during 2020 (IOC, 2020).

No documented evidence of workshops, conferences or similar events related to data rescue in SA could be traced. It is, however, assumed that unpublished meetings forming part of data rescue projects would have taken place.

No evidence of data rescue-related meetings held at the selected research institute could be ascertained via the study's empirical data collection methods.

2.6.9 Resource constraints and data rescue

As indicated in the segment on data rescue outcomes in Section 2.5, final data rescue outcomes such as public access to data, the extended coverage of data repositories and more accurate science are all linked to the data sharing activity. Data sharing and data publishing form a crucial part of most data rescue definitions and data rescue projects, and the definitions of Levitus (2012: 48), referring to 'archiving the data into an internationally available electronic database' or Brandsma (2007), stating that rescued data end up 'available to the public' are two examples of the inclusion of data sharing in the data rescue workflow.

With data sharing inextricably linked to what is frequently considered the goal of data rescue, it is well worth considering the variables which might affect the decision to share data. Many studies have in fact investigated the topic: the research of Zenk-Möltgen *et al*. (2018), Devriendt, Borry and Shabani (2021) and Hamamurad, Mat Jusoh and Ujang (2022) are examples of recent studies in this regard. Within these three studies, the role of authors' attitudes, reported past behaviour, social norms, perceived behavioural control, credit and recognition, the potential misuse of data, loss of control, lack of resources, socio-cultural factors and ethical and legal barriers were among the factors identified as affecting intentions to share data.

This study's selected South African research institute is considered as resource constrained (Patterton, Bothma & van Deventer, 2018) and it is therefore prudent to examine the findings of a study into the data sharing practices of researchers in low-resourced research environments. The research of Rappert and Bezuidenhout shed light on the data sharing practices of this demographic via a study involving two Kenyan university chemistry departments, a South African university chemistry department and a South African university chemistry/biochemistry department (2016). The study delivered thought-provoking findings regarding issues linked to the research practices and data sharing behaviours of participants working in resource-constrained environments.

The main findings linked to research activities are listed below.

- Workload and division of labour: Faculty staff generally had high teaching loads, with much of the research activities conducted by advanced postgraduate students.
- Precarious funding: Participants stated that acquiring funding for facility maintenance and improvement was problematic.
- Systemic issues: Participants at all sites mentioned infrastructure issues as daily research challenges. Regular power cuts, varying provision of backup generators, complicated and time-consuming reagent delivery, problematic sample transport options, difficulties with technical support and issues with equipment maintenance are some of the obstacles facing researchers.
- Promotion systems: Promotions were directly linked to journal paper outputs, with other forms of sharing not recognised.
- ICT provisions: All sites had internet access and computing and library facilities, however, challenges were experienced when accessing online resources. Participants mentioned power cuts, low bandwidth and variable Wi-Fi signal as daily online working obstacles.
- Professional self-promotion: Little interest was shown in the use of professional profiling sites and none of the participants had considered using social media or professional monitoring tools for research promotion.
- Securing project grants: Participants mentioned difficulties in securing project grants, including foreign assistance.

A range of data-related matters were reported by Rappert and Bezuidenhout (2016), and are listed below.

- Data collection responsibility: Research departments relied heavily on masters and PhD students for data generation.

- Low levels of data sharing: This multifaceted data aspect included strategic limitation of data to others, difficulties in gathering, curating and disseminating the data generated by students, reluctance to share data prior to publication, no sharing of unpublished data, and little sharing of data beyond project collaboration.
- Perception of data ownership: The pressures associated with research and the publishing thereof, together with the lack of support for these activities resulted in researchers often viewing data as personal property.
- Negative data sharing perceptions: Sharing of data was linked to expectations of data being scooped.

The study of Bezuidenhout and Chakauya into the data sharing practices of researchers in low-/middle-income countries supports these findings (2018). The study revealed that low-resourced environments shape data sharing activities, and that significantly more respondents were interested in using data sharing as a means of establishing future personal connections than as a means of improving research visibility. The findings also indicate that instead of data being effectively disseminated and reused, data are 'languishing' in drawers, hard drives or in cloud-based storage for many years due to issues of trust and personal connection (2018).

Based on the findings regarding participants' research and data practices, the implications for feasibility and success of data rescue projects are numerous. It is worthwhile considering the potential significance of the following aspects within the resource-constrained environment/data rescue sphere:

- availability of workforce for data rescue,
- availability of funds for data rescue,
- availability of funds for annual data repository fees (if applicable),
- suitability of available institutional infrastructure,
- feasibility of procurement of data rescue equipment and tools,
- rescue-worthiness of data (condition, availability of metadata),
- willingness to share data,
- perception of value of data rescue,
- availability of data rescue training, and
- ease of obtaining available external data rescue funding and assistance.

Some of these aspects will be further elaborated on in Section 5.9, which contains a reflection on key concepts emanating from this study.

2.7 Chapter summary

This chapter detailed the topic of data at risk and data rescue, as reported and described in documented literature. A wide range of sources were consulted, including scholarly articles, conference papers, conference posters, book chapters, blogs, relevant webpages developed by universities and university libraries, online newspapers, YouTube clips, and even writings found on professional social media platforms. The chapter could be subdivided into three distinct units.

The first part of the chapter dealt with data at risk terminology and definitions. Terminology related to data rescue were discussed, and the differences in meaning and nuance between various terms were stated. Terms clarified include data rescue, data conservation, heritage data, vintage data, vulnerable data, endangered data, and data archaeology. A discussion of topics such as the prevalence of data at risk and the factors leading to data being at risk followed. Examples of factors leading data to be at risk include catastrophic loss, data being in an outdated format, lack of metadata accompanying the data, changing priorities, sub-standard data management practices, deterioration of the media, data being displaced, and the loss of knowledge or skills pertaining to the data and/or access to the data. The first section of the chapter concluded with the identification of assessment and appraisal criteria used to identify data at risk.

The second part of the chapter was used to explore the numerous benefits of data rescue; such a discussion showed that the benefits of rescuing data entail far more than mere access to data previously considered to be at risk. Beneficial outcomes also include a better understanding of the past, clues to future conditions, science being more accurate, and the creation of new theories, publications, projects and products.

The various deliverables and outcomes emanating from data rescue projects were also considered. Improved infrastructure, job creation, community involvement, environmental literacy, impetus for other data rescue projects, and the formation of consortia are some of the tangible end products resulting from data rescue activities.

Lastly, the diverse nature of data rescue projects and data at risk formed a prominent section of the literature review chapter. The section on research disciplines involved with data rescue showed that the environmental sciences (and in particular, climatic sciences) were the most prevalent subject areas involved in data rescue. However, data rescue projects identified from literature also included data from disciplines such as the arts and music spheres, health sciences, physics, religious studies, and the socioeconomic disciplines.

Considering the geographical origins of data rescue projects supported the notion that the rescue of data is a global happening. Data rescue activities were traced on all continents, with data also collected while in space, while at sea, or from the bottom of oceans. Data rescue in SA is not a widespread practice, even though details of valuable rescue projects have been traced. The subject areas involved in these local data rescue projects comprise marine and terrestrial weather observations, and sociological data.

Data sources and formats forming part of data rescue portrayed its diverse character, with more than 20 format types traced in documented sources. While the most common formats forming part of rescue activities were paper-based data and data on magnetic tapes, formats such as strip charts, microfilm, samples, floppy disks, glass plates, slides, punch cards, seismograms, photo negatives, USB sticks, CDs and even data in a modern electronic format were mentioned in the large sample of data rescue publications reviewed.

Data rescue documentation revealed that a range of participants form part of data rescue ventures. Researchers and library and information services professionals are commonly involved, with IT experts and citizen scientists also included in a number of rescue ventures. Projects, such as DRAW, involve thousands of citizen scientists worldwide; human volunteers are required for the successful completion of the project, as handwritten paper-based data, being idiosyncratic in nature, cannot be transcribed or understood by optical scanners.

The various roles, responsibilities and contributions of the library and information services sector during data rescue formed an important part of this chapter. As one of the study’s research questions pertains to the ways in which the research library and its workforce can be involved, it was vital to investigate and report in more detail the documented participation of the library and information services sector in data rescue. This sector was found to be involved in a range of data rescue activities, including the organising of a data rescue event, locating data at risk, creating data at risk inventories, assessing data at risk, storing data, and sharing rescued data. Apart from the participation in the listed activities forming part of a data rescue project, literature also touched on the link between the LIS sector and data rescue training, the forming of data rescue interest groups by library and information services experts, and the involvement of information curators in a survey regarding data at risk. Literature also referred to information curators providing, via the aforementioned survey, a unique perspective on data at risk, and library professionals and libraries being involved in the hosting of data rescue events. Literature further described the various levels of LIS sector involvement in data rescue, and the participation in any of the four mentioned levels being dependent on the resources available to the library or library and information services professional.

Perusal of documented data rescue outputs also revealed the existence of data rescue entities and interest group, with groups being active with regard to workshop organisation, participation in conferences, and the sharing of data rescue-related information and issues by means of an online web presence. It was interesting to find that the foremost international data rescue group had undergone a name change during recent years, signalling a change in focus from data rescue to data conservation.

Manuals and guidelines are popular and helpful data rescue publications, as these outputs commonly contain data rescue workflows, frameworks, guidance, and examples of best practices. Other data rescue publications and outputs include scholarly articles, popular articles, book chapters, conference presentations and posters, blog entries, and even online video clips or channels.

The next chapter is devoted to documented data rescue workflows and frameworks. The intention of the chapter is to describe and discuss a representative sample of documented workflows. The chapter will also discuss how the information collected and insight obtained via the various examined workflows were used to create an initial Data Rescue Workflow Model to be presented for review by participants.

CHAPTER 3: LITERATURE AND THE CREATION OF A DATA RESCUE WORKFLOW MODEL

3.1 Introduction

This chapter reports on 15 data rescue workflows, models, processes and steps documented in literature, and includes a description of an initial Data Rescue Workflow Model created by this researcher after analysing published or shared data rescue literature and documentation.

As the creation of an initial Data Rescue Workflow Model is this chapter's primary deliverable, it was considered necessary to elaborate on the concept of 'workflow'. In addition, the tool used to illustrate the workflow, namely a flowchart, should also be defined and described. Both concepts are afforded a dedicated segment of this chapter and are elaborated on in Section 3.5: Creating a Data Rescue Workflow Model by means of a flowchart.

To create the initial model, it was considered crucial to examine, review and analyse relevant literature. Literature of interest entailed data rescue publications containing data rescue guidelines, data rescue models, data rescue workflows, published checklists of data rescue steps, and references to best data rescue practices. It was essential to consult a large number of data rescue publications, and involve in this analysis stage a range of data rescue projects, specifically projects where the rescue activities and steps were listed and explained. Within this data rescue scrutiny, it was therefore important to consider a range of research disciplines, data formats, age of data, size and scope of rescue projects, rescue complexity, and role-players involved. As the contents of the various data rescue literature items were viewed as crucial data during this stage, content analysis was used as a method to analyse this data. The steps forming part of content analysis, applied to rescue documentation, are also described in this chapter.

It was anticipated that by studying and analysing documented workflows, models and processes, the following would be achieved:

- a grasp of the range and diversity of workflows, models and processes used,
- an understanding of reasons why certain workflows, models and processes were used,
- an understanding of data rescue roles and responsibilities,
- insight into lessons learned by other data rescuers during their efforts,
- realising which workflows, models and processes would be applicable/not applicable in one's own institutional rescue project, and
- insight into drafting a data rescue workflow, to be tested at one's own research institute.

Adding on to the list above was the expectation of discovery of recorded LIS sector participation and contributions during data rescue projects. Furthermore, it was realised that should published outputs fail to explicitly mention the involvement of the LIS sector, the increased level of data rescue insight gained during the analysis stage would enable the identification of rescue stages and activities suited towards the participation and contributions of the LIS sector.

The remaining main sections of this chapter, each contributing towards the eventual creation of an initial Data Rescue Workflow Model, are listed below.

- Section 3.2: Introduction to data rescue workflows, models and processes reviewed and analysed
    - This segment provides an overview of the review and analysis activities performed for this research. The range of workflows, models and processes analysed is also mentioned, and the diversity regarding terminology use, complexity of projects, data formats involved, and research disciplines are included.
- Section 3.3: Data rescue workflows, models and processes
    - This segment of the chapter contains a brief description of each of the data rescue workflows, models and processes that were reviewed and analysed. Each of the analysed outputs is discussed separately under the two headings titled ‘Summary’ and ‘Contributions and limitations’.
- Section 3.4: Summary of analysed workflows, models and processes
    - This segment of the chapter provides a tabular summary of the analysed data rescue workflows, models and processes. Summarised information for each of the workflows include a URL, data format involved, research discipline involved, and the workflow’s outstanding features.
- Section 3.5: Creating a Data Rescue Workflow Model by means of a flowchart
    - As an outcome of this chapter is the creation of a Data Rescue Workflow Model, this segment discusses the concepts of ‘workflow’ and ‘flowchart’, respectively.
- Section 3.6: Implementing content analysis to create a Data Rescue Workflow Model
    - This segment of the chapter describes the implementation of content analysis to 15 published data rescue models as a method to determine the vital components of a Data Rescue Workflow Model.
- Section 3.7: Data Rescue Workflow Model: Description and characteristics
    - This segment of the chapter provides textual details regarding the initial model.
- Section 3.8: Data Rescue Workflow Model: Summary

  - This segment of the chapter provides a single-page visual representation of a summary of the initial model.
- Section 3.9: Stages of the Data Rescue Workflow Model
  - This segment of the chapter provides visual representations (i.e., flowcharts) of each of the stages of the initial model.
- Section 3.10: Chapter summary
  - This segment of the chapter provides a summary of all sections of the chapter.

The next segment of the chapter introduces the 15 different published data rescue workflows, models and processes selected for content analysis.

3.2 Introduction to data rescue workflows, models and processes

Data rescue workflows, models and processes, as documented in literature, are characterised by diversity not only in terms of wording and phraseology, but also the data formats and disciplines involved, data rescuers participating, complexity of the workflow, and the eventual outcomes. Examples of the range of terms used to describe the published data rescue project chronology and activities include 'guidelines' (WMO, 2016), 'steps' (Diviacco *et al.*, 2015), 'practical strategies' (Johnston, 2017), 'roadmap' (Brönnimann *et al.*, 2018), 'cookbook' (Kennedy, 2017), and 'model' (Brönnimann *et al.*, 2018). Additionally, while some authors describe the entire data rescue workflow, others, such as the often-recommended study by Ryan *et al.* (2018), focused on an aspect such as transcription and did not include a data rescue activity such as sharing the data to a public repository. Complexity is another aspect found to show variety between papers; the simplistic IEDRO climatic data rescue process (IEDRO, 2014) and the detailed marine species data rescue workflow (Kennedy, 2017) are examples of two publications on opposite sides of the complexity spectrum.

It was also important to realise and grasp the differences in publication objectives between different data rescue outputs. Some of the reviewed publications were produced as data rescue guidance or educational tools – the data rescue guidelines of the World Meteorological Organization (WMO, 2016) and the data rescue 'cookbook' by Kennedy (2017) are prime examples of this category. Conversely, other data rescue workflows, models and processes assessed were not published to be used as guidelines, but instead formed part of the article to give the reader basic insight into how the rescue attempt was performed, or how a hypothetical data rescue should be performed (Griffin, 2006; Thompson, 2017). As such, these 'incidental' data rescue workflows, models and processes are not as detailed or intricate as formal rescue guidelines and tend to assign a larger portion of the article to the data rescue background, characteristics of the data at risk, and rescue outcomes. Often, a data rescue article would glimpse over the actual rescue steps, as is found in the publications by Antuña *et* *al.* (2008) and Williamson *et al.* (2015). This major difference can be expected, as an article reporting on a data rescue mission has a different purpose and target audience than a data rescue document containing best practices guidelines and recommendations.

The lack of uniformity among documented data rescue workflows, models and processes posed challenges and comparing a range of different data rescue workflows was not a straightforward task. It was important, when comparing workflows, to determine which activities can be viewed as vital to all data rescue workflows and models, activities falling outside the generic rescue workflow or model, and activities unique to certain disciplines or rescue projects. The investigation into a range of data rescue publications resulted in exposure to different rescue procedures and methods, and an enhanced understanding of steps to be included when drafting a data rescue workflow for this study.

This researcher considers it important to mention that except for the workflow described in Section 3.3.11, evidence of designated or task-specific involvement of the LIS sector could not be established during the analyses of workflows. While the absence of such information does not preclude the involvement of the LIS sector, it does indicate that within the context of this study the collection of data regarding the roles and responsibilities of the research library will be indirectly derived from data obtained during the content analyses of workflows. In other words, while library and information services roles and responsibilities were found to be stated infrequently during data rescue workflows, the nature of tasks and activities will enable the matching of these data rescue tasks with library and information services skills, experience and resources when creating a data rescue workflow model.

A total of 15 different documented data rescue workflows were included in the literature analysis part of the study. The sample selected was thought to be representative of the varied nature of data rescue publications: a range of disciplines, authors, countries, publication outputs, data formats, rescue complexity, and scope of rescue can be detected when viewing the assessed outputs. It was considered vital to investigate data rescue projects that were diverse in nature in order to make better-informed decisions when creating a data rescue workflow as part of this study.

The analysed data rescue models are individually summarised below. Each of the discussed data rescue outputs features a brief description of the rescue project or publication, the data rescue steps followed or recommended, and the contribution of the study. Where applicable, limitations of the published rescue venture are also included. The descriptions of the separate outputs are followed by a table (see Table 3.16) containing a summary of the 15 rescue documentation outputs, and a short discussion on the insight gained from the 15 outputs.

3.3 Data rescue workflows, models and processes

This section features summarised discussions of 15 different data rescue workflows, models and processes scrutinised for this research. The order of discussion is in ascending order, i.e., from the oldest documented outputs to the most recent.

3.3.1 Data rescue workflow: Survey of several projects (manuscript climate data)

This summary is based on a publication by Brönnimann *et al*. (2006) and provides details regarding their published steps performed during manuscript rescue. This data digitisation workflow, published in a climatic research journal (Brönnimann *et al*., 2006), describes important data rescue steps pertaining to the digitisation of handwritten or manuscript climate data. As it only details the digitising part of the rescue workflow, it cannot be considered a complete rescue workflow. It does however contain vital information on aspects to consider, and activities to implement, when embarking on manuscript rescue.

3.3.1.1 Summary of data rescue steps

A summary of the data rescue steps of Brönnimann *et al*. is provided in Table 3.1

Table 3.1: Summary of data rescue steps (Brönnimann *et al*., 2006)

MAIN RESCUE STAGES	ACTIVITIES
Defining the requirements	• This step involves the description of data requirements based on the scientific objectives of the project • A list of quality targets, quantitative or qualitative, was created.
Locating the data	• This step entails finding out what kind of data are available, where, and in what form. • A thorough study of the historical literature, including journal articles and technical reports, is required. This activity is also extremely helpful with respect to meta-information. • It is also worthwhile searching the internet to locate the data. • In many cases the data can only be found on paper in a meteorological archive.
Assessing the data	• This step entails examining properties of the manuscript data and choosing the data source and archive. • The information that is prepared, as outlined in earlier activities, must now be compared with the manuscript data.
Preparing the archive visit and reproducing the data	• During this step, important logistical questions should be answered. • Questions include: ○ What is the length of time required? ○ Will digitising be performed in the archive, or will photocopies be made and digitisation completed elsewhere? ○ Is it necessary to bring a laptop and digital camera? • Travelling well-prepared is vital to the venture
Digitising, formatting and correcting the data	• This step entails the actual digitising of the data. • The activity can involve optical character recognition (OCR), speech recognition, or keying of entries. • After digitising, the data must normally be reformatted. • Data need to be tested and errors corrected.
Validation and description of the quality	• The description should be accurate enough for another user, with different requirements, to decide whether the data are useful.

3.3.1.2 Contribution and limitations

The steps above are described by the authors as a tentative guide (Brönnimann *et al.,* 2006: 137), and deemed to be the most important digitising considerations based on their own experience. A sobering yet honest contribution to the world of data rescue is the mention of manuscript data being very labour intensive, and often not producing any digitisation results (2006: 137).

The authors also state that it is vital, prior to project initiation, to spend time thinking about the characteristics of the data, the scientific requirements with respect to quality and coverage, the metadata, and technical aspects such as reproduction techniques, digitising techniques, and quality control strategies.

An important contribution of the study is the inclusion of recommended steps for the rescue (or in this instance, digitisation) of historic manuscript or handwritten data.

3.3.2 Data rescue workflow: Lost/endangered data (biodiversity and astronomical data)

This summary is based on a publication by Griffin (2006) and provides details regarding her published guidelines for the rescue of historic data. The data rescue workflow described here is a simplistic model containing basic steps when rescuing historic biodiversity and astronomical data, in the context of Canadian scientific research. The workflow forms part of an article published in the *Data Science Journal*, and is authored by Elizabeth Griffin, chairperson of the former RDA Data at Risk Interest Group (2006: 21–26).

The steps described pertain to the rescue of historically important data held by private citizens.

3.3.2.1 Summary of data rescue steps

A summary of the data rescue steps of Griffin (2006) is provided in Table 3.2.

Table 3.2: Summary of data rescue steps (Griffin, 2006)

MAIN RESCUE STAGES	ACTIVITIES
Request datasets in public ownership	• Requests for data at risk could be via '… thoughtful, repeated advertising …' (Griffin, 2006: 24), and can involve the local media.
Summarise the rescue project	• The project is summarised in the request for data, and in adverts. • Benefits of access to datasets would also be stated during this activity.
Involve the data owner	• Owners of the historic data are invited to visit the library/centre performing data rescue.
Copy the data	• Photocopies of data are made at the library/centre.
Return dataset to owner	• Data will usually be returned to the owner. • Alternatively, data can also be donated to the library/centre.
Key the data	• This step entails the keying of data information by a librarian.
Select the data format	• Data format is selected, and data are saved in a nationally agreed format.
Create metadata	• During this activity, all the necessary metadata are added.
Share the data	• The last step entails sending the rescued data and metadata to an appropriate data archive.

3.3.2.2 Contribution and limitations

The rescue steps described in this publication can be regarded as cost-effective and productive. Authored by a chairperson of the former RDA Data Rescue Interest Group, it provides truncated steps for data rescue projects where the data are owned by private citizens. The role of mutual trust is described, and the importance of conveying the benefits of the planned rescue venture is indicated.

3.3.3 Data rescue workflow: Low-cost rescue design (solar radiation data)

This summary is based on the data rescue steps published in a research article during 2008 describing a solar radiation data rescue project at Camaguey, Cuba. The rescue effort is described by the authors as a low-cost data rescue design (Antuña *et al*., 2008: 1507), as it made use of old, out-of-service computers, with key-entering software in FORTRAN running on MS-DOS. The article anticipated that less-developed countries would be able to make use of the low-cost project described within (2008: 1507).

3.2.3.1 Summary of data rescue steps

A summary of the data rescue steps of Antuña *et al*., (2008) is provided in Table 3.3.

Table 3.3: Summary of data rescue steps (Antuña *et al*., 2008)

MAIN RESCUE STAGES	ACTIVITIES
Transfer the original handwritten records to digital form	• Software was designed and created for data key entering; the rescue step included using old personal computers as well as state-of-the-art computers. • Two versions of the software were produced. The first version, created with FORTRAN 77, is for discontinued PCs. It runs in the original MS-DOS operating system, as well as in emulated MS-DOS versions under Windows 3.1, 95 and 98. The second version of the key-entering software was created using Borland Delphi 7 (Pascal) for running on Windows 2000 and XP operating systems. • A number of verification phases guaranteed that the keyed dataset contained exactly the information recorded in the original notebooks.
Process the digital records with software created using the appropriate algorithms	• Two 386 PCs that had been out of service were refurbished and used with MS-DOS for keying data. • Keyed datasets were stored on CD-ROMs together with the software. • For backup purposes, three copies of the CDs were made. • Processing improvements of the digitised database include replacing the nomograms, tables and plots originally used in the manual processing. • The processing software has been designed, quality controlled, and implemented

3.3.3.2 Contribution and limitations

- This data rescue effort, at the time of article publication, had keyed and checked 20 years of solar radiation data.
- The data rescue strategy used (software usable on discontinued XT-chip-based PS systems as well as early Pentium processors) meant that the rescue strategy could be implemented at a low cost by countries or organisations holding climate data records on paper and having limited funding.

- In addition, it was stated that the described rescue strategy could be adapted to other local resources and combined with funded rescue projects for expedited results.

The creation and existence of a bilingual project homepage (English and Spanish) is mentioned in the article, however, this could not be traced at the time of writing (June 2021). A learning point experienced and emanating from this published article is that future format/software obsolescence should be anticipated during the rescue, and it should be ensured that a migration plan is in place.

3.3.4 Data rescue workflow: reBIND (biodiversity data)

This summary is based on a description of the reBIND data rescue workflow, as published in a scholarly article by Güntsch *et al*. (2012). This data rescue publication provides the steps implemented during the rescue of threatened biodiversity data using the reBIND workflows. As stated in the article, biodiversity data collected during research often lack a preservation and accessibility strategy and run the risk of becoming outdated and eventually lost (2012: 752).

3.3.4.1 Summary of data rescue steps

A summary of the data rescue steps of Güntsch *et al*. (2012) is provided in Table 3.4.

Table 3.4: Summary of data rescue steps (Güntsch *et al*., 2012)

MAIN RESCUE STAGES	ACTIVITIES
Export and transformation of data	• Source data are exported and transformed into an accepted and well-documented XML-based community standard. • The transformation process is carried out using the BioCASE provider software, stated to be a widely used tool for linking up heterogeneous collection databases to biodiversity information networks (2012: 753).
Validation of the generated XML	• This step ensures that the file complies with the target format.
The validated XML-file is stored in a native XML database	• This phase also includes the creation of standardised metadata records. • Two metadata standards are applicable: the Directory Interchange Format (DIF 2010) and Ecological Metadata Language (EML). • The Dublin Core Application Profiles (DCAP 2008) allow the mixing and matching of terms from different vocabularies and thus can be used to combine the standards.
Opening the data to international information infrastructures	• The data become visible and usable to applications such as biodiversity portals and virtual research environments.

3.3.4.2 Contribution and limitations

The reBIND project provides an efficient and well-documented workflow to be used when rescuing threatened biodiversity data (2012: 752). Additional benefits are stated to be:

- its applicability to other data domains and types,
- the project making use of freely available open-source software components,
- the low cost effective workflow,
- the transformation into contemporary standardised formats,
- the fact that data are deposited into a flexible storage system, and
- the connection to relevant international data infrastructures.

Moreover, the reBIND project has published an entertaining animated online video, available in English as well as German, explaining the reBIND project (Biodiversity Informatics @ BGBM, 2011; reBIND, 2015).

While not necessarily drawbacks, the following should be kept in mind when considering the reBIND workflow:

- it is not clear whether the workflow could be applied to the rescue of paper-based data, or if it is applicable to threatened data in a digital format only, and
- the authors state that different initiatives and institutions should adopt software and workflows to build their own repositories, as the scope and purpose might differ from the vanilla version (2012: 755).

3.3.5 Data rescue workflow: IEDRO (climate and environmental data)

This summary is based on the data rescue guidelines obtained via the website of the International Environmental Data Rescue Organization (IEDRO). This data rescue workflow describes the basic steps followed by IEDRO during their data rescue projects involving climatic data (IEDRO, 2014). IEDRO is an entity organising and coordinating data rescue efforts in countries across the globe, usually involving the relevant country's meteorological services and climate researchers.

3.3.5.1 Summary of data rescue steps

A summary of the data rescue steps of IEDRO (2014) is provided in Table 3.5.

Table 3.5: Summary of data rescue steps (IEDRO, 2014)

MAIN RESCUE STAGES	ACTIVITIES
Coordinate with a country's National Meteorological and Hydrological Services (NMHS)	• IEDRO coordinates with the country's NMHS team to train staff in conducting the data rescue process. • Staff are taught how to: ○ locate the data, ○ organise the data, ○ create an inventory of the data, and ○ store the data.
Assist the NHMS to prepare data for digitisation	• IEDRO funds and delivers scanning and digitisation technologies, including computers, digital cameras, camera stands, and scanners. • IEDRO trains NMHS personnel to scan precipitation strip charts and photograph hydro-meteorological alphanumeric records. • IEDRO assists NMHS personnel in managing file organisation, file naming, and transmission of photographed and scanned images to IEDRO.
Manage the digitisation process	• IEDRO receives and manages photographed and scanned images. • IEDRO is responsible for the coordination and management of the digitisation process. • IEDRO submits the data to free and open climate research databases, such as the National Climatic Data Center (NCDC) databases, and returns the digitised data to the country's NMHS.
Assist the NMHS in using their data	• IEDRO trains NMHS in climate applications using the rescued data.

3.3.5.2 Contribution and limitations

IEDRO is one of the foremost global entities involved in data rescue, and their contributions entail funding, collaboration with nations' climatologists, training, data rescue, and promotion of rescued data or ventures. The simplistic data rescue steps found on their website can be regarded as a shortened version of more complete published data rescue guidelines, such as those of the WMO (2016) or Kennedy (2017). It states the most vital steps to be followed when rescuing historic paper-based media, and while not complete or detailed, is sufficient to give followers an idea of the gist of a data rescue project.

Training, returning data to owners, making use of free and open data archives, and data applications follow-up are additional contributions of IEDRO's data rescue stance.

The absence of recent updates on the IEDRO website can be regarded as a limitation.

3.3.6 Data rescue workflow: World Meteorological Organization (hydrological data)

This summary is based on a publication by the WMO (2014) and provides details on the organisation's description of best practices regarding the rescue of historic hydrological data. This well-known data rescue manual, published in 2014, is freely available from the website of the WMO. It features similarities with the WMO climate data rescue manual discussed under a subsequent heading (see Section 3.3.12), but due to several differences between the two manuals it was deemed to be worthy of a separate heading and discussion.

As stated in the manual's preface, the book includes a section on the rationale for hydrological data rescue, the rescue benefits, appropriate rescue methods, sound data management practices and systems, procedures for securing rescued data into the future, and storage in an international database (WMO, 2014: v).

3.3.6.1 Summary of data rescue steps

A summary of the data rescue aims of the WMO (2014) is provided in Table 3.6.

Table 3.6: Summary of data rescue aims (WMO, 2014)

DATA RESCUE AIMS
Creation of an inventory of data holdings
Rationalisation and indexing of physical records/data holdings
Identification of data holdings of genuine importance
Improving storage conditions for physical data holdings and storage media
Digitisation of time series data
Extraction of data from obsolete storage media (physical and digital)
Storage of digital data in future-proof formats and storage media
Storage of time series data within a hydrological database
Establishing procedures for the ongoing use of data archives

The manual defines the step entailing the creation of an inventory as the most crucial step in the rescue process. The necessity of an inventory was stated to be its potential in helping identify data at risk, helping to prioritise the data rescue effort, estimating the resources required, and helping with the formulation of a plan. Detailed steps for setting up an inventory, and what this entails for (i) paper data and (ii) obsolete digital data, are provided.

Apart from the listing of aims, the manual contains guidance on several data rescue activities to be performed, including the drafting of a data rescue plan. In addition, practical tips regarding aspects such as assembling a data rescue team, improving paper storage conditions, and good practice guidelines are supplied.

3.3.6.2 Contribution and limitations

The value of this data rescue document can be found in its specificity to the hydrological sciences. In addition, the manual's ease of use and range of data rescue activities included contribute to its value as a data rescue reference document. The various data rescue aims, listed within the document, are deemed to be a necessary and welcome addition to data rescue literature, forcing the data rescuer to view a data rescue project from a different angle.

3.3.7 Data rescue workflow: OGS-SNAP (seismic data)

This summary is based on a description of a seismic data rescue project published in a research article by Diviacco *et al.* (2015). The data rescue workflow describes the rescue and dissemination of substantial amounts of vintage seismic data in an internal project of the Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS). The focus of the rescue effort was the rescue of magnetic tape before degradation and scanning and converting paper data into a usable format. In addition to these aspects, the project attempted to envisage additional ways in which the data could be used in future. The importance of ensuring the data's future usage resulted in a processing path, or workflow, which includes more aspects than simply rescuing threatened data.

3.3.7.1 Summary of data rescue steps

A summary of the data rescue steps of Diviacco *et al.* (2015) is provided in Table 3.7.

Table 3.7: Summary of data rescue steps (Diviacco *et al.*, 2015)

MAIN RESCUE STAGES	**ACTIVITIES**
Media recovery	• A large quantity of data, particularly those held on old magnetic tapes, was undergoing progressive deterioration. A major effort to rescue the data was vital. • It was found easier to fix older technology problems than more recent ones, as it was easier to find spare parts to mend old tape drives than more modern versions. • The whole OGS archive was transcribed onto a RAID Hard Disk storage system that would be routinely and constantly mirrored in an independent physical location. • Similar storage hardware was installed at the OGS main site and at another research institute distant from OGS connected by a high-speed network link. This prevents critical data loss in cases of fire, electrical failures, or extreme temperatures. • In short: data owned by OGS were rescued via transcription onto secure mirrored storage systems.
Dealing with paper-based data	• The project's main data format is not paper data. • OGS does not address ancient manuscripts, although in some domains such as archaeology or natural hazard studies, scanned and geo-referenced ancient maps are common. The OGS archives contain mainly stacked seismic sections and location maps plotted on plain or tracing paper.
Conversion to SEG-Y	• Commercial and open-source software were used to convert seismic images to digital SEG-Y data. OGS had already used this type of software in another EU-funded project.
Positioning information added	• Data are georeferenced to be of use; digital positioning information was available for recent data but had to be performed manually for older data.

MAIN RESCUE STAGES	ACTIVITIES
Creation of data documentation and data intelligibility	• Concern was expressed regarding the retirement of key people with knowledge of the vintage data collection. • Concern was expressed regarding incomplete data collection details, and how this results in difficulty when trying to restore data to full intelligibility. • Incomplete data collection details result in difficulty when trying to restore data to full intelligibility. • The following tasks were undertaken: ○ digitisation of old documentation was performed in order to ensure its preservation, and ease of locating and use, ○ eccentric configurations were standardised to enable seamless ingestion by modern interpretation software, and ○ rescue efforts ensured that data documentation and metadata relate to the data, in order for data to be found and used.
Processing of seismic data	• Specific processing paths were implemented to deal with discontinuities and improve the quality of digitised seismic data restored from paper sections.
Addressing issues due to tape reading	• Magnetic tapes revealed large gaps in processed data. • Paper copies of the original processing existed, and a process of reverse processing was used to recreate missing field records.
Integration with other data	• All converted SEG-Y data, after positioning correction, were loaded, together with more recent data, into IHS Kingdom[22] .
Data were shared	• A web-based platform (http://snap.ogs.trieste.it) was developed to disseminate rescued data within the scientific and commercial communities, where all data and metadata that have been uploaded have been made available.
Adherence to standards	• The project adhered to data quality standards. • The project adhered to metadata standards.

[22]IHS Kingdom software provides the functionality needed for portfolio management from prospect to production, with interpretation to microseismic analysis and geosteering resulting in interpretation and modeling sharing of data and more decision making

Additional aspects integrated into this data rescue project include the activities listed below.

- **Data enhancement:** this was performed to ensure that data limitations connected to older technology were overcome.
- **Data integration:** this was performed to ensure that different data types could be assimilated within one data space.
- **Data discovery:** a web-based framework named SNAP (Seismic data Network Access Point) was developed that ensured end users could search, locate and preview the data.

3.3.7.2 Contribution and limitations

This project comprised several phases and activities, from restoration of data to clearance, to integration and to dissemination of the rescued data. The authors state that all the described steps are necessary, as it is only by using the data that their full value emerges. All steps must be considered and included, leading to a harmonised data space, standards, and software interoperability (Diviacco *et al*., 2015).

From an end user point of view the steps included here result in easy data discovery, access, and usage of the recovered data. This, in turn, contributes towards a collaborative environment for researchers and institutions.

The publication's most valuable contribution is its description of the recovery of older magnetic tape data, and the steps followed to make it available to the wider research community.

For the reader not familiar with magnetic tape data formats and its accompanying recovery processes, standards and databases, the descriptions can at times be difficult to understand and follow.

3.3.8 Data rescue workflow: Recovery of 'dark' data (zooplankton data)

This summary is based on a publication by Wiebe and Allison (2015) and provides details regarding their publication on the rescue of zooplankton data. This data rescue workflow describes, by means of a journal article, the rescue of dark zooplankton data collected in the 1970s and 1980s during 34 cruises to the Northwest Atlantic Ocean (Wiebe & Allison, 2015). The data were locked in notebooks and old digital file formats, and the objective of the rescue project was to deposit the data in a modern publicly available data repository.

3.3.8.1 Summary of data rescue steps

A summary of the succinct data rescue steps of Wiebe and Allison (2015) is provided in Table 3.8.

Table 3.8: Summary of data rescue steps (Wiebe & Allison, 2015)

MAIN RESCUE STAGES	ACTIVITIES
Determine that data are at risk	• This stage entails the recognition that data reside in notebooks, cruise reports, old computer files and technical reports, and are at risk.
Ensure that rescue is a team effort	• The input of the scientists who conducted the research is regarded as vital.
Comprehensive involvement of data sources	• All cruises involved in collecting data were listed, and then the data sought.
Error correction	• Errors in datasets were identified, and corrected
Keying of data	• The assembled information was typed onto a spreadsheet.
Conversion of associated tape data	• Associated environmental data were also assembled and converted.
Sharing of data	• Data from 306 tows deployed from 34 research cruises in the Northwest Atlantic, recorded in a spreadsheet, have been put online at the Biological and Chemical Oceanography Data Management Office (BCO-DMO).
Ongoing management of data	• Data management purposes required that the data be divided into several distinct groupings: one for cruise metadata, one for total biomass and euphausiid catch data, and one for environmental CTD data.

3.3.8.2 Contribution and limitations

While the intent of the article was not to provide detailed steps followed during the rescue mission, the brief description of steps provides insight into the main activities when rescuing paper and early digital data.

The emphasis on data rescue being a team effort, and the attention paid to subsequent management of the rescued data, are vital rescue concepts taken note of in this research.

3.3.9 Data rescue workflow: Scientific data at risk of loss (discipline-agnostic)

This summary is based on a publication by Downs and Chen (2017) and provides details regarding their published guidelines for the rescue of scientific data at risk of loss. The workflow described below was published in a book chapter detailing the efforts of the NASA Socioeconomic Data and Applications

Center (SEDAC) to rescue the Millennium Ecosystem Assessment (MA) collection of scientific data (Downs & Chen, 2017). The main contribution of the document is its description of issues present when rescuing data that had not been fully curated by an archive.

While not a detailed process, it states the main steps implemented during the rescue project.

3.3.9.1 Summary of data rescue steps

A summary of the data rescue steps of Downs and Chen (2017) is provided in Table 3.9.

Table 3.9: Summary of data rescue steps (Downs & Chen, 2017)

MAIN RESCUE STAGES	ACTIVITIES
Starting point of rescue effort	• This stage entails obtaining knowledge of the website closure.
Value of data on website realised	• Relevant parties realised the data value, and recognised the risk of data loss.
Assessment performed	• An initial assessment was performed to determine if data have relevance to objectives and future needs.
Reason for data rescue established	• A basic data rescue effort was implemented to enable future discovery and use of the data collection.
Data rescue group appointed	• A data rescue group was formed/appointed, consisting of scientists, representative users and other experts
Additions to the rescue effort	• The data rescue effort also included limited value-added efforts.
Data files organised	• Data files were organised into six datasets for online dissemination. • Sets contained original files and original formats, with supplementary information obtained from various sources.
Rights were established	• Authorship and dissemination rights were clarified, with the help of authors.
Data documentation issues dealt with	• This activity entails referring end users to published assessment reports for detailed information on scientific background of data, and its use in analysis.
Metadata and other outputs collated	• Relevant, associated reports and data were analysed to create a collection description, summary, and metadata record for each dataset.

MAIN RESCUE STAGES	ACTIVITIES
Analysis and quality control of data (prior to dissemination)	• Each dataset was accessed and analysed to ensure that quality was not compromised, and data could be accessed by end users. • Each dataset received was reviewed by (i) internal scientists and staff, and (ii) selected external users. • The configuration management board reviewed all comments received. • Corrections and descriptions were completed prior to public release. • Each set was archived to ensure long-term preservation.
Dissemination of data	• The website data portal stated that rescued data might not meet expectations, and that there was minimal support and documentation. • A landing page for each dataset, with a description and a DOI, was made available. • Additional webpages automatically opened when requesting data download, documentation, or metadata. • Users were required to log in to download data, and downloads were free of charge.

The chapter on data rescue stated that the following lessons were learned during the rescue effort:

- Data repositories that engage in data rescue efforts need an established selection-and-appraisal process to select the data for long-term preservation.
- A complete assessment of the envisaged data rescue should be conducted to identify the effort and resources needed.
- When considering competing priorities for limited funding, the potential value of scientific data to future scientific research should be considered.
- Alternatively, it may be worth establishing whether members of the scientific community or other role-players might be able to contribute to or support the rescue venture.
- Unlike typical data curation efforts that are conducted at scientific data centres, data rescue may require divergence from regular data curation procedures where necessary. The extent of such divergence may depend on the state of the data when it is acquired, the availability of the data producers and availability of data documentation.

- With the passage of time, the difficulty of any data rescue will inevitably increase, as metadata and data documentation become more difficult to access. It is therefore important to respond quickly when the need for data rescue has been identified.
- Early identification of candidates for data rescue and the initiation of immediate action should increase the success of data rescue efforts.
- Data rescue efforts will benefit from researchers being familiar with the data being rescued. In the rescue project described here, familiarity facilitated access to key scientists and critical information needed to document the data and determine access rights.

3.3.9.2 Contribution and limitations

The main contribution of the document is its description of issues present when rescuing data that an archive had not originally fully curated.

The following aspects mentioned in the publication can be regarded as its major contributions:

- The rescue project implemented a data rescue group.
- The rescue project performed a thorough assessment of the data and resources required prior to the start of actual rescue activities.
- Familiarity with the data was stated to be of immense value to the rescue venture, as was collaboration with scientists.
- The role of metadata and data documentation was emphasised.
- Data management was stated to be a vital component of data rescue, together with the funding of data management activities.
- An interesting activity was the creation of a website for the sharing of rescued data.

3.3.10 Data rescue workflow: LifeWatchGreece Project (biogeographic data)

This summary is based on biogeographic legacy data rescue steps described in a publication by Mavraki *et al.* (2016). This publication describes the digitisation of a series of historical biogeographic datasets based on the reports of the 1908–1910 Danish Oceanographical Expeditions to the Mediterranean and adjacent seas. The datasets digitised cover more than 2 000 samples taken at 567 stations during 1904 to 1930, in the Mediterranean and adjacent seas. The samples resulted in 1 588 occurrence records of pelagic polychaetes, fish and calcareous algae. Basic environmental data as well as meteorological conditions are included for most sampling events. In addition to the description of the digitised datasets, this article provides a detailed description of the problems encountered during the digitisation of this historical dataset, and a discussion on the value of such data.

The digitisation of the expeditions' data formed part of the activities of the LifeWatchGreece Research Infrastructure project. Mavraki *et al.* (2016: 1) state that the aim was to safeguard public data availability by using an open access infrastructure.

3.3.10.1 Summary of data rescue steps

A summary of the data rescue steps of Mavraki *et al*. (2016) is provided in Table 3.10.

Table 3.10: Summary of data rescue steps (Mavraki *et al*., 2016)

MAIN RESCUE STAGES	ACTIVITIES
Digitisation of the data	• Manual digitisation of the information on samples and their associated metadata was performed. • Independent manual digitisation of the three biological publications, including all available data on samples, sampling metadata, species occurrences, and abundances also took place.
Independent quality control of each dataset	• This stage entailed checking if the location of coordinates falls on land or outside the study area. It also checked for standardised taxon names, inconsistencies in dates, depths, locations, sample numbers, abundances, and use of gear.
Match sampling information in the biological publications against the introductory volume	• The introductory volume was used as a reference dataset. Unique sample IDs were ascribed to these samples.
Implement a second round of quality control	• A similar round of quality control activities (see earlier bullet) was performed on the integrated datasets.
Obtain missing data	• Missing data in a dataset were either derived from other datasets or from other scientific and historical publications.
Creation of metadata	• Metadata were created for each dataset.
Creation of a uniform dataset	• A uniform dataset, allowing the seamless integration of data emanating from the same core expeditions, was created. Datasets were published separately, with separate metadata and individual URLs.
Make all datasets available as Darwin Core (DwC) Archives	• Datasets were made available with a CC-Zero License via the Integrated Publishing Toolkit of the Mediterranean Ocean Biogeographic Information System, supported by the LifeWatchGreece Infrastructure.
Formatting the data	• Data had to be formatted according to the DwC schema for dissemination through global biogeographic databases (e.g., Ocean Biogeographic Information System (OBIS), Global Biodiversity Information Facility (GBIF)).
Provision of an easily usable dataset	• As the format mentioned in the previous bullet is not necessarily user-friendly for human users, a version of the datasets in a non-standardised but more easily usable format was provided as supplementary materials to the article.

3.3.10.2 Contribution and limitations

The described steps are regarded to have made a valuable contribution with regard to the rescue of biogeographic legacy data, and the reasons for this are listed below.

- The authors describe this work as an initial step towards the digitisation and harmonisation of a complex set of related data, distributed across different historical publications.
- Legacy data are stated to present a challenge when it comes to their integration into a harmonised, comparable, quality-controlled format.
- Data might be inconsistent across different publications, or even within the same publication, and obvious errors are identified during digitisation. Often, information is missing, or spread across several publications. As the original authors cannot be consulted anymore, some of these problems will remain unresolved, and currently there is no standard strategy on how to deal with such issues.
- The authors state that it was vital to have a team able to integrate knowledge and perspectives, and that experts of different disciplines were required to interpret data and reduce final dataset errors.
- Biodiversity legacy data are stated to have great value, but the rescue of such data is described as time consuming, tedious, and cannot be performed without the involvement of skilled expertise.
- The publication states that multidisciplinary working groups should be formed, as such groups will be able to develop new methods to harvest and use legacy data. The example of librarians and natural history museums initiating and facilitating access to such data, and data scientists developing tools and workflows to semi-automate data extraction, with biologists and modellers harnessing the data in global models, is mentioned.

3.3.11 Data Rescue Workflow: DataFirst (sociological data)

This summary is based on a description of apartheid era sociological paper-based data rescue, as described in a presentation delivered at the RDA/CODATA Workshop on the Rescue of Data at Risk, September 2016, in Boulder, USA (Woolfrey, 2016). Confirmation of the rescue steps were also obtained through an article on the rescue entity's website (DataFirst, 2022). The rescue project was managed by DataFirst, a prominent South African research data service entity. The rescue project involved paper records from two survey projects from SA's apartheid past, the records of which have been in storage in the decades since the surveys were conducted. As such, the data were underutilised for research, and were at risk of total loss due to physical decay.

Led by DataFirst, the rescue project involved DataFirst teaming up with several South African universities to convert the rich historical data to research-ready formats and thereby opening them up for new analyses.

3.3.11.1 Summary of data rescue steps

A 2016 workshop presentation shared with this researcher by the presenter contained a slide showing the chronological flow of the rescue project's main steps. A research article detailing the rescue project was stated to be in the pipeline; however, at the time of writing of this chapter it was not yet published, and not available for perusal.

A summary of the data rescue steps of Woolfrey (2016) is provided in Table 3.11.

Table 3.11: Summary of data rescue steps (Woolfrey, 2016)

MAIN RESCUE STAGES	ACTIVITIES
Discovery	• This step entails identifying the source and content.
Audit	• This step entails compiling a detailed inventory of records to be digitised.
Agreement	• This step entails obtaining permissions from data holders.
Capture	• This step entails converting non-digital records to digital research-ready formats.
Quality Check	• This step entails checking data against the original source, and data anonymisation.
Documentation	• This step entails tracing and digitising supporting documentation.
Description	• This step entails creating metadata containing background and usage information.
Publishing	• This step entails the online dissemination of the dataset as public use data.

3.3.11.2 Contribution and limitations

The rescue project involved survey data collected during two apartheid era surveys and contains valuable sociological information pertaining to SA's troubled past. The 1948–1950 Keiskammahoek Rural Survey provided details on experiences of family life and work after being forced into centralised residential areas and a migratory labour system, and the 1980–1981 Surplus People Project provided details of the experiences of 10 000 people forcibly removed during the apartheid era (Woolfrey, 2016).

The contribution of the project towards this study is that it is the only published example of a South African rescue project providing details of the workflow steps followed. In addition, the rescue project also showcased the importance of collaboration, in that various institutes were involved in the project. The various sources reporting on the project's success mentioned the involvement of not only DataFirst, but also other South African institutes including the Neil Aggett Labour Studies Unit, Rhodes University, University of Cape Town, and the University of KwaZulu-Natal (Woolfrey, 2016; Bernardo & Khwela, 2018; Khwela, 2018).

The fact that the rescued data were made available publicly is another contribution emanating from this South African rescue venture.

A limitation at the time of writing was the lack of more details on the rescue steps followed, the subtasks involved, and roles and responsibilities of participating parties. It is assumed that such information will be covered in a research article that is due to be published.

3.3.12 Data rescue workflow: World Meteorological Organization (climate data)

This summary is based on a publication by the WMO (WMO, 2016) and provides details regarding their published guidelines for historic climate data rescue. The WMO has published two data rescue workflows: the 2016 'Guidelines on Best Practices for Climate Data Rescue' (WMO, 2016), and the 2014 'Guidelines for Hydrological Data Rescue' (WMO, 2014; see Section 3.3.6 above). The two publications, despite having several similarities, contain major differences with regard to outlay, activities, and steps included. As a result of stated differences, both WMO workflows are discussed separately in this chapter; each under its own heading, 3.3.6 and 3.3.12, respectively.

The 2016 WMO publication is often recommended by sources other than the WMO and is seen as one of the foremost publications in the data rescue sphere. In addition, the WMO's training efforts and contribution towards curbing the loss of documents due to improper storage are mentioned by Brönnimann *et al.* (2018: 30). The WMO itself describes the climate guide as a set of 'recommended best practices' (2016: iv), and states that the rescue guidelines can be used and applied in disciplines other than the climatic sciences (2016: vii).

3.3.12.1 Summary of data rescue steps

A summary of the data rescue steps of the WMO (2016) is provided in Table 3.12.

Table 3.12: Summary of data rescue steps (WMO, 2016)

MAIN RESCUE STAGES	ACTIVITIES
Archiving of paper data	• During this stage, data are located, preserved and stored, and a holdings inventory created.
Imaging of paper-based media	• A master image inventory is created, media is imaged and validated, the inventory updated, and an image file is created for each CD/DVD.
Digitising of paper values	• This stage entails the creation of a digital data inventory, followed by the keying in of entries, and quality control of the keyed data.
Archiving of digital media	• Activities included during this step include cross-checking of printed media, images and digital data, implementing a daily backup routine, dispersing multiple copies, and refreshing and migrating the data every five years.

Additional sections of the climate data rescue manual include appendices containing:

- information related to data rescue infrastructure, equipment, supplies and personnel,
- data rescue guidance,
- prioritisation of data rescue/data,
- electronic imaging techniques,
- other digitisation methods, and
- a data rescue checklist.

3.3.12.2 Contribution and limitations

The manual's generalisability and comprehensiveness make it an essential tool when implementing a data rescue project for the first time. Other advantages of the document are the user-friendly nature of the described steps and guidelines, the detail accompanying each step, and the supplementary section of the manual describing additional data rescue tasks and activities.

Minor weaknesses and limitations of this model include the fact that it is intended for the rescue of paper data, is focused on the climate sciences, and does not provide sufficient detail on project initiation, data assessment, the concept of a data management plan, or project closure steps.

3.3.13 Data rescue workflow: RDA (discipline-agnostic)

This summary is based on data rescue features described on the website of the RDA Data Rescue Interest Group. Even though the interest group changed its name to the Data Conservation Interest Group during 2019 (RDA, 2019), and implemented a change in focus, the group had contributed to the data rescue discipline during the years leading up to the name change. The group's contributions are also mentioned in Section 2.6.5: Data rescue entities and interest groups. Despite the many valuable contributions, the RDA Data Rescue Interest Group had not published its own data rescue workflow, or detailed guidelines on its website. It is assumed that such guidelines are evidently not needed by RDA Interest Group members. Data rescue steps published by the now historic interest group can be described as brief and concise.

3.3.13.1 Summary of data rescue steps

A summary of the RDA data rescue guidelines (as opposed to rescue steps or a workflow) is provided in Table 3.13.

Table 3.13: Summary of data rescue steps (RDA, 2019)

MAIN RESCUE STAGES	ACTIVITIES
Motivation and starting point	•Various data discovery possibilities exist; some discoveries are by accident, others via dedicated searching.
Examine the physical condition of the data	•Fragility of the records needs to be taken into consideration. •Condition of data will determine the type and feasibility of data rescue.
Importance of metadata	•Availability of metadata can influence the feasibility of data rescue. •Subject expertise is often required when creating metadata. •Certain repositories require adherence to specific metadata standards.
Importance of calibration and other subsidiary files	•Availability of data documentation can influence the feasibility of data rescue. •Subject expertise is often required.
Select required digitising and scanning equipment	•A flatbed scanner, microphotometer, or digital camera are examples of digitisation equipment.
Involvement of experts	•Even though citizen scientists and volunteers can play a significant role, the involvement of archivists is often required.
Data management	•Drafting a data management plan, and adhering to it, are vital data rescue activities. •Contingency plans, secure storage, and plans for long-term preservation are of critical importance.
Make decisions regarding preservation	•It is important to consider whether all heritage data should be preserved. •One should consider the research potential and value of data, data management ability, costs associated with repositories, and faithfulness of digitisation.
Consider the repositories to be used	•Ideally, a trusted repository should be selected. •It is important to consider the costs forming part of repository use.
Education to support the need for preservation	•Rationale for data rescue needs to be marketed/promoted, and form part of education/training.

3.3.13.2 Contribution and limitations

While not providing data rescue steps, it is succinct in its inclusion of crucial data rescue aspects to be considered. Examples of these vital concepts are the references to assessment and examination of the data, role of metadata and data documentation, the significance of data management planning, the importance of selecting suitable repositories, and the crucial role of data rescue awareness training.

3.3.14 Data rescue workflow: OBIS Canada Cookbook (marine species data)

This summary is based on a publication by Kennedy (2017) and provides details regarding the rescue of marine species data. This comprehensive data rescue workflow, published in 2016 and edited in 2017, was compiled specifically for the rescue of Canadian marine species' occurrence data, and makes provision for the upload of the data to Canada's Ocean Biogeographic Information System (OBIS). The creator of this workflow refers to it as a 'cookbook'; as stated by the author, data rescue activities are often complex and difficult tasks, while a cookbook is easy to understand, the practical information therein easy to follow, and there is an absence of too much technical jargon (Kennedy, 2017). This data rescue cookbook is therefore applicable for new data rescuers as well as experienced data management teams.

The publication is freely available online on the Coastal and Ocean Information Network Atlantic (COINAtlantic) website and features detailed information about the rescue steps to be followed within each of the workflow's 12 main phases. In addition, supplementary documentation, and information regarding related rescue activities such as a metadata template, basic processing steps, a checklist, and advice on setting up a literature search project for a data rescue mission are supplied. Moreover, as this data rescue manual comes in the form of a cookbook, the author has included with each of the main phases the method and relevant rescue ingredients.

3.3.14.1 Summary of data rescue steps

The data rescue steps contained in this manual are detailed and comprehensive and require significant investment of time by the reader to study all the data rescue phases and steps that are included. With the full rescue workflow and its appendices spanning more than 110 pages, a summary of the data rescue steps is provided in Table 3.14.

Table 3.14: Summary of data rescue steps (Kennedy, 2017)

MAIN RESCUE STAGES	ACTIVITIES
Identify the sources of data (pp. 14–18)	• Define data of interest to OBIS • Compile keywords for the identified data • Identify potential sources of data relevant to the rescue project • Determine accessibility, restrictions, possibility of aggregation with other datasets, and if publication provides a summary of datasets • Flag identified datasets missing OBIS-required information • Consider a vital question: 'Does the dataset need to be rescued?' (Kennedy, 2017: 15)
Create inventory of data (pp. 19–23)	• Identify software to be used • Compile references associated with identified datasets • Review the content and standardise the inventory • Develop vocabularies • Design and revise the inventory table design
Digitise the data (pp. 24–29)	• Identify software to be used • Obtain access to the source material • Digitise the data • Realise that '… if source material is digital, then this task is completed …' (Kennedy, 2017: 24) • Review the digitised data
Describe the dataset (pp. 30–35)	• Obtain access to the source material • Create a template that can be used to gather information from primary investigators or from published information • Create a readme document or worksheet to be associated with this dataset • Consider that '… information may be stored in whatever format works for a specific project …' (Kennedy, 2017: 30) • Compile, organise and standardise notes • Consult contacts to obtain missing information • Optional: publish metadata at this time if the dataset is stuck in the processing queue

MAIN RESCUE STAGES	ACTIVITIES
Archive the data and its information (pp. 36–37)	• Ensure that source and processed datasets and associated information are stored safely • Review the digitised copy of the dataset and compare it with the original data; revise filenames if necessary • Create folders to group and organise content associated with the resource: devise a naming scheme for folders, create subfolders for readme files, data, correspondence, and processing scripts • '… identify a stable file storage location and create folders for individual resources …' (Kennedy, 2017: 36) • '… transfer content of resources to the long-term storage location …' (Kennedy, 2017: 36) • Append copies of new versions of data • Zip the archived files should storage space be an issue • '… confirm that data files are backed up on a regular basis …' (Kennedy, 2017: 36)
Create a standardised dataset (pp. 38–44)	• Append a list of Darwin Core (DwC) terms to dataset • Compare copy with the latest version of DwC terms online and update if required • Map dataset content to DwC terms • Consult with OBIS data management team and address issues • Perform document mapping • Identify unfamiliar terms added to vocabularies, new synonyms added to registers, and new areas added to gazetteers or enhanced definitions • Ensure that the outcome of this step is a cleaned, reformatted and standardised product, with quality-controlled information, updated instructions and updated worksheets
Create metadata (pp. 45–49)	• Review the readme document or worksheet to be associated with this dataset • Review the metadata and seek feedback from source, experts, and data rescue project leaders • Publish the metadata

MAIN RESCUE STAGES	ACTIVITIES
Make the data accessible (pp. 50–53)	• Upload the described dataset to a server where it can be publicly accessed • Verify the number of records uploaded • Map fields to DwC terms • Populate/review EML metadata in collaboration with the OBIS node data management team • Review resource and consult with data rescue team and data provider • Update dataset tracking status in the inventory to 'private' • Publish the resource once the review/embargo period has been lifted
Share the data (pp. 54–56)	• Share the content with global partners and make the content available from the OBIS portal • Add the content to the OBIS database and make it accessible on the OBIS portal
Promote the data (pp. 57–62)	• Upload data files to resource and verify the number of records uploaded • Map fields to DwC terms • Populate and review metadata in collaboration with OBIS node data management team • Create a data management plan and ensure that there is promotion of best practices regarding data management • Address issues such as the danger of the data being uploaded and then abandoned, and the danger that '… another group will attempt to process the same dataset …' (Kennedy, 2017: 60) • Wrap up the project, which includes consideration of recommendations, lessons learned, and thanks to parties

The remainder of the data rescue manual is devoted to activities that can complement the basic rescue steps, references and appendices.

3.3.14.2 Contribution and limitations

As stated earlier, Kennedy's cookbook is an example of a detailed workflow, with each of the main phases divided into sub-steps. As is the case with the book's main phases, sub-steps of the workflow are detailed in nature. In addition, some of the sub-steps are also split, resulting in a comprehensive data rescue manual.

The model’s minor limitations are as follows:

- the document is discipline-specific and country-specific; some of the activities and sections will be applicable to the rescue of marine/oceanographic data in Canada only,
- sections of the document include advanced activities and calculations not familiar to a novice data rescuer, and
- the manual is voluminous in size; considerable time is required to work through the described steps.

In short, this document, despite being highly discipline-specific (and at times area-specific), contains useful and detailed step-by-step data rescue instructions. An additional functional feature is the inclusion of rescue roles and responsibilities, an aspect often missed in other documented data rescue literature.

3.3.15 Data rescue workflow: Maynooth University (precipitation data)

This summary is based on a publication by Maynooth University (2017) and it provides details regarding their published guidelines for data rescue involving classroom learners. This data rescue workflow details the rescue of rainfall data collected in Ireland from 1860–1939. It is possibly the first documented instance of data rescue being performed by university students; the project involved undergraduate geography students at Maynooth University who successfully rescued 1 400 years of Irish rainfall data (Maynooth University, 2017). In addition to this novel feat, the workflow (in poster format) has been highlighted by the WMO on their data rescue pages as an exemplar of best data rescue practice (Maynooth University, 2017). The details of this workflow effort have been published in various sources: it is available as a poster at the University’s ICARUS Climate Research Centre (Maynooth University, 2017), it was presented as a talk at the 2017 meeting of the European Meteorological society (Ryan *et al*., 2017), and published as a scholarly article (Ryan *et al.,* 2018).

3.3.15.1 Summary of data rescue steps

A summary of the Maynooth University data rescue steps (2017) is provided in Table 3.15.

Table 3.15: Summary of data rescue steps (Maynooth University, 2018)

MAIN RESCUE STAGES	ACTIVITIES
Locate data at risk	• Digital images of annual rainfall sheets were obtained from Ireland's National Meteorological Services Archive.
Determine value of rescue project	• As a lead-in to the assignment, students were given a guest lecture by staff from the Irish National Meteorological Service to convey the scientific and cultural importance of the data they would be working with
File conversion performed prior to rescue	• Upon receipt of the digital images, the file format was converted from a PNG file type to a JPEG format, reducing the file size without any apparent loss of resolution. • This facilitated the distribution of images to students and avoided potential logistical issues concerning the access of large files.
Role allocation	• Each student was assigned 18 annual rainfall sheets to transcribe, and each sheet was assigned to two different students.
Double key the data	• Such double-key data entry is a widely used method of quality control to detect incorrectly keyed information. For this project, double-key entry was necessary for the allocation of student grades.
Create structured directories	• Individual directories containing 18 randomly selected annual rainfall sheets were created and distributed to each student via Dropbox along with a Microsoft Excel template for keying the data.
Use a project website for accessing and depositing data	• Students were provided a link from which they downloaded their personalised directory and performed the transcription component of the coursework. Students were given five weeks in which to complete and submit their transcribed data.
Implement rescue aids	• Several additional student aids were implemented and are summarised below. • A simple video tutorial was produced to describe the different sections of the annual rainfall sheets and to demonstrate the transcription process, and posted to Moodle, the university's online learning platform. • An automated quality-assurance check was integrated into the Excel template to generate a monthly total based on the daily values transcribed. • An online discussion forum was set up on the Moodle course page through which teaching staff could address queries raised by students. Students were invited to post questions relating to any data rescue topic. • An in-class check-in clinic was organised at the midway point of the assignment to highlight frequently asked questions from the online forum and to allow students the opportunity to raise questions in class.

MAIN RESCUE STAGES	ACTIVITIES
Upload transcribed data to online portal	• Once the assignment was completed, students compiled the transcribed files into a compressed zipped directory, maintaining a consistent file-naming convention, and then uploaded the files to the online course portal.
Conduct a post-project evaluation of tools and training	• Students identified the online discussion forum and the video tutorial as being the most useful tools in completing the assignment.
View project management as crucial to data rescue success	• The management of these resources demanded a significant investment of time by teaching staff, particularly the online discussion forum, which received more than 500 queries from students. Nevertheless, effective management of the project and student aids facilitated the development of the corrected dataset by notably reducing the propensity for errors and the amount of time required to carry out post-processing of the data.

3.3.15.2 Contribution and limitations

The workflow described here adds a valuable contribution to data rescue literature, in that it describes the ground-breaking effort entailing the involvement of undergraduate students during the rescue process. As stated by the author in a related poster detailing the rescue effort, the study also demonstrates the potential to integrate citizen science into the classroom and highlights the role that non-experts can perform during data rescue (Ryan *et al*., 2017). Despite the often-expressed queries regarding accuracy and reliability of citizen scientists during data rescue, it has been shown by Van der Velde *et al.* (2017: 127) that citizen science data are of equivalent quality to data collected by researchers. In addition, Kosmala *et al.* (2016: 551) found that data quality can even be enhanced when citizen science projects are well-managed.

It should be noted that the rescue process described in the article did not include the adding of metadata, and only involved the rescue of data in an early digital format (i.e., rainfall sheets). As such, issues such as handling of fragile paper-based data, or digitisation issues, do not form part of this data rescue workflow. In addition, at the time of article publication, the data had not been made publicly available by the students.

The concepts of citizen science, aids and feedback during the project, and double keying of values are important rescue activities taken note of in this research.

3.4 Summary of workflows

The preceding section reported on a range of data rescue workflows and highlighted and summarised the activities and steps mentioned in 15 different outputs describing data rescue's chronological flow. The detailed yet simplistic data rescue workflow of the WMO (2016), created for historic climate data, has been earmarked in this research as a primary source to consult when creating an initial Data Rescue Workflow Model. Kennedy's Data Rescue Cookbook (2017), focused on the rescue of Canadian marine species occurrence data, is another source set aside as a detailed yet user-friendly data rescue publication. While these two outputs were the main sources used when drafting the initial data rescue model, several of the other analysed publications also contributed to the understanding of aspects vital to data rescue, albeit to a lesser extent. The aspects chosen as vital to data rescue are discussed in Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics, and elaborated on in the relevant stage-specific sections included in Section 3.9: Stages of the initial Data Rescue Workflow Model.

Table 3.16 contains a summarised version of the 15 data rescue workflows, frameworks, and models reviewed and analysed. The intention of the table is not to compare data rescue projects, but to summarise the major features of each study. Of particular importance to this study is the column on the right indicating the study's unique pivotal features; many of these listed features were in some way implemented in this study's initial Data Rescue Workflow Model. The various workflows, models and frameworks scrutinised are listed according to date, from oldest output to most recent.

Table 3.16: Summary of data rescue workflows, models and processes

SHORT DESCRIPTION	DISCIPLINE	PIVOTAL RESCUE FEATURES
• **A guide for digitising manuscript climate data (refer to Section 3.3.1)** • 2006 • Journal article • https://cp.copernicus.org/articles/2/137/2006/cp-2-137-2006.pdf • Paper data	Climate data	• Provides valuable information on the digitisation part of data rescue • Provides valuable information on locating data at risk • Provides a necessary reality check by stating that not all rescue efforts (of manuscript data) are successful
• **Rescuing and recovering lost or endangered data (refer to Section 3.3.2)** • 2006 • Journal article • https://datascience.codata.org/articles/abstract/354/ • 'Non-digital' data	Various disciplines	• Provides valuable information regarding the rescue of data that is in public ownership (i.e., owned by private citizens) • The importance of trust between data rescuers and data owners is emphasised • Process described is cost-effective and productive
• **Solar Radiation Data Rescue at Camagüey, Cuba (refer to Section 3.3.3)** • 2008 • Research article • https://doi.org/10.1175/2008BAMS2368.1 • Paper data	Solar radiation data	• Describes a cost-effective rescue procedure • Project makes use of older equipment • Directed at developing countries • Rescue strategy could be adapted to other local resources, and combined with funded rescue projects for expedited results
• **Efficient rescue of threatened biodiversity data using reBiND workflows (refer to Section 3.3.4)** • 2012 • Research article • https://doi.org/10.1080/11263504.2012.740086 • Paper and early digital data	Biodiversity	• Applicable to other data domains and types • Cost effective • It involves data deposit into a flexible storage system • Offers a connection to relevant international data infrastructures
• **IEDRO: International Environmental Data Rescue Organization (refer to Section 3.3.5)** • 2014 • Data rescue steps on organisational website • https://iedro.org/ • Mostly paper data	Climate science	• Training forms a vital component of the process • Mentions role of funding, skills acquisition, and free and open repositories • Description of process is simplistic and clear • Have conducted rescue projects in many countries around the globe

• SHORT DESCRIPTION	DISCIPLINE	PIVOTAL RESCUE FEATURES
• **WMO Guidelines for Hydrological Data Rescue (refer to Section 3.3.6)** • 2014 • Online manual • https://library.wmo.int/doc_num.php?explnum_id=7891 • Mostly paper data	Hydrology	• Includes an explanation of the importance of drafting data inventories • Includes an explanation on how to draft the different data inventories • Contains clearly defined rescue steps • Includes references to rescue of early digital data
• **Data rescue to extend the value of vintage seismic data: The OGS-SNAP experience (refer to Section 3.3.7)** • 2015 • Research article • https://doi.org/10.1016/j.grj.2015.01.006 • Paper data, magnetic tape data	Seismic sciences	• Valuable description of magnetic tape rescue • Project attempted to envisage additional ways of future data use • Stated the critical importance of ensuring future reuse of rescued data
• **Rescue of dark zooplankton data collected in the 1970s and 1980s during 34 cruises to the Northwest Atlantic Ocean (refer to Section 3.3.8)** • 2015 • Journal article • https://doi.org/10.1016/j.grj.2015.03.001 • Paper data; old digital formats	Zooplankton	• Descriptions of where dark data can be located • Stipulation of rescue as a team effort • Deposit of data in a recognised disciplinary data repository • Emphasises the ongoing nature of management of rescued data
• **Curation of Scientific Data at Risk of Loss Data Rescue and Dissemination (refer to Section 3.3.9)** • 2016 • Book chapter • https://doi.org/10.7916/D8W09BMQ • Modern digital data	Environmental and socio-economic sciences	• Can be described as a case study on the issues raised by a data rescue effort from an existing archive that had not fully curated the original data • Rescue steps provided are simple yet crucial • Has a section on lessons learnt during the project • Provides a necessary reality check by stating that divergence from envisaged rescue steps may be required

SHORT DESCRIPTION	DISCIPLINE	PIVOTAL RESCUE FEATURES
• **Rescuing biogeographic legacy data: The Thor Expedition, a historical oceanographic expedition to the Mediterranean Sea (refer to Section 3.3.10.)** • 2016 • Journal article • https://doi.org/10.3897/BDJ.4.e11054 • Published historical data	Biogeographic sciences	• Mention of multidisciplinary working groups • Includes a discussion of problems encountered during data rescue • Describes the importance of quality control when assessing the data at risk • Describes how to obtain missing data (data missing from a dataset) using other sources • States that two versions of each dataset were created (standardised dataset and user-friendly dataset)
• **The Rescue of "at risk" data on forced resettlement in South Africa (refer to Section 3.3.11).** • 2016; 2022 • Presentation; web article • https://drive.google.com/open?id=0B_NJP_ik1Aj-M0Z5aG1tSFNDVmM • https://www.datafirst.uct.ac.za/services/data-rescue?highlight=WyJyZXNjdWVkIl0= • Paper data	Sociological sciences	• Example of South African data rescue project • Excellent example of the importance of cross-institutional collaboration • Public sharing of rescued data is crucial
• **WMO Guidelines on Best Practices for Climate Data Rescue (refer to Section 3.3.12)** • 2016 • Online manual • https://library.wmo.int/doc_num.php?explnum_id=3318 • Mostly paper data	Climate science	• Process includes the drafting of several inventories • Clearly defined rescue steps involving paper-based media • For use and application in disciplines other than climatic sciences • Supplementary sections hold vital rescue information • Regarded as a set of best practices for describing rescue of paper-based data
• **Data Rescue Guidelines of the RDA DATA Rescue Interest Group (refer to Section 3.3.13)** • 2017 • Guidelines on organisational website • https://www.rd-alliance.org/guidelines-data-rescue-0 • Mostly paper data	Not discipline-specific	• Emphasises the management of rescued data • Succinct description of vital rescue steps • Emphasises importance of marketing/promotion of data rescue

SHORT DESCRIPTION	DISCIPLINE	PIVOTAL RESCUE FEATURES
• **Guidelines for marine species occurrence data rescue - The OBIS Canada Cookbook (refer to Section 3.3.14)** • 2017 • Online manual • https://0fb5ebe8-ca92-4c3c-9170-ef778a77f76e.filesusr.com/ugd/cf2ff9_82e7008749294b83b31c4a8a9ecd99cf.pdf • Early digital and digital formats	Marine species	• Mentions the drafting of a project plan • A data management plan (DMP) forms part of the process • Provides a metadata template • Provides a data rescue checklist • Includes references to different data rescue role-players •
• **Integrating Data Rescue into the Classroom (refer to Section 3.3.15)** • 2018 • Journal article (also poster, meeting paper, listed by WMO as example of best practices) • https://doi.org/10.1175/BAMS-D-17-0147.1 • Transcription of electronic sheets	Rainfall and precipitation	• Emphasises the double keying of data • Describes the involvement of students • Describes the quality assessment of student input (accuracy of keying) • Rescue project is an example of 'best practices' (according to WMO)

As shown in the preceding sections, and demonstrated in the table above, the data rescue documentation reviewed and analysed covered more than a decade of rescue solutions, comprised mostly but not exclusively paper data, and represented a wide range of disciplines. While many similarities were found regarding rescue activities, rescue steps still showed differences in the way data rescue was conducted, with each uniquely contributing to the plethora of data rescue information in a distinct manner.

The main outcome of this delving into published data rescue steps is the realisation that the concept of a 'one-size-fits-all' rescue workflow is not possible. While broad generic rescue steps are found to be common in most workflows, the described features, activities and role-players are mostly unique to each rescue effort, the resources available, entity performing the rescue, data format being rescued, and research discipline involved.

As stated in the introductory paragraphs of Section 3.2., minimal evidence of designated or task-specific involvement of the LIS sector was found when analysing the workflows. The absence of such information does not mean the library and information services is not involved during data rescue, or is not able to execute data rescue activities. It simply means that establishing and recommending data rescue roles and responsibilities for the research library during this study will be indirectly derived from data obtained during the content analyses of workflows. The nature of established data rescue tasks and activities, ascertained via content analysis of published outputs, will point to potential links between the task requirements and research library skills and expertise.

Additional aspects discovered and taken note of are listed below. Where possible, these aspects will be incorporated into the initial data rescue workflow, be it in the form of activities, guidelines or outcomes.

- Data discovery is more effective, and less tedious, when original data collectors can still be consulted.
- The creation of a data rescue team is vital; data rescue is a team effort and will not easily be performed by solo researchers or research library experts only.
- The data rescue team should ideally consist of participants having combined expertise in the research discipline, ICT matters, and library and information service skills.
- The role and input of citizen scientists and volunteers should not be underestimated.
- Data at risk, or historic data may often be held by members of the public; not all data at risk are readily awaiting discovery in archives, libraries or research laboratories.
- Inadequate funding, or the non-availability of state-of-the-art data rescue equipment, is a harsh reality commonly faced by research teams.
- While most rescue projects pertain to climatic sciences, data at risk can be found in many disciplines. Data at risk are also not always historic in nature, nor do they only fit under the legacy data umbrella.

The analysis of relevant data rescue literature was performed to guide the creation of a data rescue workflow model. Scrutinising research, manuals and other outputs revealed which steps and activities can be regarded as unmissable. The analysis also indicated which steps and activities were specific to the described project and not generic in nature, or common to all rescue projects. The unique character of this study’s involved research institute was also considered, with expertise, resources and funding being important considerations.

The next section states in more detail the stages, steps and activities selected as essential components of a data rescue workflow model.

3.5 Creating a Data Rescue Workflow Model by means of a flowchart

As this chapter is concerned with the creation of an initial Data Rescue Workflow Model, it is considered crucial to elaborate on several concepts before describing the initial model. This section elaborates on the concepts of ‘workflow’ and ‘flowchart’, and how these two concepts are involved in the creation of the initial model.

3.5.1 Workflows

A workflow is described as a series of repeatable activities that need to be continued to complete a specific task (Indeed, 2021). A workflow will show how to get the work done, and comprises the sequence of rescue tasks from start to finish (Brandall, 2021). A data rescue workflow serves as a replacement for manually handled processes, thereby informing data rescue participants as to the steps and activities to be executed, and indicating to project managers the project status quo and progress made.

3.5.1.1 Definition and use

A workflow can be described in any of the following ways:

- as an orchestrated and repeatable series of tasks endeavouring to accomplishing a specific outcome (Huettich, 2020),
- as the way people get work done (Brandall, 2021),
- as work flowing from one stage to the next, whether it be through a colleague, tool, or another process (Brandall, 2021),
- as a series of tasks requiring completion in sequential order to reach a specific business objective (Monday.com, 2021), and
- it defines the work to be done, who is responsible for it, and the time that each task takes (Monday.com, 2021).

Monday.com (2021) mentions that a workflow answers the question: ‘What is the most efficient way to complete this work?’, while Huettich (2020) states that the organisation’s resources and processes need to be considered when creating a workflow. Taking these considerations into account results in workflows often being specific to the organisation or project, and the realisation that this specificity would be no less relevant in a data rescue scenario.

A data rescue workflow can therefore be defined as a series of tasks requiring completion in sequential order to ensure data at risk are converted to a common, open, modern digital format and accessible to future researchers. Taking resources and processes into account, the data rescue workflow should define each of the tasks and envisaged outcome, should indicate the responsible parties, and indicate the time required for completion of each rescue task.

3.5.1.2 Benefits of workflows

According to Lucid Content Team (2021), the following benefits are encountered when making use of workflows:

- an improvement in business operations,
- an elimination of redundant processes and activities,
- a reduction in operational expenses,
- fast responses to issues or problems, and
- an automation of processes.

In the context of this study, it is expected that the drafting of a data rescue workflow model would at a minimum be instrumental in eliminating inessential steps and enable swift reaction and feedback should problematic issues be experienced.

3.5.1.3 Workflow types

Different workflow types can be identified within the workflow realm. A good example of different workflow types are the three types mentioned by Huettich (2020), namely project workflows, case workflows, and process workflows.

Project workflows are described as an ideal tool to keep complex projects on track, and consist of a series of activities diagrammed out, illustrating which tasks need to be performed and the order of tasks to create each project deliverable. Unlike other workflows, the case workflow, according to Huettich (2020), does not occur in a sequential, orderly fashion. The case workflow is unique, as the actual progression of steps is not known at the onset of the problem, and is concerned with a unique problem that requires solving. Process workflows, according to Huettich (2020), are used to depict repetitive and predictable tasks. Within a process workflow diagram, the delineation of the tasks that need to be performed, when the tasks should be performed, and the department responsible for each of the tasks form part of the workflow.

Brandall (2021) states that a workflow can be represented either via a diagram or a checklist.

Taking the different workflow types into account, a process workflow will be created to demonstrate the initial Data Rescue Workflow Model. Furthermore, flowcharts will be used as diagrammatical representations of the rescue tasks to be performed.

The next section discusses flowcharts as a tool to demonstrate a process workflow.

3.5.2 Flowcharts

The previous section defined the ‘workflow’ concept, and described its benefits, use and types. As this study makes use of a flowchart to demonstrate the initial workflow model, it is important that similar elaboration be afforded to the ‘flowchart’ concept.

3.5.2.1 Definition and use

According to SmartDraw (2021), a flowchart is a type of diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analysing, designing, documenting or managing a process or programme in various fields.

3.5.2.2 Benefits of flowcharts

Using a flowchart to illustrate data rescue activity steps, instead of an alternative such as text-based documentation or standard operating procedures, was chosen due to the following advantages offered:

- A flowchart provides visual clarity: the tool's ability to visualise multiple processes and their sequence at a single glance is a huge benefit. According to Lynch (2019), such a representation makes it possible for stakeholders to understand the workflow while discovering which steps are unnecessary or could be improved upon. As such, it is an ideal tool to use when demonstrating an overview of a project (Lynch, 2019). The outline of the flow of information, inputs and logics in a visual manner is also mentioned by Newman (2018).
- Flowcharts are generally simple; according to Tech Differences (2019), a big advantage of using a flowchart is that it is short, easily constructed, and easy to understand. Individuals are saved the effort of reading lengthy documents, as looking at the flowchart can lead to the same amount of understanding as perusing pages of text at length. The use of flowcharts, as instantaneous communication tool and immediate tool of understanding, is an aspect also mentioned by Lynch (2019).
- Flowcharts present data in logical steps, thereby making it easy for the reader to understand the logical flow of the process.
- Flowcharts can contain a great deal of information within a single chart.
- A well-designed flowchart adhering to industry best practices can more easily be implemented into an institute's electronic workflows than a text-based document would allow.
- Flowcharts enable effective analysis: the tool's use of specific shapes to indicate the type of action required, or symbols to indicate the type of storage media used, makes it an effective analysis platform.
- With the help of a flowchart, problems can be analysed effectively.

3.5.2.3 Limitations

Despite the advantages offered by using a flowchart to demonstrate the data rescue process, use of this method may also carry the following disadvantages:

- Creating a flowchart is a time-consuming activity (Newman, 2018).
- Amendments to a flowchart can waste time and money (PlanetTogether, 2020).
- Flowcharts can be difficult to alter: a change in workflow often requires the whole flowchart to be redone, while a similar amendment in text documents might be simpler. Newman also mentioned the disruptive effect when one or more elements of a flowchart need to undergo alterations (2018).
- Flowcharts do not provide sufficient visibility into unforeseen events and circumstances that may occur when the process is applied in the real world. Newman stated that the consequences of lack of visibility can have substantial impacts in real-life scenarios (2018).
- Flowcharts are unable to portray the complex logic that animates certain modern systems and processes. This imposes limits on the flow of logic inside a flowchart and (in certain cases) may severely curtail the depth of understanding for process reviewers. Therefore, this aspect of the limitations of flowcharts poses certain problems for industrial and process designers (Newman, 2018).
- Thornton stated that the implementation of a workflow diagram might even be more difficult than the design of the flowchart (2017). He further mentioned that it would be unwise to simply post a workflow diagram and expect participants to execute correctly without any guidance.
- Many flowchart symbols cannot be typed (PlanetTogether, 2020).

In addition to the limitations stated in literature, two potential drawbacks were also envisaged when using a flowchart to demonstrate data rescue workflow. The two problematic areas, with a mitigating activity to reduce each potential limitation, are stated below.

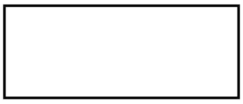
- Use of a flowchart to demonstrate data rescue workflow can present challenges should the final version of the initial data rescue workflow model be exceedingly long or complicated. To deal with this limitation, it is important that the final version of the model also features a single-paged combined summary of all the data rescue stages and main activities in each.
- The combination of a new research activity (i.e., data rescue) combined with a new workflow type (i.e., flowchart) could present comprehension issues for novice researchers or other SET-based staff who had not come across either of the concepts before. A brief introduction to

the use of flowcharts and the standard symbols might therefore be required when presenting and promoting the final model.

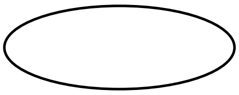
3.5.2.4 Use of flowcharts

Flowcharts tend to feature standard symbols, and the five most used symbols, according to Lumen Learning (2016) and SmartDraw (2021), are mentioned below.

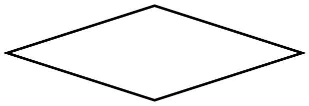
- Rectangle shapes are used to present a process, and resemble the following shape:



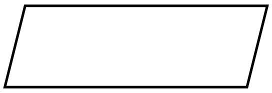
- Oval or pill shapes are used to represent the start or end, and resemble the following shape:



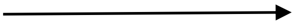
- Diamond shapes are used to represent a decision, and resemble the following shape:



- Parallelograms are used to represent input or output, and resemble the following shape:



- Lines are used as connectors and show the relationship between the representative stages:



In addition to the shapes mentioned above, flowcharts generally flow from top to bottom and left to right (SmartDraw, 2021).

After testing various online flowchart platforms, the option of custom symbols on several of the assessed platforms was also discovered. These symbols can be used by flowchart designers when in need of a shape that is not in one of the standard templates.

3.5.3 Summary

A data rescue workflow model, in flowchart format, was selected as a tool to visually portray the data rescue steps to be performed by various role-players during a rescue project. The flowchart process workflow format was used to depict the study's three different data rescue models. These three models comprise the initial data rescue model, the revised data rescue model, and the final recommended data rescue model.

The flowchart tool's logical flow of steps, colours and shapes used, and potential to portray a great deal of information within a single page made it the most user-friendly and suitable platform for illustrating the envisaged data rescue stages, activities, role-players and project deliverables.

The initial Data Rescue Workflow Model is discussed in the remainder of this chapter.

3.6 Content analysis and the creation of a Data Rescue Workflow Model

This section describes the steps followed when reviewing and analysing data rescue literature sources leading to the creation of an initial Data Rescue Workflow Model.

Prior to creating the initial model, 15 different data rescue publications, all containing references to either data rescue workflows, data rescue steps, or data rescue guidelines, were assessed. The publications were analysed to gather insight into data rescue activities, and an understanding of the crucial steps forming part of a generic data rescue workflow. This insight, understanding and exposure to tried-and-tested processes were used to create a Data Rescue Workflow Model.

With the published data rescue documentation serving as data during this stage, it was crucial that an applicable data analysis method be implemented to make sense of the volume of information contained within the 15 published items. Content analysis, described by Robson as 'the quantitative analysis of what is in the document', was selected as the method to analyse and make sense of the different rescue models, workflows and processes published (2002: 349).

Content analysis as part of the study's methodology is discussed in Section 4.13.1: Content analysis. The main steps of this method are listed in the Methodology chapter, with the crucial activities being the following (Columbia University, Mailman School of Public Health, 2019):

- selecting the level of analysis,
- selecting the number of concepts and developing a pre-defined set of categories,
- deciding whether to code for existence or frequency of a theme,
- deciding how to distinguish between concepts,
- developing rules for coding the texts, and deciding what to do with irrelevant information,
- coding the text, and
- analysing the results.

The listed steps were used as guidance when reviewing and analysing the 15 data rescue publications. Activities forming part of content analysis applied during the literature analysis are described below.

- **This researcher selected the data:** as described in Section 3.3, 15 data rescue publications covering a range of research disciplines, time periods and data formats were selected to be reviewed, systematised and analysed. rescue
- **This researcher prepared the data**: Datt and Chetty (2016) view this step as the first step in content analysis. As the data in this study have already been generated, and are present in the form of text (i.e., published articles, papers, manuals) a big part of this step has already been dealt with. The full text versions of the 15 selected published outputs were downloaded to a format that would enable digital highlighting with a coloured marker tool.
- **This researcher defined the unit of analysis**: Datt and Chetty (2016) and Zhang and Wildemuth (2009) state that a next step comprises the definition of unit of analysis. This author decided to analyse on the level of themes, sentences and phrases, as analysing on work level was anticipated to be very broad and could result in irrelevant coding. This researcher would therefore code 'scan precipitation strip charts and photograph hydrometeorological alphanumeric records' and not merely 'scan', and code '…transferred the original handwritten records to digital form. Another examples was the highlighting of 'software was designed and created for data key entering…'and not merely selecting 'transferred' or 'key entering'.
- **This researcher decided on frequency versus relation**: Columbia University, Mailman School of Public Health (2019) advises that researchers can code for existence or frequency of a concept. Although frequency of detected themes was noted, this researcher considered it more important to detect the relation of the theme to other factors, and to consider the role of data formats, scope of project, expertise of rescue participants and available resources when coding. As such, coding was made according to existence, and not frequency.
- **This researcher developed a coding framework**: after reading through all 15 published works, a flexible coding framework was developed which allowed for the adding of categories or concepts throughout the analysis process. Themes forming part of the preliminary framework were the following:
  - data rescue steps common to many of the described rescue events,
  - data rescue steps being the exception, i.e., mentioned in one study only,
  - data rescue steps identified as mandatory rescue activities, and steps forming part of most rescue projects,
  - activities forming part of pre-rescue and activities taking part post-rescue,

  - role of data formats,
  - role of research discipline,
  - role of available funding and other resources,
  - data rescue role-players, participants and stakeholders,
  - the concept of data rescue teams,
  - the role of specific skills and expertise,
  - outputs/deliverables created during rescue, and
  - outstanding or pivotal data rescue features detected.
- **This researcher coded the text; added coded segments to a group of similar segments**: Zhang and Wildemuth (2009) referred to the coding of all text as a next activity, and mentioned that it is quite likely that new themes and concepts will emerge and be added as coding progresses. Examples of coded text added to similar segments in this study include the following phrases (gathered from different data rescue publications) added to the 'data digitisation' segment:
  - 'create master image inventory',
  - 'image and validate',
  - 'digitizing data',
  - 'digitize the data',
  - 'use whatever means possible to create digital version of the dataset (EXCEL, text, or ASCII)',
  - 'OCR or manually type information from the pdf into new data file',
  - 'review and correct digitized content',
  - 'the images are digitized using the strip chart digitizer program, dual keying stations or another method chosen for delivering the best results in the shortest period of time within the project budget',
  - 'key entry',
  - 'quality-control data', and
  - 'as the information was assembled, it was typed into a spreadsheet'.
- **This researcher constructed the final narrative:** the codes and categories were used to construct the final narrative, comprising a chronological list of stages, activities and outputs that would eventually form a data rescue workflow model. The list of stages, activities and outputs resulting from content analysis are presented in Section 3.7.1: Description. The list and the chronological flow of data rescue components and activities enabled the creation of the Initial Data Rescue Workflow Model and is discussed in the next section.

Completing the content activities listed above resulted in this researcher gaining insight into data rescue best practices and an understanding of the main rescue stages and activities. The insight enabled the creation of an initial Data Rescue Workflow Model, portraying crucial data rescue steps, underlying activities and tasks, providing ancillary templates and guidance, and the vital outputs accompanying each completed stage.

The next section contains a description of the key features of the initial model.

3.7 Initial Data Rescue Workflow Model: Description and characteristics

This initial version, tentative in nature, should be viewed as an exploratory attempt at guiding novice data rescuers through the process of data rescue. The features and characteristics of the initial model are discussed in more detail in Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics, Section 3.8: Initial Data Rescue Workflow Model: Summary, and in Section 3.9: Stages of the initial Data Rescue Workflow Model.

Although this version of the model underwent various reviews and changes during the study, this initial version was the model that was distributed to Sample B to review and critique.

3.7.1 Description

After analysing the various guidelines, steps and workflows, the rescue aspects and activities mentioned below were selected as essential for inclusion in the initial Data Rescue Workflow Model:

- a project initiation phase, including data assessment, appointment of a data rescue project team, and creation of a data rescue project plan as well as a data management team,
- storage and preservation of data at risk,
- creation of data inventories,
- imaging and digitisation of paper-based data at risk,
- creation of metadata and data documentation,
- sharing of digitised data to repositories,
- long-term preservation of rescued (digitised) data, and
- project closure.

The following aspects provide additional explanatory details regarding the initial model:

- Each of the aspects mentioned in the bulleted list above would form a data rescue stage.
- Data rescue would mostly occur via moving sequentially through each of these stages.
- Despite most of the data rescue activities entailing the sequential progress through the data rescue stage, this would not always be the case.

- Non-sequential rescue movement could also occur, e.g., long-term preservation activities taking place even after project closure had been implemented and concluded.
- The model would consist of a static graphic representation of each of the rescue steps, as well as a static graphic representation of all rescue stages.
- The initial model is the version that would be distributed to Sample B for review and feedback (see Section 4.7.2 for more details on Sample B).
- The initial model entails the rescue of paper-based media, as this was the format most often included in data rescue publications, and the format at risk expected to be held by most Sample A members (see Section 4.7.1 for more details about Sample A).
- The initial model does not include detailed references to specific roles and responsibilities. It was expected that the feedback supplied by Sample B after reviewing the initial model would contain suggestions, recommendations and requests pertaining to specific role-players and rescue team members.
- Where applicable, stages of the initial model contained references to required outputs to be produced, guidelines to have recourse to, and templates to be consulted.
- It was envisaged that major amendments and revisions to this initial model would be recommended following the review and critique of the model by Sample B.
- After application of amendments and revisions to the initial model, the updated model would be referred to as the revised Data Rescue Workflow Model.
- Should feedback and critique received from Sample B be deemed insufficient, the revised model would be reviewed by Sample C (see Section 4.7.3).
- It is envisaged that major amendments and changes to the revised model will be requested and required following the review and critique of the model by Sample C.
- After application of amendments and revisions to the revised model, the updated model will be referred to as the final Data Rescue Workflow Model.

3.7.2 Summary

The initial Data Rescue Workflow Model was created based on information gained after implementing content analysis to analyse 15 data rescue publications (see Table 3.1). The initial model consists of nine stages and endeavours to portray the main stages, activities and outputs pertaining to a generic data rescue project.

It is important to state that the initial model served as a prototype, and it was anticipated and necessary that the study sample populations, who would be reviewing the model, recommend and request several changes to the model. Once the requested changes had been applied to the model, it would be referred to as the revised Data Rescue Workflow Model. Following another session of model review by a study sample, the final model would be presented and promoted. While the initial model is discussed in the current chapter, the revised model is discussed in Chapter 5: Results and Discussion, while the final model (referred to as recommended Data Rescue Workflow Model) is presented in Chapter 6: Recommendations.

This initial version, tentative in nature, should therefore be viewed as an exploratory attempt at guiding novice data rescuers through the process of data rescue. The next section (Section 3.8) provides a summarised image of the model, while Section 3.9 discusses each of the nine data rescue stages and its relevant activities and deliverables.

3.8 Initial Data Rescue Workflow Model: Summary

A summarised image of the initial model is shown in Figure 3.1.

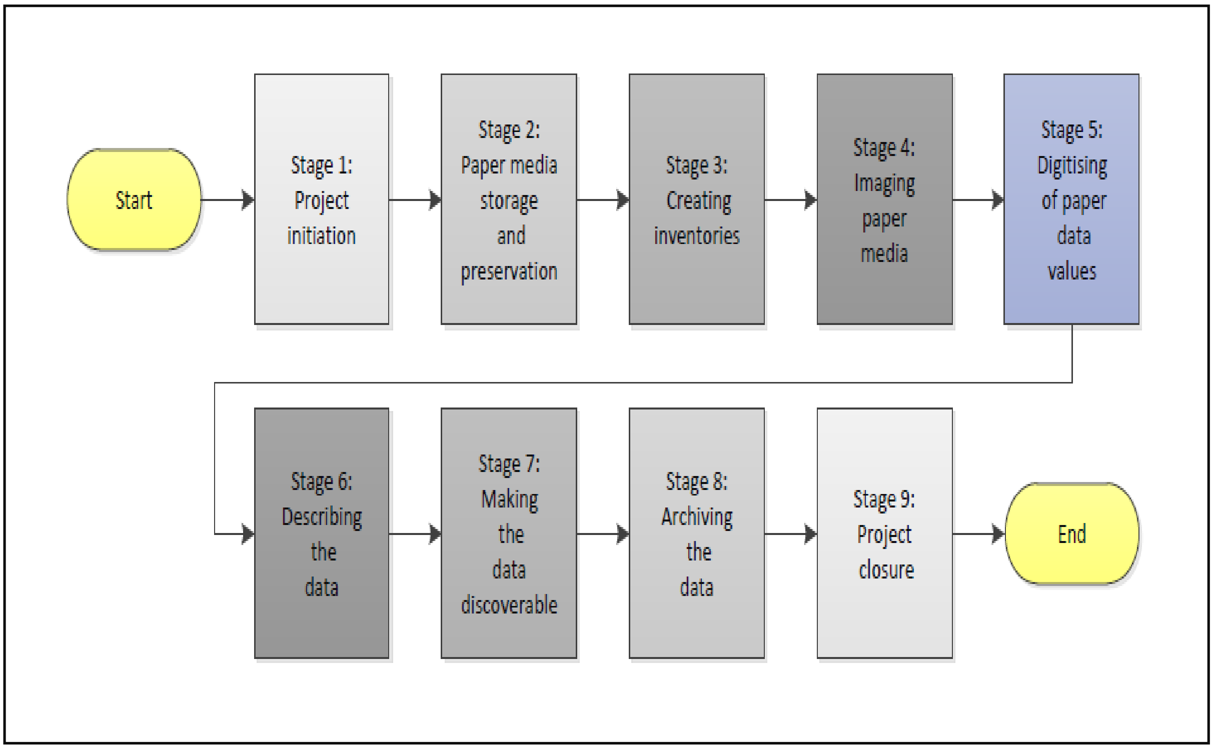

Start
Stage 1: Project initiation
Stage 2: Paper media storage and preservation
Stage 3: Creating inventories
Stage 4: Imaging paper media
Stage 5: Digitising of paper data values
Stage 6: Describing the data
Stage 7: Making the data discoverable
Stage 8: Archiving the data
Stage 9: Project closure
End

Figure 3.1: Summary of initial Data Rescue Workflow Model

As is seen in the image, the model proposes nine stages for the generic rescue of paper-based media. These stages are as follows:

1. Project initiation
2. Paper media storage and preservation
3. Creating inventories

4. Imaging paper media[23]
5. Digitising of paper data values
6. Describing the data
7. Making the data discoverable
8. Archiving the data
9. Project closure

Being a summary of the process, the image does not contain links to guidelines to be consulted, templates to be used, or outputs to be created during each stage. The summary also does not indicate where certain vital decisions are to be made. These links and references, however, feature in the images and description of the individual rescue stages.

The summary's primary objective is to demonstrate the various main stages forming part of a rescue venture, and to emphasise that rescue of data at risk entails more than transferring older data to a modern digital format. The goal of the summarised version is not to provide guidance on how to rescue data at risk. By viewing the process in its totality, researchers, or other role-players, new to the concept of data rescue are shown the stages forming part of a typical rescue process. Exposure to a summarised version also gives prospective data rescue project leaders a vague idea of the complexity of the rescue, the different role-players required, tools and equipment needed, and the time investment required for the project.

While not all data at risk are paper based, the initial model features paper as an example of data at risk. Paper was chosen for the prototype model, as many of the literature sources used when creating the model referred to paper media. This included research articles, websites of data rescue organisations and interest groups, and published guidelines. In addition, it was realised that many of the respondents who would be reviewing the initial model would also have paper-based data at risk in their research groups. In addition, the involved institute is more likely to hold equipment suited to paper data rescue, such as scanners, digital cameras, archive boxes, archive shelves, and labelling devices.

Each of the nine rescue stages are discussed in more detail during the remainder of the chapter.

[23] Imaging: the conversion of hard copy records to electronic files on digital systems

3.9 Stages of initial Data Rescue Workflow Model

This section contains clarifying details on the activities forming part of each of the stages of the initial model. Each stage section contains a visual representation of the stage (i.e., a flowchart), and also provides more details about the activities, decisions and outputs of the stage.

3.9.1 Stage 1: Project initiation

A visual representation of Stage 1 is shown in Figure 3.2. Stage 1: Project initiation.

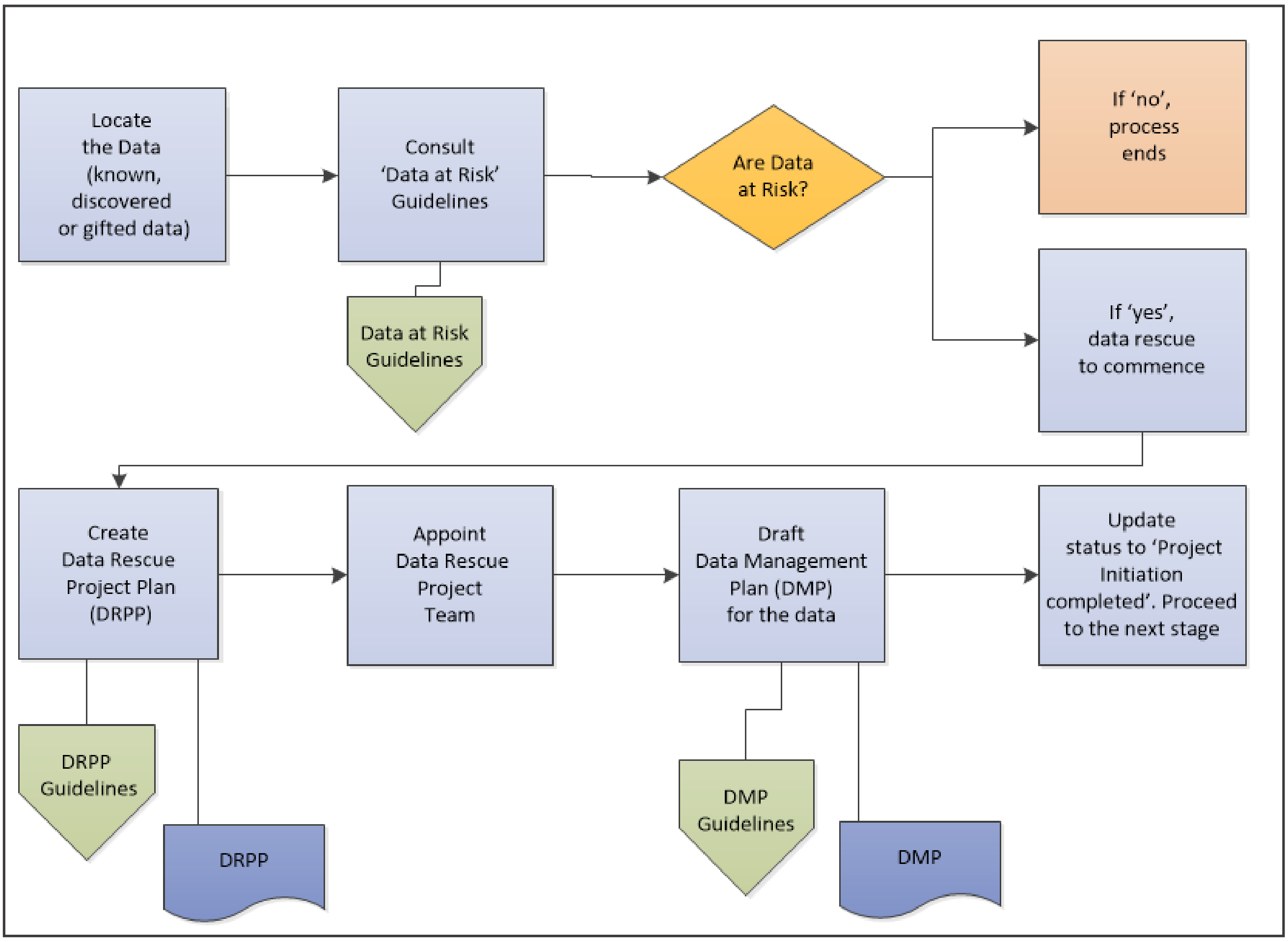

Locate the Data (known, discovered or gifted data)
Consult 'Data at Risk' Guidelines
Data at Risk Guidelines
Are Data at Risk?
If 'no', process ends
If 'yes', data rescue to commence
Create Data Rescue Project Plan (DRPP)
DRPP Guidelines
DRPP
Appoint Data Rescue Project Team
Draft Data Management Plan (DMP) for the data
DMP Guidelines
DMP
Update status to 'Project Initiation completed'. Proceed to the next stage

Figure 3.2: Stage 1: Project initiation

The first stage of the model pertains to the preparatory steps required before commencing with the rescue of data at risk. These steps are as follows:

- **Locate the data**
    - Data will be purposively found, accidentally discovered, or gifted.
- **Make a decision: are data at risk?**

  - A document titled 'Guidance on data assessment' should be consulted to assist the role-players in assessing the value of the data, and whether data rescue is a viable venture (see Appendix 10).
  - If the data are at risk, the data rescue process continues.
  - If the data are found to be not worthy of rescue, the data rescue project does not commence.
  - If data rescue is to go ahead, a data rescue project plan will be created.
- **Create a data rescue project plan**
  - Role-players should consult the document titled 'Guidance on data rescue project planning' to gain insight into the activities forming part of this step (see Appendix 11).
  - A Data Rescue Project Plan is one of the outputs created during Stage 1.
- **Appoint a data rescue project team**
  - An important part of Stage 1 entails the selection of a Data Rescue Project Team.
  - Project team members should ideally comprise a combination of discipline experts, data rescue experts, and an ICT expert. A combination of these three vital skills within the team is required.
- **Create a data management plan**
  - Stage 1 also requires the creation of a Data Management Plan (DMP) for the data involved.
  - A document titled 'Guidance on data management plans' (see Appendix 13) should be consulted for guidance.
  - A DMP is the second output created during Stage 1.

Upon completion of all the activities forming part of Stage 1, the rescuer or rescue team can proceed to the next stage, titled 'Paper media storage and preservation'.

3.9.2 Stage 2: Paper media storage and preservation

A visual representation (i.e., flowchart) of Stage 2 is shown in Figure 3.3. Stage 2: Paper media storage and preservation.

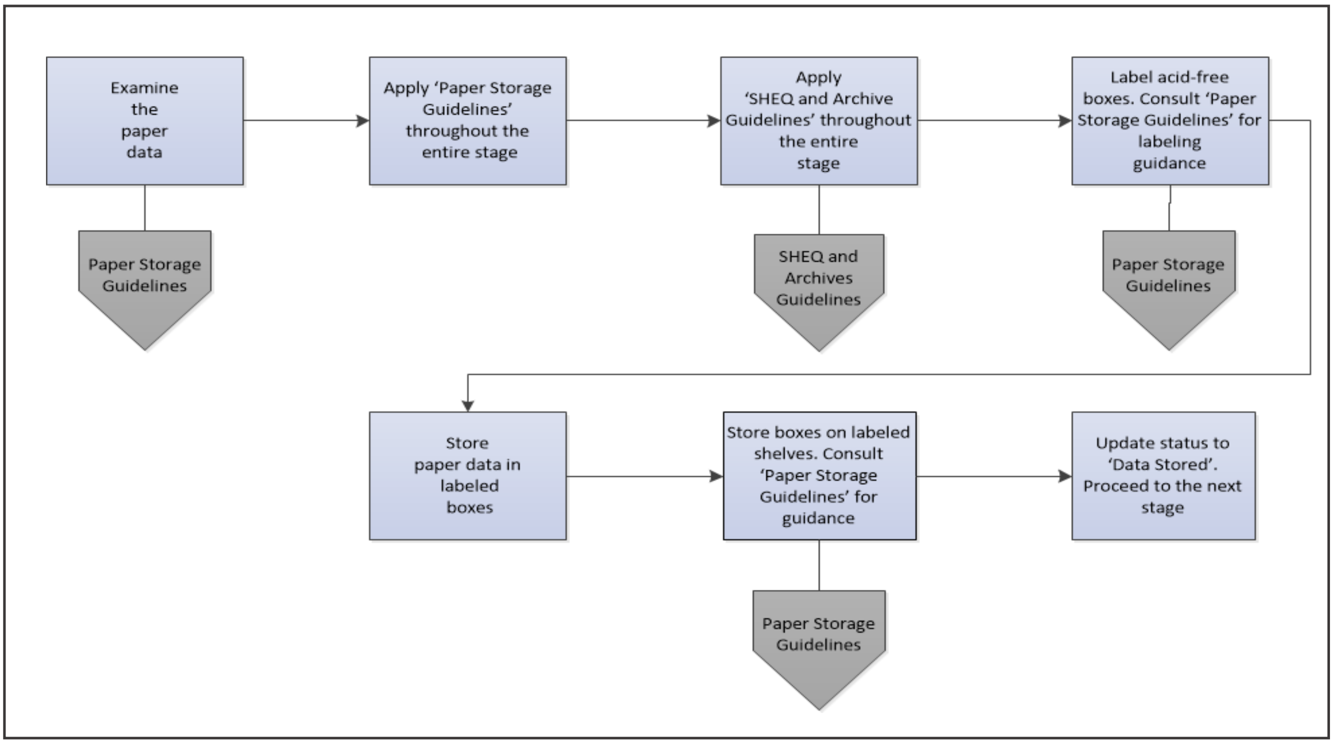

Examine the paper data
Paper Storage Guidelines
Apply 'Paper Storage Guidelines' throughout the entire stage
Apply 'SHEQ and Archive Guidelines' throughout the entire stage
SHEQ and Archives Guidelines
Label acid-free boxes. Consult 'Paper Storage Guidelines' for labeling guidance
Paper Storage Guidelines
Store paper data in labeled boxes
Store boxes on labeled shelves. Consult 'Paper Storage Guidelines' for guidance
Paper Storage Guidelines
Update status to 'Data Stored'. Proceed to the next stage

Figure 3.3: Stage 2: Paper media storage and preservation

The second stage of the model pertains to the storage and preservation of the data at risk, or the paper-based media that will be rescued. This activity should ideally be performed in a dedicated archive suited towards the storage and preservation of paper-based media. This stage does not refer to the storage and preservation of the paper-based media after it has been converted to a modern digital format.

The second stage of the initial model contains the following steps:

- **Examine the paper data**
  - During this activity, the rescuer will consult a document titled 'Guidance on storage of paper data' (see Appendix 14).
  - The document provides guidance on the handling and storage of paper media, and how to minimise media degradation.
- **Apply paper storage guidelines**
  - It is vital that the guidelines provided in the document titled 'Guidance on storage of paper data' be applied throughout the entire Stage 2 (see Appendix 14).
- **Apply Safety, Health, Environment, and Quality (SHEQ) guidelines for archives**
  - The guidance document provides details on ideal archival conditions for the storage and preservation of paper-based media.

- **Store paper in boxes, store boxes on labelled shelves**
    - Once boxes and shelves have been labelled, the paper-based media is stored in the boxes, on the shelves.

Once these steps have been completed, the data rescue team can progress to Stage 3 of the initial Data Rescue Workflow Model, titled 'Creating inventories'.

3.9.3 Stage 3: Creating inventories

A visual representation (i.e., flowchart) of Stage 3 is shown in Figure 3.4. Stage 3: Creating inventories.

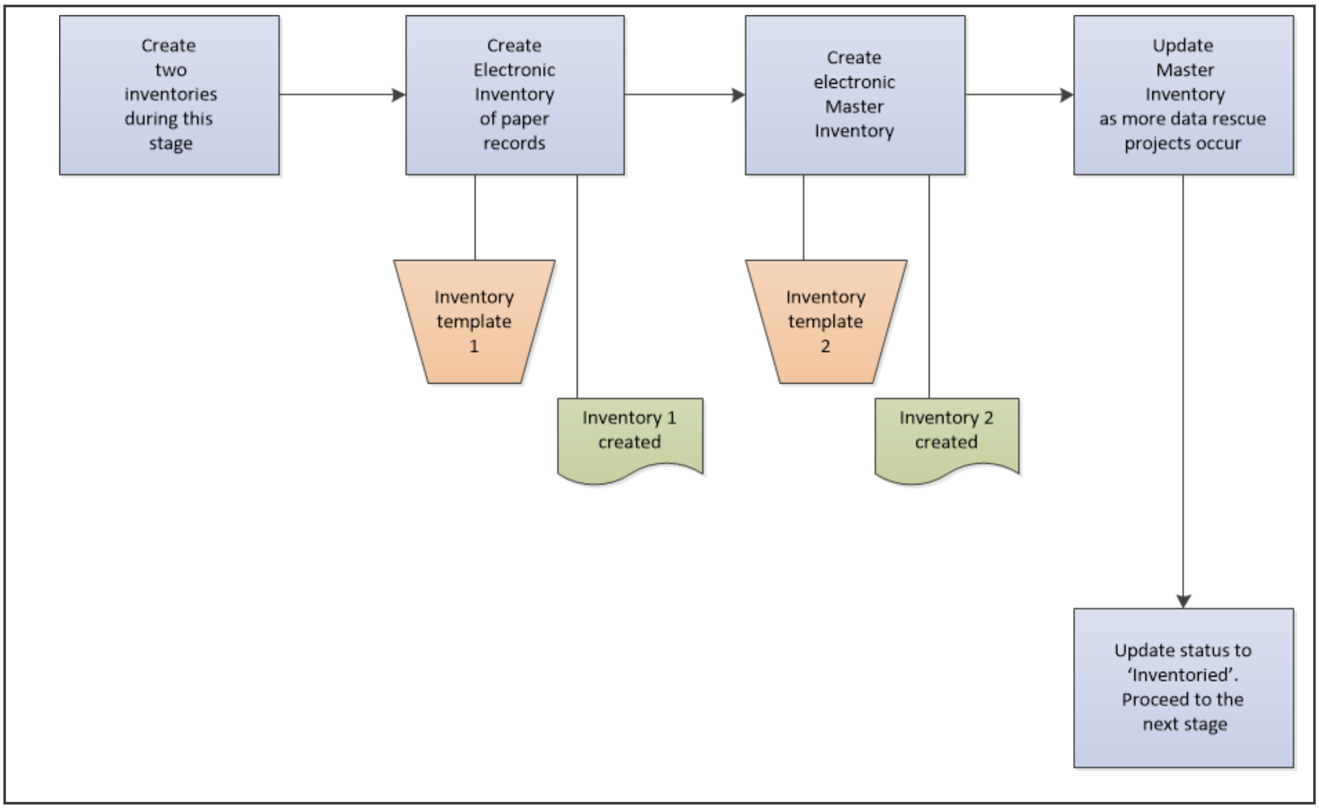

Create two inventories during this stage
Create Electronic Inventory of paper records
Create electronic Master Inventory
Update Master Inventory as more data rescue projects occur
Inventory template 1
Inventory template 2
Inventory 1 created
Inventory 2 created
Update status to 'Inventoried'. Proceed to the next stage

Figure 3.4: Stage 3: Creating inventories

The third stage of the model pertains to the creation of data inventories, thereby assisting the rescuers in making sense of the data to be rescued. The data inventories also ensure an electronic institutional record of data included in past rescue activities, and the rescue activities performed on the inventoried data. Inventories should adhere to a predetermined naming convention and be stored in a predetermined file folder and structure. Inventories should also be searchable, and not only contain the title of the data, but also the features including pre-rescued format, post-rescue format, date of inventory, inventory creator, and reference to the project and its identifying details.

The third stage of the initial model contains the following steps:

- **Two inventories are created** during this stage
  - An inventory of paper records will be created.
  - An electronic master inventory will be created.
- **Create an electronic inventory of paper records**
  - Data rescuer should consult the document titled 'Sample templates for the creation of data inventories' for guidance (see Appendix 17).
- **Create an electronic master inventory**
  - Data rescuer should consult the document titled 'Sample templates for the creation of data inventories' for guidance (see Appendix 17).
- **Update master inventory**
  - This inventory is updated as more data rescue projects occur.

After creation of the inventories, the rescue team can proceed to the next stage, titled 'Imaging paper media'.

3.9.4 Stage 4: Imaging paper media

A visual representation of Stage 4 is shown in Figure 3.5. Stage 4: Imaging paper media

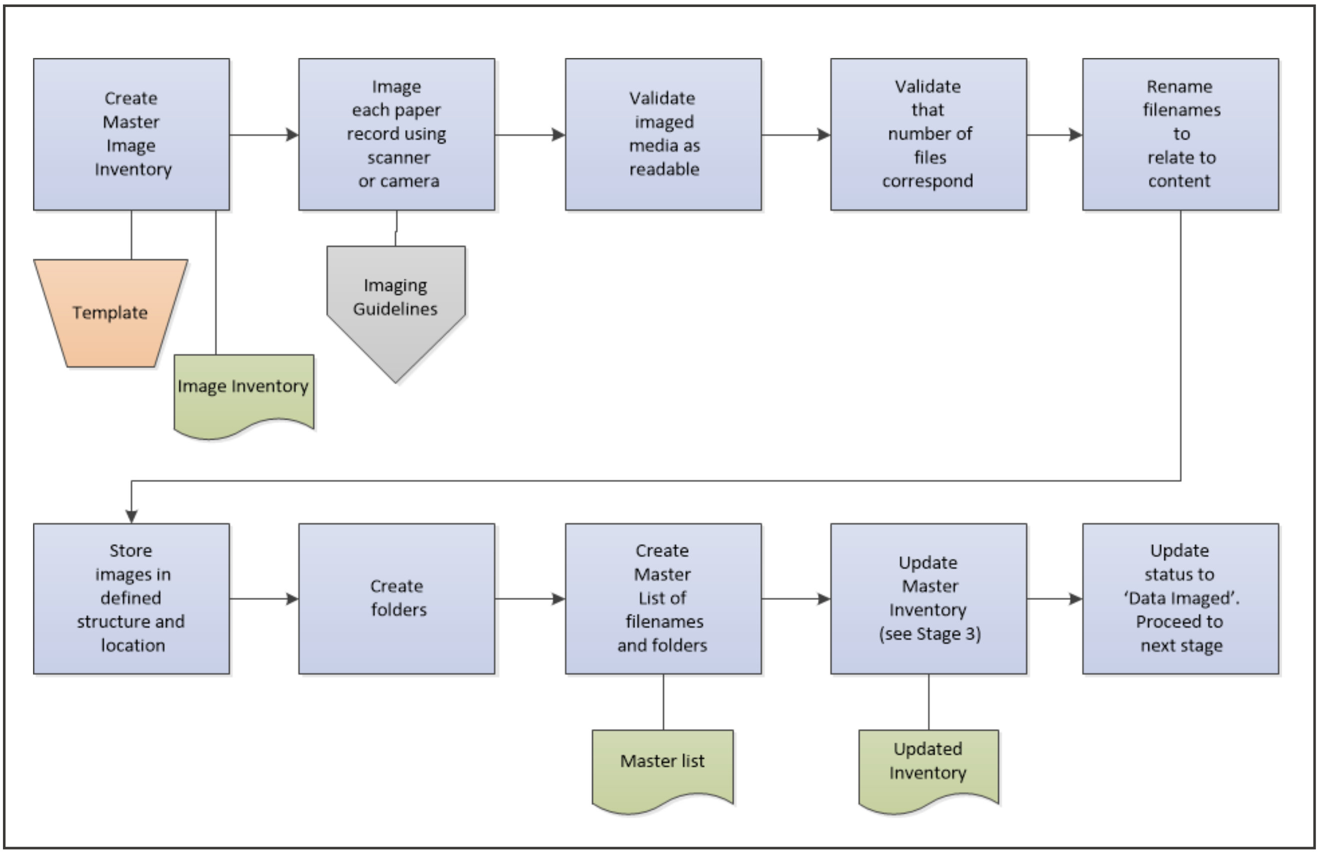

Create Master Image Inventory
Image each paper record using scanner or camera
Validate imaged media as readable
Validate that number of files correspond
Rename filenames to relate to content
Template
Image Inventory
Imaging Guidelines
Store images in defined structure and location
Create folders
Create Master List of filenames and folders
Update Master Inventory (see Stage 3)
Update status to 'Data Imaged'. Proceed to next stage
Master list
Updated Inventory

Figure 3.5: Stage 4: Imaging paper media

The fourth stage of the model pertains to the imaging of paper-based media using a scanner or digital camera. Imaging of the media results in the data being available in a modern digital format.

Steps forming part of this stage are as follows:

- **Create a master image inventory**
  - An inventory is created listing the imaged paper media.
  - A document titled 'Sample templates for the creation of data inventories' can be consulted during this stage (see Appendix 17).
- **Image each paper record**
  - Imaging is performed using a scanner or camera.
  - A document titled 'Guidance on digitisation of paper data' can be consulted during this stage (see Appendix 18).
- **Validate the imaged media**
  - This activity entails validating the imaged media, thereby comparing the imaged digital media with the paper-based media (number of files).
  - During this activity, quality control of the newly digital media can also be performed.
- **Rename files**
  - After imaging, the files in digital format will be renamed.
  - A naming convention would have been stipulated in the project plan created during Stage 1: Project initiation.
- **Store images**
  - Digital media will be stored in a predefined file and folder structure and location.
  - The file and folder structure and location would have been stipulated in the project plan created during Stage 1: Project initiation.
- **Create master list of filenames and folders**
  - A master list of file names and folders, corresponding with the files and folders dealt with in the previous bullet (see 'Store images') is created.
- **Update master inventory**
  - The master inventory, created during Stage 3, is updated to reflect that the data have been imaged.

The rescue team can proceed to the next stage, titled 'Digitising of paper data values'.

3.9.5 Stage 5: Digitising of paper data values

A visual representation of Stage 5 is shown in Figure 3.6. Stage 5: Digitising of paper data values.

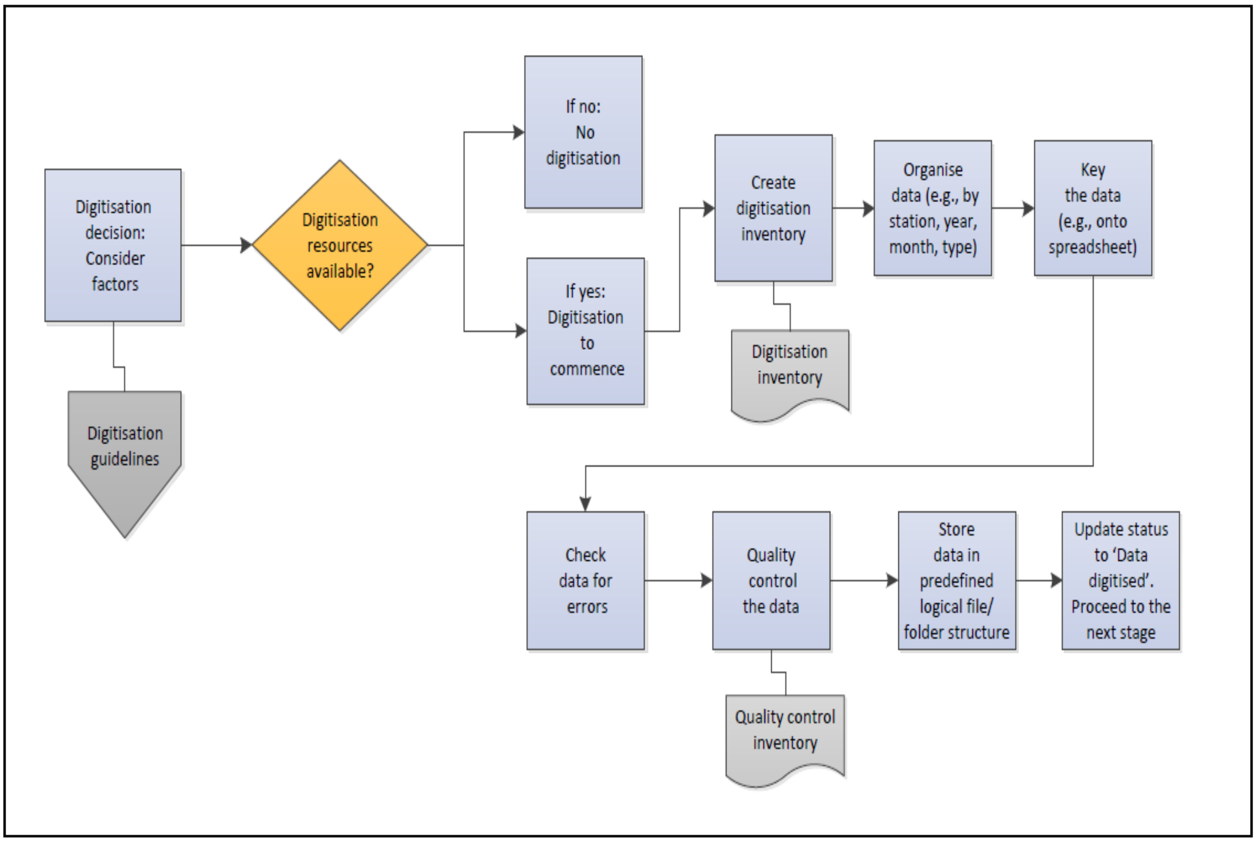

Digitisation decision: Consider factors
Digitisation guidelines
Digitisation resources available?
If no: No digitisation
If yes: Digitisation to commence
Create digitisation inventory
Digitisation inventory
Organise data (e.g., by station, year, month, type)
Key the data (e.g., onto spreadsheet)
Check data for errors
Quality control the data
Quality control inventory
Store data in predefined logical file/ folder structure
Update status to 'Data digitised'. Proceed to the next stage

Figure 3.6: Stage 5: Digitising of paper data values

The fifth stage of the model pertains to the digitisation of paper data values; in other words, the keying of tables or numerical values onto digital spreadsheets. Steps included in this stage are as follows:

- **Consider digitisation feasibility**
  - The rescue team should decide whether sufficient digitisation resources are available.
  - During this stage, the document titled 'Guidance on digitisation of paper data' can be consulted (see Appendix 18).
  - If digitisation is not a realistic venture, no digitisation activities will be performed (i.e., numeric paper values will not be keyed onto electronic spreadsheets, but instead scanned or photographed).
  - If sufficient digitisation resources are available, digitisation (i.e., keying of paper values onto digital spreadsheets) will commence.
- **Create digitisation inventory**
  - A digitisation inventory is created during this step.

- **Organise the data**
    - Data are organised, for example by station, year, month, or type.
- **Key data onto spreadsheets**
    - Numerical paper values are keyed onto digital spreadsheets.
    - Double keying is recommended to ensure minimal errors.
- **Check data for errors**
    - Keyed data are checked for errors.
    - Aspects such as correct keying, formatting, column naming, and completeness are checked.
- **Store data**
    - The spreadsheets are stored in a logical file and folder structure.
    - The predetermined file and folder structure would have been stipulated during Stage 1: Project initiation.

The rescue team can now proceed to the next stage, titled ‘Describing the data’.

3.9.6 Stage 6: Describing the data

A visual representation of Stage 6 is shown in Figure 3.7: Stage 6: Describing the data.

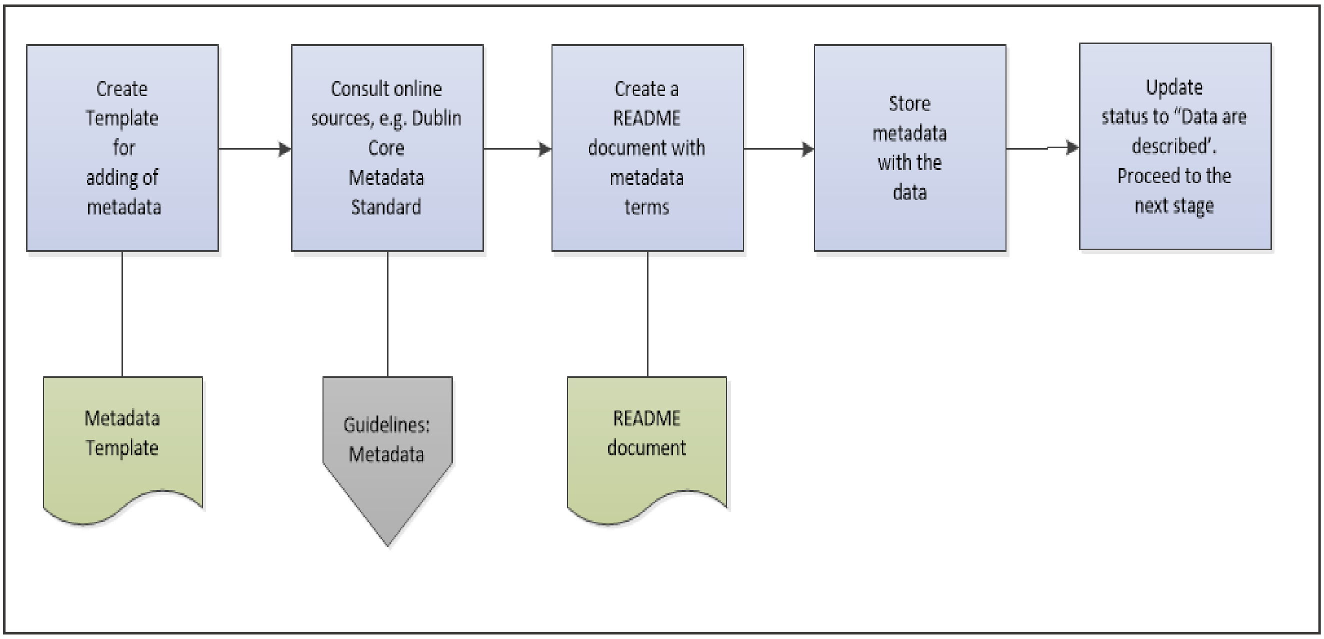

Create Template for adding of metadata
Consult online sources, e.g. Dublin Core Metadata Standard
Create a README document with metadata terms
Store metadata with the data
Update status to "Data are described'. Proceed to the next stage
Metadata Template
Guidelines: Metadata
README document

Figure 3.7: Stage 6: Describing the data

The sixth stage of the model pertains to the creation of metadata for the dataset, as well as accompanying data documentation. The steps included in this stage are as follows:

- **Creation of a metadata template** for the rescued data
  - The document titled 'Guidance on the use of metadata' can be consulted during this step (see Appendix 19).
  - The metadata standard would have been decided on and stipulated in the Data Management Plan, created during Stage 1: Project initiation.
  - Online metadata sources can also be consulted during this step.
- **Create metadata**
  - Metadata are created and saved as a readme file or similar.
  - Adherence to the metadata standard decided on during Stage 1, and stipulated in the project plan, is crucial.
- **Create data documentation**
  - This stage also entails the creation of data documentation to accompany the data. Such documentation would make it possible for secondary users of the data to make sense of the data, and to reuse the data.
- **Store the metadata and data documentation with the rescued dataset**

The rescue team can now proceed to the next stage titled 'Making the data discoverable'.

3.9.7 Stage 7: Making the data discoverable

A flowchart of Stage 7 is shown in Figure 3.8: Stage 7: Making the data discoverable.

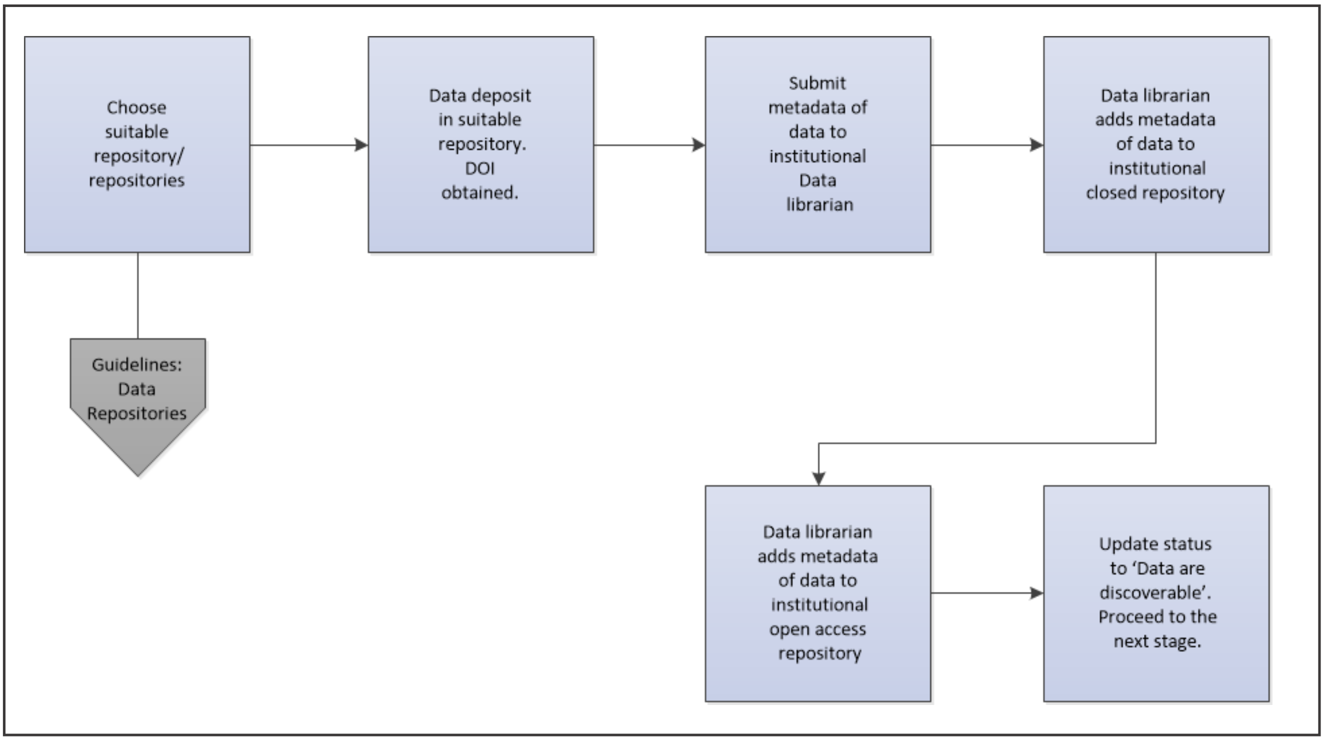

Choose suitable repository/ repositories
Data deposit in suitable repository. DOI obtained.
Submit metadata of data to institutional Data librarian
Data librarian adds metadata of data to institutional closed repository
Guidelines: Data Repositories
Data librarian adds metadata of data to institutional open access repository
Update status to 'Data are discoverable'. Proceed to the next stage.

Figure 3.8: Stage 7: Making the data discoverable

The seventh stage of the model details the way the rescued digital data will be shared with interested parties, such as fellow researchers inside the institute, and the wider external research community. The specific stage activities are as follows:

- **Select a suitable data repository**
    - The document titled 'Guidance on use of data repositories' should be consulted (see Appendix 20).
    - Ideally, data should be uploaded to an accredited discipline-specific repository.
    - In the absence of a discipline-specific repository, or when deciding not to make use of a discipline repository, an accredited generalist repository is a suitable option.
- **Upload rescued data to the selected repository**
    - The required repository steps as stipulated by the repository will be followed during upload.
    - Such steps may include registration, subscription fees, ensuring data are in a format approved by the repository, ensuring the metadata created adheres to the standard stipulated by the repository, and indicating (if applicable) the license accompanying the data.
- **Adhere to institute-specific sharing steps**
    - For the specific institute in question, data sharing will make use of institutional systems and processes and adhere to data confidentiality aspects.
    - The metadata of the rescued data are submitted to the data librarian via the institute's repository workflow system.
    - The data librarian will add the metadata of the rescued data to the institute's closed institutional repository.
    - Thereafter, the data librarian adds the metadata of the rescued data to the institute's open access repository.
    - In certain instances, non-confidential datasets will also be uploaded to the institute's open access repository.

The rescue team can now proceed to the next stage, titled 'Archiving the data'.

3.9.8 Stage 8: Archiving the data

A visual representation of Stage 8 is shown in Figure 3.9: Stage 8: Archiving the data.

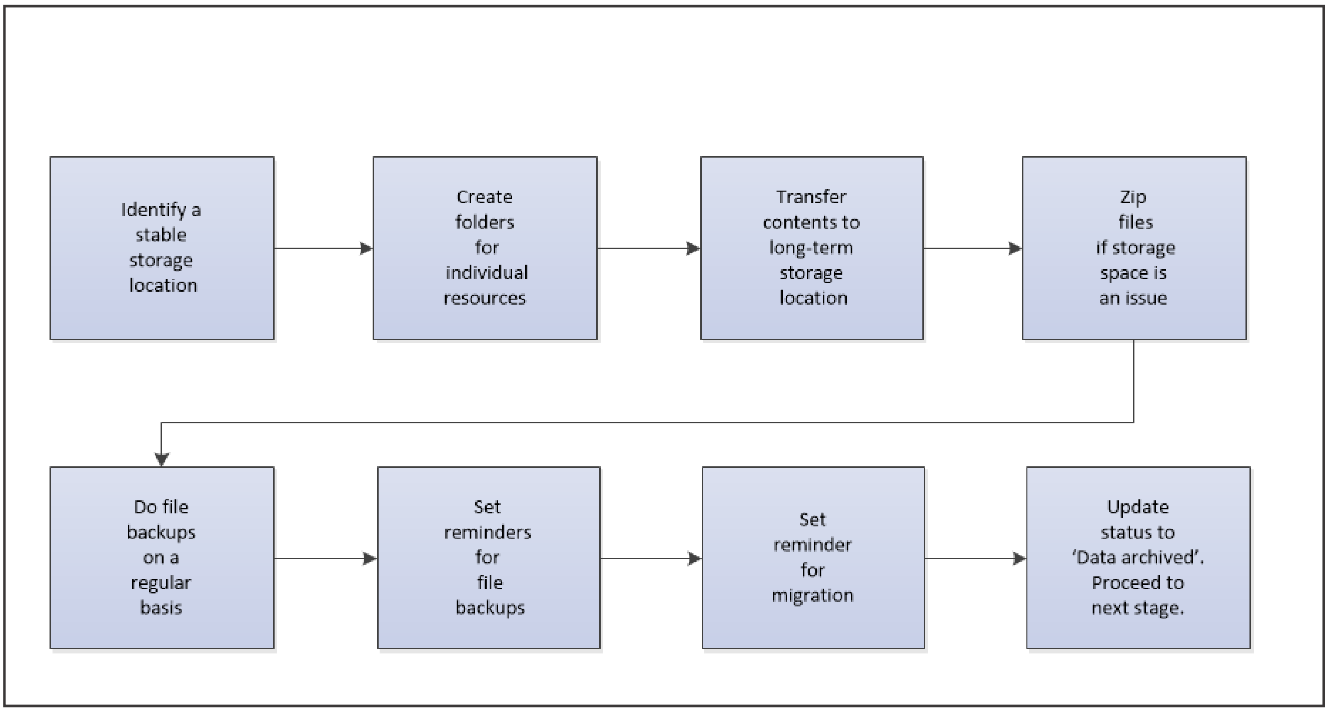

Identify a stable storage location
Create folders for individual resources
Transfer contents to long-term storage location
Zip files if storage space is an issue
Do file backups on a regular basis
Set reminders for file backups
Set reminder for migration
Update status to 'Data archived'. Proceed to next stage.

Figure 3.9: Stage 8: Archiving the data

The eighth stage of the model pertains to the archiving of the rescued data and involves activities ensuring that the data will be preserved in the long term. These activities are as follows:

- **Identify a stable storage location** for the rescued digital data.
- **Create folders** for the individual rescued datasets.
- **Transfer the contents**, with its metadata and data documentation, to the long-term storage location.
- **Files might need to be zipped** should digital storage space be an issue.
- **File backups** need to be performed on a regular basis.
- **Reminders** need to be set to ensure regular backups are performed.
- **Data will need to be migrated to an updated format** approximately every five years. Reminders need to be set of migration activities.

It is important to note that archival activities, such as data backups and data migration, need to be performed even after the project has been completed. This means that even though data have been rescued, details of the rescue venture published and shared, and a final report published, data archiving activities will continue.

The rescue team can now proceed to the final stage, titled 'Project closure'.

3.9.9 Stage 9: Project closure

A visual representation of Stage 9 is shown in Figure 3.10: Stage 9: Project closure.

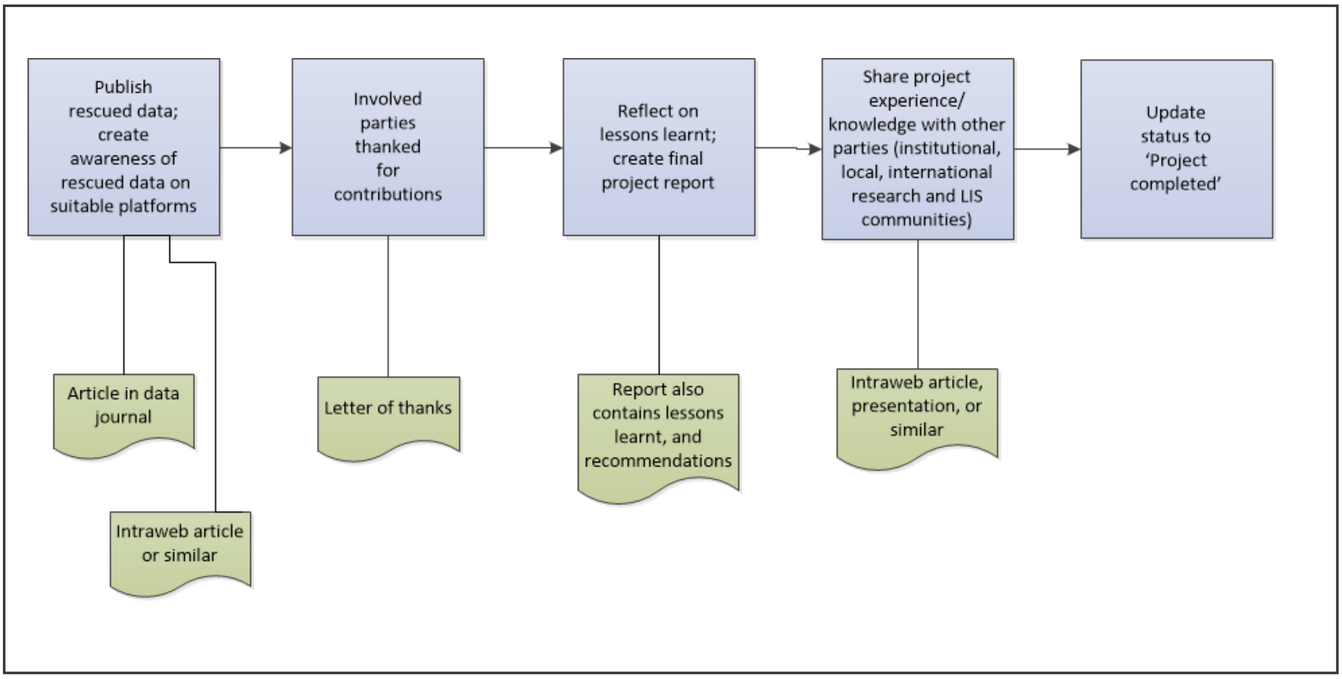

Publish rescued data; create awareness of rescued data on suitable platforms
Involved parties thanked for contributions
Reflect on lessons learnt; create final project report
Share project experience/ knowledge with other parties (institutional, local, international research and LIS communities)
Update status to 'Project completed'
Article in data journal
Intraweb article or similar
Letter of thanks
Report also contains lessons learnt, and recommendations
Intraweb article, presentation, or similar

Figure 3.10: Stage 9: Project closure

The ninth stage of the model contains details of the activities forming part of the closure phase of the data rescue project. These activities are as follows:

- **Publish details of data rescue project**
    - It is of considerable benefit to the wider research community to publish details about the data rescue project. This can be performed in numerous ways, and all options should be considered.
        - Publishing an article (in a scholarly or popular journal) about the data rescue venture is an option.
        - In addition, rescued data can be published in a designated data journal.
        - Details of the rescue project and available rescued data can also be published via an institute's intranet, or similar channels.
- **Parties to be acknowledged and thanked**
    - All involved parties should be thanked for their contribution towards the rescue project.
- **Draft a final report**
    - Drafting of a final report is paramount during this stage.
    - It is vital for the rescue team to reflect on lessons learn during the rescue project, and include these details in the report.

    - Additionally, recommendations for future data rescue efforts should also be put forward.
    - The final report is to be distributed to all team members, added to the institute's institutional repository (closed repository, also open repository if permissible), and its link included in other published outputs pertaining to the project, where permissible.
- **Share rescue experience with a range of stakeholders and interested parties**
    - The project experience should ideally be shared with a range of stakeholders and interested parties. This sharing includes, but is not limited to the following:
        - institutional sharing, in particular research staff and the research library,
        - the wider LIS community, via conferenced outputs, workshop participation, or other disciplinary meetings, and
        - any sector of academia or research dealing with data at risk, digital curation, archival activities, and repository involvement.
- **Certain rescue activities are ongoing**
    - Certain rescue activities, such as data backups and data migration, need to be performed even after the project has been completed.

This stage comprises the final stage of the initial Data Rescue Workflow Model.

3.10 Chapter summary

This chapter contains a description of the steps followed to create the initial Data Rescue Workflow Model.

Following the review of published literature on data at risk and data rescue (i.e., Chapter 2: Literature Review), a number of publications containing references to data rescue workflows, processes, steps and guidance were consulted, reviewed and analysed. Fifteen publications, comprising 14 publicly available data rescue workflows/models and one unpublished workflow, showing variety in features such as research discipline, year of rescue, data formats rescued, complexity of rescue, and modernity of rescue equipment and tools, were studied and discussed.

The initial Data Rescue Workflow Model was created based on the insight gained via the process of content analysis. It is a generic data rescue model (i.e., discipline-agnostic), is directed at the rescue of paper-based media and does not contain references to the roles and responsibilities of role-players such as library and information science professionals, researchers, ICT staff, or archivists. This initial version of the model, regarded as tentative in nature, was reviewed and critiqued by Sample B members, to provide additional feedback and recommendations that led to the drafting of a revised Data Rescue Workflow Model.

The initial model consists of nine data rescue stages. The model consists of a single-page summary, as well as stage-specific graphics for each of the nine data rescue stages. Stage completions are mostly sequential in nature, but not necessarily so. Storage and preservation of data at risk, and data archiving activities, are examples of stages that are perpetual in nature and to be monitored and executed even after project completion.

Several of the imaged stages of the model contain references to outputs created during the stage, templates to be consulted during the stage, and guidelines to be consulted during the stage. The contents of these referenced documents can all be viewed as appendices to this study.

The next chapter contains a description of the methodology followed during this research.

CHAPTER 4: METHODOLOGY

4.1 Introduction

This chapter has two distinct purposes. The primary purpose is to demonstrate how the major methodological parts work together to address the research questions of the study, while the secondary purpose entails the provision of sufficient detailed information to enable another experienced researcher to replicate the study.

Addressing the primary chapter purpose requires showing clearly how the methodological aspects, such as the choice of research approach and research design, population involved, sampling choice, instruments used, data collection and data analysis were implemented to investigate the study's research questions. These questions are:

- What are the roles and responsibilities of the research library within a comprehensive workflow for data rescue? (Main research question)
- What do the current data rescue frameworks/workflows look like?
- How do current South African workflows in data rescue compare with best practices/guidelines internationally?
- What is the current documented state of library and information services involvement with data rescue, seen in the global context?
- What is the current documented data rescue awareness within the South African library and information services community?
- What is the current documented state of data rescue involvement/participation within the South African library and information services community?
- What is the current documented state of data rescue globally and in SA?
- To what extent can the theory and practice be formalised in a model for a data rescue workflow?
- What suggestions could be made to include data rescue topics in the LIS curricula?

The choice of methodological aspects and steps will be listed, details supplied, and a rationale for the choice of procedure or technique supplied. It is crucial that the chapter contains clarity on how these procedures and techniques will enable the answering of the research questions.

This chapter is also intended as a guide should researchers wish to replicate this study. To achieve this objective, the study steps, techniques and process will be listed in as much detail as possible without the chapter resembling a manual. It is vital that researchers who wish to replicate (or evaluate) the study can obtain the required information on how data were collected and analysed.

Chronologically, the chapter consists of two distinct components: a bigger picture describing the research approach, and a section describing how the research was undertaken. The first section therefore contains a theoretical justification for the methodology chosen, while the latter section comprises the actual steps followed to address the research objectives.

Methodology aspects included in this chapter include the research philosophy, the research approach, the research design, and data collection instruments used. Thereafter, the study participants and sampling method used are discussed. The chapter also details the ethical considerations forming part of the study. The latter part of the chapter contains a description of data analysis methods used, as well as an abbreviated step-by-step outline of the data collection method. The chapter also contains, within each section, a discussion of the strengths and weaknesses of methods chosen. Limitations, delimitations and assumptions of the study are provided. The data management plan, created for this study, is attached as Appendix 22.

4.2 Research philosophy

According to Saunders, Lewis and Thornhill (2019: 130), the term ‘research philosophy’ refers to a system of beliefs and assumptions about the development of knowledge, while Crossly and Jansen (2021) describe it as the foundation of any study, as it describes the set of beliefs the research is built upon. Adding to this is the statement by Dudovskiy (2016) that the research philosophy comprises the beliefs about the ways in which data about a phenomenon should be collected, analysed and used. With this study focused on the development of knowledge in the broad field of information science, and data rescue in particular, it is necessary to describe the philosophy this research adhered to.

As stated by Žukauskas, Vveinhardt and Andriukaitienė (2018), four main trends of research philosophy can be distinguished and discussed. These four main philosophies are the realistic research philosophy, the positivist research philosophy, the interpretivist research philosophy, and the pragmatist research philosophy. Each of these portray a different worldly view pertaining to study investigations, as was confirmed by the Royal Content Research Services (2021).

The realistic research philosophy, or realism, supports a scientific journey towards the discovery of a truth rather than fiction (Royal Content Research Services, 2021), and focuses on reality and beliefs existing in a certain environment (Žukauskas, Vveinhardt and Andriukaitienė, 2018).

The positivist research approach claims that the world can be understood in an objective way and that the researcher is an objective analyst, dissociates him/herself from personal values, and works independently (Žukauskas, Vveinhardt and Andriukaitienė, 2018). Law-like generalisations form part of this research philosophy (Saunders, Lewis and Thornhill, 2019: 103), and this point of view is supported by Crossley and Jansen (2021), who state that in positivism there is only one reality and that all meaning is consistent between subjects.

The interpretivist research philosophy emphasises the influence that social and cultural factors can have on an individual (Crossley & Jansen, 2021). According to Saunders, Lewis and Thornhill (2019: 106–107), this philosophy advocates that it is necessary for the researcher to understand differences between humans, that the researcher should adopt an empathetic stance, and that the challenge of the stance is to enter the social world of the research subjects and understand their world from their point of view. Crossley and Jansen (2021) also emphasise the active role of the researcher in the study, adding that the interpretivist philosophy requires the researcher to draw a holistic view of the participants and their actions, thoughts and meanings (Crossley & Jansen, 2021).

According to Dudovskiy (2016), pragmatists recognise that there are many different ways of 'interpreting the world and undertaking research', and that it is not possible for one single point of view to provide the entire picture. Followers of the pragmatic research philosophy, or pragmatism, see knowledge as fallible (Education Studies, University of Warwick, 2017), a view supported by Dudovskiy (2016), who states that according to the pragmatic research philosophy there may be multiple realities. Saunders, Lewis and Thornhill (2019: 122) mention that pragmatists believe that a study's research question is the most important determinant of the research philosophy selected, and that one approach might be better than another for answering a particular study's questions (Saunders, Lewis and Thornhill, 2019: 110). This statement is supported by Žukauskas, Vveinhardt and Andriukaitienė, who maintain that the choice of research philosophy is mostly determined by the research problem.

The pragmatic research philosophy, highlighting the importance of using the best tools possible to investigate phenomena, is the philosophy forming part of this research study. Crossley and Jansen have stated that the intention of pragmatism is to approach a study from a practical point of view and to realise that knowledge is not fixed but instead is constantly 'questioned and interpreted' (2021). This researcher is in agreement with their stance and has held a d practical viewpoint during this study, while aiming to question and interpret findings. Additionally, the study involves an element of 'researcher involvement and subjectivity' (Crossley & Jansen, 2021), and this aspect is especially relevant when drawing conclusions around data rescue based on the study participants' responses and decisions. This study involves a practical approach and entails, as stated by Saunders, Lewis and Thornhill (2019: 607), the integration of different perspectives to assist with the collection and interpretation of data. The use of triangulation, as a component of research reliability, was implemented in this study and is evidence of different perspectives. Method triangulation and data triangulation formed part of this study and is elaborated on in Section 4.11.2.3.

4.3 Research approach

A research approach is defined by Creswell as 'plans and procedures' for the intended research and includes everything from broad assumptions to intricate details about the methods of data collection, data analysis, and data interpretation (2014: 4). A variety of research approaches are available for individuals embarking on a research project. The approach chosen, or most suited to the study, depends on many factors. Aspects such as the purpose of the research, the research questions to be answered, and resources available, all play a crucial role when deciding on a research approach (Creswell, 2014: 4).

Research methodology literature focuses on three distinct research approaches: quantitative, qualitative, and mixed methods research (Edmonds & Kennedy, 2017; Sarwono, 2022). As a lengthy discussion of the features and use of each of these approaches is not the objective of this chapter, the following table, stipulating the key features of each approach, is regarded as sufficient. The table should not be seen as a document encapsulating all features of the three approaches, but as a glimpse into the main identifying features and uses of approaches, as stated in documented sources. Listed features should also be viewed as 'generally' or 'usually', as exceptions to the rule do exist when making use of a specific approach.

Table 4.1: Characteristics of research approaches

RESEARCH APPROACH	CHARACTERISTICS AND USES
Quantitative approach	• Type of data collected: most often numbers and statistics (McCrindle, 2017: 13), numbers and values (Pickell, 2021), and data that can be counted or compared on a numerical scale (Dewitt Wallace Library, 2021). • Relies on numbers, rates and percentages typically presented in a table, gird or chart (Hesse-Biber & Leavy, 2004: 1). • Objectivity is critical (McCrindle, 2017: 13). • Research objective is to describe, explain and predict (McCrindle, 2017: 13). • Factual data are required to answer research questions (Hammarberg, Kirkman & De Lacey, 2016: 499). • Complex structured methods of analysis. Limited flexibility. Concrete framework (Pickard, 2013: 18). • Statistical relationships are identified (McCrindle, 2017: 13). • Large group of participants; randomly selected (McCrindle, 2017: 13). • Linear research design (Pickard, 2013: 18). • Allows the identification of cause–effect relationships (Leedy & Ormrod, 2005: 135). • Tests objective theories by examining the relationship among variables (Creswell, 2014: 4). • Purpose: to test hypotheses, look at cause and effect, make predictions (McCrindle, 2017: 13). • Final written report has a set structure (Creswell, 2014: 4).

RESEARCH APPROACH	CHARACTERISTICS AND USES
Qualitative approach	• Type of data collected: text and words as opposed to numbers (Hesse-Biber & Leavy, 2006: 8). • Relies on numerous forms of data (Leedy & Ormrod, 2005: 133). Collects words, images and objects as data (McCrindle, 2017: 13). • Subjectivity is expected (McCrindle, 2017: 13). • Data analysis involves the non-numerical examination of observations to discover patterns and meaning (Walwyn, 2017: 22). Findings are not arrived at by statistical procedures or other means of quantification (Strauss & Corbin, 1998: 10–11). • An approach for exploring and understanding the meaning individuals or groups ascribe to a problem (Creswell, 2014: 4). Researcher makes interpretations of the meaning of the data (Creswell, 2014: 4). Answers questions about experience, meaning and perspective (Hammarberg, Kirkman & De Lacey, 2016: 499). • Smaller group of research participants; not randomly selected (McCrindle, 2017: 13). • Examines data from various angles (Leedy & Ormrod, 2005: 133). Less linear design than for quantitative research (Pickard, 2013: 18). • Usually aspires to create a rich meaningful picture of a complex multifaceted situation (Leedy & Ormrod, 2005: 133). • Generally, does not allow identification of cause–effect relationships (Leedy & Ormrod, 2005: 135). • Requires considerable preparation and planning (Leedy & Ormrod, 2005: 134). • The 'power gap' between the researcher and the study population in qualitative research is far smaller than in quantitative research due to the informality in structure and situation in which data are collected (Kumar, 2011: 104). • Methodology might evolve over the course of the investigation of cause–effect relationships (Leedy & Ormrod, 2005: 134). • Allows the participants to raise topics and issues that were not anticipated and that might be critical to the investigation (Kuada, 2015: 57). • Purpose: to understand and interpret situations (McCrindle, 2017: 13). • The final written report has a flexible structure (Creswell, 2014: 4).
Mixed methods approach	• Involves a combination of methodologies (Pickard, 2013: 18). • Can take many forms. There is no single mixed methods design (Pickard, 2013: 18). • Core assumption: the combination of qualitative and quantitative approaches provides a more complete understanding of a research problem than either approach alone (Creswell, 2014: 4). • Different approaches and classifications of approaches exist. A well-known example is the classification by Creswell (1999: 463) into: Convergence model, Sequential model, and Instrument-building model.

This study made use of the qualitative approach, the choice of which was based on the following aspects:

- This study made use of small groups: Sample A consisted of 49 invited members, Sample B had 18 invited members, and Sample C was made up of three participants. More information on the sample selection and sample rationale can be found in Section 4.7: Sample and sampling.

- Non-random sampling was used in this study: sample participants were purposively selected, and criterion sampling and expert sampling were the techniques implemented. More information on the sample selection and rationale can be found in Section 4.7: Sample and sampling.
- Data were overwhelmingly in the form of words and text, as opposed to numbers: the web-based questionnaire, virtual one-on-one interview, the data rescue feedback, and the mini focus group session contained data in the form of words.
- Questions about activities, experiences, opinions and requirements were posed to participants: the web-based questionnaire, virtual one-on-one interview, the data rescue feedback stage, and the mini focus group session made use of questions asking participants about their data rescue experience, knowledge, requirements or opinions. Comparable questions regarding data at risk were used.
- This researcher made interpretations of the meaning of the data: a certain amount of interpretation was applied when analysing literature pertaining to data rescue steps, and when analysing the interview and mini focus group responses provided by participants.
- A certain level of subjectivity was involved: subjectivity was involved with purposive sampling, when deciding on which criteria the sample members should be selected.
- The approach allowed the participants to raise topics and issues that were not anticipated and that might be critical to the investigation. An example of such an unanticipated topic relates to the interview stage, when participants provided illuminating information regarding laboratory notebooks and data collection while doing field work.
- Considerable preparation and planning were involved: extensive preparatory work pertaining to all data collecting instruments, as well as gaining ethical approval and implementing a pilot testing of the tools, formed part of the study.
- Cause–effect relationships could not be established: the nature of the data collected, and the data analysis tools applicable meant that a clear picture of data rescue could be obtained; however, despite gaining insight into the topic, the data collected does not permit the demonstration of cause and effect.
- Findings were not arrived at by statistical procedures or other means of quantification: data analysis involved content analysis and thematic analysis; statistical analysis using numerical data was not performed. The study's data analysis methods are discussed in Section 4.13: Data analysis.
- This study was focused on creating a rich meaningful picture of the institute's data at risk, and data rescue activities, resulting in a usable Data Rescue Workflow Model.

Given all the above, the qualitative research approach was considered the most suitable research approach for the type of data that would be collected, and the research questions that would be answered. There are elements of quantification in the web-based questionnaire, and this could create the impression that the study makes use of multi-methods research. However, the quantitative nature of some of the web-based questionnaire's questions was considered crucial at that stage, as this researcher wanted to learn more about the phenomena of data at risk and data rescue at a specific stage of the research. The study is predominantly qualitative in nature.

The next section describes the research design chosen for this study.

4.4 Research design: Case study

When deciding on this study's research design, the researcher considered the research questions that had to be answered, and the type of information and data that would be required from the study participants. This study required information (and as such, data) in the form of feedback from researchers regarding the institute's data at risk, as this researcher was interested in collecting information pertaining to the current behaviours, opinions, characteristics, expectations and knowledge of researchers with regard to data at risk. The same can be said of the data rescue-related aspects: details about practices, opinions, ideas, knowledge and requirements, forthcoming from staff members (SET-based experts as well as non-SET-based experts), were collected.

Another aspect pertinent to the study was the issue of sampling, as this study investigated the data rescue (and data at risk) behaviours and attitudes of certain researchers and LIS sector professionals based at the institute. In other words, making use of a predetermined and purposive sampling method was crucial in ensuring that the correct sample(s) would be involved in the study.

While several research designs were available to the researcher (including a survey, or a quasi-experiment), the choice was made to use a case study. This research method is defined as an 'intensive study' about a 'person, a group of people or a unit', with the intent to 'generalize over several units' (Heale & Twycross, 2018: 7). Crowe *et al*. (2011: 1) state that a case study can be defined in a variety of ways, but that the central tenet of the design is the need to 'explore an event or phenomenon in depth and in its natural context'. Due to the unique nature of the case study approach, it is often referred to as a 'naturalistic' design. According to Crowe *et al*. (2011: 1), this contrasts with an 'experimental' design, where the researcher is focused on exerting control over and manipulating the variables of interest. In this study, with a focus on collecting data enabling the drafting of a workflow model to be used when rescuing data, it was decided to investigate the data at risk features, and data rescue activities of a select group of employees at a chosen research institute.

The following table demonstrates how the case study characteristics suit the features of the current study. The left column lists common features and uses of case study research, while the right column contains short notes describing the relevance of listed features to the study. While the list should not be viewed as a definitive description of the case study design, it explains the reasoning when deciding on the most suitable design for this study.

Table 4.2: Features of case study research, and applicability to the study

FEATURES OF CASE STUDY RESEARCH	STUDY IMPLEMENTATION
Use/purpose: • Allows in-depth, multi-faceted explorations of complex issues in their real-life settings (Crowe *et al*., 2011: 1; Yin, 2014: 16). • Useful when an experimental design is either inappropriate to answer the research questions posed or impossible to undertake (Crowe *et al*., 2011: 8; Zainal, 2007: 5). • Helps to understand and explain causal links and pathways (Crowe *et al*., 2011: 4). • Is a versatile form of qualitative inquiry (Harrison *et al*., 2017: 12). • Can address a wide range of questions that ask the why, what and how of an issue and assist researchers to explore, explain, describe, evaluate and theorise about complex issues in context (Yin, 2014; Harrison *et al.,* 2017: 15). • Good for describing, comparing, evaluating and understanding different aspects of a research problem (McCombes, 2022: 1). • Allows the researcher to take a complex and broad topic, or phenomenon, and narrow it down into manageable research questions (Heale & Twycross, 2018: 7). • To locate the factors that account for the behavioural patterns of the given unit (Kothari, 2004: 113).	• This study involves an in-depth investigation into the data rescue activities performed by the selected research institute. It involves describing, comparing, evaluating and understanding various aspects of the institute's data rescue experiences and needs. • The study endeavours to describe, understand and create a Data Rescue Workflow Model. • The study takes the concept of data rescue, and a Data Rescue Workflow Model, and will narrow it down to manageable research questions. • The study will locate the factors that account for the data rescue behaviours of the selected research institute.
Research disciplines: A range of disciplines make use of case study as a research method (Kothari, 2004: 113; Crowe *et al*., 2011: 1; Yin, 2014: 4; Harrison *et al*., 2017: 1).	• Case study research can be used when performing research in Information Science.
Type of research: Qualitative and quantitative methods and data can be used (Crowe *et al*., 2011: 6; Walliman, 2011: 93; Yin, 2014: 220; McCombes, 2022: 1). A case study can lead to subsequent quantitative research by pointing to issues that can be investigated (Wellington & Szczerbinski, 2007: 69).	• This study will be making use of mostly qualitative information, with some quantitative data gathered during the initial stages of data collection.
Data collection methods: A range of methods are used to collect data (Hancock & Algozzine, 2006: 24; Walliman, 2011: 94, Harrison *et al*., 2017: 12). Methods include interviews, observations, focus groups, and examining and exploring artefacts. The case study relies on multiple sources of evidence (Yin, 2014: 17, 119).	The data collection methods forming part of this study are: • Review and analysis of data rescue documented outputs • Online questionnaire • In-person virtual interview • Feedback on a proposed model • Mini focus group session
Sampling: Makes use of purposive sampling (Hancock & Algozzine, 2006: 24; TESOL International Association, 2021; 1). Kothari (2004: 116) states that sampling is often not possible. A strong case study does not require a random or representative sample (McCombes, 2022: 1). Yin states that it is desirable to refrain from referring to any kind of sampling in case studies (2014: 44).	• Three different samples will be selected during the study, all of which will involve purposive sampling (criterion sampling and expert sampling).

FEATURES OF CASE STUDY RESEARCH	STUDY IMPLEMENTATION
Assumptions: Kothari (2004: 114) states that several assumptions are made when conducting case study research, including: • Assuming uniformity in human nature • Assuming comprehensiveness of data collection	• It is assumed that the created output (Data Rescue Workflow Model) will be applicable to researchers outside of the selected institute. • It is also assumed that the various data collection stages will acquire sufficient, reliable, valid and comprehensive data.
Types of case study: Literature reveals different classifications of case study research. Classifications include: • Three types: Intrinsic (to learn about a unique phenomenon), instrumental (uses a particular case to gain a broader appreciation of an issue or phenomenon), and collective, where the study involves studying multiple cases simultaneously or sequentially to generate an even broader appreciation of a particular issue (Hancock & Algozzine, 2006: 32; Crowe *et al*., 2011: 2). • Two types: Single vs multiple case design; this simply means choosing whether the study will include just one, or several cases (Zainal 2007: 2; Adolphus, 2021: 1). • Yin (2014: 189) identifies several types of case studies: ○ Exploratory: The case study is used to define questions and hypotheses – or to test out a research procedure – for a further piece of research, such as a large-scale survey. ○ Descriptive: The case study is used to describe a particular phenomenon within its context. It can be used to expand on a particular theme unearthed by a survey. ○ Explanatory: The case study explores cause–effect relationships, and/or how events happen.	• The study will be an instrumental case study, as the study uses a particular case to gain a broader understanding of data rescue. • The study will be a single case design. • The study will be descriptive in nature, as the researcher will attempt to completely describe the different characteristics of data rescue in its context. Expanding on the aspects discovered during the literature review, literature content analysis and data collected during the study's empirical phase form the cornerstone of the study.
Main steps: The steps when using case study methodology are the same as for other types of research (Heale & Twycross, 2018: 7). Descriptions of typical steps include: • Recognition of unit to be studied, and its status • Collection of data, and examination • Diagnosis and identification of causal factors • Application of remedial measures • Follow up to determine effectiveness (Kothari, 2004: 114–115). Steps described by Crowe *et al*. (2011: 5–7) are: • Defining the case • Selecting the case • Collecting and analysing data • Interpreting the data • Reporting the findings	• The steps followed during this study will mirror the steps described by Crowe *et al*. (2011: 5–7).
Data analysis: According to Yang (2013: 5), case study data can be analysed in several ways: • Holistic (analysis of entire case) • Embedded (analysis of specific aspects of case) • Within case analysis (detailed description of each case and themes in the case) • Cross-case study followed by thematic analysis across cases	• The entire case will be analysed; the study makes use of holistic analysis.

The information contained in the table above underlines the applicability of the case study design to this study. The table and its overarching message can be summarised as follows: this study required an in-depth investigation into data rescue practices, data rescue challenges, and data rescue requirements of several samples of employees at the involved institute. Samples were purposive in nature, and included researchers and research library experts. Data collection techniques used in the study included content analysis (applied to literature), a web-based questionnaire, in-person interviews, text-based feedback supplied by participants, and a mini focus group session. Findings were used to create a Data Rescue Workflow Model, indicating the roles and responsibilities of the research library.

In addition to the nature of this study being more suited towards case study, as revealed in the table above, several key questions were also considered before deciding on the use of case study as the study design. These questions included: 'Can the required data be collected via any other research design?' and 'Is it more important that the data be obtained by means of a more naturalistic design?' The decision was made to implement case study design, being the option best suited to answering this study's research questions.

The steps adhered to in this case study closely resembled the case study steps described by Crowe *et al.* (2011: 5–7). While the steps are detailed in Section 4.9: Research process, it is worth including a concise listing of planned steps used in this case study, which are as follows:

- identify the topic (see Chapter 1: Introduction),
- establish research objectives and research questions (see Chapter 1: Introduction),
- analyse relevant literature (see Chapter 3: Literature and the creation of a data rescue workflow model),
- design a model indicating data rescue workflow (see Chapter 3: Literature and the creation of a data rescue workflow model),
- identify the study population (see Section 4.6: Research population),
- apply sampling: Sample A (see Section 4.7.1: Sample A: Web-based questionnaire recipients),
- select, design and plan data collection instruments/methods (see Section 4.5: Data collection methods and tools/instruments), which entailed:
  - a web-based questionnaire,
  - a virtual one-on-one interview schedule,
  - feedback provided after reviewing the initial Data Rescue Workflow Model, and
  - a mini focus group session,

- pilot the web-based questionnaire (see Section 4.9.6: Pilot runs of the web-based questionnaire),
- obtain managerial approval and ethical clearance for the study (see Section 4.9.8: Obtain managerial approval for data collection; and Section 4.12: Ethical considerations and ethical clearance),
- collect data: web-based questionnaire (see Section 4.9.13: Distribution of web-based questionnaire),
- analyse data (see Section 4.9.16: Data analysis of web-based questionnaire),
- apply sampling: Sample B (see Section 4.9.17: Selection of Sample B),
- collect data: virtual one-on-one interview (see Section 4.9.21: Conduct virtual one-on-one interviews),
- collect data: Data Rescue Workflow Model feedback (see Section 4.9.22: Share flowchart feedback guide),
- analyse data (see Section 4.9.26: Data analysis of interview data; Section 4.9.27: Data analysis of feedback data),
- revise and amend the Data Rescue Workflow Model (see Section 4.9.28: Revise the initial Data Rescue Workflow Model),
- apply sampling: Sample C (see Section 4.9.29: Select Sample C: Institutional experts),
- collect data: mini focus group session (see Section 4.9.32: Mini focus group session),
- analyse data (see Section 4.9.35: Analysis of mini focus group session data),
- update and finalise the Data Rescue Workflow Model (see Section 4.9.37: Update and finalise the revised Data Rescue Workflow Model), and
- present findings and make conclusions and recommendations (see Chapters 5 and 6).

The next section discusses the research methods used in documented data rescue studies.

4.4.1 Research methods used in other data rescue studies

When examining the types of research methods used during previous documented studies on data at risk or data rescue, it is important to observe the nature of the different published outputs. For the purposes of reporting on research methods found in previous data rescue studies, this researcher has not considered data rescue studies detailing a data rescue project, but has examined studies where empirical data research methods were implemented. The use of case studies, surveys, focus groups, data inventorying, artefact analysis and literature reviews is described in this section.

A high number of data rescue studies made use of **case study research**. It is also important to note that the approach is usually qualitative. A prime example of case study use during data rescue research is the research by Brunet and Jones (2011), where several case studies were implemented. The aim of the case studies was to document the potential benefits of undertaking integrated data activities and to describe some of the emerging and ongoing data rescue initiatives aimed at improving the availability and accessibility of historical climate data. Downs and Chen (2017) also made use of case study research when investigating issues raised during a data rescue effort from an existing archive that had not fully curated the original data.

A third example of case study design involves Fry's work describing the efforts of Canada's Carleton University to clean, archive, disseminate and maintain data held at the Centre for Research and Information on Canada (CRIC). CRIC had contacted and informed the university that the centre was closing down, and as a result the university accepted the task of rescuing the data about to be lost (2010b).

A fourth example refers to the study of Shiue, Clarke and Fenlon who had identified 18 data assessment factors after conducting three case studies to investigate potential issues of curating collections with the purpose of data recovery and reuse (2020). Lastly, case study research was used by Palmer, Weber and Cragin to engage with 20 participants who were active researchers in geology, oceanography and environmental science. While many more case study examples can be found pertaining to data rescue research, the five examples described above serve to illustrate the prevalent use of the design within this disciplinary sphere.

Evidence of **survey** use in data rescue research is also found. For example, Murillo stated that demographic surveys were used to gather participants' information, such as department, research area, position, years of research and age (2014). Fry (2010a) mentioned that the WMO made use of survey research when requesting information on current hydrological data rescue requirements from national hydrological services worldwide. A third example is the study by Thompson, Davenport Robertson and Greenberg who referred to the use of a survey to fully understand the data-at-risk predicament and collect information on the data characteristics and preservation plans from 43 information custodians (2014).

Interviews formed part of several data rescue studies. Palmer, Weber and Cragin mentioned that semi-structured interviews with 20 participants were used during their sequenced, multi-method data collection approach (2011). The authors elaborated on their data collection method and stated that pre-interview worksheets were used to orient participants to the authors' specific interests in their data, setting the scene for the questions that would follow in the research interview. According to the authors, the worksheet responses often provided important domain-specific information necessary for the curation process, and facilitated deep discussion on the participants' research practices. The follow-up interviews were important for probing on specific data types identified by participants as having value for other users and to document deposition requirements and curation needed to support reuse.

Schumacher and Vandecreek's study (2015) contained 56 faculty interviews, and participating individuals represented a wide range of disciplines, including specialists in the humanities, the physical sciences, the biological sciences, the social sciences, engineering and education. During the individual interviews, participants were asked to describe their professional activities, the nature of the digital information they had created in the course of their work and the types of data formats included in their body of work. They were also asked to describe how they stored and managed their data, and to identify those materials they considered to be most valuable and hence would most want to recover in the event of their apparent loss. Subsequent questions requested participants to report experiences of data loss (if any) and describe their level of confidence in the long-term availability of their digital files.

An examination of the literature showed that **focus groups** were also used, but to a lesser extent. Murillo (2014) mentioned that four one-hour focus group sessions were conducted with 14 scholars from selected scientific disciplines. The focus group method was selected due to its efficiency in generating new ideas from a group of people and being a proven method for gathering 'nuanced perspectives'.

As part of their sequenced, multi-method approach for data collection, and in addition to interviews, Palmer, Weber and Cragin employed **data inventorying and artefact analysis** to produce 'dense, high-quality' case study units of evidence (2011).

Most studies refer to the importance of the use of a literature review, and Hoffman and colleague's literature review on data rescue (2020) and the work of Downs and Chen (2017) when reviewing data rescue, data curation, rapid appraisal and knowledge retention are excellent examples.

The next section discusses the advantages of case study research.

4.4.2 Advantages of case study use

Several advantages are ascribed to the use of case study research, with benefits not only for the study, but also for the research discipline and researcher. Benefits of case study research applicable to this study are listed below.

- Case study research enables full understanding of the case being studied (Kothari, 2004: 115).

- Findings contain a real record of personal experiences (Kothari, 2004: 115); the examination of the data is most often conducted within the context of its use.
- Case study research can obtain information not possible via other methods (Kothari, 2004: 115).
- Case study research can implement one of more data collection methods, and rely on multiple sources of evidence (Kothari, 2004: 115; Yin, 2014: 17, 119; UKEssays, 2021). It also offers breadth and diversity (Zainal, 2007: 5).
- Making use of case study research can be a practical solution when a big sample population is difficult to obtain (Zainal, 2007: 5).
- Variations in terms of intrinsic, instrumental and collective approaches to case studies allow for both quantitative and qualitative analyses of the data (Zainal, 2007: 1).
- The detailed qualitative accounts often produced in case studies help to explain the complexities of real-life situations which may not be captured through experimental or survey research (Zainal, 2007: 4).
- For many case studies, no more than a few participants are required (Scott, 2013: 14).

The following advantages can be ascribed to case study research:

- Descriptive case studies can make complex science and technology projects accessible and interesting to a non-scientist audience (UKEssays, 2021).
- Case study findings can lead to subsequent quantitative research by pointing to issues that can be investigated (Wellington & Szczerbinski, 2007: 69).
- Published case studies can be of great value in teaching and learning (Wellington & Szczerbinski, 2007: 68).
- The use of case study as research design is readily complemented by the use of other quantitative and statistical methods (Yin, 2014: 22).

Case study research may also be beneficial to the researcher for the following reasons:

- Case study research can enhance the experience of the researcher, and lead to improved analytical skills (Chronopoulou *et al*., 2016; Mahdi, Nassar & Almuslamani, 2020).
- Certain case studies can prove to be an enjoyable research experience (Wellington & Szczerbinski, 2007: 68).

While case study research is linked to many benefits, certain limitations, discussed in the next section, should also be considered.

4.4.3 Limitations of case study

Despite its advantages, case study research also portrays several limitations or pitfalls. These limitations are listed below.

- Cases study findings are often criticised for not being generalisable, not being representative, not being typical, not being replicable, and not being repeatable (Wellington & Szczerbinski, 2007: 69; Yin, 2014: 20).
- Defining or bounding the case can be difficult, as many points of interest and variables intersect and overlap (Crowe *et al*., 2011: 7; Harrison *et al.,* 2017: 12).
- Case study research often involves a lot of time and requires substantial expenditure (Kothari, 2004: 116), and concerns are often raised about the unmanageable level of effort (Yin, 2014: 21).
- The volume of data, together with the time restrictions in place, may impact on the depth of analysis that is possible within the available resources (Crowe *et al*., 2011: 7).
- The large volumes of data collected may not be relevant to the study (Crowe *et al*., 2011: 7).
- The small volume of data may be too little to be of any value (Crowe *et al*., 2011: 7).
- The data collected are often not considered as significant scientific data (Kothari, 2004: 116).
- Case study data can in certain instances be subject to validity concerns, as the subject might write or reveal what he or she thinks the researcher wants. Additionally, the greater the rapport between the researcher and participant, the more subjective the study becomes (Kothari, 2004: 116).
- During a case study, the temptation sometimes exists to veer away from the research focus (Heale & Twycross, 2018: 7).
- Reporting the findings of case studies can be challenging at times, particularly in relation to the word limits for certain journal papers (Heale & Twycross, 2018: 7).

Despite these limitations and pitfalls, researchers may employ several methods to overcome certain drawbacks. Kothari (2004: 116) states that many limitations can be removed if the researcher is aware of pitfalls, and well-trained in collecting, assembling, organising and processing data. Crowe *et al.* (2011: 7) mention the importance of transparency in dealing with the case study's lack of rigour, and that case studies provide little basis for generalisation. Steps to counteract the pitfalls and enable the reader to judge the trustworthiness of the case study report are stated by Crowe *et al.* to be detailed steps pertaining to case selection, steps describing data collection, reasons for selection of chosen methods, and details regarding the researcher's background and level of involvement (2011: 7).

4.4.4 Rationale for case study use

The overarching aim of this study was to gather enough relevant information to enable the creation of a data rescue workflow model that would also indicate the involvement of the research library. To attain this goal, it was vital to obtain insight into current data-at-risk trends, data rescue activities, data rescue workflows, and the perspectives of researchers and the library and information services professionals regarding the aforementioned aspects.

This researcher opted to make use of case study research as the design and used data collection methods such as web-based questionnaires, personal interviews and a focus group. The question may well be asked whether the same information would have been generated had the study used survey research or quasi-experimental research.

Researchers make use of the **survey design** when aiming to reach a large number of people and collect a considerable amount of information. Survey research can be tailored to many types of research questions, polls or opinions and can contain multiple variables. In the context of this study and its desired outcome, making use of a survey design could potentially have involved one (or more) of the following options:

- including all 1 608 researchers at the selected institute in the study and administering an online questionnaire regarding data at risk and data rescue,
- including the selected institute's RGLs as well as expert researchers at other South African research institutes and administering an online questionnaire regarding data at risk and data rescue,
- including a range of South African research libraries and administering an online questionnaire regarding data at risk and data rescue, and
- expanding the study to include southern African countries and researchers and research library staff from those countries.

The above list is not exhaustive but indicates the different ways in which this study could have implemented a survey design instead of a case study. While some aspects of survey research might have produced suitable data, the reasons for not making use of survey research are as follows:

- responses might potentially have referred to institute-specific or in-house tools, systems and infrastructure not applicable to other institutes,
- information might potentially have referred to institute-specific or in-house tools, systems and infrastructure not understood by external parties,

- with a single data collection instrument there would have been no option for method triangulation,
- clarification of questions or responses would not have been possible,
- a low response rate might have been a potential outcome,
- such a research design would have made the building of rapport impossible, and
- the data collected would potentially be greater in volume than case study data, but would not have produced such rich data, and enabled such a detailed and thorough analysis.

Researchers make use of a **quasi-experimental design** when aiming to establish a cause-and-effect relationship between an independent and dependent variable. In the context of this study and its desired outcome, quasi-experimental methods could have been used to indicate how the introduction of a variable such as training, or exposure to data rescue practices could influence data rescue outcomes. With quasi-experiments aiming to demonstrate causality between an intervention and an outcome, their applicability to the current study would have been of limited value. The number and range of variables linked to data at risk and data rescue are too numerous to have enabled the design of an experiment where the demonstration of causality between a variable and an outcome is the crux of the research design. In addition, the problem of confounding variables – a factor other than the one being studied – is a potential outcome.

While these two methods (survey design and quasi-experimental research) were bound to have resulted in novel and thought-provoking findings, these designs would not have been the ideal way of determining what the current state of data at risk and data rescue is, how the research library should be involved, and what a data rescue workflow model should look like.

Compared with the two research designs above, use of the case study design offered the following advantages:

- it enabled the gathering of richer and more detailed information about the data rescue practices and data at risk at the selected institute,
- it yielded information that is more likely to be valid,
- it brought together multiple sources of information,
- it enabled the implementation of triangulation to enhance validity and reliability,
- establishment of rapport during the interview was an opportunity that would not have been presented if using other methods,
- familiarity with systems and procedures ensured an understanding of feedback,
- the heterogeneity existing between the selected institute and other research institutes may potentially have led to the collection of data from institutes with no data at risk, and

- the case study enabled clarification of questions and responses during several of the data collection stages.

The next section provides a background and description of the selected research institute used in this case study.

4.4.5 Background and description of selected research institute

The Council for Scientific and Industrial Research (CSIR) was established in 1945 in terms of the Scientific Research Council Act (No. 33 of 1945) of the Parliament of the Union of South Africa (Council for Scientific and Industrial Research, 2023c). It undertakes directed, multidisciplinary research and technological innovation in an effort to improve the quality of life of South Africans. Several of the institute's research innovations and scientific highlights during the past seven decades are listed below (CSIR, 2023b).

- Telecommunications: The first microwave electronic distance measuring equipment (tellurometer) was presented in 1954 and revolutionised global land surveying.
- Transport research: The Heavy Vehicle Simulator, used for accelerated pavement testing of road materials, was developed in 1965.
- Biomaterials: The bollard, an implantable expanding rivet used in conjunction with a prosthetic ligament for repair of cruciate ligaments in the knee, was invented and developed in the early 1980s.
- Aeronautics: The construction of medium-speed wind tunnels was completed during 1988.
- ICT: The institute's online services were sold during 1997 to MIH Limited to form MWEB.
- Statistical analysis: A forecasting model, used to predict the outcome of national elections based on early voting results, was presented in 1999.
- Natural product chemistry: In 2003 a benefit-sharing agreement was signed between the institute and owners of indigenous knowledge, leading to the development of a locally-produced mosquito repellent candle using the oils of indigenous plants.
- ICT: The institute developed the Digital Doorway, a computer terminal equipped with educational games and applications, design functions and access to the internet. The product was named as one of the top 50 inventions in the world by *TIME* magazine in 2011.
- Laser technology: During 2014, researchers demonstrated a large 3D printer for titanium parts needed to manufacture certain components for the aerospace industry.

The institute receives an annual grant from Parliament through the Department of Science and Innovation, which accounts for about 30% of its total income (CSIR, 2023d). Additional income is generated from contract research for the public and private sectors, locally and abroad, as well as from royalties, licences and dividends from intellectual property management and commercial companies created by the institute.

Several major changes at the institute occurred during the last seven decades, and emanate from international trends, newly-formed research disciplines, and a change in institutional strategic focus. A very recent addition to the institute's research focus entailed assisting local industry to improve its competitiveness by providing access to specialised facilities and skills. This resulted in participants having access to large-scale prototyping and pre-commercial manufacturing infrastructure, equipment, expertise and access to business and technical networks.

Adding to the factors mentioned above are expected and continual research interruptions and upheavals comprising researcher movement between research groups, resignations, retrenchments and retirements. Frequent and dynamic transitions and transformations, while crucial for institutional progression, could potentially be accompanied by less-than-ideal data-related outcomes. Examples of potentially detrimental consequences include the discarding of data when a research group disbands, knowledge of older rescue projects and associated data context leaving the institute, and researchers familiar with older data formats, readers and equipment no longer being available.

Furthermore, the institute's data management implementation was still in its infancy at the time of this study's data collection, with a data management policy drafted but not yet finally approved, data management training not yet rolled out and the institute not in possession of a dedicated institutional data repository. With the potential cumulative effects of factors influencing research continuity and the management of valuable data, the institute was considered an ideal case for use in this study, with evidence of data at risk and attempts at data rescue anticipated to be discovered. Clarification of connections between variables and concepts, such as the nature of risk factors leading to data at risk, or the obstacles experienced when rescuing data, were also expected by this researcher. Linked to these expected insights was the anticipated provision of feedback from institutional experts regarding the usability of a drafted data rescue workflow model.

At the time of the study, the selected institute's SET base comprised nine research clusters to deliver the desired research impact; the clusters were:

- Advanced Agriculture and Food,
- NextGen Health,
- Future production: Chemicals,
- Future production: Mining,

- Future production: Manufacturing,
- Defence and Security,
- NextGen Enterprises and Institutions,
- Smart Logistics, and
- Smart Places (CSIR, 2023a).

Each cluster is made up of several impact areas, which in turn consist of research groups. Research groups are managed by RGLs, with the institute having an RGL complement of 49 researchers during this study's data collection stages. The SET base, having a total of 1 608 researchers, contained 314 PhD graduates and 492 masters graduates.

The institute's research library consisted of three divisions and 18 employees. Heading the library was a portfolio manager, overseeing the records management, information services and technical services divisions. The three divisions comprised the following:

- Records management services section: A records manager, a data librarian, three repository professionals, two archives technicians, a digitisation clerk and a graduate intern.
- Information services section: A manager, two information scientists (with SET research background) and two information specialists.
- Technical services section: A manager, a systems librarian, a cataloguer, a library technician and a graduate intern.

At the time of data collection only four research library staff members were involved with data-related activities; these staff members were the portfolio manager, the records manager, the data librarian and the archives technician.

Institutional systems and platforms, anticipated to show connections to data-related matters included the following:

- Institutional closed repository: known as the Technical Outputs Database (TOdB) and which runs on the Inmagic[24] platform.
- Institutional open repository: known as ResearchSpace and which runs on the DSpace[25] platform.
- Document Management System: Groupwise DMS[26] was used up to 2019.

[24] https://lucidea.com/inmagic-dbtextworks/
[25] https://dspace.lyrasis.org/
[26] https://www.novell.com/documentation/groupwise18/gw18_guide_admin/data/adm_libdoc_overview.html

- Document Management System: Micro Focus Vibe[27] was used between 2019 and 2022.
- Document Management System: SharePoint[28] was used from 2022 onwards.
- The institute uses Oracle (also referred to as Workflow) for many activities including library indexing requests, interlibrary loan requests or requests for external research documents.
- The institute also makes use of an in-house electronic procurement system for relevant activities.
- The institute was at the time of data collection in the process of developing an institutional DMP template on the DIRISA[29] platform.

To summarise: The selected multidisciplinary research institute, established in 1945, has undergone a variety of changes during its existence. A shift in research focus, implementation of new research areas and phasing out of other research areas form part of its dynamic past. With data management still in its infancy, a dedicated data repository not yet procured and the institute displaying features of resource constraints, it was considered an ideal case for a study into data at risk and data rescue, and to involve institutional experts in the creation and evaluation of a data rescue workflow model.

4.4.6 Summary

Various options were considered when deciding on a suitable research design for collecting data and answering this study's research questions. The choice eventually came down to one of two designs, namely the case study and survey design. Case study, as research design, was chosen as it was judged to be the most suitable way of ensuring that research objectives are attained, and research questions answered. Content analysis, already described in a previous chapter (see Section 3.6), forms part of this study but was not part of considerations when deciding on a research design for the empirical stage of the study.

This section also gives context to this case study by providing background information on the selected research institute and the rationale for its selection.

Despite this case study focusing on a single selected research institute, this study also reviewed data rescue workflows, models and processes found in literature (i.e., the content analysis segment of the study). The information gathered via the review, combined with the institute-specific information collected through the case study, was anticipated to provide data that could be used to create a Data Rescue Workflow Model. The model is directed at the rescue of data at risk at the selected research

[27] https://www.novell.com/documentation/vibe4/
[28] https://www.microsoft.com/en-us/microsoft-365/sharepoint/collaboration
[29] https://www.dirisa.ac.za/

institute, but may also be of use to data rescuers not linked to the selected institute. Wellington and Szczerbinski (2007: 69) state that people reading case studies can often relate to these studies, even if they cannot generalise, and that this characteristic of the case study might be more important than the generalisation aspect.

The next section describes the data collection tools that were used during this study.

4.5 Data collection methods and tools/instruments

This section describes the data collection tools and instruments used during this study. The following aspects form part of each section:

- Description of data collection tool and technique: this part contains a detailed description of the data collection tool and its major features. Where applicable, the development of the tool is also included.
- Rationale of tool use: this entails describing why the instrument was used, and the type of data and information this research was interested in collecting.
- Advantages: the reasoning behind the choice of method and instrument is discussed; comparisons with alternative methods are also supplied where applicable.
- Disadvantages: the possible uncertainties and weaknesses when using the method or instrument are discussed.

While literature was used to assist in creating a data rescue workflow model (see Chapter 3), the study's empirical data collection tools and methods were as follows:

- a web-based questionnaire,
- virtual one-on-one interviews making use of a semi-structured interview schedule,
- feedback provided after participants had reviewed the initial model, and
- a mini focus group session.

Each of these data collection tools is discussed separately.

4.5.1 Web-based questionnaire

This section discusses the web-based questionnaire used in this study. Advantages and limitations of web-based questionnaires are also included.

4.5.1.1 Description of data collection tool and technique

The first data collection tool used in this study was a web-based questionnaire. The questionnaire was sent to Sample A (see Section 4.7.1: Sample A: Web-based questionnaire recipients). The wording of the questionnaire can be viewed in Appendix 2: Web-based questionnaire, while the wording of the cover letter and the consent form can be viewed in Appendix 3: Cover letter, and Appendix 4: Consent form, respectively.

The information collected via the questionnaire was anticipated to provide the following:

- an overview of the prevalence of data at risk within the institute,
- an overview of the prevalence of data rescue activities within the institute,
- an overview of the prevalence of data at risk within each respondent's research group,
- an overview of the prevalence of data rescue activities within each respondent's research group,
- the names of researchers/research groups who have data at risk,
- the names of researchers/research groups who have performed data rescue activities,
- the names of researchers/research groups who have accessible data at risk,
- suggested names of researchers or research groups (not forming part of Sample A) who might possess data at risk,
- suggested names of researchers or research groups (not forming part of Sample A) who might have participated in data rescue activities, and
- the names of researchers/research groups who should comprise the sample to be contacted during the next study phase.

A free online survey-generating platform, eSurv, was used to create the web-based questionnaire (eSurv, 2019). This researcher also made use of this tool during previous research. The features offered (e.g., different question formats available, data analysis, and data export options), as well as ease of use served as reasons for making use of the tool during this study.

According to Regmi *et al.* (2016: 640), there is a lack of information available to researchers designing online questionnaires. Regmi *et al.* (2016: 640–644) list six aspects as vital considerations when designing an online or web-based questionnaire; the aspects and their implementation are described below.

- Online questionnaires should have a user-friendly design and layout: the web-based questionnaire design is simple, while the questionnaire is short. A progress indicator is used, paging back and forth between questions is possible, and ample space is available for long answers.
- Participant selection for online questionnaires should be done carefully: sampling was purposive; techniques used involved expert sampling, criterion sampling, and homogeneous

sampling. More details on the sampling method used during this stage of the study can be obtained in Section 4.7.1: Sample A: Web-based questionnaire recipients.

- Online questionnaires should avoid the possibility of multiple responses by the same respondent. In this study, this was not an issue of concern, as the online questionnaire was not anonymous, and identifying details were required to select Sample B. In addition, identifying the responses linked to each research group was vital, as it would be connected to the characteristics of data at risk, and data activities performed.
- Online questionnaires should consider data management: eSurv as a web-based tool is able to export questionnaire data to PDF, CSV, and Excel formats. In addition, data are private, as the researcher's log-in details are required for viewing responses and subsequent web-survey data.
- Online questionnaires should consider ethical issues: all study components (including the questions asked, the wording of the cover letter, the wording of the consent form, and use of data) were subjected to the relevant Research Ethics Committees (see Section 4.12: Ethical considerations and ethical clearance). Ethical clearance was obtained.
- Online questionnaires should be piloted: the questionnaire was piloted and several corrections to reported aspects were made. In addition, the data export feature of the tool performed to satisfaction.

An email, informing Sample A members of the upcoming web-based questionnaire, was sent to members a week before the questionnaire was distributed. The email was the first of four contacts with Sample A; the other contacts comprised the web-based questionnaire link, a reminder email, and an invitation to partake in the next data collection stage. A subsequent email, forming the second contact with Sample A members, contained a link to the web-based questionnaire. The initial email also contained the following information:

- a basic outline of the study: its rationale, participants, sample, instruments, expected outputs and outcome,
- the reason for contacting the recipient,
- the value of the recipient's contribution, and
- the value of the study.

In addition, the email mentioned that anonymity was guaranteed, and that responses revealing confidential information would be anonymised and de-identified. More information about the email content, wording and timeline is supplied in Section 4.9.12: First contact: Email to Sample A, and in Appendix 1: Email informing Sample A about web-based questionnaire.

The web-based instrument itself consists of three sections, namely, a cover letter (see Appendix 3), an informed consent form (see Appendix 4), and the questions to be answered. The questionnaire contained eight questions, which can be viewed in Appendix 2.

Most of the tool's questions were in multiple choice format. However, several questions also had an added text box below the answer options, enabling respondents to clarify answers, supply additional information, or describe an aspect not listed in the answer choices.

The cover letter contained a request for the questionnaire to be completed within two weeks of receiving the link. Completed questionnaires and informed consent forms were submitted electronically by Sample A members. Upon accessing the tool after the cut-off date for completion, this researcher was able to export the data that had been collated by the tool to PDF, XLS, or CSV format.

4.5.1.2 Advantages

When deciding on a data-collecting tool for this phase of the study, there were a range of options to choose from. The data collected should ideally provide an overview of the institute's data at risk and data rescue activities, and as such, the following alternative tools were considered: virtual one-on-one interviews, focus group discussions, telephonic interviews, paper questionnaires, or using a web-based chat tool such as Skype to collect data and information.

The practicalities of the listed options were considered and weighed up against the web-based questionnaire. The latter tool was chosen for this phase of the study, with the main reasons being as follows:

- Web-based questionnaires can reach remote participants. This advantage is stated by Leedy and Ormrod (2005: 185) as well as Adams and Lawrence (2015: 109). The fact that the purposive sample used during this stage of the study included several researchers/research groups located in a different part of the country, made it a viable instrument. Additionally, as this study was mainly conducted while COVID-19 restrictions were in place, the researcher's travel movements, as well as those of participants were restricted, which resulted in the web-based questionnaire being an ideal tool for data collection.
- Web-based questionnaires can reach many participants in a brief period. Theoretically, there is no limit on the number of participants who can be contacted or return completed questionnaires. The cost-effectiveness with regard to time and resources is also mentioned by Regmi *et al.* (2016: 640).

- Web-based questionnaires have the advantage of being completed as and when it suits the participant. This is an advantage over telephonic interviews or one-on-one interviews, where an interview can only take place if a timeslot available to both parties is found.
- Using eSurv as a platform had the advantage of a data export function to different formats, depending on the requirements or preferences of the interviewer (eSurv, 2019).
- Using a web-based questionnaire during this study is a viable and logical platform, as all researchers forming part of the sample had access to a personal computer and internet connectivity.
- Because web-based questionnaires can accommodate visual images, and even audio and video, this makes it a flexible and multi-faceted data collection method. Many web survey platforms are also able to provide an interactive questionnaire experience, or have contingency questions resulting in the presentation of different questions to different respondents based on prior answers (Neuman, 2014: 347).
- While web surveys are often described as having the benefit of anonymity, this was not a benefit forming part of this study. This study required respondents to supply their names and research group, as linking the answers to the research group was vital for learning more about data at risk and data rescue at the institute. In addition, supplying the name was a non-negotiable step, as the responses to questions (regarding data at risk and data rescue) would be used to select Sample B members. Sample B (see Section 4.7.2: Sample B: Researchers selected for interviews) consisted of Sample A members who had indicated in the affirmative to at least one of the following:
    - their group had data at risk, or
    - their group had participated in data rescue activities.

  Sample B members were therefore 'purposively selected Sample A members' who were involved in the next data collection method, which entailed one-on-one interviews. While anonymity did not form part of the web-based questionnaire, confidentiality was guaranteed, and researchers informed that data would be anonymised and de-identified where required.
- Adding on to the previous bullet: lack of anonymity within this study can be seen as a benefit, as the researcher was able to contact respondents in instances of unclear or incomplete responses.
- Making use of an online tool can also be seen as less intrusive; the researcher had experience of previous encounters with respondents who were unwilling to participate in virtual one-on-one discussions, but quite willing to participate in a virtual electronic sphere.

While use of a web-based questionnaire offers many benefits, several limitations are also ascribed to the tool, and are discussed in the next section.

4.5.1.3 Disadvantages and uncertainties

The following potential drawbacks regarding the use of a web-based questionnaire were noted:

- Participants were unable to obtain clarification on questions or instructions unclear to them while completing the questionnaire. This aspect was minimised through the use of a pilot run before launching the tool, and by providing contact details in the cover letter accompanying the questionnaire.
- Immediate clarification on unclear, ambiguous or missing responses could not be obtained. While this is a legitimate concern when using anonymous web-based questionnaires, this study required that respondents supply their names and contact details. In cases of doubt, a simple follow-up email would have provided clarification regarding unclear responses, or incomplete questionnaires submitted.
- According to Neuman (2014: 346), web surveys have a slightly lower response rate than face-to-face interviews. This risk was minimised by simplifying the questionnaire (to contain eight questions only), and by implementing two follow-up emails. In addition, the importance of the questionnaire response and the study was stated in the cover letter accompanying the questionnaire. Sample members also received an email notifying them of the forthcoming web-based questionnaire a week before the questionnaire link was sent to them.
- While access to and completion of web-based surveys are restricted to persons having a computer and internet connectivity, this aspect was not regarded as a tool risk in this study. All researchers involved had access to computers and internet connectivity, even while working from home during lockdown.
- The fact that web-based questionnaires are cheap and easy to create and distribute is seen by Weisberg (2005: 38) to be a risk factor, as tools are then often designed without designers paying attention to quality. Three areas of concern are mentioned by Neuman (2014: 346–348), namely, coverage, privacy/verification, and design issues. These stated risk areas were overcome as follows:
    - Coverage: not considered an issue in this study, as all sample members had internet access and were purposively sampled to be invited to complete the questionnaire.
    - Privacy and verification: not considered an issue in this study, as invites were only sent to RGLs, who had to identify themselves when completing the questionnaire. Data were also kept secure, and confidentiality was guaranteed.

    - Design issues were minimised by adhering to best design practices as suggested by Regmi *et al.* (2016: 640–644). These practices include having only a few questions per questionnaire screen, having a progress indicator, making use of a simple visual appearance, the ability to move back and forth between questions, and supplying sufficient space for questions requiring long answers.
- Technical glitches and bugs were not anticipated, and it was expected that the pilot run would identify such issues and make it possible to address and correct them.
- Respondents being physically removed from the interviewer meant that the interviewer was unable to incorporate face-to-face interview benefits such as exposure to non-verbal cues. This was not anticipated to be a risk factor in this study, as this phase of data collection was mostly centred on factual feedback. As such, questions regarding participants' opinions, feelings and experiences were only incorporated during the subsequent data collection method (i.e., one-on-one interviews).

Despite the potential drawbacks of web-based questionnaires, it was deemed a highly suitable data collection method for this study, and also assured adherence to COVID-19 restrictions.

4.5.2 Virtual one-on-one interview

This section discusses the characteristics of the virtual one-on-one interview used in this study. The advantages and limitations of one-on-one interviews and virtual interviews as data collection tools are also described.

4.5.2.1 Description of data collection tool and technique

Virtual one-on-one interviews with selected RGLs were conducted to gain detailed information about data rescue practices in the institute. These interviews were conducted with RGLs who had indicated, when completing the web-based questionnaire, that:

- their research group possessed data at risk, or
- their research group had participated in data rescue activities.

As such, the selected sample, forming Sample B, was purposively selected.

Interviews were virtual in nature, as COVID-19 lockdown restrictions were in place during this data collection phase. While the original research proposal stated that one-on-one interviews would be face to face, subsequent COVID-19 lockdown restrictions resulted in this stage of the study making use of a virtual interviewing platform to collect data instead. Apart from being virtual in nature, interviews were also conducted by enabling audio only; the web-based cameras were not used. This step was thought to lessen the chances of technological glitches, freezing screens, and inferior video quality.

The interview schedule was semi-structured in nature, as it was decided not to follow a formalised and fixed set of questions. Several open-ended questions were also included, which allowed for a discussion with research group leaders rather than a simplistic question-and-answer format. While an interview schedule was used, the interviewer followed topical trajectories in the conversation which strayed from the guide but were regarded as appropriate to the process.

The responses obtained via the web-based questionnaire had already indicated that each interviewee's research group either had data at risk or had participated in data rescue activities. The objective of the virtual one-on-one interviews was therefore to gather more detailed information on the characteristics of the data, data location, rescue activities performed, rescue roles and responsibilities identified, rescue team composition, challenges experienced, and future data rescue requirements. Sample B participants, without any data rescue experience (but with data at risk), would be requested to provide feedback regarding hypothetical data rescue challenges and obstacles. Interviewees were also informed about the drafted Data Rescue Workflow Model (discussed in Chapter 3; see Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics), and were asked to study it after the interview and provide the researcher with feedback.

In short: it was anticipated that the information obtained via the one-on-one virtual interview would provide insight into the extent and nature of data at risk in the institute. It was also expected to provide an understanding of the institute's data rescue activities, projects, needs, challenges and experiences.

To obtain the information mentioned in the preceding paragraph, a series of questions were drafted and formed the backbone of the interview schedule.

Apart from these listed questions, techniques such as prompts and follow-up questions were also employed. It was anticipated that such measures would frequently be used, as the responses of the open-ended questions would pave the way for questions not forming part of the interview guide.

Examples of the open-ended questions asked included:

- 'Kindly describe the approach you used to establish whether data were at risk.'
- 'Which factors contribute to your group's data being at risk?'
- 'What challenges do you foresee should you need to rescue your group's data at risk?'

A copy of the interview schedule can be viewed in Appendix 7: Interview schedule.

The interview schedule was created as an electronic text document. A copy of the electronic version was printed prior to each interview and consulted by the interviewer as guidance during the virtual interview. Where applicable, notes were made on the printed form during the interviews. As all interviews were audio-recorded, it was not necessary to capture the responses of interviewees verbatim.

The rationale for the use of the interview was manifold; the data collected via the interviews were anticipated to provide this research with:

- detailed information about the data at risk held by the research group represented by the respondent,
- information about the value of the data at risk,
- detailed information about the data rescue activities being performed by the research group represented by the interviewee,
- feedback regarding challenges experienced or expected by interviewees during actual or future data rescue efforts,
- an opportunity to mention to respondents the initial Data Rescue Workflow Model created by this researcher and explain its intended use, and
- an opportunity to request respondents to review the initial model and provide feedback on its suitability.

The advantages and potential limitations of virtual one-on-one interviews are discussed in the next two sections.

4.5.2.2 Advantages

Using virtual one-on-one interviews as a data collection method has many benefits in the context of this part of the study. These benefits are especially pertinent when compared to a data tool used earlier in this study (web-based survey), or to questionnaires distributed via mail.

Advantages of the tool are addressed in two sections below, namely: the advantages of interviews over other data collection methods, and advantages of using the virtual option.

The main advantages of gathering data about researchers' data rescue experiences, challenges, knowledge and requirements, by means of an interview, are as follows:

- This data collection method enables the interviewer to clarify responses, probe when additional information is required, and even ask additional questions not forming part of the interview schedule. Such activities are not possible when making use of mailed questionnaires, or during web-based surveys. Adams and Lawrence (2015: 107) state that these factors increase the response rate, as well as the accuracy of responses.

- This data collection method also enables respondents to gather clarifying information from the interviewer. Whilst not impossible, gaining such feedback from interviewers is more complicated and time consuming during the completion of mailed and web-based questionnaires.
- As stated by Neuman (2014: 346), personal interviews have high response rates.
- Neuman (2014: 346) mentions that personal interviews have the added benefit of permitting the longest and most complex questionnaires.
- According to Adams and Lawrence (2015: 107), the presence of the interviewer may inspire the participant to take the research more seriously, and that this could in turn increase the accuracy of answers.
- There is opportunity to ensure that focus is kept during the interview; web-based or mailed questionnaires run the risk of being completed while the respondent is also busy with other tasks.

When looking at the virtual aspect of the interview, the following additional advantages are noted:

- Virtual interviews allow for the collection of data covering wide geographical areas (Jowett, 2020).
- Virtual interviews are often cheaper and more time-efficient than in-person interviews (Krouwel, Jolly & Greenfield, 2019).
- Virtual interviews go together with savings to the environment, as travel emissions are not generated (Krouwel, Jolly & Greenfield, 2019).
- In certain situations, sharing of problems experienced with technical glitches during interviews can be a bonding experience (Krouwel, Jolly & Greenfield, 2019: 6).
- The online interview gives participants the right to withdraw from the interview process in uncomfortable situations, simply by clicking a button (Janghorban, Roudsari & Taghipour, 2014).
- Participants may also feel more comfortable disclosing sensitive information by not being face to face and being able to participate in the interview from a familiar space such as one’s home (Hanna, 2012; Sipes, Roberts & Mulan, 2019).

While use of a virtual one-on-one interview method offers many benefits, several limitations are also ascribed to the tool, and are discussed in the next section.

4.5.2.3 Disadvantages and uncertainties

Disadvantages of the tool are addressed in two sections below, namely: disadvantages of interviews when compared to other methods, and disadvantages of using the virtual option.

Despite the range of benefits, one-on-one interviews may portray the following limitations:

- This data collection method goes hand in hand with absence of anonymity. Respondents were known by name and provided responses in the virtual presence of the interviewer. It is therefore vital that this drawback was addressed by the interviewer through ensuring confidentiality and explaining that while honesty and openness were essential to the study, any incriminating responses would be de-identified and anonymised to protect respondents from victimisation and similar outcomes.
- Time, as well as a suitable timeslot, are definite disadvantages when compared with web-based or mailed questionnaires. For a virtual one-on-one interview to take place, both the respondent and interviewer must be present, and find a gap in their schedule. While the expected length of the interview will be shared with respondents prior to the interview, extension of the session might take place. Additional time might be required if either respondent or interviewer (or both) wish to discuss an issue in more detail. It is therefore vital to establish prior to the interview whether the allocated time can be extended.
- Quality of data might be compromised if the interviewer is not experienced in this data collection method.

The virtual nature of one-on-one interviews is prone to several disadvantages and limitations, especially during a pandemic or lockdown period:

- Virtual interviews can also be a barrier to entry for low-income candidates who do not have internet connectivity, or reliable computers/laptops.
- Virtual interviews can also be a barrier to entry for participants who are not as technologically savvy to navigate interviewing software.
- Tools commonly used for virtual interviewing, such as Zoom and Microsoft Teams, were meant to be collaboration tools and are not set up for interviewing.
- Technology itself is another obstacle to virtual interviewing, with virtual interviews often prone to freezing screens, time lapses, inferior video and/or audio quality, and disconnected calls.
- Concerns have been expressed regarding the quality and content of the resulting virtual online interview data (Davies *et al*., 2020).

- It is possible for interviewees to be under additional emotional stress during a pandemic; in such situations researchers should therefore ask themselves whether the data collection stage can be postponed (Cairns, 2020).
- Being interviewed while at home could lead to concerns regarding privacy; researchers should ascertain whether participants are able to keep their participation in the research confidential from the other people they might be living/isolating with at such a time (Cairns, 2020).
- Virtual interviews can be awkward to those who had not previously participated in such types of interviews.
- Distractions during virtual interviews (ringing phone, children, pets, neighbourhood noise) are fairly common, and could lead to loss of focus.
- According to Davies *et al*. (2020), aspects such as shortened responses, less contextual information and relational satisfaction are possible limitations of virtual interviews.

Despite the potential drawbacks of the virtual interview method, it was deemed a highly suitable data collection method for this study, and one that assured adherence to COVID-19 restrictions.

4.5.3 Feedback guide and critiquing the initial Data Rescue Workflow Model

This section describes the characteristics of the feedback guide used in this study.

4.5.3.1 Description of data collection method

This data collection stage of the study entailed Sample B members using a feedback guide to examine and provide feedback on a Data Rescue Workflow Model created by this researcher.

This data collection method included the following distinct activities:

- Sample B members being sent and receiving the initial Data Rescue Workflow Model,
- Sample B members receiving the model feedback guide,
- Sample B members examining the initial model, and
- Sample B members providing feedback regarding the initial model.

As the initial model has already been discussed in detail (see Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics), the remainder of this section contains a description of the feedback guide used by Sample B to review the initial model.

The feedback guide used in this part of the study consisted of an electronic text document providing researchers with directions when critiquing the model. The guide was created by this researcher after creating the model and contained a listing of topics and aspects Sample B members needed to consider when examining and reviewing the model. As such, the feedback guide requested interviewees to critique, among other things, the model's appearance, ease of use, its main steps, the tasks and roles indicated, outputs created, and accompanying guidance.

The feedback document can be viewed in Appendix 9: Data Rescue Workflow Model feedback guide.

While the model feedback guide was structured in nature, respondents were informed that additional feedback topics, relevant to data rescue but not included in the feedback guide, were also welcome. It was anticipated that feedback provided during this stage would provide a perspective on the following information:

- an ideal data rescue flowchart,
- stages, activities and tasks viewed by researchers as impractical, superfluous, too complicated, or not understood,
- the model's outlay, shapes, colours and general appearance,
- linked guidance,
- linked templates,
- linked outputs,
- rescue roles and responsibilities,
- missing stages, activities and tasks, and
- any other topics related to data rescue.

Sample B members accessed the model via links provided in an email. The email was sent the day after each participant's virtual interview had been conducted. The initial model was uploaded to an institutional intranet site used for storage of electronic documents and files. As the model consisted of 10 pages (each comprising an A4-size image, and each of the images containing references or links to either guidance, templates or outputs), it was seen as more manageable to provide Sample B members with the links of documents to be used during the feedback process, rather than attachments to the invitation email.

The email containing the invitation to the feedback stage can be viewed in Appendix 8.

The feedback guide to be used by researchers when reviewing the model can be viewed in Appendix 9.

The model images, templates and guidance documents to be reviewed by Sample B, and to be accessed via links provided in an email attachment, mirror the model summary and stage images discussed in Chapter 3 (see Section 3.7, Section 3.8, and Section 3.9), and the documents attached as appendices to this thesis (see Appendices 11 through to 20).

As was mentioned in the email, Sample B members were free to provide feedback in any format, or by using any platform or tool that they preferred. As such, making use of emails, Skype messaging, telephonic conversation or another virtual interview were all options open to respondents. The reasoning behind the decision to give respondents free range with regard to feedback was to ease the 'adherence' burden on researchers. Respondents had already satisfied the demands of this researcher by completing a web-based questionnaire, taking part in a one-on-one virtual interview, and studying the data rescue flowchart after the interview. In addition, it was understood that feedback quantity and complexity might differ between respondents, and that it would be better to leave the feedback format up to the discretion of participants.

Advantages and benefits of this data collection method (including the feedback guide and feedback supplied) are as follows:

- The first point concerns the benefits related to the feedback obtained from participants, and has already been listed in a prior bulleted list describing anticipated information supplied via the feedback.
- Another benefit pertains to the choice of feedback format left up to each participant. The freedom accompanying this activity was anticipated to result in higher response rates, as it could be completed in the researcher's own time, by using a format of their choice. This activity's ease of implementation along with the reduced pressure to find a time suitable to both interviewer and respondents had distinct benefits over a second round of one-on-one virtual interviews.

Limitations and disadvantages of this data collection method (including the feedback guide and feedback supplied) are as follows:

- The first shortcoming relates to the feedback format not being specified; this could be detrimental to response rates, in that the freedom to choose might convey a message of non-importance. This idea was minimised by stressing the importance of feedback during the one-on-one virtual interview, during the email sent the following day (containing the feedback guide), and in a reminder email sent two weeks after the interview.
- The second drawback concerns the feedback supplied without the flowchart being tested during an actual data rescue project. While additional, or even different feedback might be obtained during future applications of the flowchart during data rescue activities, such feedback at a later stage would not be lost.

After receiving feedback, all comments, suggestions, opinions and ideas were evaluated and the recommended changes were made to the model to arrive at a revised Data Rescue Workflow Model.

4.5.4 Mini focus group session

This section discusses the characteristics of the mini focus group session used in this study. The advantages and limitations of focus groups as a data collection method are also described.

4.5.4.1 Description of data collection method

Following the feedback received from researchers via the preceding data collection method, it was considered crucial to obtain additional feedback on the initial model. Feedback received from Sample B proved to be limited and did not provide sufficient critical input to extensively revise the model. As a result, it was necessary to include an additional data collection stage. This added stage would be focused on obtaining feedback from non-SET-based institutional experts, able to provide a different yet informed perspective on data rescue and the requirements with regard to a data rescue workflow model. Such information could be obtained via various methods, including one-on-one interviews, a focus group session, or even a duplication of the preceding data collection method, but involving a different sample.

It was important that the selected data collection technique allow a free exchange of knowledge, ideas and opinions, and ideally be a session allowing several participants to discuss topical issues simultaneously. Open communication channels, involving model critique in a non-threatening environment were crucial. It was anticipated that participants would propose amendments and suggest alternative or additional data rescue stages, tasks or involved parties. Considering the prerequisites for the session, and the expected outcomes, a decision was made to implement an in-person mini focus group session with three non-SET-based institutional experts. Details of the involved sample can be viewed in Section 4.7.3: Sample C: Institutional experts.

According to Lavrakas (2008) and QuestionPro (2021), a focus group is a qualitative research method, used by a trained moderator to conduct a collective interview with a small group of participants sharing a similar characteristic or a similar concern. Adding to this is the statement by Kitzinger (1995:299) that the focus group is a form of group interview that capitalises on communication between participants to generate data. The group interview is semi-structured in nature and can be used to gather data to understand the 'attitudes, beliefs, concerns, behaviours, and preferences' of the group being interviewed (Weare, 2013).

The following characteristics are ascribed to a focus group session:

- Focus groups form part of qualitative research (Lavrakas, 2008).

- Several types of focus groups exist: single focus group, two-way focus group, dual moderator focus group, duelling moderator focus group, respondent moderator focus group, mini focus group, and an online focus group are all mentioned by Nyumba *et al*. (2018) as being identified in literature.
- Listening forms an important part of the session: interviewers listen to participants, and participants listen and respond to one another (Weare, 2013).
- Participants should be carefully selected by the person conducting the research in order to assemble a group of people who share certain characteristics, or who share concern about an issue (Lavrakas, 2008; Weare, 2013).
- The setting selected should enable participants to speak freely about the topic (Weare, 2013). In addition, the atmosphere should be non-threatening, non-embarrassing, and free from bias (Scott, 2013).
- The moderator makes use of carefully selected, open-ended questions, and engages participants in conversation about the topic at hand (Weare, 2013).
- Additional roles of the moderator are to ensure legitimate results and reduce bias in the discussions (QuestionPro, 2021).
- Focus group discussions are sometimes seen as synonymous with interviews, especially the semi-structured one-on-one and group interviews (Parker & Tritter, 2006: 25). However, focus group sessions differ from group interviews in that the former makes explicit use of group interaction as part of the method (Kitzinger, 1995: 299). As a result, participants are encouraged to talk to one another, to share anecdotes, ask questions, and comment on the group’s different points of view. QuestionPro (2021) also states that participants are encouraged to freely share opinions and to convince other participants of their ideas.
- Unlike interviews, the researcher thereby takes a peripheral, rather than a centre-stage role in a focus group discussion (Nyumba *et al.,* 2018).
- The purpose of a focus group is not to arrive at a consensus or agreement on the topic. Instead, it seeks to identify and understand participant perceptions of the discussed topic. (QuestionPro, 2021).
- Several types of focus groups exist; examples include online and dual moderator focus groups.
- Focus groups are often used to complement other mediums such as online surveys and online polls (QuestionPro, 2021).

Certain benefits as well as limitations are ascribed to the use of focus groups during research and are described in the next two sections.

4.5.4.2 Advantages

Benefits of using focus group sessions to collect data are listed below.

- The open lines of communication created across individuals, and the reliance on active participant interaction result in the yielding of data not easily gathered via other means (Lavrakas, 2008).
- Lavrakas (2008) states that well-executed focus groups offer robust insights into participants' feelings and thoughts. As a result, a richer and more refined understanding of group member perspectives on ideas and policies is obtained.
- According to Kitzinger (1995: 299), group work can assist a researcher to tap into the myriad forms of communication used daily by people. Examples of these include jokes, humour, anecdotes, arguing and teasing.
- This tool can encourage participation from people who might have shown reluctance to be interviewed in person, or people unaware that they had something of value to contribute (Kitzinger, 1995: 299).
- Focus group sessions are an ideal technique to establish the collective views of a group (Nyumba *et al.,* 2018).
- Focus groups can be used to clarify and extend findings and are often used as an add-on to other data collection techniques (Nyumba *et al.,* 2018).
- Comparatively easier to conduct, since all the target participants and the researcher are readily available in one location at the same time (Nyumba *et al.,* 2018).
- They allow the free flow of ideas, and frequently result in the production of unexpected feedback (Scott, 2013).
- According to Nyumba *et al.* (2018), the flexibility of focus group discussions makes it an adaptable tool that can be used at any stage of the research.
- Focus group discussion is a flexible technique and is adaptable at any stage of the research (Nyumba *et al.,* 2018).
- When compared to conventional techniques such as surveys or personal interviews, focus group sessions facilitate the exploration of novel or unfamiliar issues, or issues not well understood (Nyumba *et al.,* 2018).

While the use of focus groups offers many benefits, several limitations are also ascribed to the tool, and are discussed in the next section.

4.5.4.3 Disadvantages and limitations

Several limitations can be ascribed to focus group sessions and are described below.

- It can be difficult recruiting participants, as they all must show up at the same venue, on the same date, and be on time (Scott, 2013).
- It is easy to get off-topic during a focus group session, as open lines of communication between participants are encouraged, as is the free flow of ideas (Scott, 2013).
- Due to the mostly informal nature of focus groups, and the tendency to produce vast amounts of conversational data, it is often difficult to analyse the data that is thereby obtained (Scott, 2013).
- Focus groups can be time consuming to set up and implement. This is especially apparent when focus groups are large and involve several moderators who also need to be trained and briefed (Scott, 2013).
- Focus group discussions can provide depth and insight, but they cannot produce useful numerical results, and should therefore not be used where statistical data are required (Nyumba *et al.,* 2018).

Despite the potential drawbacks of a focus group session, it was deemed a highly suitable data collection method for this study, and one that could be implemented at the selected institute while adhering to the stipulated COVID-19 restrictions.

4.5.4.4 Mini focus group

As stated in Section 4.5.4.1, several focus group types exist and deciding on the format for this study necessitated a consideration of the lead-up to the focus group decision, the nature and number of potential focus group participants, the nature of the research topic and the required outcome of the session.

The decision to make use of a focus group stemmed from the limited feedback received from RGLs regarding the initial data rescue workflow model. RGL feedback lacked the necessary critical depth for a workable rescue model. It was decided that an additional group of institutional experts needed to be involved, and that this group of experts would need to provide their views on data rescue, discuss the potential role of institutional systems, discuss the potential involvement of the research library, and review the initial and amended versions of the data rescue models. Taking into consideration the study’s main research question, it was decided that this group of experts would be library and information science professionals, able to provide informed and experienced feedback during the focus group session.

While a more detailed description of this sample is provided is Section 4.7.3, it is important to note that due to the small staff complement of the institutional research library, only three research library professionals were purposively selected for the focus group session. Had more library professionals been invited, the focus group would have consisted of library professionals unfamiliar with data, archival practices, workflows, research activities and researcher-library collaborations. As such, the focus group would contain four participants only, with this researcher being the facilitator and three research library professionals providing input during the session.

It was also crucial that the session provide opportunity for all participants to be able to contribute, that the researcher be able to explain the workflow model clearly, and that each of the participants be an expert in a field or activity linked to the rescue model. One of the participants is an experienced archivist, a second participant has a history in records management implementation while the third has designed several workflow procedures linked to the automation of research records.

The decision was made to implement a mini focus group, which is a two and three person groups often showing 'undeniable group dynamic' (Methodspace, 2021). According to DJS Research, mini focus groups, also referred to as mini groups, are similar to a focus group but involve fewer participants (2022). The choice of mini focus group is supported by the view of Dell´Olio *et al*. who mentioned that the use of small groups is justified by situations where the participants have specialised knowledge and/or experience pertaining to the subject to be discussed. According to DJS Research, mini focus groups can also be used when the topic area is difficult and may need to be explained to respondents more clearly (2022).

It was anticipated that the mini focus group setting would result in 'rich and in-depth information' (DJS Research, 2022), and that participants would be able to build on each other's contributions in a number of different ways. As stated by Methodspace (2021), the dynamics of a mini group extend the data beyond what would routinely be available in a one-on-one interview.

4.5.4.5 Implementation of focus group sessions

Nyumba *et al.* (2018) described the following steps as typical of focus group planning and implementation:

- creation of the research design, which includes defining objectives, selecting and recruiting participants, and identifying a suitable location,
- data collection, which includes pre-session preparation and facilitation during the session,
- data analysis, which includes listing, coding, theme detection, pattern seeking, content analysis, and

- results and reporting, which entails sharing the findings with relevant parties and stakeholders.

Scott (2013) also mentions that the interviewer's insights and gut reaction play a key role during the data analysis stage of focus groups.

Taking all the discussed traits and steps into account, this study's focus group implementation contained the following main components:

- Research design:
  - the purpose of the mini focus group was established, the sample selected, sample members approached, and vital information regarding the study and planned mini focus group session shared with sample members,
  - a list of questions regarding the initial Data Rescue Workflow Model was drafted, and
  - ethical clearance was obtained prior to the commencement of the study.
- Data collection:
  - pre-session preparations included booking of the venue, arranging for flipcharts, catering queries, testing of audio-visual set-up, ensuring Wi-Fi connectivity, and
  - facilitating the session including demonstration of the two models, discussing the models and stages, probing, and ensuring notes were taken and opinions recorded.
- Data analysis:
  - responses, suggestions and opinions were listed and ranked,
  - key ideas and themes were coded, and
  - qualitative thematic analysis was performed.
- Results and reporting:
  - results of the study were relayed to data rescue stakeholders, and the final Data Rescue Workflow Model (drafted after the mini focus group session) considered ready for use for the rescue of the institute's data at risk.

It was anticipated that implementation of the aforementioned steps would provide this research with the following information:

- additional feedback on the initial model that had already been viewed and critiqued by Sample B members,
- insight into the opinions, ideas and concerns regarding a revised model created by this researcher after analysing the feedback received during the previous data collection phase,

- details regarding the needs, concerns and expectations concerning institutional data rescue as seen from the viewpoint of non-research staff, and
- novel, relevant and crucial feedback that could be used to update and finalise the revised model.

With the previous sections containing detailed descriptions of the study's empirical stages, the next section provides a summarised overview of the discussed methods.

4.5.5 Summary

The empirical stage of the study made use of four data collection stages to collect data. Each stage made use of a different data collecting tool or method. The data collection tools used in this study entailed:

- a web-based questionnaire distributed to Sample A members,
- one-on-one virtual interviews with participating Sample B members,
- feedback on the initial Data Rescue Workflow Model provided by participating Sample B members, and
- a mini focus group session involving Sample C members, where the initial model and the revised model were discussed.

The next stage of the chapter describes the study's research population.

4.6 Research population

This section provides more information on the participants forming part of the study, which included different participant groups in the form of the research population and different research samples. In addition to these participant groups, the unit of analysis should also be considered. Each of these groupings is discussed below.

The research population can be described as the entire set of individuals about which inference will be drawn (Pickard, 2013: 60). Welman, Kruger and Mitchell stated that a population can consist of 'individuals, groups, human products and events, or conditions' to which they were exposed (2012: 52). Adams and Lawrence (2015: 118) describe the population as groups of people, animals or archives a researcher is interested in studying, while Gardner (2021) referred to a population as the description of all of a particular group of individuals who are being studied.

For the purposes of this study, the research population was regarded as researchers and research library professionals employed at the South African research institute which formed the basis of this research. These two groups form the research population, as they comprise the sector of the institute involved with data rescue, either during the duration of the study, or in future.

Welman, Kruger and Mitchell (2012: 53) described a study's units of analysis as the members of the research population. These units can be individuals, groups, organisations, human products, or even events (2012: 53). The units of analysis of this study are individuals, belonging either to the institute's SET component or its research library division. Alternatives included research groups, or even research units such as specific projects, impact areas or research clusters within the research institute studied. The decision was made to narrow the unit of analysis to individual researcher level, as this would be the unit completing the web-based questionnaire, being invited to a one-on-one virtual interview, and the unit supplying feedback on the model.

The concept of a subpopulation is a term used by Adams and Lawrence (2015: 118) and is described as a portion of the population. Applied to the current study, the population are all researchers at this institute, while the subpopulation would be all researchers who possess data at risk. This term differs from a 'sample', which is a subset of the population meant to represent the population (see Section 4.7: Sampling and sample).

A total of 1 608 researchers and 18 research library personnel were employed by this institute during the time of data collection.

Involving the total research population in this study was impractical, as young/new staff members would most likely not be aware of the existence and storage of all data at risk in the group, or be knowledgeable about the group culture regarding the handling of data at risk, or be fully informed about prior data rescue activities. In addition, aspects such as time constraints, scheduling conflicts, participation unwillingness and cost made surveying the research population an unrealistic and unfeasible activity. The same argument holds true for the research library of the institute: it would be impractical and unwise to involve library and information service professionals who had no experience of data rescue-related activities, or with library and information service activities forming part of a data rescue project. To ensure a more feasible data collection process, a research sample, obtained via a process termed sampling, was used. The samples and sampling activities are discussed in the next section.

4.7 Sampling and sample

According to Welman, Kruger and Mitchell, it is impractical and uneconomical to involve all members of a population; for this reason, researchers rely on and make use of a sample (2012: 55). Pickard supported this idea by stating that a sample is used because it is usually not possible, or not practical, to include the entire research population in one's study (2013: 59). The sample is therefore the subset of the population meant to represent the population (Adams & Lawrence, 2015: 119). The process by which a sample is obtained is called sampling.

Sampling can be divided into two broad categories: probability or random sampling, and non-probability or non-random sampling. The defining difference between these categories is that the former makes use of random selection techniques to obtain a sample, while the latter does not rely on random selection. Each of these two broad categories contain various sampling methods, and the choice of method depends on several factors.

This study made use of non-random sampling, which means that predetermined categories were made use of when deciding on sample membership.

The following table briefly states the most prominent features of non-random sampling, and includes short notes on each feature's use in this study (where applicable).

Table 4.3: Non-random sampling and its relevance to this study

NON-RANDOM SAMPLING FEATURES	RELEVANCE TO STUDY
Non-random sampling is also referred to as purposive sampling (Pickard, 2013: 59), as well as judgment sampling, selective sampling, or subjective sampling (Dudovskiy, 2019).	This study makes use of several of the terms listed on the left.
The researcher relies on own judgment when choosing members of the population to include in the study (Dudovskiy, 2019; Laerd Dissertation, 2012). Palys described purposive sampling as a 'series of strategic choices', pertaining to the 'whom, where and how' of the planned research (2008: 697).	Own judgement was used when deciding which participants would be included in each sample.
Leedy and Ormrod stated that this type of sampling is purposeful in that the researcher selects individuals who will yield the most information about the topic being investigated (2005: 145). Moreover, in purposive sampling personal judgment needs to be used to choose cases that help answer research questions or achieve research objectives (Dudovskiy, 2019).	Personal judgement was used when deciding which participants would best answer the study's research questions.
Non-probability sampling is commonly used in qualitative research; Pickard stated that as a simplistic rule of thumb qualitative research tends to make use of purposive sampling, while probability sampling techniques form part of quantitative research (2013: 59). Leedy and Ormrod also mentioned that qualitative researchers are intentionally non-random when selecting their data sources (2005: 145), while Hesse-Biber (2010), referring to the logic of qualitative research as concerned with an in-depth understanding of generally small samples, stated that qualitative research often makes use of purposive sampling.	This study is qualitative in nature (see Section 4.3: Research approach) and makes use of non-probability sampling.

NON-RANDOM SAMPLING FEATURES	RELEVANCE TO STUDY
According to Pickard (2013: 54), the choice of sampling technique relies on the research design, and it is important to make correct sampling choices to fulfil the research goals. Adding to this is the statement by Dudovskiy (2019) that purposive sampling is used when the study strives to understand or gain insight into a case.	This study focused on gaining a better understanding of data at risk, and data rescue, at the identified research institute. To reach these objectives, the decision was made to choose non-random sampling as a technique.
Several advantages are ascribed to making use of purposive sampling: Dudovskiy (2019) mentioned the cost-effective and time-effective aspects, and that it may the only appropriate method when there are limited primary data sources available.	This study implemented purposive sampling as it was deemed to be the only appropriate method, with limited primary data sources available to the researcher. Primary data sources were required to have data rescue knowledge and experience.
The purposive sampling method may prove to be effective when only limited numbers of people can serve as primary data sources due to the nature of the research design and aims and objectives (Dudovskiy, 2019).	The objective of this study was the creation of a Data Rescue Workflow Model. As such, only a limited number of institutional experts would be able to provide input on such a model, applicable to data rescue at this institute.
A disadvantage of purposive sampling is its inability to generalise findings (Dudovskiy, 2019).	This study is interested in the data rescue knowledge, perceptions and habits of researchers at a particular research institute; generalising the findings to researchers worldwide is not the main objective. As such, purposively choosing a sample, based on certain characteristics of interest, was vital.
As stated by Hesse-Biber (2010: 70), it is not unusual to make use of more than one purposive sampling techniques within the same qualitative study. Laerd Dissertation (2012) supports this concept by claiming that more than one type of purposive sampling may be used during qualitative research.	This study made use of several different non-random sampling techniques to obtain three different samples investigated during different phases of the study. Criterion sampling, expert sampling and convenience sampling were techniques implemented during this study.

Within the broad grouping of non-random sampling, distinct categories exist based on how the non-random sampling criteria are applied. Examples of these sampling techniques include criterion sampling, homogenous sampling, expert sampling, and convenience sampling (Laerd Dissertation, 2012). Non-random sampling techniques relevant to this study are briefly discussed below.

Criterion sampling is a method described by Palys as a result of searching for cases that meet a certain criterion (2008: 697). This study implemented criterion sampling, which was a method used to select the three different samples used in this study. Each of these three samples is discussed in detail during the next section (See Section 4.7.1: Sample A, Section 4.7.2: Sample B, and Section 4.7.3: Sample C).

In brief:

- Sample A consisted of research group leaders (RGLs) at this institute. The criterion used in this instance was job level; researchers not currently employed as RGLs were not selected for Sample A.

- Sample B consisted of Sample A members who had indicated, via the study's web-based questionnaire, that their research group had data at risk, or had participated in data rescue activities. The criterion therefore (via web-based questionnaire responses) was that the group had data at risk or had data rescue experience.
- Sample C consisted of non-SET-based institutional experts with extensive experience in research library workflows, research library tasks, researcher/library collaborative projects, and archival activities. The criterion was research library experience, and in particular experience regarding research library activities which could form part of a data rescue process and workflow.

Homogenous sampling is described by Etikan, Musa and Alkassim (2016: 3) and Dudovskiy (2019) as a type of purposive sampling where the researcher focuses on individuals who share similar traits or specific characteristics. The idea is to focus on this precise similarity and how it relates to the topic being researched (Etikan, Musa & Alkassim, 2016: 3). Laerd Dissertation described homogeneous sampling as a sampling method where units are selected based on having similar characteristics, with these characteristics being of particular interest to the researcher (2012). This study did not make use of homogenous sampling, even though the application of the selection criterion involving Sample B did result in Sample B members portraying homogenous characteristics. It can even be stated that selecting Sample B members by means of homogenous sampling might have resulted in the same Sample B obtained via criterion sampling.

Expert sampling is described by Laerd Dissertation (2012) as a type of purposive sampling technique that is used when research needs to glean knowledge from individuals who have **particular expertise**. This expertise is often required during the **exploratory** phase of qualitative research (Laerd Dissertation, 2012), or to investigate new areas of research, or when determining whether further study of the topic is feasible (Etikan, Musa & Alkassim, 2016: 3). According to Etikan, Musa and Alkassim (2016: 3), this sort of sampling is also useful when there is lack of observational evidence, or when research is expected to take a long time to provide conclusive results. This study was exploratory in nature and thus made use of expert sampling in the following manner:

- Researchers not forming part of sample A who were named by Sample A members (during completion of the web-based questionnaire) to have data at risk, or to have participated in data rescue activities, were invited to form part of Sample B. These nominated researchers were seen to be experts with regard to the topic investigated and selecting them for Sample B inclusion was an application of expert sampling.
- Expert sampling was also used to select Sample C: sample members were deemed to be institutional experts who would be able to understand data rescue, be able to review and

critique two versions of the Data Rescue Workflow Model, and provide recommendations for a workable Data Rescue Workflow Model for use at this institute.

Convenience sampling (also known as 'Haphazard sampling' or 'Accidental sampling') is a type of non-probability or non-random sampling where members of the target population who meet certain practical criteria, such as easy accessibility, geographical proximity, availability at a given time, or the willingness to participate are included for the purpose of the study (Etikan, Musa & Alkassim, 2016: 2). In this study, convenience sampling played a part during the selection of Sample C, as it would not be ideal to select a potential Sample C member who would not be able to attend the in-person mini focus group session, due to being on leave or due to geographical impossibility.

As seen from the paragraphs above, this study primarily made use of criterion sampling as a purposive sampling technique. However, the selection of each of the three different samples also involved the use of more than one purposive sampling method. These methods will be discussed under the separate sections devoted to each of the three samples.

In order to understand the rationale behind the necessity of these three samples featured in this study, it is considered necessary to give a brief description of the steps in the research process involving the three samples. The main participant-reliant steps of this study are:

1. create an initial Data Rescue Workflow Model,
2. establish in broad terms the state of data rescue at this research institute via the completion of a web-based questionnaire by selected researchers (i.e., Sample A members),
3. gain detailed information regarding data rescue and data at risk at this research institute by means of virtual one-on-one interviews with selected researchers (i.e., Sample B members),
4. demonstrate the initial Data Rescue Workflow Model to Sample B,
5. request feedback regarding the initial model from Sample B members,
6. revise the initial model based on insight gained after Sample B interviews and Sample B feedback, thereby creating a revised model,
7. demonstrate the initial model and the revised model to Sample C members; also request their input regarding the revised model that was created based on Sample B feedback, and
8. evaluate all feedback and amend the revised model to produce the recommended Data Rescue Workflow Model.

The three sampling activities preceded steps 2, 3 and 7, respectively. More details are provided in the list below.

- Sampling prior to step 2 entailed selecting a group of scientists deemed to likely have come across data at risk in their research groups.
- Sampling prior to step 3 involved selecting researchers who had indicated that they had data at risk in their research groups.
- Sampling prior to step 7 entailed selecting institutional experts who would be able to review the initial and revised versions of the Data Rescue Workflow Model.

Each of these three samples, and the purposive sampling method used to obtain the sample, is discussed below.

4.7.1 Sample A: Web-based questionnaire recipients

This section explains in more detail the sample of researchers selected and who were invited to complete a web-based questionnaire. The questionnaire would be used to establish, in broad terms, whether the respondents' research groups had data at risk in their research groups, and/or whether they had participated in data rescue activities at all. The selected group of researchers is referred to as Sample A. The purpose of Sample A was to have a group of experienced researchers who would:

- provide a broad overview of the presence of data at risk in the institute,
- provide a broad overview of the presence of data rescue activities in the institute, and
- enable the selection of a sub-sample (Sample B), based on Sample A's responses.

Sending the web-based questionnaire to the entire research population was not a feasible nor sensible option, due to the following reasons:

- The institute in question has more than 1 800 researchers and including all of them would be neither practical nor ideal. It is also highly unlikely that such a venture would be approved by institutional management.
- Involving more than one person per research group could lead to duplicate information, or even conflicting data per group.

A decision was therefore made to implement non-random sampling, which entailed using pre-determined variables to select members of Sample A. Sample A was a workable number of researchers who would be able to provide information assisting in establishing an overview of data at risk and data rescue at the institute. The following aspects were considered when deciding on Sample A eligibility:

- Researchers were required to be knowledgeable about data at risk or data rescue activities at the institute. As such, emerging researchers would not be involved in this study.

- It was sensible to involve experienced researchers in charge of certain research divisions within the institute, rather than involving all experienced researchers.
- When deciding on research divisions, three options were available: clusters, impact areas, or research groups. Research groups are sub-divisions of impact areas, while impact areas are sub-divisions of clusters. The decision was made to choose research groups as the research sub-division, as these groups would be more likely to deal with data belonging to a distinct discipline, rather than a combination of disciplines, as would be the case with impact areas or clusters. For more information on clusters, impact areas and research groups, consult Table 4.4: Full population (all research groups) and Sample A (indicated in bold).

Taking the above into account, Sample A was selected as follows: Sample A comprised the institute's RGLs. Purposively selecting this sample entailed the use of **criterion sampling**; the criterion chosen was job level, with researchers currently employed as research group leader being the determining job level. This meant that all RGLs at the institute, and RGLs only, were selected for inclusion in Sample A.

While criterion sampling was the purposive sampling method used to obtain Sample A, it might also be argued that **expert sampling** played a role in the selection of Sample A. This type of purposive sampling is defined by Laerd Dissertation (2012) as a sampling technique used when research needs to glean knowledge from individuals having particular expertise. It was assumed that Sample A members, overseeing a research group, would indeed be experts in their field, and likely also have data rescue experience. The selection of Sample A members, based on their research experience and assumed expertise, was therefore an example of expert sampling being implemented.

While it can also be argued that homogenous sampling was used to select Sample A members, the fact that RGLs were working in difference disciplines and using different data formats and research techniques disqualified this statement. As such, Sample A members were not a homogenous group, but did share a common feature, i.e., researcher job level.

The steps implemented to select Sample A are described below.

- Permission was obtained from the institute's Human Resources Division to supply this researcher with a list of names of RGLs, and their relevant research groups.
- After obtaining permission, a name list was requested and supplied.
- Managerial approval for involving RGLs in the study was obtained from cluster management.
- Ethical approval was obtained from the University of Pretoria and the research institute where the study would be taking place.

These steps are discussed in more detail in Section 4.9: Research process.

It is important to mention that data management implementation at this institute was still in its infancy at the time of the study. With a data librarian only recently appointed, and the data management procedure still under institutional review at the start of this study's data collection stages, it was anticipated that even experienced researchers, such as those comprising Sample A, might be lacking data management (and data rescue) skills and experience.

The number of researchers who were purposively chosen to form Sample A came to 49. The selected institute had 109 research groups at the time of the empirical data collection phases (see Table 4.4) but not all groups had RGLs; additionally, some RGLs were leading more than one research group. This resulted in the number of members of Sample A being smaller than the number of indicated research groups.

Sample members were contacted via email and informed that they had been selected to be included in Sample A. The wording of this introductory email, distributed to sample members two weeks prior to circulation of the web-based questionnaire, can be viewed in Appendix 1. In brief, it informed sample members of the objective of the study, the sample and how sampling was done, and gave a brief description of the web-based questionnaire to be distributed in a fortnight's time. The email also mentioned that managerial approval had been given, and that ethical clearance had been obtained.

A summary of all the research groups at the selected institute (i.e., comprising the full population) and the research groups comprising Sample A (shown in bold) is portrayed in Table 4.4.

Table 4.4: Full population (all research groups) and Sample A (indicated in bold)

RESEARCH CLUSTERS	IMPACT AREAS	RESEARCH GROUPS
Advanced Agriculture and Food	Agromanufacturing	• Enterprise Creation for Development • **Agroprocessing** • Food Safety Programme • **Precision Agriculture** • SANBio
Defence and Security	Aeronautics Systems	• Aeronautics Innovations • Experimental Aerodynamics and Facilities
	Landward Systems	• **Applied Detonics, Ballistics & Explosives** • **Detonics, Ballistics & Explosives Laboratory** • **Rapid Operational Solutions** • **Tactical Vehicle Mobility**
	Technology for Special Operations	• **Airborne and Landwards Special Operations** • **Integrated and Maritime Special Operations**
	Radar and Electronic Warfare	• **Radarware** • **Radar Software** • **Imaging Radar Techniques** • **Surveillance Radar Techniques** • **Digital Electronic Warfare** • **Electronic Warfare Testing and Evaluation** • **Electronic Warfare Technology Demonstrators**
	Optronics Sensor Systems	• **Infrared Electronic Warfare** • **Opto-Mechatronics Engineering** • **Video and Image Processing**
	Integrative Systems	• **Command and Control** • **Integrated Capability Management** • **Systems Engineering and Enterprise Architecture**
	Information Security Centre	• **Cyber Warfare** • **Cyber Security** • **ID Authentication** • **Network and Data Security**
Future Production: Chemicals	Biomanufacturing Technologies	• Technology Demonstration • **Bioprocessing Technologies** • Biorefinery Industry Development Facility • Biomanufacturing Industry Development
	Centre for Nanostructures and Advanced Materials	• **Advanced Polymers and Composites** • **Advanced Metals Processes** • Nano Emulsions and Delivery Systems • Nanostructured Materials for Sensing Applications • **Materials Characterisation Facility** • Nanomaterials Industrial Development Facility • Nano-Micro Device Manufacturing
Future Production: Manufacturing	Industrial Sensors	• Industrial Systems • **Sonar** • **Human and Societal Systems**
	Advanced Materials Engineering	• Design Testing and Commercial Services • **Advanced Casting Technologies** • **Powder Metallurgy Technologies**

RESEARCH CLUSTERS	IMPACT AREAS	RESEARCH GROUPS
Future Production: Manufacturing (continued)	National Laser Centre	• **Bio-Photonics** • **Novel Lasers** • **Laser Enabled Manufacturing** • National Programmes
	Centre for Robotics, Autonomous Systems and Future Production Systems	• **Industrial Robotics** • **Inspection and Mapping** • **Future Production Systems**
Future Production: Mining	Mining Testing and Training	• Rope and Mechanical Testing Lab • Self-Contained Self Rescuer Lab • Air & Dust Lab • Kloppersbos
	Mining and Minerals Resources	• Infrastructure Automation
Nextgen Enterprises and Institutions	Networked Services and Application	• Cloud and Network Architecture and Services • Spectrum Access and Management Innovation • **Digital Audio-Visual Technologies**
	Operational Intelligence Impact Area	• **Design and Optimisation** • Data Science • Geospatial Modelling and Analysis
	e-Government	• Spatial Information Systems • **Software Architectures and Solutions** • **Technology Implementation, Monitoring and Evaluation** • ICT Interoperability
	Emergency Digital Technologies for 4IR	• **Advanced Internet of Things** • Artificial Intelligence and Augmented Reality • Distributed Ledger Technology
Nextgen Health	Medical Devices and Diagnostics	• Veterinary and Molecular Diagnostics and Vaccines • **Human Molecular Diagnostics and Omics Technologies** • Pharmaceuticals Development
	Synthetic Biology	• **Bioengineering and Integrated Genomics** • Companion Diagnostics and Array Printing • Synthetic Nanobiotech & Biomachines
Smart Mobility	Transport Infrastructure Engineering	• **Accelerated Pavement Testing** • **Advanced Materials Testing Laboratories** • **Pavement Design and Construction** • **Coastal Engineering and Port Infrastructure**
	Transport System Operations	• **Transport Management Design and Systems** • Transport Network Asset Management
Smart Places	Equitable Built Environment	• **Urban and Regional Infrastructure Planning** • Sustainable Human Settlements
	Functional Building Infrastructure	• **Infrastructure Innovation** • Construction Materials • Strategic Infrastructure Management
	Sustainable Ecosystems	• Biodiversity and Ecosystem Services • Coastal Systems • Waste Beneficiation • Environmental Management Services • Earth Observation

RESEARCH CLUSTERS	IMPACT AREAS	RESEARCH GROUPS
Smart Places (continued)	Holistic Climate Change Impact	• Ocean Systems & Climate • Climate and Air Quality Monitoring • Climate Services • Alliance for Collaboration in Climate and Earth Systems Science (ACCESS)
	SMART Places Hosted Programmes	• National Cleaner Production Centre (NCPC) • National Foundry Technology Network (NFTN)
	Energy Centre	• **Energy Supply Systems** • Energy Industry • Energy Autonomous Campus • Hydrogen SA • **Energy Systems Optimisation** • **Electro-Chemical Energy Technology**
	Water Centre	• **Hydroscience** • **Integrated Water Analytics and Solutions** • **Integrated Water Infrastructure Services**

(Source: Du Plessis, 2020)

To summarise: Sample A was selected using criterion sampling; the criterion used was the institute's research post level. Sample A members were invited to complete a web-based questionnaire.

4.7.2 Sample B: Researchers selected for interviews

This section describes the sample selected and interviewed about their research group's data at risk and/or data rescue practices. The section also describes the sampling method used to obtain the sample. The sample is referred to as Sample B.

The previous sampling section (Section 4.7.1: Sample A) described the web-based interview phase involving RGLs; results of the questionnaire were used to obtain an overview of data at risk and data rescue at this institute. A next step entailed gaining in-depth information on data at risk and data rescue. This was done by involving researchers who had data at risk and/or had performed data rescue activities in the past (as indicated on the web survey). It made sense to create a sample for this activity; involving all researchers or even all RGLs would be impractical and not approved by institutional management. Apart from these issues, involving all researchers or even all RGLs would be a superfluous activity in many instances, as the results of the questionnaire had already indicated that certain research groups did not have data at risk, and had never undertaken data rescue. These factors necessitated the purposive selection of a sample of researchers who, based on their questionnaire responses, formed Sample B. Sample B was the group of RGLs able to supply in-depth information about data at risk at this institute.

Selection of Sample B was done after receiving completed web-based questionnaires from Sample A respondents (refer to Section 4.7.1 for more information on Sample A). Sample B members, who were a subgroup of Sample A respondents and selected for involvement in the interviewing phase, were chosen based on web-based questionnaire responses. Sample A respondents were selected for inclusion in Sample B if indicating via the web-based questionnaire to have:

- accessible data at risk in their research groups, and/or
- participated in data rescue activities.

The reasoning behind the selection of Sample B was to have a group of researchers who were able to provide detailed information on the institute's data at risk or data rescue activities.

The sampling technique used was non-random in nature, in that members of Sample B were purposively selected. Sample B members were chosen based on their expected knowledge and experience about data at risk or data rescue and were able to provide insight into this activity as practiced at the institute. This decision is supported by the statement of Laerd Dissertation (2012), which declares that the main goal of purposive sampling is to focus on particular characteristics of a population of interest who will best enable a researcher to answer the research questions.

The application of non-random sampling entailed the concurrent application of criterion sampling and expert sampling. To clarify:

- **Criterion sampling:** The criterion applicable was the indication via the web-based questionnaire responses that the respondent's research group either had data at risk, or had participated in data rescue activities.
- **Expert sampling:** Sample B members were also selected on their expertise of the subject being investigated. The web-based questionnaire included a question on suggestions for data rescue experts within the institute (not included in Sample A); these suggested researchers were included when Sample B was selected.

Selection of Sample B entailed the following steps:

- Approval for approaching RGLs was attained prior to contacting Sample A; this approval was also valid for Sample B.
- Web-based questionnaire data were analysed; research groups with data at risk or with data rescue experience were sent an email inviting them to the next phase of the study.
- In addition, non-sample A members who were nominated by Sample A respondents to have data at risk, or data rescue experience (when answering Question 8; see Appendix 2: Web-based questionnaire) were also invited to the interview phase of the study.

Following the analysis of the web-based responses, selection of Sample B members was made. Respondents were chosen to form part of Sample B after indicating that they had data at risk and/or had participated in data rescue activities. The number of researchers selected to form Sample B came to 18. Sample B membership was also considered for non-RGL-level researchers whose names had been mentioned by Sample A respondents when answering question 8 of the web-based questionnaire: 'If you are aware of any employee, research group, impact area, or cluster who might have data at risk, or have participated in data rescue activities, kindly provide contact details below.' In addition, Sample B membership was considered for researchers who had been nominated as proxy by an emailed Sample B member. These nominated researchers, although being neither Sample A members nor RGLs, were also included in Sample B based on their suggested and relevant data at risk/data rescue expertise.

Sample B members were contacted via an emailed invitation to participate in a virtual one-on-one interview. The wording of the respective emails can be viewed in Appendix 5 and Appendix 6.

To summarise: Sample B was selected using criterion sampling; the criterion used was the indication of Sample A members via the web-based questionnaire that a research group had data at risk or had performed data rescue activities. Expert sampling and convenience sampling were also techniques applied to a lesser extent. Sample B members were invited to participate in a virtual one-on-one interview.

4.7.3 Sample C: Institutional experts (not SET based)

This section describes the sample selected to be included in a mini focus group session regarding data rescue at the institute, and to critique the initial Data Rescue Workflow Model previously distributed to Sample B. The sample would also review a revised Data Rescue Workflow Model, drafted after receiving feedback from Sample B. The section also describes the sampling method used to obtain the sample. The sample is referred to as Sample C.

The main objective when purposively selecting Sample C was to collate a small group of institutional experts, not forming part of the SET base, who would be able to:

- critique the initial model,
- critique the revised model, and
- provide data rescue input and recommendations from the perspective of a research library professional.

Three experienced institutional research library employees were selected and approached. The respective Sample C members possessed the following characteristics:

- Member 1: more than three decades of library and information services experience, with institutional workflows involving library, SET staff and ICT responsibilities, institutional repository experience, and project management being particular areas of expertise.
- Member 2: three decades of experience in both disciplinary research as a scientist, and library and information services expertise in user training, publishing advice, and information scientist activities.
- Member 3: close on a decade's experience in records management, archival roles, activities and procedures, plus the drafting and editing of institutional workflows.

Selected Sample C members were contacted via email and informed of the study's objectives. In addition, the need for additional input from research library experts was stated, emphasising the need for a sample different to Sample B (i.e., this sample would not be SET based) to review an initial Data Rescue Workflow Model. Other details in the email included:

- a description of the data collection method (i.e., mini focus group session),
- dates and times available at the institute's research commons, the venue to be used as a mini focus group venue, and
- an assurance that COVID-19 regulations would be adhered to.

After acceptance of the invitation and indication of dates suitable to each sample member, a suitable venue was booked. Sample C members were forwarded copies of the initial Data Rescue Workflow Model and the revised model. Sample C members were free to review the documents, though assurance was given that it was not a pre-requisite to mini focus group attendance.

When looking at how Sample C members were chosen, it can be said that the selection of the sample was achieved via **criterion sampling**. All Sample C members were experienced research library professionals, were familiar with certain research library activities enhancing their understanding of certain data rescue activities (e.g., archival experience, project management experience, SET–library collaboration), and had managerial experience.

At the same time, due to the traits possessed by Sample C members, expert sampling as a technique was also present.

Summary: Sample C was selected using criterion sampling and expert sampling concurrently. Sample C members were selected based on attributes such as library and information services experience, and familiarity with activities which would promote an understanding of data rescue activities (e.g., archival experience, project management experience, and dealing with SET-based staff).

4.8 Assumptions, limitations and delimitations of study

Simon and Goes (2013) stated that few elements are as confusing to doctoral learners as assumptions, limitations and delimitations, and that many students have difficulty understanding and differentiating between these elements. The confusion between limitations and delimitations is also mentioned by PhDStudent (2019a). Before addressing the prevalence of these three elements within this study, it is crucial to define them, and stipulate the difference between them.

Study assumptions are the necessary element in research proposals; they are required for the study to be conducted (Simon & Goes, 2013). Common study assumptions include:

- believing that respondents answered questions honestly (Simon & Goes, 2013),
- accepting that a sample is representative of the population (Adu, 2017), and
- believing that participants had a sincere interest in participating in the research and did not have other motives, such as impressing a supervisor, or obtaining better grades (Wargo, 2015).

Study limitations are weaknesses related to decisions made in a study (Adu, 2017), and are difficult to contain. The fact that limitations are weaknesses mostly out of the researcher's control is mentioned by PhDStudent (2019b), Studylib (2019) and William and Mary School of Education (2019). Common study limitations include:

- the nature of self-reporting (Studylib, 2019),
- unknown factors that could result in participant bias (Wargo, 2015),
- time constraints (Studylib, 2019), and
- limitations to the generalisability of the findings (William and Mary School of Education, 2019).

Study delimitations are factors defining the parameters of the investigation (William and Mary School of Education, 2019). They are choices made by the investigator and describe the boundaries that the investigator has set for the study (Studylib, 2019). Stating the delimitations narrows the scope of the study (Creswell, 2012), and ensures that the study does not become impossibly large to complete (PhDStudent, 2019b). Study delimitations often include the following research aspects:

- research populations chosen (PhDStudent, 2019b),
- literature that will not be reviewed, and the reasons for this (Studylib, 2019),
- activities not performed (Studylib, 2019), and
- research questions chosen (Adu, 2017).

Assumptions of this study are as follows:

- Honesty of Sample A: It was assumed that Sample A members provided honest answers when completing the web-based questionnaire.
- Honesty of Sample B: It was assumed that Sample B members provided honest answers when participating in the virtual one-on-one interview.
- Honesty of Samples B and C: It was assumed that Samples B and C members provided honest answers when providing feedback on the Data Rescue flowchart.
- Reason for participation: It is assumed that all respondents participated in the study due to interest in the project, or wanting to contribute, and not because they were being coerced, or wanted to impress superiors.
- Understood the concept: It was assumed that respondents understood the concept of ‘data at risk’ and ‘data rescue’ after being given definitions of each.
- Insight of RGLs: It was assumed that RGLs had data knowledge and were aware of data and data-related activities in their research groups.

Limitations of this study include the following:

- Time constraints: A time limit placed on the study duration meant that the researcher was not able to interview participants who would only be available on a later date or accept submitted web-based questionnaires or data rescue model feedback supplied long after the cut-off date for each phase of the study.
- Response rate: Response rate with regard to completion of the web-based questionnaire, agreeing to participate in a virtual one-on-one interview, and supplying data rescue flowchart feedback were factors beyond the control of this researcher.
- Data rescue experience of respondents: The degree of data experience and understanding of data rescue activities (and as such, feedback supplied with regard to the data rescue flowchart) were aspects not under control of this researcher.
- Disciplines within the institute: The scientific disciplines within the institute, and how they might have affected the responses given, and data rescue model feedback supplied (i.e., specificity vs generalisability) were aspects beyond the control of this researcher.
- Data collection tools come with their own set of limitations and these have been detailed during the sections devoted to each instrument (see Section 4.5: Data collection methods and tools).

Delimitations are as follows:

- Sampling methods used: The purposive sampling techniques used meant that only RGLs were involved during the selection of Sample A. While the assumption was made that the RGLs were the people most knowledgeable about data in their group, the possibility exists that either a more experienced, or even a junior member of the research group had better data rescue insight and experience.

- Population: Only researchers and research library professionals were included in this study. The institute's support services (e.g., Human Resources, ICT or Finances) were excluded from this study. This decision is because the support services, unless busy with research themselves, do not collect or work with what is commonly referred to as the institute's 'research data', and this study involved the rescue of such data.
- Institute selected: Only researchers from this institute, and data belonging to this institute were involved.
- The web-based questionnaire as data collection tool: The delimitations related to this tool (choice of tool, length, questions, format, appearance) are all delimitations of the study.
- The virtual one-on-one interview as data collection tool: The delimitations related to this tool (choice of tool, length, questions, format, appearance) are all delimitations of the study.
- The Data Rescue Workflow Model feedback as data collection tool: The delimitations related to this tool (choice of tool, length, questions, format, appearance) are all delimitations of the study.
- Not testing the data rescue model: It can be argued that not testing the data rescue flowchart (both the prototype and the recommended final version) is a delimitation in this study.
- Research questions: The choice and number of research questions related to this study are delimitations.
- Literature analysis: The range and selection of documented sources analysed are study delimitations.

4.9 Research process

This section describes the chronological steps followed during the study. The described steps start with the literature analysis leading to the creation of an initial Data Rescue Workflow Model, then continue with the steps leading to the creation of a revised version of the model, and end with the description of the steps leading to the creation of a recommended/final Data Rescue Workflow Model. A diagrammatical summary of the steps followed is supplied at the end of this section.

4.9.1 Content analysis of published literature

The research process commenced with an in-depth investigation into published outputs on research data at risk, and data rescue. Following a conventional literature review of scholarly articles, popular articles, relevant websites, online manuals/books, online tutorials, and slideshows on data at risk and data rescue, selected data rescue publications were examined and analysed to arrive at a point where it was possible to draft a data rescue workflow model. A process of content analysis was applied to 15 published or shared data rescue outputs; the selected publications and activities are discussed in Section 3.2: Data rescue workflows, models and processes. Insight obtained via content analysis enabled this researcher to draft an initial Data Rescue Workflow Model.

4.9.2 Create an initial Data Rescue Workflow Model

Following the investigation into documented literature on data rescue, the initial model was created. This model was designed to be a combination of best practices as discovered in documented sources, as well as the integration of activities, systems and tools unique and specific to this particular research institute.

The model contained an imaged summary of the nine main stages, as well as nine images detailing each of the data rescue stages individually. The individual diagrams also contained references to documents, drafted by the researcher, to be used as guidance throughout each stage. In addition, several stages also referred the data rescuer to pre-designed templates, thereby assisting novice rescuers in their tasks.

After creating the initial model, and prior to data collection activities involving researchers, a meeting was held with institutional ICT staff, familiar with business processes and workflows. Following this meeting, small cosmetic changes were made to the initial model.

This initial model is the version referred to and used during the study, right up to the phase involving Sample C. Prior to the mini focus group session held with Sample C members, and following analysis of data obtained from Sample B via one-on-one interviews and data rescue model feedback, the initial model was amended to incorporate learnings and shortcomings of the first draft of the model.

The initial draft of the model (see Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics) is therefore the version that was created after the literature analysis, and is the version demonstrated to Sample B participants and critiqued by Sample B following the virtual one-on-one interviews.

4.9.3 Select empirical data collection tools/methods

In addition to the use of documented output to guide the creation of the initial model, four distinct data collection tools (or methods) were implemented in this study. These data collection tools (or methods) were selected to collect the data required to answer the study’s research questions. The data collection tools are listed below.

- A web-based questionnaire: this method assisted in gathering data providing an overview of the institute’s data at risk and data rescue activities performed. Forty-nine RGLs (also referred to as Sample A) were involved during this stage. The tool is discussed in Section 4.5.1: Web-based questionnaire.
- Virtual one-on-one interviews: this method provided detailed and specific information regarding data at risk at the institute, data rescue activities performed, data rescue requirements, and data rescue challenges experienced. The sample involved here (referred to as Sample B) was selected based on their web questionnaire responses. The tool is discussed in Section 4.5.2: Virtual one-on-one interview.
- Feedback on an initial Data Rescue Workflow Model: interviewed participants (Sample B) reviewed the initial model created, and input received resulted in the drafting of a revised model. The feedback steps and tool as data collection method is discussed in Section 4.5.3: Feedback guide and critiquing the initial Data Rescue Workflow Model.
- Mini focus group session: Research library experts (also referred to as Sample C) were involved during this session, and reviewed both the initial and revised model. Data collected during this stage resulted in the drafting of a final version of the Data Rescue Workflow Model. The tool is discussed in Section 4.5.4: Mini focus group session.

4.9.4 Design a web-based questionnaire

A next step entailed the creation of a short web-based questionnaire. This tool, being a data collecting instrument, is discussed in detail in Section 4.5.1: Web-based questionnaire. The questionnaire contained eight questions and the responses were anticipated to provide an overview of the institute’s data at risk and data rescue activities performed. Responses would also indicate which institutional research groups has data at risk, had performed data rescue, or were in possession of data rescue documentation. The questionnaire would be distributed to members of Sample A. The questionnaire was designed using the eSurv platform.

To summarise: a short web-based questionnaire, containing eight questions, was distributed to members of Sample A of this study. The questionnaire contained questions concerning data at risk and data rescue, and included a cover letter as well as a space to indicate that consent was supplied by the participant (see Appendices 2, 3 and 4).

4.9.5 Design the interview schedule

After creating the web-based questionnaire, the next step entailed the creation of the interview schedule to be used during the virtual one-on-one interview stage of the study. This tool, being a data collecting instrument, is discussed in detail in Section 4.5.2: Virtual one-on-one interview.

The interview schedule was semi-structured in nature, with frequent deviations from the question list and order and the need to address topics not foreseen being anticipated. Questions pertained to the nature of data at risk held by the group, factors causing data to be at risk, data rescue activities performed, and actual or envisaged data rescue challenges and obstacles. Prompts were also used during the interview, as were questions not forming part of the original interview schedule. The interview schedule was used during the virtual one-on-one interviews involving Sample B.

The intention of the interview schedule was to create a set of questions that would provide this research with information regarding the data at risk held by each interviewee's research group, data rescue activities performed, data rescue challenges experienced or envisaged, and an introduction to the initial Data Rescue Workflow Model.

To summarise: an interview schedule was drafted and was used during the virtual one-on-one interview stage of this study. The schedule was semi-structured in nature (see Appendix 7).

4.9.6 Pilot runs of web-based questionnaire

It was considered crucial to conduct a pilot run of the web-based questionnaire to test the following aspects:

- accessibility of the web-based questionnaire,
- clarity of interview preamble,
- clarity of questions and answer options,
- workability of online consent form,
- workability of different question formats,
- workability of text box options,
- workability of questionnaire submission,
- workability of data export,
- spelling and grammar, and
- any other issues emanating from the cover letter, consent form, or actual questionnaire.

Testing the web-based questionnaire involved sending the link of a duplicate version of the web-based tool to 12 pilot participants, based at the involved institute. It was vital that the pilot participants did not form part of the future research samples who would be involved in the web-based questionnaire phase of the study. Participants were requested to complete the questionnaire online, and to submit the completed questionnaire electronically via the web-based tool. In addition, pilot participants were also requested to note any unclear or ambiguous details, instructions or questions, and inform the researcher of any difficulties experienced when accessing, completing and submitting the web-based tool. Participants were also requested to study the cover letter and consent section accompanying the questionnaire and provide evaluative feedback accordingly.

4.9.7 Incorporate pilot run learnings into instruments

Feedback received from respondents during and after the pilot run was evaluated and applied to the study instrument where relevant.

Necessary changes included the following:

- correction of spelling errors,
- a question allowing multiple responses was found to only accommodate a single answer; this was corrected,
- a notification of survey completion was added,
- consistency in formatting of managerial email addresses (forming part of the cover letter) was applied, and
- inaccurate managerial contact details (forming part of the cover letter) were corrected.

4.9.8 Obtain managerial approval for data collection

After creating the study instruments (see Section 4.5: Data collection methods and tools), and prior to approaching study participants, managerial approval was obtained from the research institute where the study participants were based.

The process of managerial approval contained the following steps:

- A meeting was held with the Portfolio Manager of the institute's Information Services unit (the unit where this researcher is based) to discuss the study, the instruments to be used, and the participants who would be involved.
- An official email was drafted by the Portfolio Manager and sent to all research cluster directors, informing them of the planned study involving researchers in their unit. The letter contained the outline and objectives of the study, the name of this researcher, and the required involvement of the RGLs in their cluster. In addition, the letter stated that a response

(from cluster management) was required, and that lack of response after a two-week period would indicate that permission was granted to proceed with the study.

- In addition, a letter from this researcher's Group Leader (the person to whom the above-mentioned Portfolio Manager reports) was obtained, giving permission for the researcher to involve institutional staff in the research project, and to conduct research on the institute's premises (see Appendix 23: CSIR Permission Letter).

The managerial approval letters formed part of the documents submitted when requesting ethical approval for the study.

4.9.9 Obtain ethical clearance

The next step entailed obtaining ethical approval from relevant authorities to proceed with the intended study. Ethical approval was required to be obtained from the following authorities:

- The Committee for Research Ethics and Integrity of the Faculty of Engineering, Built Environment & Information Technology, University of Pretoria, and
- The Research Ethics Committee of the institute where the researcher and study participants were based.

A detailed description of the ethical approval steps can be found in Section 4.12: Ethical considerations and ethical clearance.

4.9.10 Obtain a list of names of Research Group Leaders (RGLs)

The participants involved in the four data collection phases of this study, namely Samples A, B and C, (see Sections 4.7.1, 4.7.2, and 4.7.3) were based at the research institute where the researcher is employed. A list of names of all RGLs at the institute was requested via email from the Human Resources division.

The email contained the following information:

- a short outline of the study,
- a request for the list of RGL names,
- a request for the name of the research group of each RGL,
- contact details of the researcher's line manager,
- proof of permission obtained from the cluster directors (see Section 4.9.8: Obtain managerial approval for data collection) was attached to the email request,
- proof of ethical clearance (from both the research institute and the University of Pretoria) was attached, and

- contact details of the researcher were attached as signature to the email.

Once the list of names was supplied, the document could be studied and Sample A selected and contacted.

4.9.11 Selection of Sample A

Sample A involved non-random sampling (criterion sampling) and included all RGLs employed at this institute. The criterion in this instance was researcher job level; Sample A included all RGLs at this institute, and only RGLs. The number of researchers who were purposively chosen to form Sample A came to 49. Although the selected institute had 109 research groups at the time of the empirical data collection phases (see Table 4.4: Full population (all research groups) and Sample A (indicated in bold)), not all groups had RGLs, while some RGLs were leading more than one research group. This resulted in the number of members of Sample A being smaller than the number of indicated research groups.

For more information on sampling, see Section 4.7: Sampling and sample.

4.9.12 First contact: Email to Sample A

An email was sent to Sample A members informing them about the study, the upcoming web-based questionnaire, the involvement of Sample A, and the crucial nature of their contribution. This email was sent a week before a second email was sent; the second email included a link to the web-based questionnaire.

The introductory email contained the following information:

- a short outline of the study,
- the recipient was informed of the web-based questionnaire to be sent to them a week after this communication,
- brief details about the questionnaire (objective, scope, estimate of completion time) were supplied,
- guarantee of confidentiality during the study was ensured,
- the option to opt out at any time was mentioned,
- permission obtained from cluster directors was mentioned,
- permission obtained from Human Resources management to provide a list of names of RGLs was mentioned,
- ethical clearance obtained from both the research institute and the University of Pretoria was mentioned, and

- contact details of the researcher were attached as signature to the email.

The wording of this email, sent to 49 RGLs, can be viewed in Appendix 1: Email informing Sample A about web-based questionnaire.

4.9.13 Distribution of web-based questionnaire

A week after first contact (see Section 4.9.12: First contact: Email to Sample A), an email containing a link to the web-based questionnaire was distributed to Sample A members. Apart from the link to the questionnaire, the email also contained the following information:

- a brief outline of the study,
- assurance of confidentiality,
- a statement underlining the value of each respondent's contribution to the study,
- contact details of the researcher,
- contact details of the researcher's line manager, and
- contact details of the researcher's group manager.

The web-based questionnaire was accessed by clicking on the link in the email. The questionnaire consisted of the following three parts:

- The first part was a cover letter containing a brief outline of the study, brief details about the questionnaire, a statement mentioning that confidentiality is guaranteed, a statement mentioning that respondents had the option to opt out at any time, a statement mentioning that permission from cluster directors had been obtained, a statement mentioning that ethical clearance had been obtained from the University of Pretoria as well as the research institute, and contact details of the researcher.
- The second part was a paragraph containing a statement of consent; respondent had to click a 'Yes' option to indicate that the study terms were read, and that consent was given.
- The third part contained the questions to be answered by the recipient; a total of eight questions were included.

Following the introductory email, one of the RGLs nominated a proxy to be involved in the remainder of the study on behalf of the research group. Upon receipt of this information, a copy of the email described in this section was sent to the proxy RGL, and the proxy was invited to complete the web-based questionnaire.

The email containing a link to the web-based questionnaire was followed by a reminder email, sent two weeks after the email described in this section. The reminder email is briefly described in the next section.

4.9.14 Reminder

Two weeks after distributing the email with a link to the web-based questionnaire, Sample A members (see Section 4.7.1: Sample A: Web-based questionnaire recipients) were sent a reminder email. This email contained the following:

- a brief description of the previous email sent,
- a brief description of the study,
- a brief statement pertaining to the value of each respondent’s contribution, and
- a link to the web-based questionnaire.

4.9.15 Final contact

Two weeks after the third contact reminder, a final email was sent to non-responding Sample A selectees. The email was short and concise, and stated that this would be the final contact made with Sample A members. A link to the web-based questionnaire was supplied one final time. In total, 49 RGLs and one proxy RGL were contacted. Twenty-three completed or partially completed questionnaires were submitted by closing date. In total, 22 completed questionnaires and one partially completed questionnaire were received.

4.9.16 Data analysis of web-based questionnaire

Following the questionnaire cut-off date, the spreadsheet function of the web-based tool was accessed, and data cleaning was also performed. Where applicable, text data were summarised and collated to enable the creation of relevant figures and tables. A note was made of missing, unclear or ambivalent responses, and linked to the relevant RGL. Where data were deemed to be crucial, these RGLs were contacted to obtain clarification. Spreadsheets were exported to the hard drive of this researcher’s laptop, stored securely, and backed up. In total, 23 questionnaires were received, comprising 22 fully completed questionnaires, and one partially completed questionnaire.

4.9.17 Selection of Sample B

After analysing the responses received via web-based questionnaires, a next group of RGLs (i.e., a subgroup of the respondents of Sample A) was selected to form part of the subsequent data collection phase. This group formed Sample B. The next phase entailed virtual one-on-one interviews with

Sample B members; more information on this sample can be viewed in Section 4.7.2: Sample B: Researchers selected for interviews.

Sample B was a purposively selected sample, with sample members chosen based on the following criteria: RGLs whose web-based questionnaire responses had indicated that their research group had data at risk, or that their research group had previously participated in data rescue activities.

Sample B also comprised institutional researchers indicated by Sample A to have been involved in data rescue. The request for indication of names was one of the questions posed to Sample A and formed part of the web-based questionnaire.

4.9.18 Contact Sample B members

After selecting the RGLs forming part of Sample B, an email was sent to Sample B members, requesting their participation in virtual one-on-one interviews.

The single-page email contained the following information:

- recipients were thanked for their web-based questionnaire participation and response,
- recipients were informed that based on information supplied on the web-based questionnaire, they had been selected to form part of Sample B,
- a brief description of Sample B, and what their participation would entail,
- a brief description of the planned interview: the time required, and range of topics to be covered, and
- a request for a date and time to conduct a virtual one-on-one interview with the email recipient.

The wording of the respective emails can be viewed in Appendix 5: Email to Sample A respondents regarding interview, and Appendix 6: Email to non-Sample A respondents regarding interview.

4.9.19 Follow-up email

Two weeks after Sample B selectees were first contacted, a follow-up email was sent to the sample members who had not yet responded to the email. They were again informed of the study objectives, the importance of their input, and requested to supply an interview date and time for conducting the interview.

4.9.20 Schedule interviews

Upon receiving a date and time from responsive Sample B selectees, interviews were scheduled and confirmed. Interviews were scheduled as soon as individual responses were received. Due to COVID-19 lockdown restrictions, all interviews were virtual in nature.

Hour-long virtual interviews were scheduled with each of the respondents via the institutional online scheduling tool. In addition, an email was sent to interviewees 24 hours before the interview, reminding them of the scheduled interview.

It was fully anticipated that scheduled interviews could be cancelled after scheduling, due to more pressing RGL meetings or other engagements. In such cases, an alternative date was sought, and an interview scheduled.

4.9.21 Conduct virtual one-on-one interviews

A virtual one-on-one interview was held with willing Sample B respondents. In total, eight interviews were conducted. Interviews lasted between 25 minutes and 60 minutes; the length of interviews were mostly dependant on the data at risk held by the interviewee's research group and/or the rescue activities performed. Data rescue challenges anticipated or experienced also formed a large part of several interview sessions.

Interviews were recorded by two devices: the virtual interview platform, and the researcher's mobile phone recording function. Permission for recording each interview was obtained from each RGL at the start of the session.

The semi-structured interview schedule was consulted during this interview. Being semi-structured, strict adherence was not vital, and use was made of prompts, clarification requests, and questions not anticipated prior to the session. In addition, several respondents also posed data rescue-related questions.

The latter part of each interview entailed a brief discussion of the initial Data Rescue Workflow Model created. Each interviewed RGL was also requested to review and critique the model in their own time, and informed that an electronic copy of the model, and feedback guide, would be emailed to them following the interview.

The wording of the semi-structured interview schedule can be viewed in Appendix 7: Interview schedule.

4.9.22 Thank respondents, and request feedback on data rescue model

An email thanking the respondent for their participation in the virtual one-on-one interview was sent to each participant during the week following the session.

The email also included the following details:

- a brief description of the data collection stage,
- a paragraph detailing that informed consent is implied when providing feedback, and
- an attached feedback guide, with applicable links, to be used when examining and critiquing the initial Data Rescue Workflow Model.

The Data Rescue Workflow Model and the feedback required were briefly mentioned to respondents during the latter stages of the previous week's interview. The feedback guide contained links to additional documents to be studied, including imaged versions of each of the nine data rescue stages, guidelines to be consulted during rescue, and templates to be used during rescue.

The email reiterated the previous week's request to supply feedback regarding the model and stated that the attached feedback guide should be consulted during the critiquing process. The email also requested that feedback be supplied within a month after the email was sent. Contact details were also supplied.

A copy of the wording of the email and the guidance document can be viewed in Appendix 8: Email requesting Data Rescue Workflow Model feedback, and Appendix 9: Data Rescue Model feedback guide.

4.9.23 Transcription of interviews

Each of the eight audio-recorded interviews was transcribed, which entailed manually transferring (i.e., typing) the recorded audio data onto an electronic text-based tool (i.e., Microsoft Word). Transcription was done solely by this researcher, and did not involve transcription services, interns or colleagues.

Based on previous interview transcription experience, it was estimated that four hours of transcribing was required for one hour of interviewing time. It should be mentioned that transcription was not done verbatim, but instead captured the essence and main points within each answer. Each of the transcribed interviews was emailed to the applicable RGL, who was asked to review the transcription for correctness.

4.9.24 Review, code and organise interview data

Transcriptions were scrutinised, followed by thematic analysis of the data (see Section 4.13 for more details). Data rescue themes and patterns were identified, followed by coding applied to relevant sections. All findings, in their organised, segmented and assigned categories, were collated into manageable text-based documents.

4.9.25 Collate and summarise flowchart feedback

All feedback was supplied via email, except for one instance where a virtual interview was requested by the RGL. Feedback supplied in the latter instance was transcribed to a text document.

A feedback guide was distributed, however, most of the feedback supplied did not show adherence to the guide but was presented in free form. Whilst feedback themes had been pre-identified, additional patterns and themes were detected when reviewing received feedback.

Coding was applied to the feedback received, and applicable paragraphs, quotes, sentences or words used by respondents highlighted and assigned to identified model topics or themes.

All responses and suggestions, including those not applicable to the Data Rescue Workflow Model, were included in the results and discussions section of the thesis.

4.9.26 Data analysis of interview data

Following the steps described in Section 4.9.24, the organised, segmented and assigned interview data were analysed.

4.9.27 Data analysis of feedback data

Following the steps described in Section 4.9.25, the organised, segmented and assigned feedback data were analysed.

4.9.28 Revise the initial Data Rescue Workflow Model

After reviewing the Data Rescue Workflow Model feedback, and interview data applicable to a data rescue model, several amendments were made to the initial model. Changes and additions comprised the following:

- an indication of tasks to be performed by specific positions within the institute, for example SET staff, research library personnel, ICT personnel, Communications personnel, or Waste Services,
- changes were made to the outlay of the model, and the colours and shapes used,

- the realisation that a data rescue model should provide for more than generic paper rescue activities; rescue activities linked to other formats were included in the revised model,
- the realisation that a data rescue model should cater for full rescue as well as partial rescue, as the lack of data rescue resources may result in valuable data not being destroyed, but rescued as far as available sources allow, and
- the model was revised to differentiate between full rescue and partial rescue and include activities applicable to each of these rescue types.

All amendments and additions are discussed in more detail in the results section of the study (see Section 5.6: Revising the initial Data Rescue Workflow Model). This version of the model is also referred to as the revised model.

4.9.29 Select Sample C: Institutional experts (not SET based)

Results of the data obtained via the eight one-on-one interviews and initial model feedback provided by four members of the interviewed group brought to light that the volume and depth of information received from Samples A and B were not sufficient to finalise the model and were deemed inadequate for the purposes of a study of this nature. Additional input, recommendations and feedback from dissimilar sources were required to lead this researcher to a workable and detailed version of the Data Rescue Workflow Model. Although much of the previous received feedback could be implemented, it was clear that additional recommendations and input regarding the requirements for a workable data rescue workflow model were crucial.

As a result of the above, Sample C was selected. The sample comprised three institutional library and information services professionals based at the institute's research library and regarded as experts in their various fields. Purposive selection, by making use of criterion sampling, was employed. Experts selected were deemed to be authorities in the following fields:

- Sample C, Member 1: more than three decades of library and information services experience, with institutional workflows involving the research library, SET staff and ICT responsibilities, institutional repository experience, and project management being particular areas of expertise.
- Sample C, Member 2: a combined three decades of experience involving disciplinary research as a scientist, and library and information services expertise in user training, publishing advice, and information scientist activities.
- Sample C, Member 3: close on a decade's experience in records management, archival roles, activities and procedures, plus the drafting and editing of institutional workflows.

While the mini focus group sample was small, it was deemed as a sufficient size for this case study; a view supported by Scott (2013: 14), who states that many case studies require no more than a few members.

4.9.30 Contact Sample C members

Three institutional experts, deemed to be experts in either research or a research library, were earmarked to form Sample C. Sample C candidates were contacted and informed of the study objectives and the need for additional information regarding the creation of a usable data rescue model. All three selected research library professionals accepted the invitation to take part in a mini focus group session.

4.9.31 Schedule mini focus group session with Sample C members

A suitable date and time for a mini focus group session was obtained. Selected sample members indicated acceptance and approval for the use of the institute's research commons being used as mini focus group venue. Booking of the venue was carried out, with part of the acceptance of the booking including an undertaking of strict adherence to COVID-19 safety measures.

While a virtual focus group session was an option, the loosening of COVID-19 lockdown restrictions at the time made it possible to use a suitable indoor venue on the grounds of the research institute.

4.9.32 Mini focus group session

The session was held on the premises of the research institute. All COVID-19 safety protocols were adhered to. The room used for the session, providing seating for 17 people, allowed ample space for social distancing between four people. Each participant was provided with a flipchart should the need arise to make use of drawings or diagrams to demonstrate a point of view, idea or suggestion.

Data at risk, data rescue activities, the initial Data Rescue Workflow Model, and the revised model were explained to participants. After explaining and demonstrating both models and their various stages, a series of open-ended questions were posed to participants regarding both models. The reasoning behind the questions was to initiate discussions between participants about the two rescue models, and in order to become aware of Sample C's views of the models. The researcher took note of the different points of views pertaining to rescue stages, activities indicated, roles and responsibilities described, terminology used, and shapes, colours, and other diagrammatical objects used. Notes were taken, and any outputs created by participants (e.g., flipchart drawings) recorded or photographed.

The mini focus group session covered, discussed and debated all stages of the models.

4.9.33 Thank Sample C

Participants were subsequently thanked for their valuable contribution towards the study, and in providing information that would be of value when finalising the Data Rescue Workflow Model.

4.9.34 Feedback collated, coded, segmented and organised

Following the mini focus group session, notes were reviewed and coded. Coded feedback was transferred to a document containing pre-identified categories. Categories included but were not limited to:

- physical appearance of the model,
- model terminology and instructions,
- data rescue stage,
- data rescue activities, and
- roles and responsibilities.

Novel and unexpected themes or patterns emanating from the mini focus session were identified and added to the document as additional categories. Information pertaining to the novel themes and patterns was incorporated into the document.

4.9.35 Analysis of mini focus group session data

Following the steps described in Section 4.9.34, the organised, segmented and assigned data were analysed.

4.9.36 All feedback relevant to data rescue and model categorised

At this point of the study, feedback had been given by Sample B, after reviewing the initial Data Rescue Workflow Model. Feedback had also been provided by Sample C after viewing the initial model and the revised model. With all feedback already coded, segmented and organised, an additional step entailed collating all model-related feedback into one document.

This collated document would be used when drafting the final version of the model.

Drafting the final version of the model therefore entailed reviewing all feedback obtained from Samples B and C, and applying this to the relevant segments of the model.

4.9.37 Update and finalise the revised Data Rescue Workflow Model

Drafting the final version of the model was preceded by a review of the relevant feedback added to the collated feedback document mentioned in Section 4.9.36: All feedback relevant to data rescue and model categorised.

The final version of the model was created. The model is presented and described in Chapter 6: Recommendations.

4.10 Summary of methodology process: diagram

A diagrammatical summary of the research process steps is displayed in Figure 4.1: Summary of methodology process.

4.11 Validity, reliability, and transparency

This section reports on the methods implemented to ensure that the methodology applied and reported findings demonstrate validity, reliability and transparency. The section also explains the reason for exclusion of discussions on replicability and reproducibility, two commonly used criteria for evaluating quantitative research.

4.11.1 Validity

According to Mohajan (2017: 58), validity is concerned with what an instrument measures and how well this is done, while Price, Jhangiani and Chiang state that the term refers to the extent to which the scores from a measure represent the variable they are intended to (2015: 99). The encapsulation of the essence of validity as ‘the degree of truthfulness’ of the research is aptly conveyed by Mohajan (2017: 71).

Price, Jhangiani and Chiang mention three basic kinds of validity, namely face validity, content validity and criterion validity (2015: 99–100). These criteria are described below.

- Face validity refers to the extent to which the various data collecting tools appear to measure what they claim to measure, based on face value. In the context of this study, face validity would be the apparent extent, evaluated informally, to which the web-based questionnaire, the interview schedule, the feedback schedule, and the lines of discussion during the focus group session were measuring the prevalence of data at risk at the institute, factors leading to data being at risk, data rescue activities performed, data rescue challenges experienced, and requirements for a data rescue workflow model.

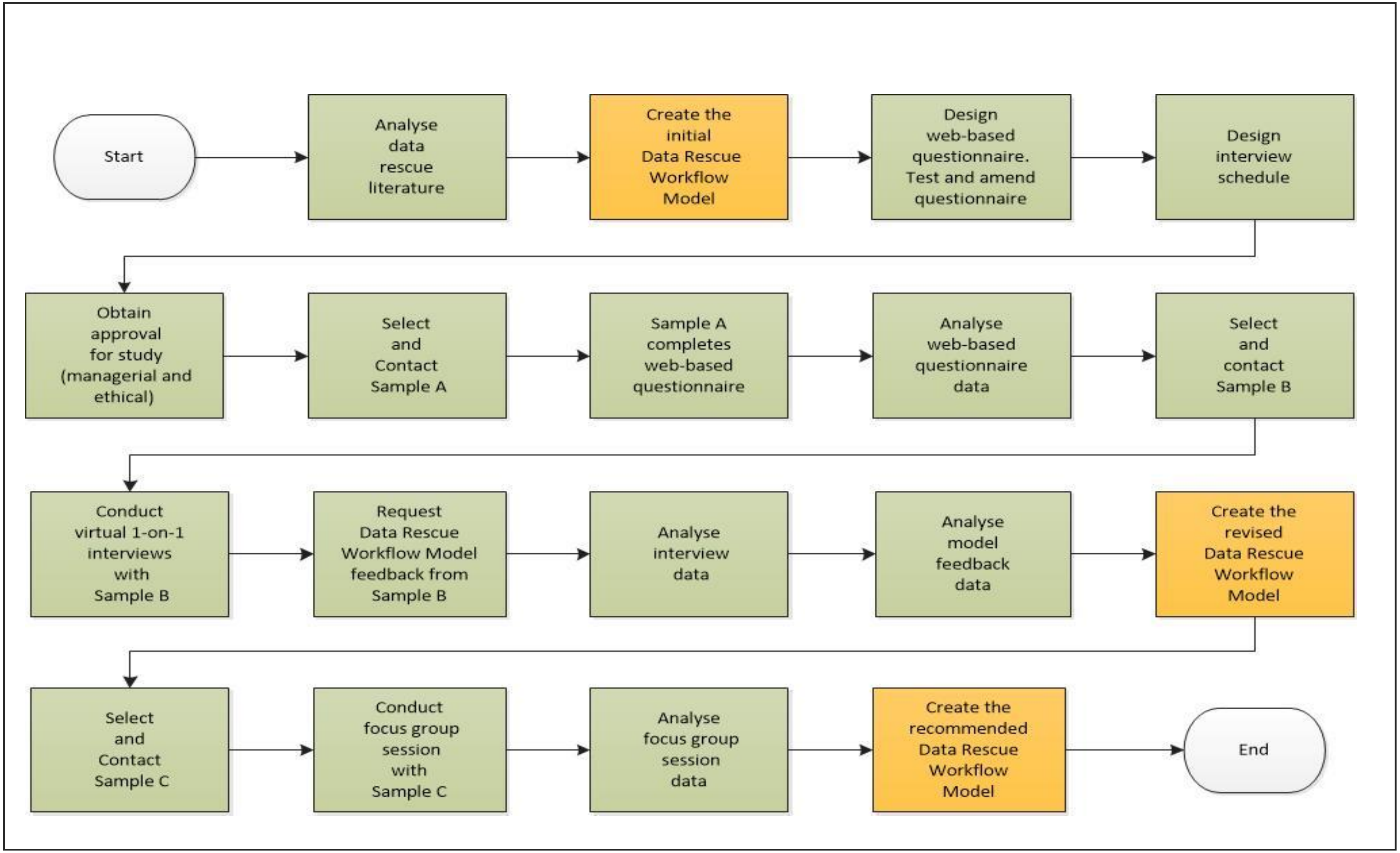

Start
Analyse data rescue literature
Create the initial Data Rescue Workflow Model
Design web-based questionnaire. Test and amend questionnaire
Design interview schedule
Obtain approval for study (managerial and ethical)
Select and Contact Sample A
Sample A completes web-based questionnaire
Analyse web-based questionnaire data
Select and contact Sample B
Conduct virtual 1-on-1 interviews with Sample B
Request Data Rescue Workflow Model feedback from Sample B
Analyse interview data
Analyse model feedback data
Create the revised Data Rescue Workflow Model
Select and Contact Sample C
Conduct focus group session with Sample C
Analyse focus group session data
Create the recommended Data Rescue Workflow Model
End

Figure 4.1: Summary of methodology process

- Content validity refers to the extent to which a measure 'covers' the construct of interest. When applied to the current study, construct validity would require conceptually defining aspects such as 'data at risk', 'data rescue practices' and 'a data rescue workflow'. For this study to adhere to construct validity, definitions of 'data at risk, 'data rescue' and 'data rescue workflow' would need to be made, followed by the careful checking of the measurement against the conceptual definitions.
- Criterion validity is the extent to which people's scores on a measure are correlated with other variables/criteria that one would expect them to be correlated with. Applied to the current study, one can assign criterion validity to a measurement tool when it shows a negative correlation with a criterion expected to negatively correlate with, e.g., prevalence of data at risk, and adherence to data management best practices.

4.11.1.1 Measures to improve validity

Relevant literature describes numerous ways to improve the validity of research, and the list below contains examples of mitigating activities.

- A valid study contains a clear definition and operationalisation of the study's goals and objectives (Mohajan, 2017: 77).
- For a study to be valid, the assessment measure should match the study's goals and objectives (Mohajan, 2017: 77).
- A valid study ensures the absence of 'troublesome' wording and difficult terminology (Mohajan, 2017: 77).
- An audit trail should be kept of the research performed: keeping a record of all the research-related activities and data (e.g., the raw interview and audio-recordings) can increase research validity (Robson, 2002: 176; Kriukow, 2018).
- Prolonged involvement of the researcher in the study can lead to an increase in trust (Robson, 2002: 172; Creswell, 2013: 250), thereby increasing the validity of research.
- Triangulation as a validity strategy entails the triangulation of the data using different instruments of data collection, and methodological triangulation through mixed methods approach (Robson, 2002: 174; Creswell, 2013: 251; Kriukow, 2018).
- Peer debriefing such as opportunities to present and discuss one's research at its various stages provides the researcher with valuable feedback, criticism and suggestions for improvement (Robson, 2002: 174; Creswell, 2013: 251; Kriukow, 2018). Debriefing can take place internally within the institute or externally at workshops or conferences.
- Member checking: testing the emerging findings with the research participants can lead to an increase in research validity (Robson, 2002: 174; Creswell, 2013: 252; Kriukow, 2018).

- Negative case analysis: analysing cases or data that do not match the patterns emerging from the rest of the data can lead to an increase in validity (Robson, 2002: 173; Creswell, 2013: 251; Kriukow, 2018).
- Clarifying researcher bias: when the researcher comments on past experiences, biases, prejudices and orientations that have likely shaped the interpretation and approach to the study, the reader has a better understanding of the researcher's position, biases or assumptions (Creswell, 2013: 251).
- Rich, thick description: when the researcher supplies a detailed description of the participants or the setting, the reader can transfer the information to other settings and determine whether the findings can be transferred (Creswell, 2013: 252).
- External audits: Creswell states that having an external consultant assess the accuracy of the research process and the research product is a way to increase research validity (2013: 252).

The next section discusses the measures taken to ensure that this study demonstrates validity.

4.11.1.2 Validity in this study

The previous section contains a description of the ways in which a study can address and demonstrate research validity. The table below indicates the ways in which this study has implemented validity measures.

Table 4.5: Validity strategies in study

VALIDITY STRATEGY	PRESENCE OF STRATEGY IN CURRENT STUDY
Clear definition of study goals	The study's objective and sub-objectives are described clearly.
Assessment measure to match the study's goals and objectives	Study objectives, research questions, and their attainment in this study form part of Chapter 6: Recommendations.
Elimination of troublesome wording	Data collection instruments were reviewed by peers before being administered to participants. Amendments were made where suggested. The web-based questionnaire underwent a pilot run.
Audit trail	Research-related documentation is included with the thesis as appendices. The study's data and data documentation have been uploaded to the open access repository of the University of Pretoria.
Prolonged involvement of the researcher	With the data collection stages of this study taking place over several months, evidence of prolonged researcher involvement is evident. Communications with participants via several emails over the span of the study's empirical phase also added to the extended researcher involvement, leading to an increase in trust from participants.
Triangulation	The study employed four different data collection instruments and thus involved triangulation of the collected data.
Peer debriefing	This researcher was able to discuss the study and its methodology with several peers on an ongoing basis. Valuable feedback was received, and changes were implemented where deemed necessary.
Member checking	RGLs were supplied with transcriptions of their own one-on-one interviews. Participants were encouraged to indicate errors or omissions.
Clarifying researcher bias	The researcher's bias towards study participants, emanating from knowledge pertaining to the current state of data management at the selected institute, is mentioned during the results chapter of the study. This declaration provides the reader with information to make a decision regarding this researcher's biases or assumptions regarding the topic.
Rich thick description	The detailed nature of this study's methodology chapter, and its description of the setting and the participants, provide evidence of ample and detailed description.

Creswell, in his summary of methods that address the threats to data validity (and included in the list in Section 4.10.1.1), recommends that qualitative researchers implement at least two methods in any given study (2013: 253). This researcher agrees with this stance and is confident that the validity strategies listed in the table above provide evidence of the study's validity.

4.11.2 Reliability

According to Mohajan (2017: 67), reliability indicates the extent to which a study is without bias, and that many researchers prefer to use the term 'dependability' instead of reliability. No matter which term is used, it is the degree to which an assessment tool produces stable and consistent results, and is free from errors (Mohajan, 2017: 67). Robson describes reliability in qualitative studies as mostly a matter of 'being thorough, careful and honest in carrying out the research' (2002: 176), while Price, Jhangiani and Chiang (2015: 96) state that reliability refers to the consistency of a measure.

Lincoln and Guba (1985) viewed a study's trustworthiness as an important indicator of its reliability, and proposed that four criteria be used to evaluate the worth of a study. These four indicators of trustworthiness are credibility, transferability, dependability and confirmability. According to the Applied Doctoral Center (2022), credibility refers to whether the data collected are representative of the phenomenon under study, while transferability can be defined as the extent to which the findings are transferable to other situations (Applied Doctoral Center, 2022). This concept should not be confused with generalisability, as transferability is concerned with whether the findings are applicable to similar contexts or individuals, and not to broader contexts.

A third concept linked to trustworthiness is dependability, which is a concept referring to the in-depth description of the study procedures and analysis to allow the study to be replicated. The final concept related to trustworthiness relates to confirmability, and is described by Applied Doctoral Center as the steps taken to ensure that the data and findings are not due to the participant and/or researcher bias (2022).

Measures to improve study reliability and in particular trustworthiness are touched on in the next sections.

4.11.2.1 Available measures to improve reliability

According to Mohajan (2017: 77), threats to reliability in qualitative research involve the following:

- lack of clear and standard instructions,
- not all alternatives are provided to participants,
- the questions to participants are not presented in the proper order,
- the measurement instruments describe items ambiguously so that they are misinterpreted,
- the questionnaire is too long or hard to read, and
- the interview takes too much time.

Creswell (2013: 253) states that the following aspects can increase a qualitative study's reliability:

- the researcher obtaining detailed field notes using good-quality recording equipment,
- the recordings being transcribed,
- transcriptions to indicate the 'trivial, but often crucial, pauses and overlaps',
- further coding being done by staff and analysts who do not have knowledge of the expectations and questions of the principal researcher, and
- using computer programs to assist in recording and analysing the data.

The next section lists the measures taken to ensure that this study demonstrates evidence of reliability and trustworthiness.

4.11.2.2 Available measures to improve trustworthiness

Credibility: According to the Applied Doctoral Center (2022), credibility can be assured through multiple perspectives throughout data collection. Stahl and King (2020) mention that the various processes of triangulation is one method to promote study credibility; in other words, several sources of information or procedures can be used to repeatedly establish identifiable patterns. Participant validation and rigorous techniques can also be used to increase credibility (Applied Doctoral Center, 2022).

Transferability: A way to achieve transferability is to ensure that the findings from multiple data collection methods are thickly described (Applied Doctoral Center, 2022). The Farnsworth Group (2022) elaborate on this idea by referring to the provision of adequate details on the site, participants and methods or procedures used to collect data during the study. Adding to this is the view of Stahl and King stating that transferability is only possible when a thick description provides a rich enough portrayal of circumstance for application to others' situations (2020).

Dependability: Achieving this trustworthiness criterion is ensured through rigorous data collection techniques and procedures and analysis that are well documented (Applied Doctoral Center, 2022). An inquiry audit using an outside reviewer or a study committee can be seen as an indicator of dependability. The Farnsworth Group states that dependability starts by tracking the precise methods used for data collection, analysis and interpretation and providing adequate contextual information about each piece (2022). By doing this, the study could theoretically be replicated by other researchers and generate consistent results. Stahl and King refer to peer debriefing or peer scrutiny as a dependability method, and are of the opinion that being aware that the work and the products from the work are to be inspected by a peer would cause the researcher to be careful with what is recorded as fact and what is set aside as researchers' interpretive comments about the data (2020).

Confirmability: Confirmability of qualitative data is assured when data are checked and rechecked throughout data collection and analysis to ensure findings would likely be repeatable by others. Confirmability can be enhanced via the use of an audit trail (Stahl & King, 2020), triangulation, member checking of the data, and practicing reflexivity to confront potential personal bias (Applied Doctoral Center, 2022).

4.11.2.3 Reliability and trustworthiness in this study

The previous section contains a description of the ways in which a study can address and demonstrate research reliability and trustworthiness. The tables below indicate the ways in which this study has implemented the concepts in question. Table 4.6 lists the reliability strategies used in this study.

Table 4.6: Reliability strategies used in study

RELIABILITY STRATEGY	PRESENCE OF STRATEGY IN CURRENT STUDY
Clear and detailed instructions	The methodology chapter and its attachments provide evidence that clear and detailed instructions were supplied to participants at the start of every data collection phase, and in the cover letters accompanying each of the data collecting instruments or methods.
Alternatives are available	The inclusion of open-ended questions in all data collection instruments is evidence of alternatives being available to responding participants.
Questions in proper order	Questions were in proper order. The interview schedule was reviewed by external parties, and no obvious limitations reported. This researcher was aware that questions should follow a logical flow, not jump between topics, and commence with questions about the present before querying about the past or future. The last questions allowed participants to add anything they deemed suitable or usable, and to provide impressions of the interview.
Clear instrument description of items	Descriptions of items on measurement instruments were clear, non-ambiguous and not prone to misinterpretation. Measurement terminology and instructions were reviewed by peers.
Length and complexity of questionnaire	The study's web-based questionnaire consisted of eight questions only, while the interview schedule contained five main categories of questions. Both tools were reviewed by peers and not judged as being too long or too complicated.
Questionnaire completion time	It is estimated that the web-based questionnaire took 20 minutes to complete. The study's one-on-one interviews lasted between 20 and 45 minutes.
Recording equipment	Recording equipment used during the one-on-one interviews were of good quality and supplemented by a second recording instrument used during the interviews.
Interview transcription	Interviews were transcribed verbatim, and the transcriptions sent to interviewees for review and correction. No errors or omissions were reported by interviewees.

Table 4.7 lists the trustworthiness strategies used in this study in order to comply with credibility, transferability, dependability and conformability requirements.

Table 4.7: Trustworthiness strategies used in study

TRUSTWORTHINESS STRATEGY	PRESENCE OF STRATEGY IN CURRENT STUDY
Credibility	• Used multiple perspectives throughout data collection process • Made use of methods triangulation • Made use of data triangulation
Transferability	• Findings from multiple data collection methods were thickly described • Adequate details on the site, participants and methods or procedures used to collect data during the study were provided
Dependability	• Data collection techniques and procedures are well documented • A certain level of peer scrutiny was used in the study
Confirmability	• Data were checked and rechecked throughout data collection • Triangulation was implemented • Researcher reflected on data collection and data analysis to confront potential personal bias

With triangulation being part of trustworthiness, and used during this study, it was considered crucial to include within this section a visual representation of the various triangulation activities implemented in this research. Figure 4.2 displays the different triangulation methods implemented in this study, and comprise data triangulation and methods triangulation.

Methods triangulation is the use of multiple methods to study a situation or phenomenon. According to UNAIDS, the core strength of methods triangulation is its potential to expose unique differences or meaningful information that may have remained undiscovered with the use of only one approach or data collection technique in the study (2010: 22). Bhandari states that it is a useful method because one avoids the flaws and research bias that come with reliance on a single research technique (2022). In this study, the use of methods triangulation comprised the data supplied by RGLs during the web-based questionnaire and during the in-person interviews.

A second triangulation technique implemented in this study comprise data triangulation, a technique described by Bhandari (2022) as the use of multiple data sources to answer the research questions. Data collection can be varied across time, space or different people, and has the advantage of resulting in data that are more generalisable to other situations. As stated by UNAIDS (2010: 21) data triangulation makes it likely that the data will be drawn from a much more diverse set of sources and this diversity ensures a more expansive look at the situation. In this study, data triangulation involved the use of data (i.e., data rescue model feedback) from different expert samples. The data collected from two different groups were used to revise the data rescue model that had been reviewed by the experts from the different groups.

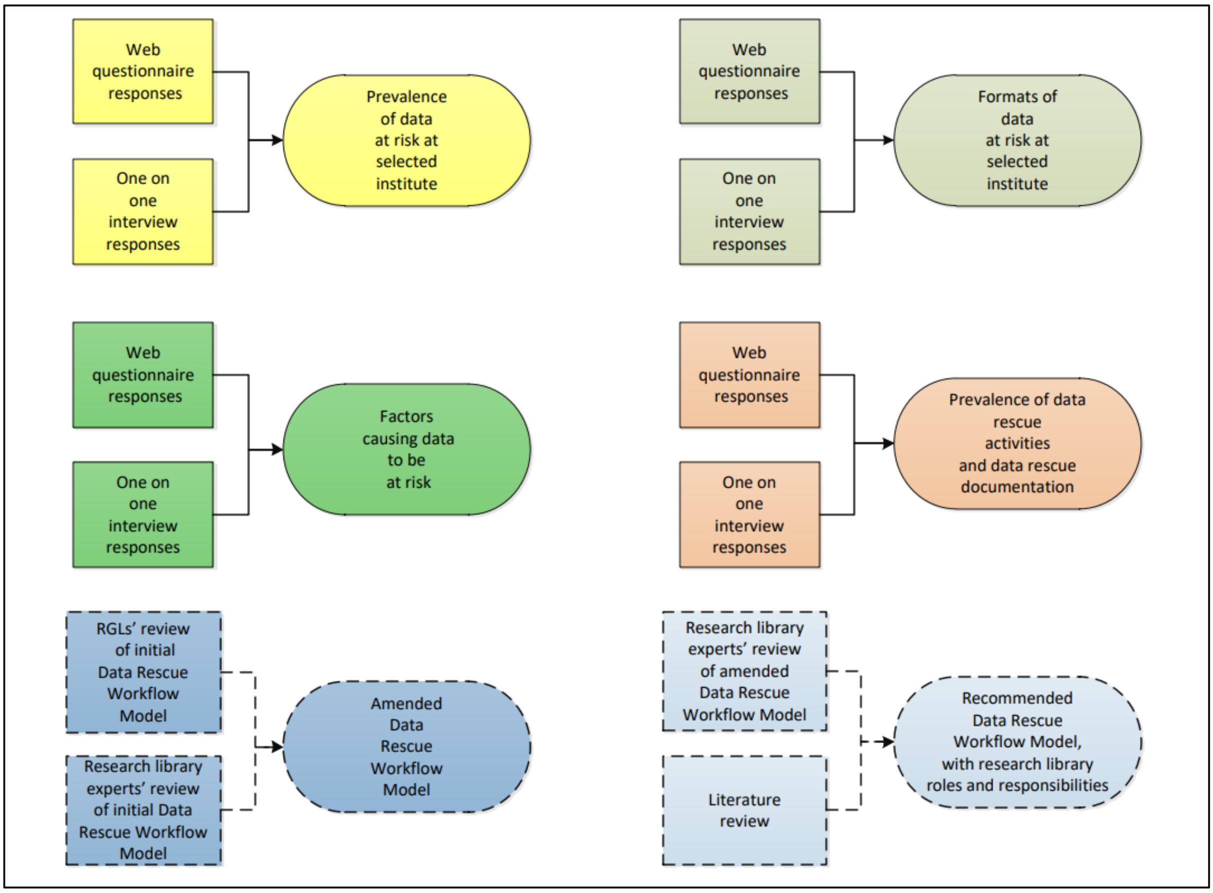

Web questionnaire responses
One on one interview responses
Prevalence of data at risk at selected institute
Web questionnaire responses
One on one interview responses
Formats of data at risk at selected institute
Web questionnaire responses
One on one interview responses
Factors causing data to be at risk
Web questionnaire responses
One on one interview responses
Prevalence of data rescue activities and data rescue documentation
RGLs' review of initial Data Rescue Workflow Model
Research library experts' review of initial Data Rescue Workflow Model
Amended Data Rescue Workflow Model
Research library experts' review of amended Data Rescue Workflow Model
Literature review
Recommended Data Rescue Workflow Model, with research library roles and responsibilities

Figure 4.2: Use of triangulation in this study

The next section contains a short discussion of replicability and reproducibility and how it relates to this study.

4.11.3 Replicability and reproducibility

An examination of literature pertaining to replicability in qualitative research shows that several terms are often used in the same sentence as the term in question. Aguinis and Solarino (2019: 1291) stated that their searches have included the terms quality, transparency, reproducibility, trustworthiness and rigour, while Corti mentioned that terms such as verification, replication, reproducibility, transparency, integrity and restudy all form part of the same 'debate space' (2018).

Adding to the above is Karcher (2019), who refers to a terminology 'crisis' in qualitative research. Karcher also provides the following definitions as clarification between the main terms in question:

- **Reproducibility** refers to the extent to which use of the same data and the same methods will produce the same results.
- **Replicability** refers to the extent to which using the same method but with different data (or sample) will arrive at the same results.

- **Transparency** refers to the extent to which all information necessary to fully evaluate a study is provided.

Seadle and Rügenhagen (2018) mentioned that it is understandable that ambitious researchers avoid doing replications, especially for qualitative research. The authors stated that when attempting qualitative research replications, the risk of failing is high. Furthermore, the authors were of the opinion that the only benefit linked to the successful replication of a qualitative study is the message that the original study was executed well.

Seadle and Rügenhagen (2018) and Seadle (2018) also referred to the difficulty in applying replication to qualitative studies and pointed out that replication entails recreating the exact conditions of the original study, a condition that is often impossible in the real world. They emphasised that qualitative research normally does not generalise about results beyond the community involved in the samples, and that the research question is often limited and relevant to a specific context.

Karcher (2019) agrees with the above and argues that reproducibility and replicability are based on a positivist view of science (e.g., testing a hypothesis in a lab), whereas qualitative research is likely to involve interpretive research methods where the study conditions are difficult to recreate. He suggested that the rigour and quality of qualitative research should not be judged based on reproducibility or replicability, but by the transparency of the research.

This study is a case study involving a unique group of participants at a unique research institute, with data collected at a specific point in time. Exact study conditions would be difficult to recreate, and it would not be wise to attempt the reproduction of findings or expect another case to deliver identical results. Based on this, it was decided to delve into the transparency features of this study. This measurement of rigour is discussed in the next section.

4.11.4 Transparency

As mentioned in the previous section, transparency refers to the extent to which all information necessary to fully evaluate a study is provided (Karcher, 2019). Karcher also referred to three types of transparency, namely:

- production transparency: information on how data are collected or generated, such as research questions, sampling, and subject recruitment,
- analytic transparency: documentation process of qualitative data preparation and analysis leading to the conclusions of a study, and
- data access and taking into consideration access conditions.

The extent to which these three categories of transparency are present in the current study are addressed in Section 4.11.4.2.

The next two sections include a listing of various methods traced in relevant literature describing transparency strategies. The transparency measures forming part of this study are also described.

4.11.4.1 Measures to improve transparency

Aguinis and Solarino (2019: 1291–1306) published a list of 12 transparency criteria which can be used by researchers to gauge or increase the transparency of a qualitative study, and which are outlined below.

- Kind of qualitative method: researchers should be explicit about what specific kind of qualitative method has been implemented (e.g., narrative research, grounded theory, ethnography, case study, or phenomenological research).
- Research setting: researchers should provide detailed information regarding contextual issues pertaining to the research setting (e.g., power structure, norms, heuristics, culture, and economic conditions).
- Position of researcher along the insider–outsider continuum: the researcher should provide detailed information regarding the researcher's position along the insider–outsider continuum (e.g., existence of a pre-existing relationship with study participants, or the development of close relationship during data collection).
- Sampling procedures: the researcher should be explicit about the sampling procedures used.
- Relative importance of the participants/cases: the researcher should be explicit about the contribution that key informants made to the study.
- Documenting interactions with participants: the study should document interactions with participants and specify which types of interactions led to the development of a theme.
- Saturation point: the researcher should identify the theoretical saturation point and describe the judgment calls the researcher made in defining and measuring it.
- Unexpected opportunities, challenges and other events: the researcher should report what unexpected opportunities, challenges and other events occurred during the study, how they were handled, and implications thereof.
- Management of power imbalance: the researcher should report and describe whether power imbalance exits between the researcher and the participants and how it was addressed. The use of endorsement from a prestigious institution, self-acquaintance, and asking sensitive questions form part of this factor.
- Data coding and first-order codes: the researcher should be clear about the type of coding strategies adopted.
- Data analysis and second-order codes (or higher): the researcher should state how the data were analysed.

- Data disclosure: the researcher should make raw materials available, e.g., transcripts and video recordings.

Aguinis and Solarino (2019: 1306) emphasise that these criteria should not be applied rigidly to all qualitative research, as not all of them apply to every situation and type of qualitative study. Overall, this researcher agrees with the opinion of Aguinis and Solarino (2019: 1306), who state that when a larger number of the above criteria are met, a study will display a greater degree of transparency. Conversely, the absence of any item, according the Aguinis and Solarino (2019: 1306), does not mean the study is not worthy of publication and recommendation.

Karcher (2019) emphasised the importance of access control when it comes to the sharing of qualitative data, and in so doing increasing the study's transparency. In a related study, Kapiszewski and Karcher (2021) mentioned the qualitative data sharing strategies available; these strategies and their implementation in this study are described below.

- **Preregistration:** The authors stated that researchers should specify a research project's rationale, hypotheses, design, and plan for data generation and analysis before initiating data collection (2021: 286). This activity is identical to the creation of a data management plan (DMP) and ensures that the researcher has a timestamped record of the original research and analysis plan, as well as changes made during the research process. The DMP is described by Kapiszewski and Karcher as a low-cost way to demonstrate research rigor without overloading a publication with methodological details (2021: 286–287).
- **Methodological Appendices:** According to Kapiszewski and Karcher (2021: 287), the creation of supplementary material that discusses how an author collected, generated and analysed data can advance transparency in several types of research.
- **Annotation** also can help scholars to achieve transparency, and can 'elucidate data generation or analysis, make explicit the link between a source and a claim in a published text, or discuss other aspects of the research process' (2021: 287). The authors further stated that the sharing of extended excerpts can facilitate transparency when it is not possible to share the underlying data.
- **Provide access to data:** The authors explained that this strategy intersects with some of the other activities, such as appendices or annotation (2021: 288). They also mentioned that 'achieving transparency' does not require that researchers share all their data, but that cautious decisions regarding sharing be made. Issues to consider include copyright, whether sharing consent has been obtained, and applicable ethical and legal constraints.

The next section discusses the measures taken to ensure that this study demonstrates evidence of transparency.

4.11.4.2 Transparency in current study

The previous section contains a description of the ways in which a study can address and demonstrate research transparency. The table below indicates the ways in which this study has implemented transparency measures.

Table 4.8: Transparency measures in study

TRANSPARENCY METHODS	PRESENCE OF METHODS IN CURRENT STUDY
Indication of qualitative method	Details of the case study as qualitative method are supplied in Section 4.4: Research design: Case study.
Research setting	The methodology chapter contains details of the research settings (multi-disciplinary SET-based research institute), its range of research disciplines, and the characteristics of participants.
Position of researcher	The position of the researcher within the institute, and in relation to the study participants are elaborated on during the methodology chapter. This researcher had no prior relation with any of the RGLs involved in the study. This researcher had professional working relationships with the LIS experts.
Sampling procedures	The sampling methods used in this study are described in Section 4.7: Sampling and sample. Purposive sampling was the method used in this study, for all samples.
Importance of the participants	The importance of participants was emphasised, and establishing the prevalence of data at risk at the institute, data rescue activities performed, and data challenges experienced would not have been possible without participant involvement. The reason for selecting RGLs and LIS experts as participants was also explained during the methodology chapter.
Documenting interactions	All interactions were added as appendices to the study or formed part of the data of the study. Data were uploaded to the data repository of the University of Pretoria.
Saturation point	This researcher considered a saturation point with regard to data collection being reached during the focus group session of the study. Prior data collection methods had not produced sufficient data, resulting in this researcher extending the empirical phase of the study to include LIS experts in the creation of a Data Rescue Workflow Model.
Unexpected opportunities and challenges	Challenges were included in the methodology chapter and are discussed in Section 4.15: Methodology challenges.
Management of power imbalance	With many of the participants (RGLs) being in a managerial position at the institute, it was vital that permission be obtained from Human Resources for a name list, from the relevant Group Manager to conduct research at the institute, and from each of the institute's research clusters to approach and involve RGLs. All stated permissions were obtained, and the fact shared with participants.
Preregistration	A DMP was drafted prior to the study and added as an appendix to the methodology chapter.
Methodological appendices	All documentation and explanatory records relevant to the study method have been included as appendices.
Annotation	Data documentation will be attached to the data when uploading the data to the repository of the University of Pretoria.
Data access (while considering sharing limitations)	Anonymised data collected during the various stages of the study will be uploaded to the open access data repository of the University of Pretoria. Informed consent has been obtained and the data are free from legal restrictions.

4.11.5 Summary

This section briefly described the concepts of validity, reliability, replicability, reproducibility and transparency in qualitative research. The section also contained brief mentions of the ways in which the researcher has addressed the concepts discussed. An investigation of relevant literature revealed that transparency is a more reliable indicator of scientific rigour, with the lack of generalisability of qualitative research and the unique sample invalidating the inclusion of replicability and reproducibility. The researcher is confident that the study has considered and implemented the concepts of validity, reliability and transparency and that trust and confidence in the findings are bound to be established.

4.12 Ethical considerations and ethical clearance

Ethical clearance from two separate authorities was required in order to proceed with the part of the study dealing with participants and the collection of data. These two entities, tasked with looking at the study's research objectives and methodology, and ensuring that the dignity, rights and safety of participants would be protected, were the University of Pretoria and the institute where participants were employed.

4.12.1 Clearance from the University of Pretoria

This study was one of the requirements for completion of a Doctor of Philosophy degree in Information Science at the University of Pretoria, which is the degree registered for by this researcher. According to University of Pretoria regulations, ethical clearance of all research that include humans (or animals) is required before data collection can commence. This meant that details of the study were required to be submitted to the relevant Faculty Committee for Research Ethics and Integrity, before sending the email containing the web-based questionnaire link to Sample A members. Prior to applying for ethical clearance from the Faculty Committee, all required faculty documentation also had to be submitted to the Departmental Committee, who reviewed the documents and provided feedback.

Required documentation included:

- copies of consent forms that would be used,
- copies of the questionnaire schedule, interview schedule and feedback form that would be used during data collection,
- institutional managerial permission allowing the researcher to conduct the study at the institute,
- approval letter from the university's Department of Information Science, stating that the research proposal had been accepted, and

- a signed declaration by the researcher.

All documents were submitted online.

As this study is not health related, the university ethical clearance process was a one-step process only, compared to health study clearance, which is more complicated and consists of more steps.

Prior to submitting the application, the researcher familiarised herself with the University Code of Ethics for Scholarly Activities, as well as the Policy and Procedures for Responsible Research, as requested on the application webpage.

Ethical clearance was obtained on 23 June 2020, under Reference number **EBIT/114/2020**, and the researcher informed of the outcome via email. The approval document is attached as Appendix 1.

4.12.2 Clearance from researcher's institute

Several groups of researchers and institutional experts were involved as participants in this study. These participants were all based at and employed by the same research institute where the researcher was employed. According to the institute regulations, research conducted by staff members must follow the institute's research ethics policy.

According to the research ethics webpage of the research institute, studies involving and impacting human subjects, animals, genetic manipulation of pathogenic microorganisms, and environmental studies are subject to ethical evaluation. As this study involves human subjects being surveyed, via both web-based questionnaire and virtual interview, obtaining ethical clearance from the institute's Research Ethics Committee (REC) was a requirement before proceeding with data collection.

According to the institutional REC (Mohapi, 2019), the normal procedure for obtaining ethical clearance from the REC involves the following steps:

- The study's principal investigator preparing a new application by completing the relevant application form found in the intranet.
- Electronically submitting completed forms through the REC website 28 days before the REC review meeting.
- The REC reviewing the submission.
- The outcome of the review can be one of four possibilities:
  - approved,
  - approved with provisos,
  - modifications required, or
  - deferred.

Studies are also classified according to risk, and examples of medium- and high-risk projects are supplied on the institutional intranet site. This study is classified as containing medium risk.

This study, where research forms part of a degree, was not required to go through the institutional REC review process as described above. Instead, the review and approval obtained from the university where the qualification is registered (i.e., University of Pretoria) was submitted to the institutional REC for 'noting' (Mohapi, 2019).

Upon receiving an email from the university indicating that ethical clearance had been granted, the institutional REC was supplied with a copy of the letter. The institute's REC was also provided with a copy of the university's ethics application document to show the contents of the ethics application, and the documents provided as part of the ethical clearance process. In addition, the institute's REC was also provided with a copy of a letter signed by the researcher's group manager, providing permission to involve institutional staff members in the study (see Appendix 23: CSIR Permission Letter).

A return email from the institute's REC stated that this research project was exempted from the institute's REC review process, as it was not conducted by an employee forming part of the institute's SET base. Permission was thereby given to proceed with the research project in line with the university's REC approval, and with permission from the researcher's general manager (already supplied).

4.13 Data analysis

The data analysis phase of the study involved analysing data collected during four distinctly separate stages of data collection. In addition, the literature review phase of the study can also be regarded as a stage where information was gathered that would assist in drafting an initial Data Rescue Workflow Model. The remaining four data collection stages were:

- data collected through the web-based questionnaire completed by Sample A (see Section 4.5.1: Web-based questionnaire),
- data collected via virtual one-on-one interviews with Sample B (see Section 4.5.2: Virtual one-on-one interview),
- data collected via Data Rescue Workflow Model feedback (see Section 4.5.3: Feedback guide and critiquing the initial Data Rescue Workflow Model), and
- data collected during a mini focus group session held with Sample C (See Section 4.5.4: Mini focus group session).

Data collected were qualitative in nature and the objective of data analysis was to identify the themes present in all collected data. Based on these aspects, the study implemented content analysis and thematic analysis as data analysis methods. According to Vaismoradi, Turunen and Bondas (2013: 398), these two theme-searching approaches are often used interchangeably, and their similarities result in researchers experiencing difficulties when having to choose between them. The most prominent similarities, as described by Costa (2020: 6), are listed below.

- Both approaches use codes and coding as a method of analysis.
- Both approaches often involve participants being asked questions, which are often itemised in interview questions.
- Questions asked to participants are mostly linked to the study's research questions and hypothesis.
- Messages derived from the data collection tools form the focus of the study.

Despite these similarities, differences between the two approaches exist, with the main difference being the opportunity for quantification of data. Vaismoradi, Turunen and Bondas stated that the measurement of frequency of various categories and themes is possible in content analysis (2013: 398).

The next two sections discuss each of these two methods, followed by a summary of the study's research questions and applicable data analysis methods.

4.13.1 Content analysis

Content analysis refers to the process in which presentations of behaviour or qualitative data from self-reports are analysed (Tutor2u, 2021). In its most common, quantitative form, the method assumes that the texts are phenomena to be examined and provide the units of data collection (Neuendorf, 2019: 212). According to Costa (2020: 8), content analysis as a methodology is applicable to all texts such as written speeches, newspaper articles, documents, reports, digital media, pictures and audio-visual content.

As stated by Tutor2u (2021), the strength of the content analysis approach lies in the offering of a method to analyse a variety of forms of data, including media and self-report. Tutor2u (2021) also mentions that the subjective identification of suitable themes and codes is a limitation of the method, resulting in the conclusions lacking 'scrutiny or objectivity'.

The course of action followed during this study when conducting a conceptual content analysis, based on the steps described by Columbia University, Mailman School of Public Health (2019) are indicated below.

- The researcher decided on the level of analysis, and whether it would be on the level of word, word sense, phrase, sentence or themes. Word sense was selected.

- The researcher decided on the number of concepts to code for and developed a pre-defined set of categories or concepts. The researcher decided to allow flexibility to add categories through the coding process and concluded that it was not necessary to stick with the pre-defined set of categories.
- The researcher decided to code for frequency of a concept while also evaluating all existing data rescue concepts mentioned in published workflows or models.
- Text was coded using highlighted colours and linked comments made possible by the MS Word text tool.
- Results were analysed.

The next section provides a similar abridged discussion of thematic analysis as applied in this study.

4.13.2 Thematic analysis

Thematic analysis offers a means of identifying patterns or themes in a dataset, and for describing and interpreting their meaning and importance (Braun, Clarke & Weate, 2016: 192). The flexibility of thematic analysis means it can be used with a wide range of different research designs and data collection methods, and there is no ‘ideal’ data type (Braun, Clarke & Weate, 2016: 195). Costa (2020: 9) agrees with the above, and states that the purpose of thematic analysis is solely to generate themes for drawing conclusions.

Costa further states that the method is applicable to most forms of qualitative research (2020: 9), while Tutor2u (2021) mentions that it is an alternative to content analysis when converting qualitative data into quantitative data. Once data are transcribed (where necessary), data are reviewed repeatedly so that the researcher can identify trends in the meaning conveyed by language (Tutor2u, 2021).

Neuendorf mentions that the frequency of occurrence of specific codes or themes is usually not a main goal of the analysis (2019: 212), and that the text itself is the data, while codes emerge inductively from the texts (2019: 212).

Braun, Clarke and Rance (2015: 188–189) mention that thematic analysis comprises six phases; these phases have been implemented in the current study in the following manner:

- The researcher ensured familiarisation with the data (text and transcriptions) and identified items of potential interest.
- Initial codes were then generated that identified prominent features of the data relevant to answering the research questions. Codes were consistently applied to the dataset, and codes were collated across segments of the dataset.

- Searching for themes was done by examining the codes and collated data to identify broader patterns of meaning.
- Themes were reviewed and the potential themes were applied to the dataset to determine if they describe a narrative that can answer the research questions. Themes were refined, combined or discarded.
- Themes were then defined and named, and each theme analysed.
- A written report was produced by combining the analytic narrative and data segments.

The steps above refer to the study's thematic analysis activities, and not to the content analysis steps. Different data were used during content analysis and thematic analysis respectively, and Table 4.9 portrays the different datasets and activities.

Braun, Clarke and Weate suggest that six interviews be the minimum sample size for thematic analysis, but add that the suggestion does not consider the specifics of the research questions and research design (2016: 196). In the current study, data collection stages involving thematic analysis were the one-on-one interviews (eight respondents), feedback stage (four respondents) and the mini focus group session (three respondents).

4.13.3 Summary

The table below lists each of the study's research questions and indicates the data sources and data analysis that were used to answer these questions.

Table 4.9: Data sources and data analysis for answering of research questions

RESEARCH QUESTION	DATA SOURCE(S)	ANALYSIS METHOD
Sub-question 1: The current data rescue frameworks/workflows	Literature analysis	Content analysis
	One-on-one interview	Thematic analysis
Sub-question 2: The current South African (SA) workflows in data rescue vs international best practices	Literature analysis	Content analysis
	One-on-one interview	Thematic analysis
Sub-question 3: The current documented state of library and information services involvement with data rescue in the global community	Literature analysis	Content analysis
	Mini focus group session	Thematic analysis
Sub-question 4: The current documented state of data rescue awareness in the SA library and information services community	Literature analysis	Content analysis
	Mini focus group session	Thematic analysis
Sub-question 5: The current documented state of data rescue involvement in the SA library and information services community	Literature analysis	Content analysis
	Mini focus group session	Thematic analysis
Sub-question 6: The current documented state of data rescue globally and in SA	Literature analysis	Content analysis
	One-on-one interview	Thematic analysis
Sub-question 7: Formalising the theory and practice in a workflow model for a data rescue	Mini focus group session	Thematic analysis
	One-on-one interview	Thematic analysis
	Data Rescue Model feedback	Thematic analysis
	Mini focus group session	Thematic analysis
Sub-question 8: Inclusion of data rescue topics in the LIS curricula	Literature analysis	Content analysis
	Web-based questionnaire	Thematic analysis
	One-on-one interview	Thematic analysis
	Data Rescue Model feedback	Thematic analysis
	Mini focus group session	Thematic analysis
Main research question: What are the roles and responsibilities of the research library within a comprehensive workflow for data rescue?	Literature analysis	Content analysis
	Web-based questionnaire	Thematic analysis
	Virtual one-on-one interview	Thematic analysis
	Data Rescue Model feedback	Thematic analysis
	Mini focus group session	Thematic analysis

As shown in the table above, this study overwhelmingly made use of content and thematic analyses as data analysis techniques.

4.14 Data management plan

Following the approval of the study proposal, and prior to data collection, the researcher drafted a DMP for this study. As the data librarian at a research institute, the researcher was well informed with regard to the benefits of such a plan. According to the Research Data Service of the University of Edinburgh (2019), the use of a DMP results in the following benefits:

- provides impetus for researchers to consider how they will manage and share data (before project begins),
- establishes framework and resources to support the data,
- encourages better time management,
- results in lower research costs,
- assists in finding one’s data, and

- allows or validates published results.

The DMP was created making use of the online tool of the Digital Curation Centre[30]. The tool is known as 'DMPonline' and enables any researcher to freely access the platform to create a plan. The tool also provides tailored guidance and examples, and once created can be exported to various formats and shared as the plan creator sees fit.

The DMP drafted for this study can be viewed in Appendix 22: Data management plan.

4.15 Methodology challenges

The main methodological challenge of this study was the introduction of COVID-19-related lockdown restrictions, an event coinciding with the submission of documents for ethical clearance. As a result, the ethical clearance submission documents had to be amended before being submitted, in order for the planned methodological activities to adhere to all restrictions.

The envisaged in-person one-on-one interviews had to be changed to virtual one-on-one interviews. Knowledge of recommended virtual interview tools had to be obtained, and an understanding of best practices regarding virtual interviews as a data collection method was vital.

4.16 Summary

This chapter presented details regarding the methodological steps followed during the study. It outlined the rationale for the qualitative approach and explained why a case study design was used. The chapter clarified the samples chosen, and sampling techniques used.

All four empirical data collection methods were described; these methods comprise a web-based questionnaire, virtual one-on-one interviews, feedback regarding a newly created initial Data Rescue Workflow Model, and a mini focus group session. These four methods were selected as they were deemed the most suitable tools for the collection of data about the institute's data at risk, data rescue experience and data rescue activities performed, and data rescue obstacles encountered or anticipated. In addition, data collection methods provided insight into the requirements of the institute regarding a usable Data Rescue Workflow Model.

The chapter also presented the chronological outlay of the data collecting steps. The data collection chronology entailed the following main steps:

- the distribution of a web-based questionnaire to the selected institute's RGLs (Sample A),
- the analysis of the web-based questionnaire responses; selection of Sample B based on web-based questionnaire responses,

[30] Digital Curation Centre (DCC): an internationally recognised centre of expertise in digital curation with a focus on building capability and skills for research data management. See http://www.dcc.ac.uk/

- virtual one-on-one interviews conducted with Sample B participants,
- following the interviews, feedback on an initial Data Rescue Workflow Model supplied by Sample B participants,
- a revised model created after studying Sample B feedback,
- a mini focus group session held with institutional research library experts (i.e., not SET based) forming Sample C; the session entailed discussions around the initial Data Rescue Workflow Model, and the revised model, and
- all feedback evaluated for inclusion in the recommended Data Rescue Workflow Model.

The process of obtaining ethical clearance from two entities was described. Data analysis steps were listed. Assumptions, limitations and delimitations of the study were also included in this chapter.

In addition, the chapter included a reference to the DMP drawn up prior to data collection; the plan is attached as Appendix 22. The chapter concluded with a description of the methodological challenges experienced.

The next chapter represents the findings of the study. Results of the various empirical data collection methods are provided, together with a discussion on the significance and implications of the results obtained.

CHAPTER 5: RESULTS AND FINDINGS

5.1 Introduction

This chapter presents the notable findings of the study in a non-evaluative, unbiased, organised manner. The chapter also reveals how the findings relate to the research questions of the study, and how the research objectives are addressed throughout the reported findings.

To reach the listed objectives, several data collection undertakings were performed. Four distinctly separate but vital data collection activities (a web-based questionnaire, virtual one-on-one interviews, feedback on a Data Rescue Workflow Model, and a mini focus group session) were completed; each of which contributed in a different manner towards reaching the study's research objectives and answering the research questions.

5.2 Chapter outline

This chapter, containing the results of the study and a discussion of findings, consists of four main sections. Each section made use of a different data collection tool to collect data, which along with the information gleaned from the data collected were essential in meeting the study's research objectives and answering the study's research questions. The four main sections of this results chapter, with the respective data collection tool/method used, are:

1. **results of a web-based questionnaire involving Sample A:** providing basic information on the institute's data at risk, and data rescue experience (see Section 5.3),
2. **results of one-on-one interviews with Sample B**, providing in-depth information on the institute's data at risk, and data rescue practices (see Section 5.4),
3. **results of feedback supplied by Sample B**, after they had reviewed the **initial Data Rescue Workflow Model** (model discussed in Section 3.7, Section 3.8 and Section 3.9), are provided in Section 5.5, and
4. **results of a mini focus group session held with Sample C**, after focus group discussions about the **initial Data Rescue Workflow Model**, and the **revised Data Rescue Workflow Model** (see Section 5.7).

Detailed discussions on the findings emanating from each of the data collection stages are provided in the remainder of the chapter.

5.3 Results: Online questionnaire

5.3.1 Introduction

This portion of the chapter contains the results emanating from the responses supplied by research group leaders (RGLs) when completing an online questionnaire pertaining to data at risk and data rescue. The results of the short questionnaire provide a broad glimpse into the prevalence of data at risk in the institute, the formats involved, the research disciplines involved, and locations of the data at risk. In addition, the results indicate the incidence of data rescue activities within the institute, and the availability of data rescue documentation.

The main objective of this part of the study was to provide the researcher with a sample of RGLs who would be involved in the next phase, which entailed in-depth one-on-one interviews regarding their data at risk, and/or their data rescue practices.

The RGLs of the research institute were invited to complete a short web-based questionnaire. The questionnaire contained eight questions related to the group's data at risk, and data rescue activities performed by the group. Respondents also needed to state their names and research group, as these details were required should the respondent form part of the sample of the study's next data collection phase.

A detailed description of the web-based questionnaire can be found in Chapter 4, Section 4.5.1: Web-based questionnaire.

This section contains and elaborates on the responses pertaining to the following:

- sample and responses,
- research disciplines of respondents,
- prevalence of data at risk,
- formats of data at risk,
- description of data at risk,
- location of data at risk,
- data rescue experience,
- data rescue documentation,
- shareability of data rescue documentation,
- suggestions of follow-up names, and
- any additional comments and concerns.

Each of the listed topics is discussed, with discussions containing responses, trends (if applicable), explanations, and references to relevant scholarly work.

When considering the study's research objectives, the results gained via the online questionnaire contribute towards the following:

- establish how data rescue is currently performed,
- determine who is currently involved in data rescue, and
- establish current data rescue perceptions, needs and challenges.

5.3.2 Questionnaire sample and response rate

A name list obtained from the research institute's Human Resources Department indicated that the institute had 49 RGLs in its employment at the time of the list request. Although the institute had 109 research groups at the time of the study, not all the groups had designated RGLs, while several RGLs were also responsible for the management of more than one research group. These factors resulted in the questionnaire sample (Sample A) comprising 49 RGLs in total.

Twenty-three completed or partially completed questionnaires were submitted by closing date. In total, 22 completed questionnaires and one partially completed questionnaire were received.

One RGL stated that a proxy (designated by him) would be a more informed data rescue respondent, and requested that the proxy, also affiliated to the same research group, be invited to complete the questionnaire.

One RGL submitted a partially completed questionnaire. Despite this questionnaire not containing responses to all the questions, it was regarded as admissible, as stated in the questionnaire cover letter.

The table below indicates the sample size and response numbers for the online questionnaire.

Table 5.1: Online questionnaire sample and response

RGL CATEGORY	NUMBER OF QUESTIONNAIRES
RGLs in institute	49
RGLs emailed	49
Questionnaires received	23
Fully completed questionnaires	22
Partially completed questionnaires	1
Questionnaires completed by proxy researcher	1*

(* Questionnaire completed by proxy was included in the total of 22 fully completed questionnaires)

5.3.3 Research disciplines of respondents

Indicating the research discipline was one of the questions of the online questionnaire. Results showed that the 23 respondents were affiliated to a wide and varied range of research disciplines. The table below indicates the research areas connected to questionnaire respondents. The column on the right indicates the research area, while the left-hand column is the wider cluster (see Section 4.7.1, Table 4.4, for details of all research group areas at the institute).

Table 5.2: Research groups of respondents

CLUSTER	RESEARCH GROUPS OF RESPONDENTS
Advanced Agriculture & Food	Agroprocessing
Defence and Security	Command and Control
	Cyber Security
	Digital Electronic Warfare
	Electronic Warfare Technology
	ID Authentication
	Integrated Capability Management
Future Production: Chemicals	Advanced Polymers and Composites
	Bioprocessing Technologies
Future Production: Manufacturing	Advanced Casting Technologies
	Future Production Systems
	Human and Societal Systems
	Industrial Robotics
	Laser Enabled Manufacturing
	Sonar
Nextgen Enterprises and Institutions	Advanced Internet of Things
	Digital Audio-Visual Technologies
Nextgen Health	Bioengineering and Integrated Genomics
	Human Molecular Diagnostics and Omics Technologies
Smart Mobility	Accelerated Pavement Testing
	Advanced Materials Testing Laboratories
	Coastal Engineering and Port Infrastructure
	Pavement Design and Construction
	Transport Management Design and Systems
	Infrastructure Innovation

A total of 25 research disciplines were included in the 23 responses. One of the respondents represented three research groups: the Accelerated Pavement Testing group, the Advanced Materials Testing Technology group, and the Pavement Design and Construction group.

5.3.4 Prevalence of data at risk

Question 1 of the online questionnaire required respondents to indicate whether they had data at risk within their research group.

5.3.4.1 Findings

The results indicated that most research groups had data, described by the group's RGL, to be at risk.

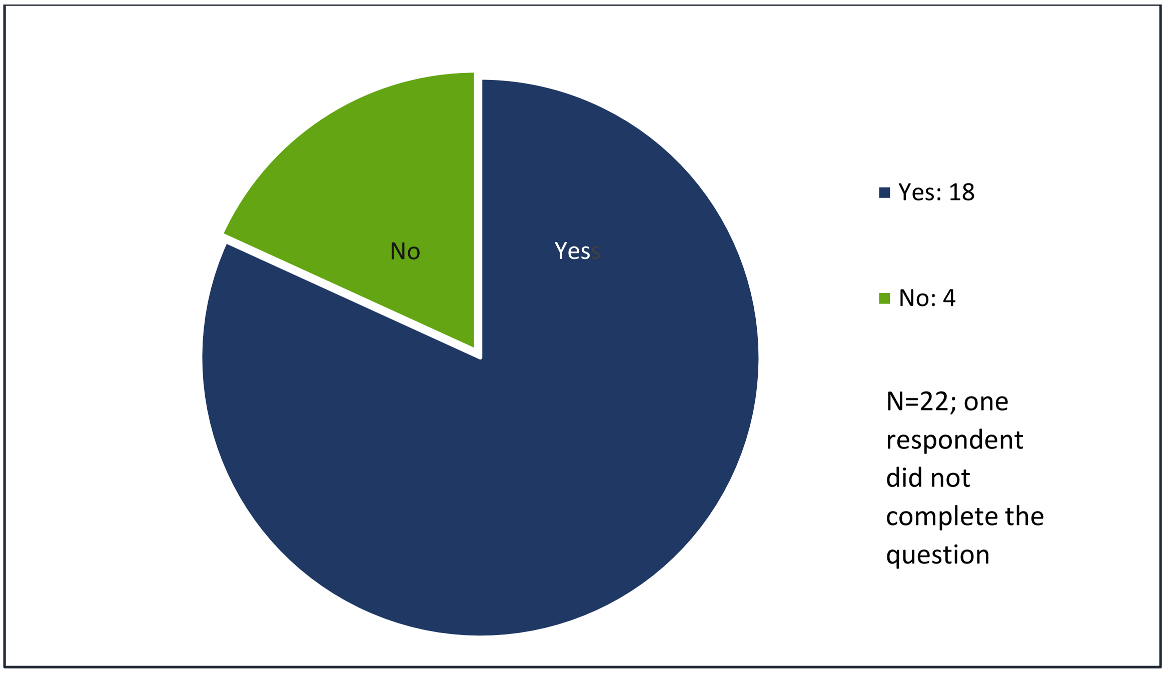


Figure 5.1: Data at risk

5.3.4.2 Discussion

The most significant finding with regard to data at risk is the revelation that most respondents (72%) had data at risk in their research groups. This finding was not entirely unexpected; previous research into the data management practices at the same institute (Patterton, 2014; Patterton 2016; Patterton, Bothma & Van Deventer, 2018) had revealed that both experienced and emerging researchers had shown low adherence to good data management practices. With low observance of best data management practices, particularly with regard to metadata, safe storage locations and a preservation strategy, data are bound to end up being at risk, damaged, of no use to future parties, or lost. In addition, the data management procedure had at the time of the data collection stages of the study not yet been approved, data management plans were not yet mandatory, and institution-wide data management awareness training had not yet commenced.

In the absence of institutional policies and services with regard to data management, it came as no surprise that most of the responding RGLs stated that their group had data regarded as at risk of being lost or damaged.

5.4.4.3 Implications

Aspects emanating from the findings and discussion above, and resultant implications, are provided below.

- It is an unfortunate reality that data at risk are likely present in most of the institute's research groups, including groups that had declined to participate in the study.

- Newer groups, for example the 'Advanced Internet of Things' research group, might not yet have data at risk.
- Type, format, scope and time period of data at risk are bound to differ between groups and are most likely linked to certain disciplines.
- The volume/scope of data at risk is likely to increase daily, unless research practices are changed, or data rescue interventions are implemented.
- Data at risk were found to involve the institute's older data as well as modern data. A discussion around the age of data at risk appears later in this chapter (see Section 5.4.5.3: Discussion).
- Data found to be at risk, and of value to others, should ideally be rescued. Data rescue and its accompanying activities are discussed later in this chapter.
- The prevalence of data at risk at this institute underlines the listed aspects below.
    - There is a need for data rescue.
    - There is a need for a data rescue workflow model, as such a model is currently not available.
    - Data at risk factors, and data rescue obstacles/challenges should be examined, and steps implemented to reduce these challenges and the eventual need for data rescue.
    - It is assumed that non-adherence to best data management practices can contribute to data being at risk.
    - There is an urgent need for data management awareness training, and rollout of data management procedures, services, tools and systems.

Many of the implications linked to findings regarding data at risk prevalence are further discussed in the Recommendations chapter of the study (see Section 6.4.1).

Scholarly comparison regarding data at risk prevalence

It is important to note that the findings with regard to prevalence of data at risk at this institute build on existing evidence regarding the pervasiveness of such data. Various entities, scholarly publications and other outputs, focused on data at risk, are listed below.

- Interest and Task Groups focused on data at risk: these include the former CODATA 'Data at Risk' Task Group[31], the former RDA Data Rescue Interest Group[32], and the recently formed RDA Data Conservation Group[33]. The objectives and outputs of these groups are discussed in the literature review (see Section 2.6.5: Data rescue entities and interest groups).

[31] https://codata.org/initiatives/task-groups/previous-tgs/data-at-risk/
[32] https://www.rd-alliance.org/groups/data-rescue.html
[33] https://rd-alliance.org/groups/data-conservation-ig

- I-DARE[34], IEDRO[35] and ACRE[36] are all examples of organisations involved with the rescue of valuable data at risk of being lost. I-DARE, for example, provides a single point of entry on the status of past and present worldwide data rescue projects, data to be rescued, and best methods and technologies for rescuing data.
- The World Meteorological Organization (WMO) has published two separate guidelines on how to rescue data at risk (WMO, 2014; WMO, 2016).
- The Data Rescue initiative in the USA refers to a movement among scientists, researchers and other concerned parties to preserve primarily government-hosted datasets, often scientific in nature, to ward off their removal from publicly available websites (Schlanger, 2017).
- The plethora of data rescue studies referred to in the literature review confirm the abundance of data at risk in research environments. Examples of data at risk and their resultant rescue include the rescue of climate data at risk (Brönnimann *et al*., 2018), the rescue of physics data at risk (Curry, 2011), the rescue of health data at risk (Thompson, 2017), and the rescue of religious data at risk (ESA, 2018).

Research disciplines and data at risk

With respondents managing research in a range of subject areas (see Table 4.4, Section 4.7.1), and data at risk already identified to be a prevalent feature at the institute, the range of disciplines having data at risk was an expected outcome.

These findings also correlate with scholarly and other publications on data at risk: studies about data at risk in the social sciences (Khwela, 2018), the socioeconomic sphere (Downs & Chen, 2017), health sciences (Thompson, 2017), physics (Curry, 2011) and arts and humanities (Libraries+ Network, 2017) are indicative of the range and scope of such vulnerable data.

The implications of this diversity of disciplines and subject areas are numerous, and are listed below.

- It seems unlikely that a generic version and template for data rescue guidelines, procedures, and a workflow model will suffice. Different disciplines would have different requirements, standards, equipment and ontologies. While a generic template could be drafted and made available, it is probable that future rescue efforts by different disciplines will lead to different versions of this study's initial Data Rescue Workflow Model. It is anticipated that future versions or revisions of the model will cater for different research disciplines.
- A data rescue workflow model should consider the differences between various research disciplines or subject areas. This is especially relevant in rescue activities such as metadata

[34] https://www.idare-portal.org/
[35] https://iedro.org/
[36] http://www.met-acre.org/

creation or selection of a discipline-specific data repository. Different disciplines have different metadata standards, and different repositories recommended or used for data upload.

- It is not only vital that a data rescue model accommodate different disciplines and subjects, as training with regard to data rescue should also incorporate these differences.
- It is estimated that the input and involvement of subject specialists will be required should discipline-specific data rescue models and guidelines be drafted.
- Further investigation into data at risk and different research disciplines might reveal that data risk factors are discipline specific.
- Further investigation into data rescue challenges and different research disciplines might reveal that data rescue challenges are discipline specific.

The next section features the study findings regarding formats of data at risk.

5.3.5 Formats of data at risk

Question 2 of the online questionnaire required respondents to indicate the formats of the group's data at risk.

5.3.5.1 Findings

A wide range of formats was indicated, with paper and modern electronic formats being the most common, followed by early digital formats and physical samples. The prevalence of the different formats, as found in the 25 research groups, is presented in the graph below.

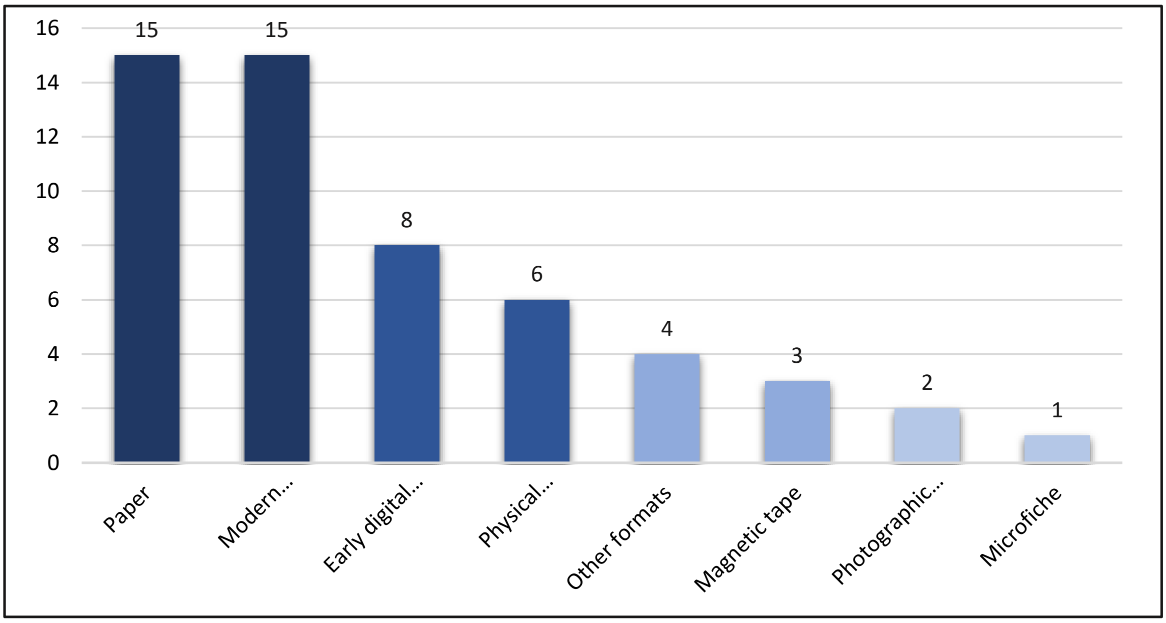


Figure 5.2: Data at risk formats

Four research groups indicated that they had ‘other data’, which was a response option available should the group have data at risk not belonging to one of the listed format options. These data formats were described by respondents as:

- audio tape (one respondent),
- film: 8mm and 16mm formats (one respondent), and
- video tape (one respondent).

While not format-specific, respondents also listed the following research-related outputs to be at risk:

- contract information (one respondent),
- data collected through the deployment of concept (one respondent),
- data collected through the deployment of technology demonstrators (one respondent),
- design documentation on paper of [*sensitive discipline*] laboratory systems (one respondent),
- design environments captured on specific tools in specific versions of Windows, which is ‘required for logistics support as our systems are located the world over and is maintained for decades’ (one respondent),
- equipment user manuals (one respondent),
- programming data on equipment that had to be ‘removed due to space restrictions’ (one respondent),
- ‘project and contract information’ (one respondent),
- project work, including the use of supplied data, but also the generation of new data (one respondent),
- reference material (one respondent),
- reports on technical work (one respondent),
- software code (one respondent),
- source code, including the ability to run old versions of software applications (one respondent),
- standard operating procedures (one respondent),
- technology demonstrators (one respondent),
- various course notes and books in paper format (one respondent), and
- work authorisations (one respondent).

5.3.5.2 Discussion

As indicated earlier in Table 4.4 in Section 4.7.1, this institute conducts research in a diverse range of subject areas. Research groups and their encompassing subject areas range from the natural environment to the built environment to defence-related research. While most research groups and projects are found to be SET based, the institute also conducts research featuring aspects of the social sciences. This multitude and heterogeneity of the institute's research disciplines automatically result in a range of data formats and types used during research. Invariably, the diversity in data types and formats result in data at risk also featuring a range of data formats and data types, as was revealed in the study's results chapter.

These findings regarding a range of data types at risk are mirrored by the findings of the literature review into data at risk. Examples of the diversity of data at risk include the study of photographic slides rescue (Wippich, 2012), rescue of cartridges (Curry, 2011), physical samples rescue (Hills, 2015), rescue of photographs, photoplates and negatives (Wippich, 2012; Thompson, Davenport Robertson & Greenberg, 2014), analogue video rescue (Barnard Library & Academic Information Services, 2019), the rescue of video games (Janz, 2018), and punch cards rescue (Krotz, 2011).

It must be clarified that the statement 'all data formats were found to be at risk' does not translate into all data formats automatically being at risk in all research groups. It merely reveals that all formats listed as answer option were indicated by at least one RGL as a format forming part of the group's data at risk.

A significant finding is that the most common data at risk format was found to be paper-based data, while data in a modern electronic format was stated to be the second most prevalent. Early format digital data were not found to be a common risk format at this institute; the same was revealed for magnetic tape data and plates.

The profusion of paper-based data at risk within this institute builds on existing evidence emanating from scholarly data rescue projects, where paper data were also found to be one of the data formats featuring most frequently during data rescue. Examples of paper-based data at risk include the studies by Brandsma (2007), Wippich (2012) and I-DARE (2019).

However, the findings at this institute also show modern electronic data to be this institute's second most frequent format at risk, and early digital formats as the third most frequent format at risk. This is in contrast with scholarly publications where magnetic tape data (e.g., Curry, 2011) and early digital format (e.g., Thompson, Davenport Robertson & Greenberg, 2014) are the formats following paper-based data in frequency or prevalence. The reasons for these findings and discrepancies are hypothesised below.

- Paper-based data are often mentioned due to its use and prevalence in historic or older projects.
- Paper-based data are prone to a multitude of risk factors, including fire, flood, insect damage, rodent damage, temperature extremes, humidity, acidity of paper, natural degradation over

time, being misplaced, loose papers stored out of sequence, and degradation due to skin oils on ungloved hands when handling paper.

- Most data rescue scholarly publications report on rescue of older data. Such data, especially historic in nature, were paper based: maps, ships' logbooks, weather charts and diaries are examples of such data.
- It is fair to state that the frequency of respondents mentioning modern electronic data as at risk is indicative of such data's frequent generation and abundance at this institute. Many modern electronic data records, stored on the previous document management system (DMS) (replaced in 2019) are irretrievably lost due to being corrupted. Some of the data can be recovered at considerable financial expense. Apart from the expressed dissatisfaction with the former DMS was respondents' disappointment with the new DMS (used from 2019 onwards), which was described as slow, complicated and cumbersome. The DMS was also mentioned as a risk factor by one of the RGLs; this factor is discussed in Section 5.4.7: Data at risk: Factors. These recurring and serious issues with modern electronic data have most likely contributed to it being mentioned by several RGLs as a data risk factor.
- Data rescue of modern electronic data have not featured in the literature consulted for this research. Current rescue trends and scholarly outputs were found to focus on the recovery and preservation of paper-based or early digital data.
- The fact that the findings of the study did not reveal early digital data to be a common risk format was surprising. The suspected reasons for this are listed below.
    - It could be a question of 'out of sight, out of mind'. It is likely that these data formats are stored away from researchers' offices and laboratories, and hardly ever encountered or accessed.
    - The institute had undergone various restructuring initiatives prior to the study's data collection stages; it is likely that newly formed research groups might not have any older data in early digital formats.
    - Another possibility is that the group's early digital data might not be at risk, but still in good condition and accompanied by working data readers.
    - Adding on to the above point: the group's early digital data might be perceived to be not at risk.
    - Lastly, respondents might have stated that the data are not at risk as it is not worthy of being rescued (i.e., data have no value), or a duplicate of the data is available elsewhere.

The next section touches on the implications emanating from findings into formats of data at risk.

5.3.5.3 Implications

The various implications resulting from the study's findings into formats of data at risk are listed below.

- Parties dealing with research data, including researchers, technicians, research managers, funders, ICT support and research library personnel should be cognizant of the fact that all data formats have the potential to be at risk of either damage or loss.
- Risk factors for different data formats are bound to differ, e.g., the likelihood of paper data damage by insects or humidity is greater than it is for modern digital data. In a similar vein, fires, floods or earthquakes are greater risk factors for paper-based data than for data saved in cloud storage, or on an institutional server.
- This specificity of data risk factors with regard to data formats should be considered and included when drafting a data management plan, a risk management plan, or teaching data management to relevant parties.
- As was stated in the earlier section on data at risk and research disciplines, a data rescue workflow model should be inclusive of a range of data formats. The differences between formats with regard to data rescue would need to be addressed and explained; rescue aspects such as estimated cost (in time, money and human resources), type of rescue equipment required, skills required, storage space, storage locations and storage devices should feature in data rescue guidance.
- The possibility exists that it is likely that not all data formats at risk can be rescued. In addition, some data at risk might require considerable effort and collaboration to rescue the specific data age, data format, or volumes at risk. A good example of such problematic data rescue is the study describing the locating of misplaced 40-year-old moon dust data, and the efforts required to trace and restore a machine capable of reading the data (Modine, 2008).

5.3.6 Description of data at risk

Question 3 of the online questionnaire required respondents to provide a brief description of the research group's data at risk.

5.3.6.1 Findings

Examples of the institute's data at risk as indicated via the online questionnaire are listed below. The exact discipline and research details have been removed to ensure anonymity of responses.
Examples include:

- biometric data in modern electronic format,
- genetically engineered physical samples stored in a chemical-specific tank,

- historical data pertaining to specific aspects of SA's built environment, dating from the 1970s,
- historical work done in the bio-scientific sphere,
- materials sciences data emanating from the 1980s and 1990s (paper format and early digital format),
- materials sciences data generated after 2000 (electronic format),
- social sciences data emanating from the 1990s; also used during an RGL's doctoral studies research 30 years ago, in which focus groups were used (audio and video data), and
- social sciences survey data (paper format).

In addition, the problematic and often overwhelming issue of data at risk was detected from some of the responses. Direct quotes illustrating this situation are listed below.

- 'Not possible to describe all the data; it would be a very significant activity.'
- 'Human knowledge lost because of accidents, COVID-19, and people leave the organisation. The manpower is very limited in the organisation due to simple economics.'
- 'Group dissolved. Information difficult to find.'
- 'During the lockdown for example, we were (almost) unable to top up, which would've cost us five to six years of research.'
- 'Some of this data have been lost due to restructuring of groups and individuals. Continuance on projects from members that have left the [*institute*] is problematic.'
- 'Sometimes we get requests from external interested partners and then half of the hard copies and documents are not properly stored and therefore the opportunity can't be rekindled.'
- 'A lot of the knowledge of historic data lies with former employees outside the [*institute*] who are mainly retired personnel.'

5.3.6.2 Discussion

The reader should ideally read the findings forming part of this section (Section 5.3.6.1) in conjunction with the findings presented for data at risk formats (Section 5.3.5.1) and findings linked to location of data at risk (Section 5.3.7.1). The combined reading of the sections will provide a better overall view of the institute's data at risk as reported by study respondents.

Results of this section indicated that the data at risk cover many different projects, disciplines, time periods and objectives. This is not an unexpected finding, especially when considering the multitude of research disciplines linked to respondents (see Table 5.2) as well as the fact that the institute had been involved with research for more than 70 years (see Section 1.2).

The implications of the findings are briefly stated in the next section.

5.6.3.3 Implications

The major implications of the web survey findings into data at risk descriptions relate to such data being inextricably linked to the institute, as such data feature in a range of research disciplines, research groups, time periods and formats. In addition, the concerns mentioned by respondents reveal that the issue of data at risk is manifold, with data loss, inability to trace data, and inefficient external client relations examples of the practical effect of data at risk.

In this study, and following the web-based questionnaire findings, the established interwoven nature of the institute's data at risk has resulted in the following steps:

- Data at risk is an aspect to be investigated further and features in the one-on-one interviews forming part of the next empirical data collection method.
- Data at risk is a topic addressed in the Recommendations chapter; Section 6.4.1 (discussion around awareness of data at risk), Section 6.4.2 (discussion around the correct handling of data at risk), and Section 6.4.3 (discussion around data rescue awareness) are three of the sections addressing the fact that data at risk are a cause of concern at the institute, and that reparatory steps are required.

The next section deals with the location of data at risk, as revealed via the web-based questionnaire.

5.3.7 Location of data at risk

Question 4 of the online questionnaire required respondents to indicate the current location of the group's data at risk.

5.3.7.1 Findings

The following locations were mentioned as locations for the storage of the group's data at risk:

- archives,
- data centre of the institute; institute's IT server,
- 'electronic platforms'; 'electronic storage devices',
- GroupWise (the institute's previous document management system; it was replaced in 2019),
- hard drives in offices,
- shared official computer infrastructure referred to as the H-drive, the I-drive, or the K-drive,
- laboratory (for physical samples),
- offices still in use, locked offices, empty offices, photocopy rooms, storerooms, desk drawers, a locked safe, locked cupboard, and filing cabinets in basements (all these locations were listed for paper data),
- personal computers,

- personal hard drives,
- personal storage devices,
- research group's database,
- secure location that cannot be mentioned (top secret and secret data), and
- transferred to the Digitisation Unit of the University of [*location*] for data rescue; rescued data uploaded to a discipline repository.

The statements below refer to data location, and indicate the wide range of locations used, as well as the general lack of standardised storage procedures.

- 'Some of it is captured on [*institute's*] systems as reports but there are a lot of actual data that are still on personal computers, or electronic storage devices. The data for the [*machine*]: stored in filing cabinets in the basement of building [*number*]. [*Machine*] data: on researchers' computers and personal storage devices.'
- 'Majority of the data is on the I-drive with other data stored in bound files in empty offices scattered around my building. A lot of the knowledge of historic data lies with former employees outside the [*institute*] who are mainly retired personnel. Whenever I require information on such data for the [*research area*] research group for example, I have to contact the former RGL or former employees who may know where to look for it. This is especially applicable to paper data.'
- '… there are storerooms full of unindexed stored materials, and institutional memory loss.'

The next section contains a discussion of the findings regarding the locations of data at risk.

5.3.7.2 Discussion

The significant yet expected finding with regard to the location of data at risk revealed the multiplicity of data at risk locations. These locations were not only multitudinous in number, but also displayed variety in their range. Physical samples were situated in a laboratory, early digital data on old computers or on 'stiffies and floppies', and modern digital data on laptops, on servers, on portable devices, and in cloud-based storage. Paper-based data showed an even bigger diversity of locations, with respondents mentioning archives, boxes, desks, cabinets, laboratories, and sundry storage areas within the group's workspace.

The literature review of this study did not investigate the location of other studies' data at risk. This was not a factor explicitly investigated in many of the studies, even though locations were mentioned by several authors. Storage locations include a researcher's desk (Wyborn *et al*., 2015: 106), an archive without a data reader (Mayernik, 2017: 4), library holdings and databases (Brönniman, 2018), private computer (Brönniman, 2018), various archives and museums (Brönniman, 2018), and 'stored haphazardly' (Curry, 2011: 694). In addition, other studies mentioned data on degenerating tape (Brönniman, 2018), photographic plates, in books, or on magnetic tapes with neither format information nor metadata available (Griffin, 2015).

Guidelines and project reports by several international data rescue organisations include references to the random, diverse and non-secure location of data at risk. Examples of such references are listed below.

- A report by the WMO stated that older paper-based data were often stored in basements, sheds, and other 'undesirable' facilities (2018: 45).
- A WMO report on a rescue project in Uzbekistan stated that locked away papers, not available to researchers, were stored in an archive (2017).
- An I-DARE guide to data rescue best practices warned that data are often located in unknown locations, or in unexpected places (Wilkinson *et al.*, 2019: 4, 8).
- A report on the data rescue activities of the EUMETNET Portal stated that there are often uncertainties regarding the storage location of older data, and uncertainties regarding its actual existence (Auer, Chimani & the EUMETNET Data Rescue Expert Team, 2014).
- ACRE (2019) maintains that historic data are often lost, misplaced, stored in deteriorating archives, and even held by private citizens.

The implications resulting from the discussion above form part of the next section.

5.3.7.3 Implications

The multiplicity of locations of data at risk goes hand in hand with the diversity of data at risk formats and disciplines. Possible reasons for the selection and use of data at risk locations are provided below; the listing of potential reasons is followed by a short description of implications emanating from these reasons.

Ease, convenience and availability

- Researchers are bound to find it convenient to have their data on their laptop, together with other project outputs such as reports, project files, correspondence and notes.
- Placing paper-based data in a box and storing it out of sight, in a cupboard in the office, is another example of ease and convenience of storage.
- Using a small, portable, inexpensive and available USB device is an easier option than purchasing more secure and costly storage. Such purchases could often take several weeks to arrive if institutional procurement steps are to be followed.

Cost

- Using a readily available storage box lying about for paper-based data stashing saves on project expenses.
- Similar cost savings are incurred when making use of storage locations or devices already at hand. Project costs are lowered when using portable USBs, available laptop storage, network storage, or leaving the data on the machine used for collecting data.
- In addition to the above, costs are also saved initially when not making backups of huge datasets that cannot be accommodated by conventional digital storage devices.

Research culture

- The common practices of the group are thought to have an influence on selected storage locations. A research group or RGL who had previously not considered secure data locations is unlikely to change these practices unless compliance to best practices is enforced, or awareness of data at risk is improved.

Lack of institutional guidance on data storage and data location

- Institutional guidance and policy around secure data storage are bound to affect data location usage compliance. The locations and devices used for data storage when complying with suggested practices are bound to be more secure in nature.

Implications of these practices, especially when eyeing the way forward for data rescue and data management, are stated below.

- Several respondents have admitted to making use of non-secure data storage options and locations.
- Establishing the reasons for making use of non-secure options is cause for future investigation and addressing.
- It is unclear whether researchers are always aware that certain storage locations are not secure, and whether they are aware of more secure options.
- It is anticipated that the recent approval of the institute's data management procedure will increase the use of more secure data storage options, as the use of institutionally approved systems is mandated.
- The current data storage location selections underline the need for data management policy, procedure and guidelines.
- The current data storage location selections also underline the need for increased data management awareness and training within the institute.

- Data management training and awareness sessions should include aspects around data storage. Realisation of the importance of secure storage media, and its role in minimising the eventual need for data rescue, should eventually form part of the institute's research culture.
- It is likely that a research group with a history of non-adherence to best storage practices would have valuable data that the group is not aware of, or data at risk without relevant parties being aware of the fact. These research groups would benefit from dedicated searches of cabinets, laboratories, archives and other storage spaces to discover and locate such data.

The next section of the chapter deals with the findings regarding the data rescue experience of respondents.

5.3.8 Data rescue experience

Question 5 of the online questionnaire required respondents to indicate whether the research group had ever performed data rescue activities.

5.3.8.1 Findings

Responses regarding the prevalence of data rescue performed within the respective research groups are indicated in the figure below.

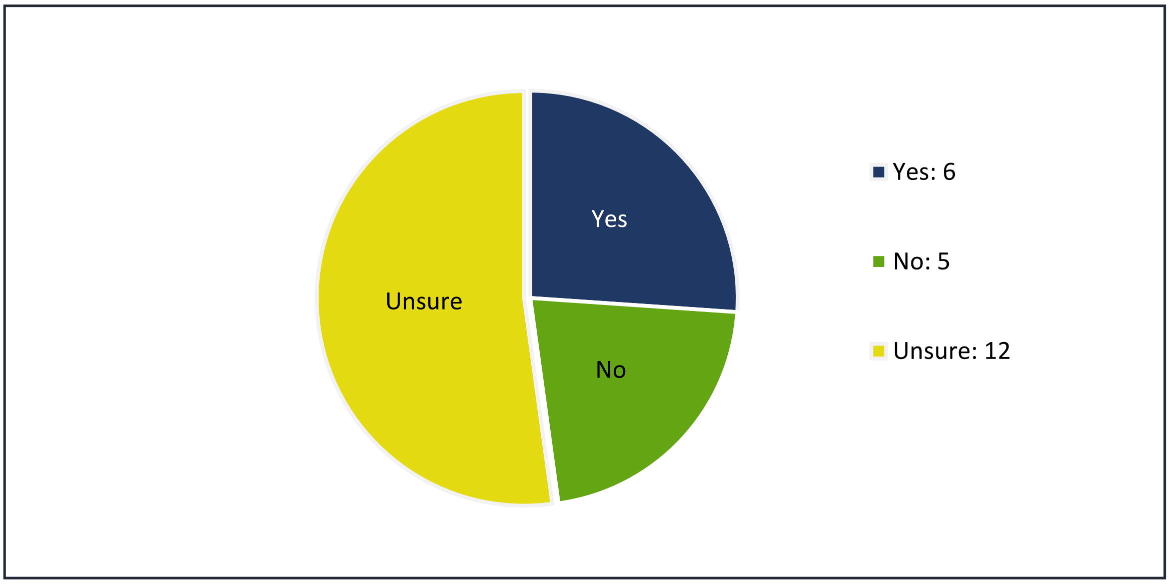


Figure 5.3: Data rescue experience

The question regarding data rescue activities performed by research groups produced the admissions listed below.

- Six groups had performed some data rescue activities in the past.
- Five groups had never performed any data rescue activities.

- Twelve RGLs (or their proxy) indicated that they were unsure whether the group had performed such activities in the past.

The next two sections comprise a discussion of findings regarding the data rescue experience of respondents, and a brief description of resultant implications.

5.3.8.2 Discussion

Significant findings related to survey responses are listed below.

- Most research groups had never performed any data rescue activities, with only a quarter of respondents stating that some data rescue activities had taken place.
- The number of respondents indicating that they were unsure whether data rescue had been done is troubling; however, it is not clear from the responses whether some of the RGLs were newly appointed managers (and not familiar with the group's activity history), or whether data management activities were not shared with all members of the group, and with managers.

These results were unexpected. Even though this researcher was not surprised by the prevalence of data at risk in the institute (see Section 5.3.4: Prevalence of data at risk), the overwhelming absence of data rescue projects and activities was not an anticipated outcome of the web survey. The expectation was that many of the research groups would have somehow attempted or organised data rescue initiatives, ranging between rudimentary and structured in nature. The unanticipated nature of this finding is further exacerbated by the fact that most of the subsequently interviewed parties indicated that their data at risk was of value and needed to be shared.

5.3.8.3 Implications

The reasons for the largely absent data rescue scenario within research groups was not clear from the web survey data, as the reasons or challenges were not requested from survey respondents. Subsequent one-on-one interviews provided an opportunity to delve into data rescue efforts and gain a better grasp of problems faced by research groups. These data rescue challenges are discussed in detail in Section 5.4.10: Data rescue obstacles and concerns.

As the one-on-one interviews involved a different sample than the group involved with the web survey, different results with regard to data rescue activities were anticipated. The sample invited to be interviewed was made up of only RGLs who had indicated on the web survey to have data at risk, or to have performed data rescue, and the interview responses with regard to data rescue were expected to provide a different picture than the web survey findings.

The next section deals with the results and findings regarding data documentation based on the online questionnaire responses.

5.3.9 Data rescue documentation

Question 6 of the online questionnaire required respondents to indicate whether the research group had documentation in its possession related to the group's data rescue activities.

5.3.9.1 Findings

Responses to the question about data rescue documentation revealed the findings listed below.

- Two RGLs stated that they had some form of data rescue documentation in their group.
- Seventeen RGLs remarked that they had no data rescue documentation.
- Three RGLs were unsure about the existence of data rescue documentation within the group.

The existence of data rescue documentation, as revealed via the web-based questionnaire, is portrayed in the figure below.

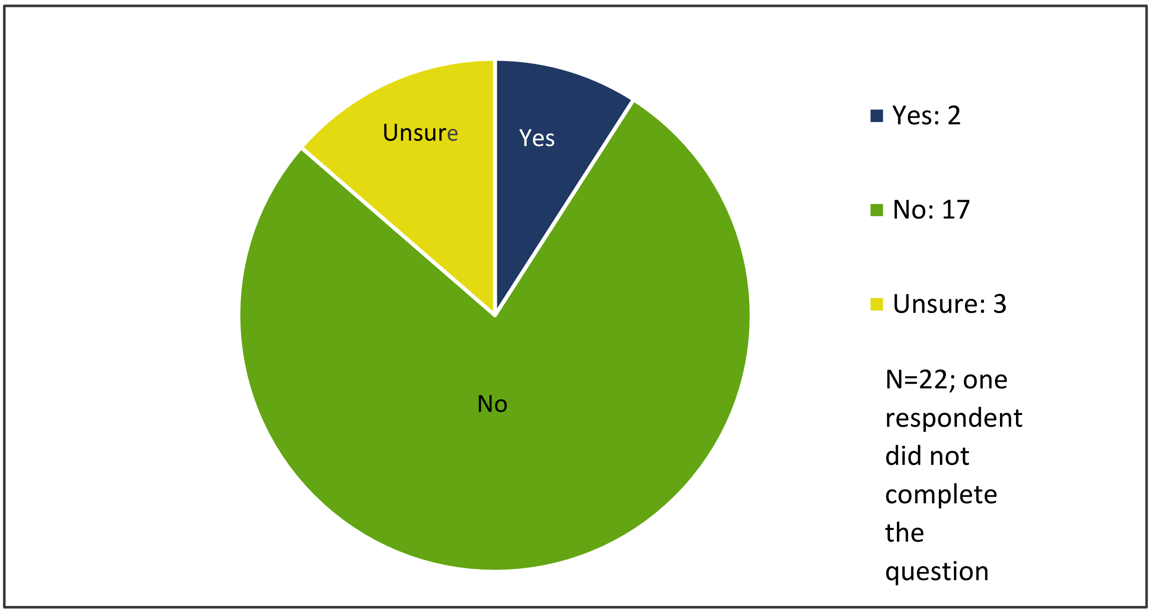


Figure 5.4: Data rescue documentation

One of the respondents posted the following elaborative details regarding their group's data rescue documentation:

> 'I have known of data being lost in the past, especially during [*research area*] testing both locally and internationally. As such, we do not have a formal stand-alone document but more of a best practice principle guideline that all personnel are trained

> with when they either measure or store data. Basically, we teach them how to label data files accurately and we also have a general rule of thumb that suggests we store all active data in three separate locations at any given time during a test; e.g., one onsite using [*institute*] campus hardware, one infield and one on the cloud.'

5.3.9.2 Discussion

Prior to the analysis of findings, a preponderance of documentation on data rescue had been expected. It was anticipated that several groups would have drafted standard operating procedures, guidelines, or a documented record of rescue activities performed/to be performed. However, data rescue was found to be a rare activity within the institute, with the resultant lack of documentation part of this situation.

In addition, the data rescue documentation indicated as being shareable for this research was in fact not documentation specifically pertaining to the rescue of data. The documentation contained storage and backup guidelines, and details about the group's outsourcing of data rescue to the University of [*location*].

Lack of data rescue documentation points to the need for a data rescue workflow model: even though most research groups were found to have data at risk, not one of the responding groups made use of a data rescue model, had data rescue guidelines, or had created a document containing data rescue best practices.

With the absence of such documentation going hand in hand with a lack of data rescue activities, it was not surprising that no respondents admitted having documentation for either planned rescue activities, or for urgent rescue activities required for continuation of a research project.

5.3.9.3 Implications

The implications of a lack of data rescue documentation are listed below.

- Results of the web-based questionnaire failed to show proof of the existence of data rescue documentation within the selected institute.
- Data rescue is not regarded as an activity potentially forming part of research projects, and the necessity of having emergency data rescue guidelines to ensure project continuation was not anticipated by any respondents.
- The lack of data rescue documentation, coupled with the prevalence of data at risk, indicate the need for such documentation to be created and distributed.

- In the absence of data rescue documentation, researchers resorting to less-than-ideal rescue practices is an unfortunate possibility.

Steps to be taken to address the lack of data rescue documentation are put forward in Chapter 6: Recommendations. Section 6.4.1 (discussion around data at risk awareness) and Section 6.4.2 (discussion around the correct handling of data at risk) mention the vital role of data documentation and its necessity in ensuring that current and future data do not end up being at risk, and that the data documentation accompanying historic data be viewed as valuable and as such stored with the data.

5.3.10 Shareability of data rescue documentation

Question 7 of the online questionnaire required respondents to indicate whether their data rescue documentation could be shared with this researcher.

Two respondents indicated that their group had data at risk, but only one of those RGLs stated that the documentation could be shared with this researcher.

5.3.11 Suggestions of follow-up names, groups, areas or clusters

The final question of the online questionnaire required respondents to suggest names of research clusters, research impacts areas, research groups or researchers who might have performed data rescue activities in the past. Suggested entities would need to be based at this research institute. Supplied details, tantamount to snowball sampling[37], enabled the researcher to approach additional subjects and involve them in this study.

5.3.11.1 Findings

The entities indicated below were suggested by questionnaire respondents to be potential sources of data rescue information.

- **Senior or principal researcher** within the same research group: three respondents provided the name of another researcher.
- **Other impact areas:**
  - One of the respondents named two other impact areas, as well as the names of relevant data-involved researchers within those areas.
  - One of the respondents named an impact area to be a possible source of data rescue information.

[37] https://research-methodology.net/sampling-in-primary-data-collection/snowball-sampling/: this sampling method involves primary data sources nominating other potential primary data sources to be used in the research. It is a sampling method often used when research subjects are rare or difficult to find.

- **Vague suggestion:** one of the respondents indicated that there might have been a data rescue effort during 2016.
- **No suggestion:** five respondents declared that they were not aware of any data rescue projects or entities, with an additional respondent using the term 'unsure', and another the term 'n/a'. In addition, eight respondents did not answer the question.

Email invitations, similar to the study invitation letter sent to RGLs (see Appendix 1), were sent to suggested contact persons. None of the contacted parties accepted the invitation to participate in the study.

5.3.12 Additional comments and concerns

A final feature of the web-based questionnaire comprised a text box allowing respondents to add any suggestions or comments related to data at risk, or data rescue. Many of these comments are not linked to any of the previous questionnaire headings and are accordingly mentioned below. These comments provided more details regarding the institute's data at risk issues, and data rescue challenges and practices encountered. All relevant remarks and information will be used when revising the initial Data Rescue Workflow Model.

5.3.12.1 Findings

Findings related to data storage and backups, and data at risk are briefly discussed in this section.

Data storage and backups

- Five respondents commented that all their data had backups:
  - One of the respondents stated that their research group uses [*institute*] databases plus research group databases for double safety.
  - One of the respondents stated that most of the group's data are in electronic format with backups.
  - One of the respondents stated that modern databases with double systems are used for current data storage.
  - One of the respondents stated that their electronic data are stored on [*institute*] drives, with backups on personal hard drives.
  - One of the respondents stated that their research group had used several servers (the group server as well as institute servers) in parallel with their group's configuration system.
- The following responses were in contrast with the above bullets regarding backups:
  - An RGL commented that some of the group's data 'sit' with individual researchers on external hard drives, and that the data had 'little backups'.

  - An RGL commented that the group's contract stated that data with a 'secret' rating were required to have a single copy only and had to be stored in a secret location.
  - An RGL commented that their historic data were stored in an archive and were no longer regarded as important to the research group.
- One of the RGLs mentioned that their research group had documentation on how to store data, including the data storage period: 'We have documentation and procedures how to store data, and for how long.'

Data at risk

- One of the respondents stated that their data stored on external hard drives are most at risk.
- A comment read: 'Last year we had a big problem tracking old documents on the GroupWise DMS[38] system. Numbers not correlating to the correct documents, documents disappearing, access changes, etc.'
- It was mentioned that handover is done via Workflow[39] but that the quality is not great.
- It was stated that the group practiced inconsistent data handling and storage of historical data on [*research area*]-related research projects that have been undertaken by various members of the team over the years.
- It was mentioned that some data have been lost due to restructuring of groups and individuals.
- It was mentioned that continuance on projects from members who had left the [*institute*] is problematic.
- A direct comment read as follows: 'Old paper trails are no longer traceable.'
- One of the respondents stated: 'Data storage on digital formats is incomplete.'
- One of the respondents mentioned the risk of not being able to find relevant older data and records after being approached by an interested external party. Failure in locating the required records had resulted in the collaborative project not being initiated.
- A comment about knowledge held by former employees read as follows: 'A lot of the knowledge of historic data lies with former employees outside the [*institute*] who are mainly retired personnel. Whenever I require information on such data for the [*research group*], I have to contact the former RGL or former employees who may know where to look for it. This is especially applicable to paper data.'
- The issue of big data being at risk was mentioned via the following comment: 'During [*research area*] tests, numerous instruments are installed, all of which record huge quantities of data over periods of up to nine months per test. The [*research group*] historical archive is therefore

[38] DMS: The selected institute's data management system used up to 2019.

[39] The selected institute's electronic request system (involving the library, finances, procurement, human resources).

> a huge database dating back more than 30 years, but it is not stored in any database currently. As of 2020, I have a big data project which is looking at hopefully developing a database for such data, but it is a huge task.'

A discussion of these findings is presented in the next section.

5.3.12.2 Discussion

The responses linked to the web-based questionnaire's last question can be divided into two broad categories, namely comments linked to data storage and backups, and comments linked to data at risk. While it was evident that many of the respondents ensure that data are backed up, a small number of RGLs supplied brief details about their group's lack of backups, and even subpar data storage methods. The responses pertaining to data at risk revealed many concerns regarding the security of the data generated by the group, with issues such as troublesome institutional systems, non-ideal data storage devices used by group members, inconsistency in a group's data management practices, institutional restructuring and group disbanding, and retirements of research experts cited as factors contributing to data loss. The range and nature of comments posted also indicated the need for the next data collection phase, comprising in-depth one-on-one interviews with selected RGLs regarding their data at risk, data rescue practices, data rescue requirements, and data rescue challenges.

The next section lists several implications resulting from the findings presented in this section.

5.3.12.3 Implications

The findings related to the questionnaire's final question reveal a culture of subpar data management activities. Issues such as less-than-ideal storage locations, inferior backup practices, incorrect handling of big data, inconsistency in data-related practices, poor handover, knowledge departure from the institute, and unsatisfactory restructuring underline the culture of non-adherence to best practices regarding data management.

Factors mentioned above have resulted in an increase in data at risk, inability to locate data, and potential negative effects on client and funder relations/contractual opportunities. While several respondents have indicated that data backups are ensured, this being but only one aspect of data management accentuates the urgent need for increased awareness regarding data at risk, data management, and data rescue practices. The findings linked to this last question of the web-based questionnaire also indicate the importance of the next data collection stage, entailing in-depth discussions regarding data at risk and data rescue practices with selected RGLs.

The various recommendations forming part of Chapter 6 are put forward to address the worrying implications of the institutional status quo regarding data at risk.

5.3.13 Summary of questionnaire findings

Results of the web-based questionnaire revealed that data at risk was a common feature at the investigated research institute, with more than half of the respondents' research groups having data at risk. Such data cover a wide range of projects, disciplines, formats and time periods.

Results also showed that data rescue is not commonly performed at the institute, with only six of the 23 respondents having performed such activities.

While the goal of the web-based questionnaire was not to establish the reasons for data being at risk, or to investigate data rescue challenges experienced, the brief responses submitted via the questionnaire revealed data rescue concerns. Concerns about poor handover, inferior data management practices, unsatisfactory infrastructure and other institutional challenges were detected via the responses. It was anticipated that the briefly mentioned areas of concerns would be elaborated on during the next data collection stage, and the extensive list of challenges emanating from the next stage proved this anticipation to be correct.

A total of 18 of the 23 RGLs (or their proxy) indicated having either data at risk or having performed data rescue activities. These RGLs would form the next sample of the study and were invited to participate in the next data collection phase – the one-on-one interviews.

5.4 Results: One-on-one interviews

This segment of the chapter contains the results of one-on-one interviews with selected RGLs of the institute.

5.4.1 Introduction

Following the results of the online questionnaire, selected RGLs were invited to form part of the study's next data collection phase, which took the form of virtual one-on-one interviews. RGLs who had indicated 'yes' to one of the following online questionnaire questions were invited to take part in the interview stage of the study:

1. Does your research group have data at risk?
2. Has your research group ever performed data rescue activities?

Eighteen RGLs met the requirement as described above by stating that their group had data at risk. Six of the 18 RGLs also indicated that their group had previously performed data rescue activities. These 18 RGLs were invited via email to participate in the interview phase of the study.

Eight RGLs responded in the affirmative to the interview invitation. All eight interviewees had data at risk in their research group, while only three of the eight RGLs had previously been involved with data rescue activities. The table below contains a summary of the main web-based questionnaire responses supplied by the eight RGLs who had accepted the interview invitation.

Table 5.3: Summary of online questionnaire responses of interviewed RGLs

INTERVIEW RESPONDENT	DATA AT RISK	RESCUE EXPERIENCE	DATA RESCUE DOCUMENTATION	WILL SHARE RESCUE DOCUMENTATION
Respondent 1 (RGL1)	Yes	Yes	No	n/a
Respondent 2 (RGL2)	Yes	No	No	n/a
Respondent 3 (RGL3)	Yes	No	No	n/a
Respondent 4 (RGL4)	Yes	Yes	No	n/a
Respondent 5 (RGL5)	Yes	No	No	n/a
Respondent 6 (RGL6)	Yes	Yes	Yes	Yes
Respondent 7 (RGL7)	Yes	No	No	n/a
Respondent 8 (RGL8)	Yes	No	No	n/a

The invitation email, in addition to reiterating the objective and sub-objectives of the study, also described the key features of the interview. Interviews were to be comprehensive in nature, and it was anticipated that the interview findings would provide in-depth information regarding the institute's data at risk and data rescue activities performed. The email also included the details listed below.

- The interview would be a one-on-one format and involve this researcher and the selected RGL.
- Due to lockdown restrictions, interviews would be conducted virtually only.
- The letter stated that Skype was the preferred virtual interview platform.
- The interview would be semi-structured in nature.
- The interview length was expected to be a maximum of 60 minutes.
- The interview would preferably need to take place during September 2020.
- Interviewees would be able to indicate the interview starting time.
- The virtual interview would be recorded.
- Web camera use was not mandatory, as only audio data were vital to the study.

Interviews were conducted during September 2020, with all RGLs agreeing to use the Skype platform as the virtual interview tool. Interviews lasted between 20 and 45 minutes. Interviews were recorded by the Skype recording tool, with the voice recorder on a mobile phone also being used as backup. Following the interview, the Skype recording was saved to a cloud-based platform.

Each interview was transcribed, and the transcription emailed to the applicable interviewed RGL. Interviewees were requested to study the transcription and point out errors, and to supply the researcher with clarifying interview-related information, should it be required. None of the interviewed RGLs indicated any transcriptions errors, and no additional interview-related information was provided.

All participants indicated approval of the transcriptions. Recordings were destroyed after approval of the transcriptions.

5.4.2 Interview schedule

Interviews were semi-structured in nature, which means that a few predetermined questions were asked, while the rest of the questions were not planned. This resulted in different RGLs being asked different questions, topic-wise as well as in the number of questions posed. Questions asked and prompts made depended on responses to previous questions; an RGL who had performed data rescue in the past, for example, would be asked different questions than a respondent who had never attempted data rescue. The outline of the semi-structured interview schedule can be viewed in Appendix 7.

5.4.3 Rationale of data collection tool

When considering the study's research objectives, the results gained via the one-on-one interviews were expected to contribute towards the following:

- establish how data rescue is currently done within the research institute, or how it would be done hypothetically (as not all interviewees had data rescue experience),
- should data rescue not have been done, respondents would be requested to describe the hypothetical or envisaged data rescue steps,
- determine who is currently involved in data rescue, and their roles,
- establish current data rescue perceptions, needs and challenges within the institute,
- establish anticipated data rescue needs and challenges (should respondents not have rescued data before),
- gain a better idea about how library and information services professionals can be involved in data rescue, and
- provide valuable information to be used when amending the initial Data Rescue Workflow Model.

5.4.4 Data analysis, and themes emanating from interviews

Themes identified after analysis of the data were as follows:

- general description of the research group's data at risk, such as data formats, data scope and data location,
- value of the data at risk, including usefulness of the data, and possible interested parties,

- data sharing and data publishing,
- factors causing data to be at risk,
- data management activities in the group,
- data rescue activities performed, including data loss experience,
- obstacles and challenges in the path of data rescue, and
- any other aspects emanating from the interviews, deemed to be relevant to data at risk and data rescue.

Each of these themes are afforded a separate heading, and the applicable responses of each of the interviewed RGLs are included with each theme.

Responses have been anonymised, and feedback containing direct identifiers, such as respondent name, subject area, and project/contract details, have been removed. Names of RGLs are replaced with a numbered pseudonym, e.g., RGL1. The pseudonym assigned remains constant throughout this chapter. These pseudonyms also mirror the pseudonyms used previously in Table 5.3.

The layout of the interview results section of the chapter follows the sequence of identified themes listed above.

5.4.5 General description of data at risk

As was expected, and mirroring the findings of the web-based questionnaire, responses revealed that data at risk residing with research groups are diverse in nature, with different formats, scope and location being mentioned by respondents.

The table below summarises data at risk formats, location and scope (volume) for each of the interviewed respondents:

Table 5.4: Summary of data at risk

RESPONDENT	DATA AT RISK FORMATS	DATA SCOPE AND LOCATION
RGL1	• Different formats in group • Paper data from pre-2000 most at risk • Early digital data	• Seven steel cabinets of paper data • Early digital data on old computers • It is difficult to find specific data
RGL2	• Physical samples	• Physical samples are housed in a storage tank
RGL3	• Floppy disks from the 1970s • Hard copy files • Various electronic formats	• Hard copy data fills two offices, data are kept in boxes • Electronic data: experiments run for 24hrs, 7 days a week, 3–9 months; huge quantities of data • Electronic data are on I-drive
RGL4	• Modern digital formats (some at risk) • Hard copy data • Hard copy laboratory notebooks • 5% of data generated via software; expired software license	• Massive data amounts • Several terabytes of data
RGL5	• Paper data (mid 1980s to 2000) • Early digital data • Modern digital data	• 15–20 cabinets in storage area in the lab • Some electronic data storage is not secure (CDs, stiffies, USB devices) • It is difficult to find specific data
RGL6	• Old format audio and video data • Paper data • Data date from 1990s	• Not applicable • Data have been rescued and deposited in discipline repository (see Section 5.4.9)
RGL7	• Electronic data • Paper data are NOT at risk	• Top secret paper data are couriered to client
RGL8	• Electronic data only • Data are images and video • Mostly 'common' format e.g., MS Office • Some proprietary formats	• 100–500 MB of data • External hard drives, computers, servers

As shown in the table above, all interviewed respondents indicated that their research group had data at risk. Data at risk showed variance within groups; the same was found for data at risk between research groups.

The findings, a discussion of findings, and resultant implications are discussed in the next sections.

5.4.5.2 Findings

Findings emanating from the interviews, regarding data at risk, are listed below.

- Most research groups had more than one format of data at risk within their own group.
- One of the groups only had one data format at risk.
- Paper-based media and modern digital media were the most common formats identified to be at risk.
- Physical samples being at risk was not a format commonly stated to be at risk, although it was stated by one RGL.

- Data at risk also varied between historic in nature (i.e., from the 1970s) to current data stored on non-secure devices, or not backed up.
- Various data volumes and scope were revealed: responses included many files of paper-based media and several terabytes of modern electronic data.
- Data locations showed variance as well: desks, laboratories, offices and storage rooms were mentioned for storage of paper data.
- Data in modern electronic formats were stated to be stored on laptops, hard drives, archives and servers, while older digital media were housed on 'stiffies and floppies', on magnetic tape, and on computers no longer in use.

A discussion of findings is presented in the next section, and is followed by a section touching on implications emanating from the findings regarding data at risk.

5.4.5.3 Discussion

Even though the age of data was not a survey question, nor a topic investigated via the subsequent one-on-one interviews, the year of approximate data collection featured prominently in the questionnaire responses and during the one-on-one interviews. The time periods mentioned varied between the 1970s and 1980s, to data collected before the year 2000 (see Table 5.4), to very recently collected data. In short: data at risk were found to be historic in nature, as well as new and current. Data at risk, based on the findings of this study, can therefore not be described as a time-specific phenomenon.

These findings are not unexpected, as previous investigations regarding this institute's data management practices pertaining to current data had revealed low adherence to best storage and metadata practices (Patterton, 2014; Patterton, 2016; Patterton, Bothma & Van Deventer, 2018). Making backups of data, however, was found to be a regular practice, but this alone does not ensure that data are not at risk of loss or damage.

The same can be said for older data, and the fact that such data were also found to be at risk. As the institute does not have a history of established and implemented healthy data management practices, it can be deduced that older data would also not have been treated in a way conducive to good data management. Older and fragile paper-based data would probably not be treated with gloves, and not filed in acid-free boxes; magnetic tape would not have been stored in an ideal environment; digitisation of older media would not have been a regular activity. In addition, multiple organisational restructuring events, closing of projects, and the data-related effects of resignations and retirements would have all increased the likelihood of this older data being at risk.

The findings agree with the time periods briefly investigated when conducting a review of literature for this study. While most studies described the rescue of historic data, efforts to deal with the conservation of recent data were also forthcoming. The literature review revealed that many projects rescued data older than the 'older data' mentioned by this study's respondents: rescue projects involving sea-level measurements from the 19th century (Caldwell, 2003), weather data from the first half of the 20th century (Kaspar *et al*., 2015), fish sounds data from 1940–1980 (Rountree, 2002), and solar radiation data dating from 1966–1996 (Antũna, 2008) are examples of such ventures. With regard to the rescue of modern electronic data: in order to include the rescue of modern electronic data, the RDA Data Rescue Interest Group decided during 2019 to broaden its scope and undergo a name change (Hills, 2019). The focus change (the rescue of not only paper-based and early digital data) and name change to the RDA Data Conservation Interest Group are testament that both historic data and modern data can be at risk.

5.4.3.3 Implications

The implications of these findings overlap with many of the implications mentioned earlier in the chapter (Section 5.3.4: Prevalence of data at risk):

- All data, irrespective of its age, may potentially end up being at risk. Data at risk does not involve historical data, or data in a non-modern format only.
- It makes sense that the lapse of time, undoubtedly resulting in institutional changes, can add to the odds of data being at risk. Research groups disbanding, institutional restructuring, and researchers (involved with the data) resigning or retiring increase the chances of data being lost, misplaced, damaged, or no longer understood. In addition, this institute also had several research campuses across the country in earlier years, and many of these remote sites are no longer part of the organisation. It is currently unclear where the data generated at remote sites are housed, what the condition of the data is, and whether the institute still has the research expertise to understand the context of such data.
- Older data also go hand in hand with data degradation and fragility, equipment replacement, and the need for format migration.
- Despite the deluge of factors increasing the odds of older data being at risk, modern data present its own set of problems as well.
- In many disciplines, modern data generation goes hand in hand with extreme volumes and huge velocity as its characteristics. With increased data volumes comes increased data loss risk, especially if such data cannot be backed up when generated. In disciplines such as climate science, a dataset often exceeds several terabytes in size, making backups and even secure and reliable storage a costly activity.

- Another aspect related to the voluminous attributes of modern data pertains to the need for file naming conventions, coupled with structured file and folder organisation. Absence of these practices is estimated to contribute to inability to locate or make sense of recently generated data.
- Newer data in newly created groups, or with newly appointed researchers and managers may be at risk of loss if all parties, especially those responsible for strategy and risk management, are not familiar with the discipline's data.
- Another aspect unique to current research and modern data is that much of the data created while connected to the network are automatically backed up daily by ICT. This can result in fake assurance that the data are safe from loss; backing up data is only one part of data loss minimisation and cannot ensure that data are not at risk of loss or damage.
- Despite this bleak picture painted for modern data and its risk of loss, it is important to realise that not all modern data end up damaged or lost, and in need of rescue. Some data can be regenerated with minimal effort and time, while other data are only needed once and not again. These are aspects that will feature when discussing the factors determining the need for data rescue, and the factors determining the prioritisation of the rescue of certain data (see Chapter 6: Recommendations).
- In a similar vein to the aspects mentioned in the above point, not all older data, despite its seeming value, will need to be rescued. Investigations prior to rescue, when assessing which data are in need of rescue, and devising a priority list, might reveal that the data are no longer of interest to others. The assessment activities could possibly also reveal that a duplicate of the data, digitised and accessible, already exists elsewhere.
- While data from all time periods can be at risk, the nature of risk factors, as described above, differs between old and modern data.
- Recommendations emanating from the date-specific risk factors, and how these should be addressed, will be put forward in the Recommendations chapter of this study (see Chapter 6).

Factors leading to data being at risk are discussed in a subsequent section of this chapter (see Section 5.4.7: Data at risk: Factors).

5.4.6 Data value and data sharing

This section of the results chapter contains information supplied by interviewed RGLs regarding the perceived value of their group's data. This section also describes each of the involved research groups' data sharing activities and pertinent aspects, such as the need for data anonymisation, or the use of public repositories.

Details regarding data value and data sharing often overlap, hence the headings in this section will be respondent-wise, as opposed to activity-wise. As such, respondents will be listed numerically (RGL1 through to RGL8), with the following data sharing-related aspects, where applicable, being discussed under the RGL headings:

- the usefulness of data over the long term,
- details about parties who have shown interest in the data, or who might benefit from using the data,
- actual and future data sharing,
- required preparatory data sharing steps, as well as data sharing obstacles or any relevant aspects, and
- data repository use.

Information about data value and data sharing is inextricably linked to making decisions about data rescue, as it reveals details about whether data rescue is worthwhile, and how rescued data should be treated, stored and shared. Table 5.5 contains a summary of the discussion that follows.

5.4.6.1 Data value and data sharing: RGL1

This section contains RGL1's statements regarding the research group's long-term value of data, the parties potentially interested in the data, and the group's data experience and practices.

Long-term value of data

RGL1 stated that their group had received queries regarding their data, more than 10 years after the data had been generated. The group's data were also connected to intellectual property (IP) issues, with the funder still maintaining an IP register more than a decade after the original project. The respondent also stated that the funder frequently tracks the older data and related documentation.

Interested parties

The group's data shareability was described as:

> '… a mix. There are complete public consumption type of data. My specific group, I think the highest classification we have is restricted, from a legal perspective.'

RGL1 mentioned that while their data were of 'immense value to researchers', much of the data housed by the group were not owned by the group. Data ownership is therefore the determining factor with regard to this group's data being shared with other parties. The issue was described as follows:

> ‘But the problem is data ownership. We are custodians of the data; we don’t own the data. So especially the ones that I say are ‘sensitive’. The data would be of immense value to researchers, especially people in [*research discipline*], people in [*three research disciplines*]. There will definitely be wide interest. Because our data, in this specific case, is extremely structured … well-described in terms of metadata, highly structured, so that is an extremely useful dataset. But we don’t own that data.’

Data sharing experience and practices

The respondent mentioned that the data were shared with the funders, and the funders only. RGL1 also remarked that the funder still frequently tracks the older data and related documentation. Data were not shared with other parties as the research group was the custodian of the data only, and not the owner.

RGL also commented on the sensitivity and confidentiality concerns related to their data:

> ‘If you look at data that is sensitive in nature, we still have a lot of that, and it is very current. So we have data that from a POPI perspective, a privacy perspective ... sensitive. As well as from a criminal justice viewpoint: sensitive. I suppose you could argue it could actually be, from a safety and security perspective, it could also be sensitive. Meaning that if someone knows you’ve got the data, they might actually go to illegal extent to try and get rid of it.’

The respondent also stated that there could be some instances where data have been desensitised, and that these data could be published as a DOI item. However, RGL1 did reiterate that the group’s data are generally too sensitive to be uploaded to a public repository:

> ‘I suppose there is [sic] some cases where the data has been desensitised and especially any individual, I suppose there are some datasets that we could publish as a DOI item but generally not.’

5.4.6.2 Data value and data sharing: RGL2

This section contains RGL2’s statements regarding the research group’s long-term value of data, the parties potentially interested in the data, and the group’s data experience and practices.

Long-term value of data

The respondent, while being uncertain about the exact parties who would benefit from using the data, remarked that their data would not be discarded:

> 'I don't think so, at this point. You never know; I certainly won't throw it away. It might come one day that someone would say, "Well you took the images, we'd like to apply a different analysis on the data", which is a common theme for a lot of the data. So we would never throw that away. And it does not cost us anything to store that. So it's possible, but it's unlikely.'

It should also be added that the data described in the paragraph above are not the group's data at risk, but data that are backed up and published. The RGL's data at risk are data in physical sample format. The long-term value of the data was not stated.

Interested parties

RGL2 did not mention the exact parties who would be interested in viewing the published data. The respondent also claimed that the data were not sensitive and could be shared.

Data sharing experience and practices

The respondent stated that the data had been published with an article and was freely available to external parties.

5.4.6.3 Data value and data sharing: RGL3

This section contains RGL3's statements regarding the research group's long-term value of data, the parties potentially interested in the data, and the group's data experience and practices.

Long-term value of data

RGL3 remarked that the data at risk have long-term value, with data requested often even after projects had ended. In addition, the respondent also stated the following:

> 'It's always good to keep the data and track back if you want to access the data.'

Interested parties

RGL3 stated that their group's data have definite value to other parties:

> '… people like [*organisation*], Department of [*government entity*] definitely … all the research projects that we do are actually funded by them. The previous RGL now actually works

> for the [government department]; he is the person who actually stored all the data. Every now and then when he wants something or is looking for something, he tells me where to find a file or that sort of thing …’
>
> ‘And another person as well is the University of [*location*]. The current dean there as well at the Department of [*discipline*].
>
> So every now and then I get a call from either the [*government department*] or [*university department*] or someone from [*organisation/funder*], just looking or asking a question and I have to go spend a couple of hours looking for stuff.’

Data sharing experience and practices

The RGL remarked that data are often shared with researchers in the same discipline. Researchers requesting data are from government departments, discipline entities and university departments. RGL3 also stated that data are never shared without permission from the client/funder, and that the data are not shared blindly with just anyone.

When asked about the suitability of data being uploaded to a public repository, the respondent mentioned the following:

> ‘That is what we are looking at at this moment. Something like that, where we can put it on a platform, and someone can actually access it from wherever they are. That’s the exact kind of thing we are trying to look at.’

5.4.6.4 Data value and data sharing: RGL4

This section contains RGL4’s statements regarding the research group’s long-term value of data, the parties potentially interested in the data, and the group’s data experience and practices.

Long-term value of data

According to RGL4, the group’s older data have value, and have been requested by prospective clients/ funders during discussions regarding new contracts.

Interested parties

RGL4 remarked that an external party might show interest in the data, but due to data being proprietary, data sharing would be prohibited.

Data sharing experience and practices

Data sharing was not done, as data had proprietary restrictions. In addition, RGL4 stated that locating older data, required when discussing future contracts with new funders, proved to be problematic.

> 'We've had some queries in the past year or three where they actually enquire about a project, and they would be interested in funding work in that field. But then you can't access older data for continuance. And that's the challenge. Not specifically for the raw data itself but in terms of … picking it up where we left off, basically.'

This RGL also mentioned that the group's data are often connected to trade secrets, and that intellectual property rights are involved. Data are therefore confidential, not shareable, and not suited to discipline repository deposit.

5.4.6.5 Data value and data sharing: RGL5

This section contains RGL5's statements regarding the research group's long-term value of data, the parties potentially interested in the data, and the group's data experience and practices.

Long-term value of data

RGL5 remarked that due to the nature of their discipline, data need to be kept for several decades:

> 'In [*branch of science*], especially when you are working with [*research discipline*] and so on, it is quite critical to ensure we keep all our data and information documented in a way that is easily accessible, now and in the future.
>
> And not just in five years, maybe even 20 years, especially for [*research discipline*] … for 20 even 30 years.'

In addition to the comment above, RGL5 also mentioned the following with regard to research data in general:

> '… data can still be used beyond completion of the project.'

Interested parties

The respondent remarked that their group had data that could be shared with the public, as it were data emanating from government-funded projects.

The group also had data from privately funded research or funded by [*discipline*] entities. This data, as well as the data that had intellectual property rights connected to it, are not openly available and can be shared with funders only.

Data sharing experience and practices

According to RGL5, the shareability of data in their group is not uniform, and depends on the project funders, and whether intellectual property rights are at stake. RGL5 stated the following:

> 'It depends on the project. Then a lot of government programmes, because it is government money, it is public money … we have to share it with everyone, with the public. Unless the government says no, we want to keep it secret. Like when you develop a patent and people won't steal that idea and then leave the country. So it depends. If it's generic data that can be used more widely, then yes. But if it's got IP linked to it: that you won't share, even if it's public funding. There's a balance.'

5.4.6.6 Data value and data sharing: RGL6

This section contains RGL6's statements regarding the research group's long-term value of data, the parties potentially interested in the data, and the group's data experience and practices.

Long-term value of data

When asked whether the data had long-term value, the respondent replied in the affirmative. RGL6 expanded on the issue as follows:

> '… that's very useful information if one can do a longitudinal study in other townships, or in that township again. And you can see how [*subject area*] changed over time … so yes, I do think that there is value to that.'

Interested parties

Data have been uploaded to a discipline repository and are therefore used by researchers and members of the public.

Data sharing experience and practices

RGL6's group had handed a portion of their humanities-related data at risk to the digitisation unit at the University of [*location*]. Digitised data were then deposited into a South African discipline repository focused only on resources related to this subject area. As such, data were accessible to others, while being curated by a professional entity experienced in managing [the discipline] data, and in managing a [*discipline*] repository. The unique DOI assigned to the data ensured that the data of RGL6's research group were citable and immutable.

RGL6 also revealed that the audio data were anonymous, as it had been obtained during focus group discussions.

5.4.6.7 Data value and data sharing: RGL7

This section contains RGL7's statements regarding the research group's long-term value of data, the parties potentially interested in the data, and the group's data experience and practices.

Long-term value of data

RGL7 described their different data retention periods and stated that some project data were kept for five years, while other project data were kept for 10–15 years, depending on the level of secrecy. Apart from the secret project data, normal research data were also generated, but such data were only retained for approximately two years.

The reason for not keeping data longer is related to this group's research discipline, and the objective of the data. The respondent stated that the immediate distribution and application of data was an important activity in the group, as the specific data lose relevance after a while. The respondent stated that there was 'no point' in keeping data for longer than required.

Interested parties

Although not having shared data previously, RGL7 stated that they had recently had a data request from a researcher stationed at the University of [*location*]. The respondent further stated that sharing of the data was not possible, as the data had not yet been anonymised. The group had been in talks with an entity enabling data storage and resultant visibility of the data to external parties, but the project did not materialise as the group was unable to anonymise the data. The participant mentioned that the group's research data were therefore currently only for researchers' personal projects.

When asked about other potential interested parties, RGL7 stated as follows:

> 'I know there's a lot of universities … I know [*name*] University in the past. And [*university*] in the past. They've also requested similar information, but we could not give them the data because we could not anonymise it to that level.
>
> So there are universities that are interested, as they have researchers that also want to investigate [applied discipline]. But it is sort of a trade-off: how do you anonymise it without

> losing the actual value of the data itself? You are literally changing how the data is structured. That is actually a problem that we are trying to address.'

Data sharing experience and practices

RGL7's group had not yet shared data with other parties. The group might share data in future, after its anonymisation had taken place. The issues with regard to the deposit of data in a public repository, and data anonymisation, were described as follows:

> 'Certain things can. But before we can, we have to anonymise the data. If it gets out that [*revealing statement with regard to failure of specific discipline*], it won't go down well with stakeholders. It is one of those touchy subjects, especially now with POPI and all of that. Until we have an anonymisation process in place, we probably can't.'

RGL7 also described their discussions with [*organisation*], a national data deposit provider located on the premises of the institute. Using this data deposit tool would enable the group to share their data, have a DOI assigned to it, and lead to ease of citation, and easier finding of the data. However, the tool proved not to be ready to accommodate the group's data; the RGL commented as follows:

> 'So they have the storage capacity; they have the network capacity, the infrastructure is there. They just need approval from the larger [*organisation*] consortium. I think they need consensus from all of them: this is the process if anyone wants access, say for instance on archives now, [*university*] wants access, this is the process they must follow, and this is the data they are allowed to have access to. Don't think those processes have been ironed out yet on their side.
>
> There are too many stakeholders, there's chaos with that. Other than that, it's a nice idea. But the practicality is not really there yet.'

5.4.6.8 Data value and data sharing: RGL8

This section contains RGL8's statements regarding the research group's long-term value of data, the parties potentially interested in the data, and the group's data experience and practices.

Long-term value of data

RGL8 remarked that some of their data at risk were still being used, while other data, specific to a certain project, were no longer of much use. However, the respondent also stated:

> '… with research one can always go back …'

thereby confirming that their data have long-term value.

Interested parties

The respondent listed researchers, government entities, and institutions making use of [*subject*] data as parties who would be interested in the data.

Data sharing experience and practices

Currently, data were only being shared with funders. RGL8 described the data, being biometric in nature, as sensitive. It would therefore need to be anonymised before being placed in a discipline repository, as sharing of raw biometric data could enable identity fraud.

The respondent also stated that data deposit into a repository would probably have financial implications for the group, and costs might be prohibitive. RGL8 further mentioned the possibility of data not being suitable for repository deposit due to size restrictions on datasets.

5.4.6.9 Summary

The table below provides a summary of aspects discussed within this section. All the RGLs remarked that their data have long-term value; parties interested in the data varied but were most commonly other researchers and funders. Not all research groups share data, and even fewer groups make use of a data repository for data sharing and curation.

Table 5.5: Data value and data sharing

RGL	ARE DATA VALUABLE?	INTERESTED PARTIES	ARE DATA SHARED?	DATA REPOSITORY USE
RGL1	Yes	• Funder/client • Researchers	• With funder only	• Data are not owned by research group • Group is data custodian only
RGL2	Yes	• Researchers	• Yes	• Data are published
RGL3	Yes	• Government departments • Discipline entities	• Yes	• Frequent sharing with others • Data repository is in pipeline
RGL4	Yes	• Researchers • Funder/client	• No	• Some data are proprietary • Anonymisation is required • Difficult to locate data
RGL5	Yes	• Funder • Discipline entities	• Some data are public • Some data are not shared	• Some data have IP implications to consider
RGL6	Yes	• Researchers • Public	• Yes	• Some anonymisation required • Encryption required • Data in discipline repository
RGL7	Yes/No	• University researchers	• Not yet	• Anonymisation required • Some data are top secret; not shared
RGL8	Yes	• Colleagues	• No	• Data can contractually not be shared • Biometric data are sensitive

The next two sections comprise a brief discussion on the findings regarding data value and data sharing, followed by a listing of implications emanating from the discussed findings.

5.4.6.10 Discussion

All interviewed RGLs indicated that their group had data of interest to other parties. A range of parties were mentioned, including colleagues, university researchers, other researchers, other research entities, funders/clients, government departments, and the public. Despite interested parties featuring prominently in responses, the sharing of data was not an activity practiced by all interviewed RGLs. Added to this was the infrequent use of data repositories, with factors such as non-institutional data ownership, contractual obligations, sensitivity of data, anonymisation requirements, encryption issues, and intellectual property concerns playing a part in the non-usage of data repositories.

The discrepancy between data being of value to other parties, and data not readily uploaded to a data repository is cause for concern.

5.4.6.11 Implications

The study's one-on-one interviews with eight RGLs have revealed that the institute's data are of value to other parties, but that limited steps are implemented to make such data available to others. With valuable data not shared, and not being curated by means of a dedicated data repository, it is evident that lack of best sharing practices is part of the institutional research culture. Awareness around the links between data value, data sharing, data at risk and data repositories need to be ensured, and systems and services put in place to enable such activities. With current data (i.e., data currently being generated) also not uploaded and shared via repositories, more than the institute's historical data are being put at risk.

Steps to address the data sharing status quo are presented via several recommendations in Chapter 6; the sections recommending increased awareness around data at risk (Section 6.4.1), awareness around data rescue (Section 6.4.3), the mitigation of data rescue challenges (Section 6.4.5), the promotion of the data rescue model (Section 6.4.6), and the implementation of a data rescue project at the institute (Section 6.4.14) are examples of suggested directions for the institute to follow in order to talk to the indicated value of institutional data.

The next section features the factors putting data at risk as identified via the study's one-on-one interviews with eight RGLs.

5.4.7 Data at risk: Factors

This section lists and describes factors contributing to data being at risk of loss or damage. It also lists risk factors contributing to difficulty in accessing data. Data risk factors discussed in this section are factors mentioned by interviewed RGLs, during their one-on-one interviews.

5.4.7.1 Findings

All aspects listed as risk factors are included in this section, even if listed by one RGL only, or only applicable to certain data types. A total of 34 factors were identified and these are addressed below.

1. Data location is unknown

Uncertainty about the location of data, or inability to find the data, was listed by several RGLs as a factor putting data at risk.

RGL 1 stated:

> '… and at the time we did not realise that an important part of the work he did was not logged into our software repository …'

while RGL3 mentioned the following:

> 'Right now, if [*funder/client*] or anyone has to ask me "I want this data that was done 20 years ago", I have no idea where I would find it … all the data is there, I just don't know if it's in

> the files, the hardcopy files, the I-drive or … I have no idea where I would find it.'

RGL3 further stated that the inability in tracing the data has proven to be problematic for the group.

RGL5 also touched on this aspect and revealed that locating data was a problem when first joining the institute, and that it is still a problem. The same respondent revealed that the recent passing of a colleague had accentuated this problem, as data searched for could not be found.

Regarding the difficulties encountered when searching for data after a colleague has left the institute, RGL5 mentioned that:

> 'You've got to go and dig through their computer. People do experiments and it's not recorded in a fashion where, as a manager … I cannot get access to it easily.'

2. Being unaware that data are at risk

RGL1 mentioned a situation where group members had not realised that an ex-colleague's data were not added to the group's repository, and thus at risk.

3. Data being in an old format

According to two RGLs, data can be at risk when the data are in an older format. RGL1 mentioned that some data backups were still on magnetic tape, and that the magnetic tape kept on failing. This resulted in the format being considered too impractical for backup use, even though it was 'proper Dell equipment'.

RGL6 had a similar point of view and mentioned that data were housed on older formats such as 'stiffies and floppies'.

4. Lack of metadata

A single respondent (RGL1) stated that data can be at risk when the data have no metadata, as was the case in the respondent's research group.

It should be stated, however, that not all research groups lacked metadata. RGL2, for example, remarked that metadata had been added to all their data.

5. Lack of backups

Four respondents (RGL1, RGL2, RGL3 and RGL7) indicated that lack of backups had resulted in their data being at risk.

RGL1 described how their new electronic data were not always backed up, due to the nature of field work, and not having a backup device at hand. The same respondent also stated that data were at risk when different machines in their group are not synchronised. The following two scenarios were described:

> 'If our software developer has not synchronised with our server [laptop gets stolen during the day], then you sit with the problem that the source code developed that day is not synchronised with the server.'
>
> 'If they should lose their laptop or their laptop crashes, they should have a sufficiently late enough copy of their work. Although some people use platforms such as Google Docs, or shared folders, cloud-based folders, but if there was not a last synchronised step, you might lose the last unsaved work.'

RGL3 declared it likely that their paper data have no backups, as it had not been converted to a digital format.

> 'That's how I understand it. I am not sure if they've converted it. All the old data, I am not sure that someone would have taken the time to convert it electronically. I can't imagine that. I would imagine there is only one set of paper data.'

RGL7 agreed with the above and mentioned that there have been several instances of data loss, due to the absence of a 'real backup process'.

6. Unable to save data to a secure location immediately

According the RGL3, the group's data are at risk as their current workflow did not allow for the immediate storing of data in a secure location. The set-up was described as follows:

> 'A lot of the time we actually have to collect the data on site, and then once a week someone brings all the data from the site to the [*institute*]. Either by bringing the entire computer, or by driving to the [*site*]. Sometimes I drive to the site and copy it myself. And what can happen: a computer can crash, or something can go wrong, either the guy on site forgets to save the data properly, or that sort of thing. A USB can get lost, for example. A lot of our data is quite susceptible to getting lost in

> the process and can have very devastating effects on our projects. If you can imagine: we are testing seven days a week, 24 hours. Just one week of data lost is a lot of data.'

7. Data volume/rate of data generation

RGL7 mentioned how data volume, and the rate of data creation, could be a data risk factor. The RGL stated that the group's data storage options were limited with regard to space, and that they were only able to keep three months' work (i.e., one set of data).

8. Failure to read the data

RGL1 mentioned that the non-availability of data readers constitutes a data risk factor. Included with this risk factor are aspects such as many newer laptops no longer equipped with a DVD reader, or small hard drives no longer having working data readers. The respondent elaborated as follows:

> 'And then I switched to small hard drives, 500 gig. And went through a period where this was used for backups. And then later PCs stopped reading those disks. It's pretty much like that problem most people sit with: you've still got DVDs with data on it. Try to find a working DVD player in our building … It's quite a challenge. All of the laptops issued by the [*institute*] these days, very few have a DVD reader in. I've got stacks of backups on DVDs. External readers don't always work … So the reader was kept in a proper space, and so forth. At some point those readers just stopped working.'

9. Software obsolescence

RGL1 mentioned that software becoming obsolete can be viewed as a data risk factor. The respondent described how obsolescence had resulted in some of the data no longer being available:

> 'And one of the things that happened: the software used to create backups became obsolete. At one point I could not recover at all data that was on tape.'

10. Electronic protocol not compatible with new machines

One of the respondents (RGL1) stated that the electronic protocol the group had been using was not compatible with new machines.

11. Equipment compatibility issues

RGL4 mentioned equipment incompatibility as a data risk factor and stated that the group had lost data as some of their laboratory equipment was not compatible with the network systems.

The same respondent also mentioned that some of their equipment was 'sensitive', and at risk of being infected by viruses. To combat this risk factor, the group had ensured that such equipment would not be connected to the institute's ICT network.

12. Data value not recognised

One of the respondents (RGL1) described how not being cognizant of the value of the data is a risk factor:

> 'Any data can be debated to be relevant or useful … I suspect there's a danger that someone might not realise the value of the data.'

13. Data curation is a time-consuming activity

The fact that data curation is a time-consuming activity was mentioned by one respondent (RGL1); this time-consuming characteristic (resulting in less-than-ideal data management practices) could lead to data being at risk.

14. Data curation is a demanding and arduous activity

The fact that data curation is a laborious activity was acknowledged by one respondent (RGL1).

15. Data curation is a costly activity

Data curation was described by one respondent (RGL1) as an expensive activity.

RGL2's statement regarding the cost of a backup medium for physical samples also supports the idea that costs can contribute to data being at risk:

> 'It would be tricky. Because I cannot find anyone to do it. But I guess someone would charge us in the region of … if I am not mistaken it is R25 000 once off, with R5 000 a year. But that's only for one sample and we have about 100 samples.'

16. Lack of handover accompanying resignations

Several RGLs highlighted the fact that lack of handover prior to a colleague resigning constitutes a data risk factor.

RGL1 mentioned that their group had experienced data loss due to not being informed where an ex-colleague's data were located. RGL4 and RGL5 agreed with this statement, with the former mentioning that data loss is usually not a problem if the research is active, as the involved parties can be consulted regarding data location.

RGL5 remarked with the following:

> 'When some key staff members resigned, it was difficult to find information. Mentioned in staff meeting yesterday that we need to start fixing these problems, especially in our environment. When people started to leave you would find data just missing. It's not filed in appropriate places, and you never find it eventually.'

RGL7 stated:

> 'For some of our projects, we have a wiki page with the information, how you access the information etc. But it's not a formal process; some people do it, some people don't. We have had the issue where people resign or leave the [*institute*] and we lose all that data. There is no handover process. So we have lost data like that as well.
>
> It's usually when people resign; there is no process for them to hand over their data. It's [a] personal responsibility to maintain access to the data via our wiki pages.'

17. Equipment failure

According to RGL2, running out of storage medium for their unique physical data samples was a major problem in their group:

> 'So that actually nearly happened during lockdown. Liquid nitrogen has to be topped up once a month, and because we weren't allowed in the lab there was a period where we could not get in. So they had to allow me to get in to top it up. If that goes, then the temperature goes up from minus 192 degrees. If they rise to anything like minus 40 then they become degraded.'

18. No data management plan

RGL2 mentioned that lack of data management planning would put data at risk. The respondent stated the following:

> 'And they … I don't think there was a data management plan.
>
> And their data … I think it is just sitting on someone's laptop.
>
> I would definitely consider that to be data at risk.'

19. Lack of strategy or contingency plan

Several respondents indicated that not having a proper data curation strategy, or contingency plan, could put data at risk.

RGL2 remarked as follows:

> 'I would definitely consider that to be data at risk. I mean, they are not irresponsible individuals. I am just saying there was not a contingency plan, for that, I don't think was made. So they might be able … or maybe they did, I don't know.'

RGL4 stated that their group does not have a data management framework, while RGL4 alluded to the risk associated with not having a data curation strategy in the group:

> 'Again, each researcher is responsible for projects. At the end of the year, I have to collate all of that data. Whether it be in the form of a report, or numbers, or whatever. I am trying to encourage my young people to dump it onto our project files, on the I-drive. So when you finish your experiment, take all the raw data, and just dump it into a folder. So you know it's safe, and it's not just on your computer. I constantly get, "Oh I lost all my data; I need to redo it". It's expensive.'

In the same vein, RGL8 declared that the group had never had a proper strategy for the safeguarding of data.

20. Stored in sub-standard locations/media

RGL5 declared that the storage of data in less-than-ideal locations, or on substandard media, could be data risk factors. The respondent said that difficulty in finding data of an ex-colleague eventually revealed that the data files were stored in inappropriate places such as 'CD or stiffies' and are often never found. This RGL also put forward an envisaged solution, by stating that there should be automatic storage or backups for everything once a researcher switches on a computer.

21. No backup device, or no backup solution

RGL2 emphasised the vital role that readily available backup devices, and quality backup material, play in limiting data being at risk.

> 'We have genetically engineered [*sample*] that are sitting in liquid nitrogen, and we don't have backups of those … we have just been struggling to find a company that would be willing to take [*backups*] on. So we could either find a second liquid nitrogen tank, so that's a possibility.'

22. Non-secure storage/loss of storage device

Several respondents mentioned the non-secure nature of certain storage media and portable storage devices, as factors putting data at risk.

RGL5 stated that a hard disk, containing lots of data, had been stolen. The respondent further mentioned:

> 'That is a risk with the hard disk. It's also a risk with a laptop. That's one of the reasons I said when I plug in my laptop every morning, I would like it to be dumped in some folder on [*institute's*] database. So that I know it's safe.'

RGL6's comments agreed with the above, and mentioned that even though not a secure option, the group commonly makes use of external hard drives, as the devices offer large volumes of data storage space.

The details supplied by RGL8, regarding encouraging young researchers in the group to 'take all the raw data and just dump it in a folder … so you know it's safe, and it's not just on your computer', support the previous comments.

23. Lack of understanding of basics of data management / not applying good data management principles

Several respondents mentioned the importance of adhering to good data management principles as a factor in limiting data being at risk.

RGL3 declared:

> ‘A lot of the guys working there are not very technical people in terms of data management. They are normal field technicians. They don’t really know how to upload it to the cloud, or how to copy it to a USB, and that sort of thing. They just have been trained to put it onto a computer.’

RGL7 described their data management activities as ‘a sort of hotchpotch’ that is slapped together for a project, and mentioned that once a project is completed everything is dismantled for the next project.

24. Uncertainty regarding data backups, or digitisation of paper data

Uncertainty about whether data have backups, or have been digitised, was mentioned by one of the respondents to be a data risk factor. RGL6 revealed not having performed correlations between electronic data and hard copy data, and not knowing which paper data had a digital equivalent.

25. Data only really understood by one person in group

The fact that data are often only understood by the group’s data specialist, or the data collector/creator, was mentioned as a data risk factor by several respondents.

RGL3 declared that despite being a research group leader, the group’s raw data would probably not be understood by said RGL. It was also stated that the group had a dedicated data specialist who was responsible for collecting data, processing data, and cleaning the data. Only once these steps have been completed will the data be sent to the RGL for interpretation.

RGL5 also commented on the fact that certain data in the group might only be understood by the original data collector:

> ‘Also depends on equipment you are using. If I’m using for instance a data logger where I am just logging temperatures, it is up to me when I download the data. I put a name for it, a date, and so on. So that dataset now gets linked to that, so I know what the data is. But somebody else looking at it may not understand. So that could be at risk.’

RGL6 underwrote the ideas above, by declaring that it is unlikely that anyone else in the group would understand this RGL’s data should the RGL not be available.

26. Accessibility to others in research group

One of the respondents (RGL7), when asked about the accessibility of the group’s data to all group members, stated that this would depend on the specific project. RGL7 also mentioned that some projects had a wiki page with all the information, and details on how to access the data. The RGL further stated as follows:

> 'But it's not a formal process; some people do it, some people don't. We have had the issue where people resign or leave the [*institute*] and we lose all that data. There is no handover process. So we have lost data like that as well. We have only ever had one person die while at the [*institute*] so … that's not a common thing for us. We have had at the [*institute*] in the last 15 years, I don't think any other … It's usually when people resign; there is no process for them to hand over their data.
>
> It's [a] personal responsibility to maintain access to the data via our wiki pages.'

27. Data damage and catastrophic data loss

Three RGLs emphasised the possibility of catastrophic data loss, and how such an event can affect data rescue.

RGL3 stated that while the group's paper data are still safe and in good condition, one never knows what could happen. The respondent further mentioned that should there be 'a fire or something', the data will be destroyed.

RGL5, when discussing paper data, alluded to the possibility of 'fish moths eating away at the paper'.

RGL4 believed that non-archived hard copy data, and hard copy laboratory notebooks, are at risk of being lost or damaged. The respondent elaborated on this risk factor as follows:

> '… so it might lie in an office or in some corner … and sometimes you cannot even find the raw data for that … and that also happens, that the hard copies … we don't have a framework (maybe you guys have) on what to do with completed lab books … there's no formal local archiving system, or maybe a department that can collect them and hold then for safekeeping for 10 years or what not … unless it is archived it is always a challenge …'

28. Storage space

One of the respondents (RGL5) stressed the importance of having enough storage space and commented that the cabinets used for the storage of all their paper-based data were taking up substantial space.

29. No data management continuation when project stops

RGL4 highlighted the importance of continuation after a project is completed, or when a project gets put on ice. The respondent described the difficulties and ramifications of not being able to locate data after a project has been completed.

30. Issues with previous document management system (GroupWise)[40]

According to RGL4, problems experienced with the institute's previous document management system, known as GroupWise, contributed to data being at risk. The respondent described their recurring problems with the system, and the sporadic nature of the inaccessibility of certain documents on the system. The RGL stated that some documents stored in the system had gone missing, or that access to the document was suddenly denied.

31. Issues with current document management system (Micro Focus Vibe)[41]

According to RGL4 and RGL5, many problems were experienced with the institute's new document management system (Micro Focus Vibe):

> 'I still think there's a lot of teething problems with Vibe. Many teething problems with Vibe; it's big and it's bulky and it's slow and it's very frustrating. I don't like Vibe. Because it takes me 10 minutes to open a file. To reload a file. Because it had to enter into the cloud, then it accesses … you can go and have some coffee and come back and then it might have opened … so in that sense it's very frustrating.'

RGL5 explained the frustration with the document management system as follows:

> 'It's been like that ever since I've joined here. There's a plan to put something in place; we've got the DMS, but that's quite difficult to … it's not a straightforward thing. Researchers find it too difficult, so they just don't bother.'

[40] The document management system used by the institute until 2019
[41] The document management system used by the institute since 2019

32. Difficulty in getting systems approved

Difficulty in getting managerial buy-in for new systems was cited as a factor leading to data being at risk. RGL5 stated that this has been a problem since 2002 (when the RGL started working there), and explained as follows:

> '… it's an uphill struggle first to convince managers to set up systems, because there seems to be, how can I put it … when we want to set up something … say from executive or higher level they are putting something in place, and nothing gets done …'

33. Lack of infrastructure, backup systems, and processes

Two RGLs stated that lack of infrastructure, systems or processes could lead to data being at risk. RGL7 believed that lack of infrastructure could lead to data being compromised. In agreement with this was RGL5's description of the lack of backup systems as 'problematic'. This RGL further stated the following:

> 'There must be something simpler to backup researchers' computers. I've logged in now, I've not been here for a while so there is a lot of additional data on there … I'm working at home. Can the data be dumped so we know it's safe? If my computer gets stolen, I know it's safe?'

34. Nature of electronic data

One of the respondents (RGL8) emphasised how the nature of electronic data is a risk factor. The respondent also mentioned the following:

> 'But generally, as you know, electronic data can have issues. And data can get lost. And mostly we have external hard drives. These can have breakages, or easily disappear, or be misplaced. Or get corrupted somehow.'

5.4.7.2 Summary table

Table 5.6 provides a summary of factors that may cause data to be at risk, as provided by RGLs during the one-on-one interviews.

Table 5.6: Summary of indicated risk factors

	RESEARCH GROUP LEADERS							
RISK FACTORS	RGL1	RGL2	RGL3	RGL4	RGL5	RGL6	RGL7	RGL8
Lack of backups	X	X	X				X	
Lack of handover/colleagues resigning	X				X		X	
Non-secure storage/loss of device					X	X		X
Data only understood by one person			X		X	X		
Data damage and catastrophic data loss			X	X	X			
Lack of strategy or contingency plan		X		X				X
Data location is unknown	X		X		X			
No infrastructure, systems, processes					X		X	
Issues with current DMS				X	X			
Lack of understanding of basics of RDM			X				X	
Data curation is a costly activity	X	X						
Data being in an old format	X					X		
Uncertainty regarding data backups						X		
Nature of electronic data								X
Data volume/rate of data generation							X	
Stored in sub-standard locations/media					X			
Lack of storage space					X			
Accessibility to others in research group							X	
Difficulty in getting systems approved					X			
Issues with previous DMS				X				
No RDM continuation post-project				X				
Equipment compatibility issues				X				
No immediate secure data storage			X					
No data management planning		X						
No backup device or solution		X						
Equipment failure/no maintenance		X						
Being unaware that data are at risk	X							
Lack of metadata	X							
Data curation is a demanding activity	X							
Data curation is time consuming	X							
Data value not recognised	X							
Incompatible electronic protocol	X							
Software obsolete	X							
Failure to read the data	X							

The table above indicates, in order of prevalence, the factors putting data at risk. Lack of backups was the risk factor most frequently mentioned by respondents, with half of the eight respondents citing it as a risk factor. Risk factors mentioned by three of the eight respondents include lack of handover, non-secure storage, data only understood by one person in the group, catastrophic data loss, lack of a contingency plan, and not knowing where the data are located.

This researcher regarded the individual listing of identified risk factors as important, but has also taken note of the fact that identified risk factors can be placed into broader risk-related categories. As such, the 34 listed risk factors can be placed into the following wider headings:

- risk factors related to institutional policies and procedures,
- risk factors related to ICT matters,
- risk factors related to equipment, systems and platforms,
- risk factors related to researcher behaviour, and
- risk factors that can be described as project-specific dilemmas, often entailing lack of resources such as time, expertise or manpower.

A degree of overlap can be detected (i.e., a risk factor can be placed into more than one broad category); such overlap is indicative of the ingrained and at times complicated nature of issues placing institutional data at risk. The next section, comprising a discussion of risk factors and the resultant implications, will examine the risk factors by way of broader categories. Comparisons with scholarly findings are also made.

5.4.7.3 Discussion

The findings as documented in the preceding section indicate the multitude of factors mentioned by RGLs as contributing towards data being at risk. A total of 34 risk factors have been identified; the number of factors is troublesome and greater than anticipated. These risk factors are not only numerous, but also display diversity in factors. Risk factors touch on aspects such as human behaviour, ICT matters, and institutional policy. With many of the risk factors sharing similarities in nature and root cause, it was possible to place the listed factors into fewer and broader categories. In addition, several of the listed risk factors fit into more than one of the broader categories, and a great deal of overlap can be detected. These broader divisions, containing instances of the respondents' mentioned data risk factors, are listed and expanded on below.

Factors related to institutional policies and procedures

This category can be viewed as a category encompassing many, if not most of the root causes of factors causing data to be at risk. Institutional policies and procedures (and lack thereof) can indirectly influence the categories of (i) researcher behaviour, (ii) recommended or implemented equipment and systems, and (iii) ICT-related services. These three categories have also been identified as broader categories for data being at risk and appear further down this section.

The aspects listed below are examples of risk factors mentioned by respondents, and falling under the banner of institutional policies and procedures:

- lack of data processes (this includes data management processes and processes affecting research activities, and data being at risk),
- getting systems approved: the approval procedures and accompanying red tape result in delays in acquiring, implementing and using required equipment and systems,

- displeasure expressed by respondents with both the previous and current DMS support the argument that institutional policies influence data risk further down the line, and
- lack of strategy.

With the institute's data management procedure approved soon after completion of this study's data collection stages, and the recent appointment of a data librarian, it is anticipated that several of the risk factors will be addressed and minimised in the near future.

Factors related to ICT

Factors belonging in this group include the storage being substandard, protocols not compatible with equipment, the nature of electronic data, lack of infrastructure, DMS issues, storage space, data damage, no available backup device, loss of storage, and substandard storage.

Reasons for the preponderance of factors in this category are listed below.

- Insufficient funding: funding issues would affect the storage infrastructure and devices that can be purchased; this would also affect backup locations. This issue is most likely not research group specific but institute wide.
- Systems purchased, implemented and recommended by the institute (e.g., the document management systems) without considering the possible needs of all researchers and their data can lead to hesitancy in, or failure of uptake. Alternatively, a DMS purchased at the time might have met all earlier research needs but turned out to be an impractical system as time passed.
- In the absence of data management topics on risk management plans (e.g., storage, backup, data volume), ICT-related issues are bound to contribute to data being at risk.
- In the absence of data management plans (DMPs), ICT-related issues are bound to contribute to data being at risk. DMPs typically state the data volume, data storage and backup locations, infrastructure required, and long-term preservation activities linked to a planned research project.

Factors related to researcher behaviour

This category is two-pronged in nature. On the one hand, it includes behaviour patterns based on non-compliance, and on the other, it includes activities or practices unwittingly performed while being oblivious to best practices. Examples of the former are poor handover upon resignation or retirement, only one person in the group understanding the data, or storing data on sub-standard devices. By contrast, practices such as not creating a DMP, or trusting the institutional DMS to be secure, are examples of instances that could be changed with increased understanding of data management best practices.

Reasons for the existence of these factors are estimated to be linked to the aspects described below.

- Data management and its compliance is a new practice at this institute. At the time of writing, a drafted data management procedure had been newly approved by the institute. A research library professional responsible for creating data management awareness had only been appointed in 2019. This situation resulted in a research environment, present since the institute's inception, where data management was not formalised, not marketed, and not subject to monitoring and compliance. With data management awareness still in a transitory phase at the institute, less than ideal data practices were still commonplace, and this inevitably resulted in data being at risk.
- Sub-standard data management practices would over time lead to such behaviours forming part of the institutional research culture.
- It is highly probable that random, sudden or unpredictable events can result in data being at risk. Sudden resignations or departures are examples of such events and can lead to a situation where no research group member is able to access the data or understand it. Even so, envisaging risk factors prior to a project should form part of every project, as without these in place, sudden data loss due to inferior research behaviours can occur.
- There is bound to be a link between ease and convenience of data handling or data management behaviour, and the resultant risk of data loss. This aspect was also discussed in Section 5.3.7: Location of data at risk. Examples of convenient data handling include storing electronic data to one's laptop only, making use of a USB stick that happens to be nearby, or stashing paper-based data in an unmarked cardboard box in an unused office.
- The line between inferior data management practices due to deliberate non-compliance, and practices due to ignorance, is often blurred. For this reason, the vital roles of known, distributed, explained and implemented institutional policy and procedures, related to data, are discussed in the Recommendations chapter.

Factors related to equipment, systems and platforms

Findings of the study revealed that stated risk factors such as software obsolescence, equipment incompatibility, data damage, equipment failure, substandard storage, insufficient storage space, and failure to read the data are all related to equipment or infrastructure issues.

Reasons for this category featuring such a multitude of risk factors are supplied below.

- Lack of adequate funding is suspected to be a substantial contributor of risk factors linked to equipment. Cost can be a vital determinant when deciding whether to make use of secure off-site paper data storage options, or to digitise paper, or simply store the paper in an unused office nearby.

- It is suspected that policy regarding recommended platforms and systems may also contribute to it being a risk factor. Institutional recommendations are linked to institute-wide usage of a system such as the previous DMS, where substantial volumes of data were later found to be inaccessible or corrupted. Another example is the institutional recommendation that the new DMS (2019 onwards) is to be used for data storage; however, respondents in this study have already indicated their displeasure with the system, and expressed doubts regarding its speed, or it being a user-friendly platform.
- The absence of DMPs in earlier years of research can be linked to use of equipment that turned out to be sub-standard, faulty, or not ideal for data storage. Drafting such plans and having them reviewed by other parties, such as funders, the rest of the research group, or collaborators, could have resulted in the detection of planned use of non-ideal equipment, systems and platforms.
- It is noticeable that many of the equipment-related factors are also linked to ICT-related factors. This emphasises the importance of clear and open communication channels with ICT, and prudent reporting and feedback when errors are detected, or problems experienced. Lack of input/involvement of research sectors prior to institutional purchasing and rollout of equipment can also be regarded as a risk factor.
- The institute has a dedicated asset management division performing regular asset verification and asset disposal tasks. This is an auditing requirement, and ensures that asset devaluation is up to date, and insurance cover is correct. However, it is vital that older and seemingly ‘outdated’ equipment be discarded with caution to ensure that such equipment is not required to read older format data at a later stage.

Project-specific dilemmas

Factors belonging under this category tend to be sporadic and unusual in their nature and prevalence. They are deemed to be project specific as they often go hand in hand with the unique circumstances linked to the project and may dissipate as a risk factor once the project is concluded. Examples of such risk factors include not being able to save data to a secure location immediately (field work), not being able to make immediate backups of data (field work), or the rate of data generation (including data volumes) connected with the project.

Reasons for the existence of project-specific ‘random’ risk factors are listed below.

- In many instances, funding is a determining factor when it comes to having access to more secure mobile storage, or being able to store vast amounts and velocities of data.
- As was mentioned by at least one of the respondents, training field workers to upload data to a cloud-based environment while still based in the field can lessen the risk of data loss.

- In some instances, the non-mitigating nature of risk factors should be seen as a research reality. Steps to lessen the chances of data loss include making backups of data as soon as possible, prioritising data which are irreplaceable, and ensuring available off-site data management options have been considered. Even so, catastrophic data loss or criminal events can occur, but the chances are minimised through data management best practices.

Comparison with scholarly findings

Factors listed by RGLs as contributing towards data being at risk should also be compared to the global risk factors as reported in this study's literature review (Chapter 2). For the bigger part, data at risk factors discussed in the literature review tend to mirror the data at risk factors listed by the respondents involved in this study.

A comparison of this institute's reported risk factors and scholarly evidence has revealed the similarities between data at risk factors identified via the interviews, and factors mentioned in the study's Literature review chapter (see Section 2.3). Even though the terminology differs, the parallels between the two sectors are obvious:

- 'not knowing where the data are' (RGLs) vs 'missing/displaced data' (Literature review, Section 2.3.6),
- 'unaware data are at risk' (RGLs) vs 'lack of awareness' (Literature review, Section 2.3.8),
- 'data in old format' (RGLs) vs 'outdated format/media' (Literature review, Section 2.3.4),
- 'lack of metadata' (RGLs) vs 'metadata/documentation issues' (Literature review, Section 2.3.10),
- 'lack of backups' (RGLs) vs 'data in one location only' (Literature review, Section 2.3.14),
- 'data value not recognised' (RGLs) vs 'perceived data value' (Literature review, Section 2.3.7),
- 'lack of handover' (RGLs) vs 'loss of human knowledge/skills' (Literature review, Section 2.3.3),
- 'lack of strategy' (RGLs) vs 'archiving/preservation policy' (Literature review, Section 2.3.13),
- 'DMP' (RGLs) vs 'sub-standard data management practices' (Literature review, Section 2.3.15),
- 'lack of RDM training' (RGLs) vs 'lack of awareness' (Literature review, Section 2.3.8),
- 'storage substandard' (RGL) vs 'sub-standard data management practices' (Literature review, Section 2.3.15),
- 'data damage' (RGLs) vs 'catastrophic data loss' (Literature review, Section 2.3.2),
- 'software obsolescence' (RGLs) vs 'outdated format/media' (Literature review, Section 2.3.4), and
- 'failure to read data' (RGLs) vs 'outdated format/media' (Literature review, Section 2.3.4).

In addition, certain risk factors mentioned by RGLs, while portraying a slight difference in nuance, show resemblance to the literature review's listed factors. Good examples fitting into this category are as follows:

- 'not sure whether data have backups' (RGLs) vs 'sub-standard data management practices' (Literature review, Section 2.3.15),
- 'data not being accessible to others' (RGLs) vs 'sub-standard data management practices' (Literature review, Section 2.3.15),
- 'data only understood by one person' (RGLs) vs 'sub-standard data management practices' (Literature review, Section 2.3.15),
- 'getting systems approved' (RGLs) vs 'government funding, administrative policy, and changing priorities' (Literature review, Section 2.3.11), and
- 'data management continuation' (RGLs) vs 'government funding, administrative policy, and changing priorities' (Literature review, Section 2.3.11).

Similarities between the findings resulting from interviews and the literature review findings can be seen as proof of the significance of these findings. Recommendations regarding factors putting data at risk are submitted in Chapter 6.

Despite the similarities found and described above, the interview findings of this study have also uncovered several factors not found via the literature review. Many of these factors can be described as institute-specific, and are listed below.

- Data volume: respondents have indicated that the huge volumes of data generated constitute a risk factor. With immense volumes come difficulties in (i) adding metadata, (ii) ensuring data documentation is sufficient, (iii) finding secure storage space or location, and (iv) creating backups. Failure in having a hold over data, due to its volume or scope, can put data at risk.
- Lack of storage, or storage medium running out: this risk factor is related to the previous factor, and refers to inadequate data storage, irrespective of data volumes. One of the respondents mentioned the danger of running out of storage medium for valuable, sensitive physical samples, and the financial implications such a loss would have.
- No backup device: this risk factor links to the previous two factors, and refers to the fact that data stored in one location only is regarded as a risk factor. In this instance, not being able to make backups of data, either due to the remoteness of the data collection location, size of data, or other logistical issues, result in the data being at risk.
- Protocol not compatible, and equipment incompatibility: these two risk factors refer to RGLs suspecting that data can be at risk due to a protocol not being compatible with other machines, or incompatible research equipment.

- Efforts to mitigate data at risk factors are time consuming, costly and demanding: several respondents have stated that lack of time, lack of funding, and insufficient manpower can result in data being at risk. It would appear to be a factor of particular concern at this institute and is most likely related to the institute's recent restructuring activities and freezing of non-vital positions.
- Lack of infrastructure: at the time of data collection, the institute in question did not have (i) a data management platform, (ii) a data repository, or (iii) an approved institutional data management plan template. Lack of vital data-related infrastructure was mentioned to contribute to data being at risk.
- Equipment failure: loss of data due to equipment malfunction was mentioned as a risk factor. Equipment applicable to this risk factor include laboratory equipment generating the data, field work equipment, data storage devices, and data backup devices.
- Issues with the institute's DMS: several respondents referred to the scenario where the institute recommended the DMS for storage of records, even though the DMS was revealed to not be an ideal platform. The previous DMS (pre-2019) was linked to corruption of records, inferior search facility, and inaccessibility of many items. The current DMS (2019 onwards) is described by respondents as slow and not user-friendly.

Factors listed above were mentioned by this study's respondents and did not feature strongly within the literature viewed for this research. While not suggesting that these factors are never experienced by researchers elsewhere around the globe, it is prudent to examine the reasons for this institute's researchers mentioning these risk factors. It would seem that the common denominators with many of the risk factors (stated above) amount to the institute's lack of suitable infrastructure, systems, platforms and servers. Lack of management support and buy-in are suspected to be related to stated issues. The absence of required data management foundations, coupled with insufficiencies in systems and storage options, have potential dire consequences for valuable data. These issues will be discussed in more detail in the Recommendations chapter.

5.4.7.4 Implications of findings regarding data at risk factors

Implications emanating from this institute's data at risk factors are listed below.

- While a data rescue model and accompanying guidelines can assist with much of the data at risk, it is important that the factors leading to data being at risk be acknowledged and addressed.
- Not all factors leading to data at risk can be mitigated or eliminated with ease. Examples of such factors include funding issues, lack of resources, and software obsolescence.

- Data at risk resulting from aspects mentioned in the previous bullet can be prioritised for data rescue (if still possible), with these rescue projects flagged for collaboration, equipment sharing, equipment requests, and consideration of the involvement of external parties.
- There are data at risk factors deemed to be reversible, improvable, or having the potential to be eliminated. Examples of such factors include lack of a data management/rescue strategy, lack of data management/rescue processes, or making use of substandard data storage. These factors should be attended to, with possible actions entailing the inclusion of 'data at risk' in data management training sessions and ensuring 'data at risk' forms part of data management awareness materials and guidelines.
- The link between an institutional data management policy, data management procedures, and the resultant data being at risk is obvious.
- The following questions can be asked in this regard: 'Are structures in place where data compliance can be monitored? Are data addressed in the project plan? Is the topic of "data becoming at risk in future", part of every project plan?'
- Many of the behavioural factors leading to data being at risk are connected to good research practices and ethical research behaviour.
- Of particular importance is the creation of a DMP at the start of every research project. Such a plan would ideally adhere to data management best practices and detail the envisaged data storage locations, data backup practices, access arrangements, metadata activities, and long-term preservation strategy. Adhering to the DMP in conjunction with considering the project's risk management plan are vital steps in minimising future data at risk.
- Mitigating activities listed above emphasise the role and need of a data manager (either institutional or within the research group), and data management awareness training for research group members.
- The fact that ICT participation, systems, platforms and infrastructure were frequently mentioned as data risk factors demonstrates the need for close and clear communications with ICT. Involvement of, and collaboration with ICT during the mitigation of data at risk factors cannot be disregarded.
- Despite all aspects listed here, it is part of the research reality that risk factors will always be present. Limited manpower, limited skills, budget cuts, equipment obsolescence and equipment failures are as much a part of the research environment as are best research practices and approved procedures.

This section of the results chapter discussed the most significant findings related to data at risk, how these findings compare with other studies on the same topic, and the implications of the data at risk findings.

Data at risk were found to be a common feature at this institute, with most respondents indicating that their group had data at risk. The number of factors mentioned is a cause for concern, and greater in number than was anticipated. In addition, data at risk were stated to be in a range of formats, emanating from various time periods (1970s up to early 2021), stored in a variety of locations, and present in a range of disciplines. These findings also tend to mirror the scholarly output findings, thereby supporting the idea that data at risk is a common and global phenomenon.

While respondents have mentioned 34 factors in total, these factors can also be merged into broader categories, thereby portraying common root causes. Categories of risk factors have been identified as institutional, behavioural, ICT, equipment/systems/platforms, and project-related (often random) in nature. Institutional policy and related issues (e.g., strategy, procedures, approved and recommended systems) seem to have either a direct or indirect link to most risk factors stated by respondents. While many factors are beyond a researcher’s control, this is not the case with all risk factors. Several factors are rooted in non-compliance, ease of practice, and lack of foresight. It is also suspected that many of the risk factors listed are also present in research groups that had declined to participate in the interview stage of the study.

With data at risk found to be prevalent both locally and abroad, and risk factors being numerous as well as varied, it is not a phenomenon that should lead to the shifting of blame, or to shaming or scapegoating. The prevalence of data at risk is estimated to be strongly connected to the institute’s lack of institutional data management policies and procedures, and that newly approved policies linked to data management and records management are likely to have a considerable influence on data at risk.

Data at risk are prevalent at the selected institute, and steps should be taken to alleviate potential risk factors. Steps and activities to deal with the current situation will be put forward in Chapter 6: Recommendations.

5.4.8 Data management activities

This section details the main data management activities performed by the interviewed RGLs, as revealed via the interviews. Establishing a detailed and comprehensive list of data management activities performed within each of the relevant research groups was not the goal of the interview, and did not form part of the interview schedule. However, activities listed below were mentioned by interviewees and provide a glimpse into the group’s data management awareness and practices. These data management activities are only stated briefly, to highlight factors which could contribute to data being at risk.

5.4.8.1 Findings

A total of seven major data management activities, and a number of applicable minor activities, featured in the interviews and are listed below.

1. Data management awareness and implementation

Evidence of varying degrees of data management awareness and implementation in the respective research groups, and across the institute, surfaced via the interviews.

RGL1 stated that data management is not prioritised, and explained that researchers were under much pressure to deliver on external contracts, resulting in data management not receiving precedence.

In contrast with the above is RGL6's statement, offering an example of the implementation of data management in the research group:

> 'There's a lot of work being done there to make sure that everything is secure, that backups happen correctly, that all the metadata is there that needs to be there, etc.'

RGL3 revealed that their group had members who were battling with the application of good data management principles and ascribed it to many group members not being technically minded in terms of data management. RGL3 added that efforts had been made to improve the situation by implementing international technology transfer programmes and teaching researchers how to manage their data.

2. Storage and backups

Remarks about data storage activities and data backups featured in the comments of most RGLs. As was the case with the previous point, practices varied between research groups. Described practices ranged between exemplary data storage and backup practices, and subpar activities resulting in an audit finding indicating non-conformity.

RGL2's group displayed excellent storage and backup practice when applied to their research images, as stated below:

> 'Going back to the images I've got on my laptop: we've set up a very strict pipeline. So there's the unprocessed images, and the processed images, and also backed up on G-drives. We haven't needed to rescue because we have all those backups.'

The activities of RGL3's group also show adherence to good data storage/backup practices:

> 'We've been trying slowly to teach the guys to upload data via cloud, that sort of thing. We have a small protocol that we use: we try to copy it on multiple places at the moment. But it's still prone to actually getting lost in the process. That's my opinion at least.'

RGL4 ('… make multiple copies of data, to try and minimise the risk of losing the data …') and RGL6 ('… we use Google Drive … so we have a backup …'*)* also illustrate at least partial adherence to the idea that there should be several copies of the group's data.

In contrast with the above, RGL5 reported that their group had received a negative audit finding after inspection, due to not adhering to ISO standards. Details of their data storage and backup practices are stipulated below:

> 'Currently we've set up a data logger and we just dump all the data onto a little stiffy disk (correction: memory stick) and then to Excel and then we process it. What we try to do is save everything on the I-drive if it is [*institute*] related. And if you need to use it remotely, then we add it to a hard disk and take that/carry that with. This was a finding for us basically, for ISO as well ... that we don't have a structured backup system.'

3. Data management planning

While some of the RGLs revealed that they were familiar with the concept of a DMP, the common use of such a plan could not be established via the information supplied during the interviews. Interviews showed that only one respondent (RGL2) had made use of such a plan, and had incidentally contacted this researcher, prior to the commencement of the study's online questionnaire phase, to discuss DMPs.

RGL2 further mentioned the dilemma of being approached by a contractor due to previous research, but being unable to provide a full updated DMP for the past research.

According to RGL7, their group was familiar with the concept of a DMP, but had never created one at the start of any of their projects. When asked whether their group had drafted a DMP, the respondent stated that they had created recommendations for certain infrastructure projects, but not implemented a DMP yet.

4. Metadata

Availability of metadata is a vital component of data rescue; the absence of metadata could render a dataset unusable, or open to misinterpretation. As such, enquiring about the addition of metadata to the group's data formed part of the interview schedule. Responses regarding the topic revealed that there is large variance between the different research groups around metadata.

Two of the RGLs declared that their group creates metadata for all data. RGL6 stated that '… there's a lot of work being done there to make sure that … all the metadata is there that needs to be there …'. RGL7 explained that their group often adds metadata to data, and to network data especially. The respondent mentioned that this is seen as a form of enrichment of the data.

RGL8 stated that metadata were created for some of the data, and that a small amount of data documentation was available. The respondent described the metadata and data documentation as not really mature in nature, and not complete.

The remainder of the RGLs commented that metadata were hardly ever applied to data, if at all. RGL1 described their metadata addition as 'very little', while RGL4 revealed that metadata are most likely not added to data, as every team has their 'own methods and means with their own shortcomings or not'.

The responses of other RGLs showed a similar trend, with RGL5 claiming:

> 'Also depends on equipment you are using. If I'm using for instance a data logger where I am just logging temperatures, it is up to me when I download the data. I put a name for it, a date, and so on. So that dataset now gets linked to that, so I know what the data is. But somebody else looking at it may not understand. So that could be at risk.'

Adding to the above was RGL3 commenting the following regarding metadata practices:

> 'No, definitely not. We label the data in terms of date. We have specific file names in the way we record it … I would not say it's metadata, but just a way of labelling files and tracking, being able to track your files.'

5. Digitisation of data

Only one of the RGLs was found to have digitised data at risk; this activity entailed the conversion of older digital data to a modern digital format. This activity, while forming part of data management, is one of the main data rescue steps; the digitisation details as described by RGL6 will be mentioned in more detail in Section 5.4.9: Data rescue activities and experience.

6. Data deposit in a repository

Only one of the interviewed RGLs mentioned that their data were uploaded to a public repository after being digitised. RGL6 stated that the group was working towards distributing data via public repository and ensuring that all data are curated.

RGL3 described how they were in the planning stages for developing a repository, and had been approached by a contractor to develop a database for the discipline's historic data. The group was currently busy with the planned activity.

RGL3 also stated that the research group had investigated depositing data in the data deposit tool of [*organisation*]. However, due to the sensitive nature of the group's data, the tool could not be used for data deposit.

7. Data sharing

While several RGLs declared that their data could be of interest to others, only two respondents admitted sharing their data with parties other than the client/funder of the data.

RGL6 described how their data were handed to an external party for digitisation and uploaded to a discipline-specific public repository, thereby enabling sharing of the data. The respondent also mentioned that the repository assigned DOIs to the data, thereby allowing the data to be identified and accessed with certainty.

RGL2 mentioned that their data (not the group's data at risk) were published with a recent research article, and that publishing data is a common form of data sharing in the linked discipline.

8. Other data management issues emanating from interviews

Several additional aspects, related to data management, emanated from the interviews.

- The public repository used by RGL6 for data deposit was in the process of applying for the CLARIN Core Trust Seal.
- Raw data were never discarded. (RGL3)
- Some sensitive data and equipment were never connected to the institute's network. (RGL4)
- The group's data are only understood by the data collector/creator, or the group's data scientist; this was also mentioned as a factor leading to data being at risk. (RGL1, RGL3)
- There are capacity issues when it comes to data management. (RGL1)
- The institute lacks standardised data management processes. (RGL1, RGL4, RGL7)
- There is often not enough time to manage data properly. (RGL1)

The next section contains a discussion of the listed data management activities and touches on the linked implications.

5.4.8.2 Discussion and implications

Findings emanating from this section support the findings reported in earlier sections, i.e., that data management at the selected institute is still in its infancy. Application of data management practices is random, ad hoc, inconsistent, and varies between groups. Current data practices linked to modern data, coupled with past treatment of historic data, have resulted in vast amounts of institutional data being at risk. It is anticipated that the recent appointment of an institutional data librarian and the approval of an institutional data management procedure will result in an improvement of the status quo.

In addition to the envisaged positive outcomes related to the data librarian and data management procedure, a series of recommendations (see Section 6.4) have also been put forward to deal with the effects of subpar data management practices, the current state of data at risk, and the absence of data rescue practices.

5.4.8.3 Summary

The preceding section described the data management activities mentioned by respondents during the interview sessions. The findings also contain details of data management activities detected when analysing the results of the web-based questionnaire. As was revealed via previous research into the institute’s data management practices (Patterton, 2014; Patterton, 2016; Patterton, Bothma & Van Deventer, 2018), the practices are not yet mandatory, do not yet form part of the institute’s research culture, and their implementation is up to individual researchers. Despite these negative findings, several data management practices do take place on a regular basis, with preparations for sufficient storage space, and backing up of data, being at the forefront of these regular practices. It is likely that lack of adherence to best practices regarding data management is mostly rooted in this not being a priority, rather than researchers not being aware of best practices.

With the institute’s data management procedure approved soon after completion of the study’s interview stage, an increased uptake of data management practices is anticipated.

5.4.9 Data rescue activities and experience

This section contains details of the data rescue activities performed by RGLs. This includes activities and practices when conserving data at risk, as well as efforts to recover electronic data that could not be found, accessed or read. This section also includes details of data loss where data recovery and rescue were not successful.

Each of the data rescue activities mentioned by interviewed RGLs will be briefly discussed under its own heading. A total of 10 data rescue activities were identified during the interviews and are discussed below. A section tabulating each group's data loss experiences follows the data rescue activities discussion.

5.4.9.1 Findings

1. Ensuring digital storage space is sufficient

RGL2, when asked whether data rescue had ever been performed, replied as follows:

> 'No, I don't think I've ever done data rescue. It was never needed. We haven't needed to rescue because we have all those backups.'

However, the interviewee also described removing data from a computer, and saving it to a portable hard drive to make space for new data on the computer. The velocity of data generation in their group meant that the computer hard drive filled up on a weekly basis.

2. Retrieving data after hardware failure

RGL4 stated that their group had attempted data rescue after a computer had crashed and then made use of the institute's ICT services to retrieve some of the data on the damaged hard drive.

3. Retrieving data after data breach/ransomware incident

RGL7 declared that their group did not have data rescue projects as such, but that rescue of data was performed 'ad hoc' as the need arose. The assistance of network forensics was required for many of these rescue efforts. This RGL's response, when asked about the group's data rescue activities, indicated that data rescue was usually performed due to data breach or a ransomware incident.

4. Retrieving data after document management system failure

RGL4 detailed the problems experienced with the institute's previous DMS, and how efforts were made to retrieve data saved on the system.

5. Keeping sensitive data offline

RGL4 mentioned that their group keeps much of their sensitive data, and many data storage devices, offline. The respondent further explained that due to data sensitivity, it was vital that 'viruses or something' be prevented from infecting the equipment and the data residing on it. To ensure that they are not infected, certain equipment were not connected to the institute's network.

6. Creating and adhering to a Data Management Plan

RGL2's comments regarding the creation of a DMP prior to a project (see Section 5.4.8.3: Data Management Plan) can be viewed as a data rescue activity. Stipulating how data will be managed during the project, having the DMP approved by funders, and adhering to the DMP during the project, are steps taken to ensure that data are managed well, and that the risk of data loss is minimised.

7. Adding metadata to data

As was described in detail in the section about data management activities and metadata practices (See 5.4.8.4: Metadata), three of the RGLs remarked that metadata were added to their datasets.

RGL6 stated that their groups always added metadata to data, while RGL7 referred to metadata creation as an activity that was performed often. RGL8 mentioned that some of their datasets had metadata added to it.

8. Ensuring data have backups

As was discussed in Section 5.4.8.1, regarding 'Storage and backups', the creation of data backups are common in most research groups. Adherence to consistent and standardised backup practices was found to vary between research groups.

RGL3's comment regarding data backups captures the general backup sentiment revealed through interviews, and shows the link between previous data loss, and resultant data backup:

> 'I don't know what happened exactly, but I know that they … before they never used to back up regularly. And I think one of the experiments, I think they lost almost two months' worth of data if I remember properly.
>
> … And I think that [*data loss*] is what's prompted the practice of just copying it [*all the data*] to three separate locations at any time …'

RGL6 remarked that their group had done a lot to ensure that data are secure, and that the backing up of data is performed correctly.

9. Uploading data to a repository

Uploading data to a data repository is regarded as a vital step in limiting the loss of data. Two of the interviewed RGLs discussed the issue of data repositories, and remarked that they had either ensured that data were uploaded to a repository, or were in the process of creating such a database/repository.

The data of RGL6's group had been uploaded to a public discipline-specific repository by an external party, with the repository assigning a DOI to the data. The assignment of such an identifier, coupled with the deposit of data in a repository, are important steps in ensuring that data are not lost.

In addition, envisaged repository-related rescue activities for this group included the events below.

- A large, funded project, involving vast amounts of audio data, was mentioned to be in the pipeline.
- Discussions were underway to make use of [*data deposit organisation's*] repository for the deposit of the project's data.
- The idea was to automate the system, leading to seamless uploads to a data repository.

RGL3 provided details of planned data repository use and explained that the group had recently discussed the development of a discipline-specific repository while finalising the details of a new [*funder*] project. Using an existing platform and repurposing it or making it suitable to their type of data also formed part of the discussions.

The objective of the planned repository was to provide access to all historic [*research area*] data and was regarded as a key project. At the time of the interview, the research group was awaiting approval for the project to commence.

RGL8 had also approached [*data deposit organisation*] as RLG8's group had anticipated to make use of their data deposit tool, but due to the sensitive nature of data the group could not be accommodated. In addition, the group had top secret data; data with a high security rating may also not be uploaded to [*organisation's*] data repository.

10. Outsourcing the rescue of data at risk to an external party

RGL6 had identified humanities-related data to be at risk (audio and video data in old formats) and subsequently made use of an external party to rescue said data. This collaborative effort therefore involved institutional researchers, external digitisers, other experts in the same humanities-related subject area, and external discipline repository staff. The external party performed a range of data rescue activities, including:

- digitising data to a common, open, modern format,
- adding metadata to supplement the metadata already added by RGL6's group,
- uploading the dataset to an open access discipline repository, and
- assigning a DOI to the dataset.

The data rescue effort as described by RGL6 is the only example of 'complete' rescue of data at risk established via the personal interviews. Even though the rescue effort was outsourced, it was the only instance of a research group identifying data at risk, having it digitised, and uploading the data to a discipline-specific data repository where the digital version of the data would be accessible to more than the group members. The rescue effort adhered to all the main data rescue steps forming part of this study's initial Data Rescue Workflow Model (see Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics).

11. Data loss experience

Most of the interviewed RGLs had either experienced data loss themselves or had knowledge of someone in the research group who had suffered such a fate. A summary of the respondents' data loss experiences is provided in the table below.

Table 5.7: Data loss experiences of RGLs

RESPONDENT	DATA LOSS	DETAILS
RGL1	Unsure	• Stated that some reports might refer to data that are not available anywhere • Stated that some data might be on a computer that 'will never boot up again in its life' • Stated that the indexing/metadata, created 20 years ago, might not be understood by current researchers
RGL2	No	• Absence of data loss was ascribed to an organised backup system
RGL3	Yes	• Has not personally experienced data loss • Members of research group have lost data • Stated that data loss was from more than one project • Data loss entailed losing two months' worth of data
RGL4	Yes	• Unable to retrieve corrupted documents on the DMS • Laboratory notebooks have been damaged or misplaced • Hard copy reports were damaged by burst water pipe • Data loss after hard drive crashed • Data loss due to equipment/network incompatibility issues
RGL5	Yes	• Has not lost data personally, but members of research group have: 'I constantly get *"Oh I lost all my data, I need to redo it."*'
RGL6	Yes	• Group had performed complete data rescue of data at risk • Other data on stiffies/floppies are no longer accessible
RGL7	Yes/No	• Data breach was experienced, but data could be retrieved
RGL8	Yes	• Had not lost data personally, but members of research group have: 'I know that people who have lost their computer, like their computers crashed, then it becomes a problem. Especially if it's not backed up.'

The next section lists general findings with regard to data rescue activities.

12. Overview of data rescue activities

In-depth interviews with RGLs revealed that only one group had participated in a structured data rescue project, involving vital rescue steps such as identifying the data, digitisation, and repository upload. This project not only involved the research group, but also external parties including a university digitisation unit and a national discipline-specific data repository. As such, it is apt to describe this project as being the result of collaboration and outsourcing.

Interviews identified several data rescue activities, as summarised below.

- Data were retrieved with the assistance of ICT after data breach or a ransomware incident.
- Groups ensured digital storage space was sufficient; this was most commonly done during the project and not as a planned pre-project activity.
- Data were retrieved after hardware failure, as the data were deemed vital for project continuation and completing research.
- Data were retrieved after document management system failure or after data had become corrupted while stored on the document management system. The retrieval of such corrupted data was a common event and mentioned by several RGLs.
- Sensitive data were kept offline, as per funder/client stipulation in the project plan.
- Although a rare activity overall, certain research groups had created, and adhered to, a DMP.
- Despite not performed by all groups, adding metadata to data was a prevalent activity in certain groups.
- Ensuring data have backups was common among interviewed parties, and mostly entailed the automatic and daily backups performed by ICT.
- Some groups uploaded some of their data to repositories; this was not a data management activity routinely performed by any of the RGLs' groups.

5.4.9.2 Summary

The table below provides a summary of the data rescue activities performed by the respective research groups involved in the interviews.

Table 5.8: Summary of data rescue activities performed

	RESEARCH GROUP LEADERS							
DATA RESCUE ACTIVITIES	RGL1	RGL2	RGL3	RGL4	RGL5	RGL6	RGL7	RGL8
Uploading data to a repository			X**			X		X*
Adding metadata to data						X	X	X
Ensuring data have backups			X			X		
Ensuring sufficient digital storage space		X						
Creating and adhering to a DMP		X						
Retrieving data after DMS failure				X				
Keeping sensitive data offline				X				
Outsourcing data rescue						X		
Retrieving data after data breach/ransomware incident							X	

(* attempted, but not completed / ** in pipeline)

The above table contains the various data rescue activities mentioned by respondents, and is listed in order of prevalence.

As indicated in the table above, the two most common data rescue activities involved the upload (or planned upload) of data to a repository, and the adding of metadata to data. Other data rescue activities have only been mentioned by one respondent each, apart from 'ensuring data have backups', a rescue activity stated by two of the eight respondents.

In general, many researchers immediately referred to modern data when questioned about data at risk and data rescue. However, one of the RGLs described how data in older formats were identified and rescued, with data rescue steps resembling the main steps described in this study's initial Data Rescue Workflow Model. An interesting feature of this rescue project entailed the outsourcing of most of the data rescue activities to an external party. The end product, comprising data in a modern electronic format, was an exemplar of a successful data rescue venture: older data converted to a common, open, modern digital format, accompanied by metadata, and made accessible to the public by uploading it to a dedicated discipline repository.

As shown in Table 5.7, many of the interviewed RGLs had experienced data loss, either entailing the loss of data they had collected themselves, or data belonging to the research group.

Recommendations emanating from the findings regarding data rescue activities will be presented in Chapter 6: Recommendations.

5.4.9.3 Discussion and implications

Significant aspects emanating from the above are stated below.

- Apart from one research group reporting on their formalised data rescue project, complete and structured data rescue initiatives were not prevalent at this institute.
- Adding to the above: only one of the eight interviewed RGLs provided details of a structured data rescue project, involving the identification/location of data at risk, digitisation, adding metadata, uploading it to a repository, and assigning a DOI to the data to ease sharing and citation.
- The above project involved data in an early digital format (audio data) and emanated from a research project conducted in the late 1990s to the early 2000s.
- An additional important finding regarding the above group's data rescue project was the outsourcing of the entire activity to a third party. After identifying the group's data at risk (and relevant discussions with the third party), the data were handed over for digitisation into a common, open, modern digital format. The data were indexed and uploaded to an accredited disciplinary repository.

- The remainder of interviewed RGLs' groups had been involved with data rescue to a much lesser extent; this entailed discrete activities such as recovering electronic data, making data backups, or adding metadata to data.
- It is important to note that, while each of these data rescue-related activities listed in the previous bullet is important, the activity on its own does not ensure that the data are rescued, or no longer at risk.
- Data rescue activities frequently involved the rescue of corrupted modern electronic data on an institutional system or server; this was a common and recurring occurrence among several research groups.
- Data rescue often involved the assistance of ICT staff and seemed to not be an activity that could be successfully performed by members of the research group only.
- None of the reported data rescue activities (including the data rescue project) had accompanying documentation or guidelines supporting the rescue activities.
- The affirmative answer with regard to having data documentation was found to consist of:
    - data storage guidelines, or
    - data backup guidelines.
- The absence of data documentation accentuated the absence of best practices guidelines with regard to data rescue, and the absence of documents to be consulted when drafting a data rescue workflow model.
- Frequently performed data rescue activities at this institute share the commonalities stated below.
    - Rescue activities were often urgent in nature, and critical to the completion of the research project.
    - Rescue activities were sporadic in nature, and mostly occurred on an ad hoc basis.
    - Rescue activities did not involve a data rescue project plan, procurement of new equipment, or budgetary allocations.
    - Rescue activities often required the participation and assistance of institutional ICT experts.
    - Rescue activities often entailed the recovery of recent data in a modern electronic format.
    - Rescue activities rarely involved the rescue of pre-digital data, or early format digital data.
    - Rescue activities were mostly completed without the involvement of the institutional library and information services, data management staff, archivists, or repository professionals.

  - Rescue activities often involved data that were stored on researcher laptops, institutional servers, or systems.
  - The nature of rescue activities was not identical throughout the institute, as not all groups participated in the same rescue activities. Implementation of rescue activities seemed to be at the discretion of each research group.

Taking all the above into account, there appears to be no long-term strategy when it comes to the handling of data at risk held by research groups. Reasons for data rescue not being prioritised, and the long-term preservation of data not being considered, are estimated to include the aspects listed below.

- The number and range of data rescue problems and challenges experienced by researchers are cause for concern, and are most likely hindering many data rescue efforts. These challenges are described in more detail in Section 5.4.10: Data rescue obstacles and concerns. Judging by the number and range of challenges listed, it would appear that data rescue at this institute is a venture that requires researchers to surmount various obstacles.
- There have never been any data rescue discussions, data rescue awareness creation sessions, or data rescue events presented at this institute. The concept of a structured data rescue project appears to be novel and unknown to interviewed parties.
- The absence of any available institute-related data rescue models, guidelines, or best practices document is a direct result of the status quo alluded to in the previous bullet.
- Data management implementation at this institute is still in its infancy. With a data librarian only recently appointed, the data management procedure still under institutional review at the start of this study’s data collection stages, and the absence of a dedicated institutional data repository, the absence of structured data rescue projects is disappointing yet not altogether unexpected.
- Previous data management surveys had revealed low levels of data management adherence (Patterton, 2014; Patterton, 2016; Patterton, Bothma & Van Deventer, 2018).
- Organisational restructuring, forming of new research groups, disbanding of old research groups, and appointment of new RGLs can affect prioritisation of activities.
- It is also possible that some data at risk are not earmarked for rescue as the data can be regenerated. Additionally, a copy of the data may already exist elsewhere. In these instances, the data are not truly ‘at risk’, as its loss or damage does not imply that it cannot ever be regenerated.

Recommendations emanating from findings into data rescue activities are presented in the final chapter, with Section 6.4.1 (Create awareness around data at risk), Section 6.4.2 (Encourage the correct handling of data at risk), Section 6.4.3 (Create awareness around data rescue), Section 6.4.5 (Address and mitigate data rescue challenges), Section 6.4.6 (Promote the data rescue model) and Section 6.4.14 (Launch a data rescue project at the selected research institute) containing suggested steps to deal with the current lack of data rescue activities, and resultant absence of data rescue knowledge, insight and experience.

5.4.10 Data rescue obstacles and concerns

This section provides details regarding the obstacles and challenges experienced by interviewed parties while performing or planning to perform data rescue. RGLs who had not performed data rescue were asked to state the potential challenges in a hypothetical data rescue scenario. In addition to obstacles experienced or envisaged during data rescue, concerns related to data loss, and data rescue, are also listed in this section. A total of 26 data rescue obstacles have been identified via the interviews and are each discussed below.

5.4.10.1 Findings

Each of the data rescue obstacles is afforded its own heading and inclusion of brief details related to the obstacle. The findings section is followed by a discussion of identified data rescue obstacles.

1. Data rescue involves a great deal of manual work

One of the respondents (RGL6) stated that a major data rescue challenge would be the labour and effort required to perform the required tasks.

2. Time required to rescue data

RGL6 mentioned that the amount of time required to perform data rescue constitutes a challenge. The respondent stated that data rescue will be time consuming, as would the time required to understand the indexing, to create a new index, to add proper metadata, and to scan paper-based data.

RGL4’s comments agreed with the above, and described the challenge of time as follows:

> ‘As custodian, if they want it on an electronic platform, they have to do it themselves, and most people don’t bother. Because of the time it takes to go and scan every document and every page, some people do it, some don’t. It’s up to that person.’

RGL3 concurred with the comments above and stated that the conversion of data to an electronic format would be a huge challenge in terms of archiving it onto a digital platform, and would take a long time to execute.

3. Determining whether data rescue is worthwhile

One of the respondents (RGL1) described how determining the value of data, and the value of data rescue, can be seen as data rescue challenges. The respondent stated that it is vital to ascertain whether the planned data rescue activity is worthwhile, and whether the data had much value.

4. Not knowing what has already been digitised or rescued

One of the interviewees mentioned that uncertainty about whether data had already been digitised can be a data rescue challenge. RGL1 stated:

> 'Not knowing whether it is already digitised, and some of it is actually partially digitised … so that is not always very clear. And that is a very important aspect: when you stand in front of a file cabinet, full of paper-based and hard copy data, you don't have an understanding which parts of it has got an electronic equivalent.
>
> And also, the other side: if you find the electronic equivalent, by looking at it, do you have any clue whether there is a paper copy somewhere?
>
> So just that aspect is tiring and confuses the system.'

5. Cost

Several of the interviewed RGLs mentioned that the cost of data rescue is a definite challenge, or an aspect that needs to be considered before embarking on a rescue venture.

RGL2 stated that backup devices are required for data rescue and that the cost of these devices, such as external hard drives, would not be prohibitive. In contract with this was RGL7, who mentioned that lack of funding would constitute a data rescue obstacle for their group, as the group did not have a separate fund for collecting or maintaining data.

RGL8 also mentioned that finances would be regarded as a data rescue obstacle in their group, while RGL4 mentioned that data rescue would have an impact on resources and finances.

6. Problematic ICT assistance

One of the respondents (RGL4) described how the assistance provided by ICT services is regarded as a data rescue obstacle when it comes to electronic data.

> 'Even with ICT services, sometimes I find there is not the skills to know what to do. There is a standard operating procedure,

> how you might be able to rescue data, but if that doesn't work, they cannot help you. And I find that frustrating because I know there's multiple options that you can do and try and rescue data (on an electronic platform that is).'

7. Equipment/hardware/software required

According to two of the interviewed RGLs, the equipment required for data rescue, and to read rescued data, are data rescue challenges.

RGL8 mentioned the issue of a bespoke sensor required to read some data formats and stated that such a sensor would need to be purchased when dealing with data in an uncommon format.

RGL6 described how software can be a data rescue challenge:

> 'I guess software. As I say, one can just put it through the scanner and scan it to pdf, but that won't be very useful. So I think one would need software then to convert back to Word or something. I guess, I don't even know what would be useful for people.'

8. Manpower

Two interviewees (RGL6 and RGL4) mentioned the issue of manpower when asked to discuss data rescue obstacles. RGL4 stated that restructuring had resulted in the group running on skeleton staff.

9. Finding a suitable platform/having institutional infrastructure

Two respondents (RGL3 and RGL7) remarked that the following issues are data rescue challenges:

- deciding on a suitable data repository,
- finding a suitable repository, and
- lack of institutional data rescue infrastructure.

RGL3 detailed the issue as follows:

> 'The first thing would be: finding a good platform. I know in the [*funder*] proposal that we had done, a lot of the team members were talking about developing our own database. I don't think that would be good, if I understand from the comments of the team. They said it would be good if we could use an existing platform and just repurpose it or make it suitable to our sort of data.'

10. Limitations of current configuration system

RGL7 stated that while their group had a tried-and-tested configuration system, the system was unable to handle secret data, or large quantities of data. This limitation was described as a data rescue obstacle.

11. Limitations of [*organisation's*] data deposit platform

Two of the RGLs described how the current institutionally recommended data deposit tool [*name of service/tool removed*] was not able to handle their group's classified data. RGL7 supplied details of the group's problems regarding anonymisation and encryption, and how the current institutionally recommended data deposit tool was unable to handle classified data. While the data deposit tool has the infrastructure, network capacity and storage capacity, the approval from the larger [*organisation*] consortium was still needed before such data would be accepted by the platform. According to the respondent, the lack of consensus among stakeholders was the cause for the delay in the platform being able to accommodate the group's data.

RGL6's comments agreed with the above; the respondent stated that their data storage challenges relate to the same data deposit tool:

> 'And one of the challenges we got to in the discussion with [*data deposit organisation*]: they require encryption, because voice data has biometrics in it, right? So we would need to encrypt our data which brings with it certain implications for accessing that data and processing that again. And concerns about degradation of quality and so on and so on. Those are some of the challenges that we have …'

12. Systems are not easy, fast, safe

According to RGL5, data management and data rescue are hindered by the institute's lack of user-friendly, secure and high-speed systems:

> 'When we want to set up something … say from executive or higher level ... they are putting something in place, and nothing gets done. It's been like that ever since I've joined here. There's a plan to put something in place; we've got the DMS[42], but that's quite difficult to … it's not a straightforward thing. Researchers find it too difficult, so they just don't bother.'

[42] Groupwise DMS: The institute's document management system; used up to 2019.

RGL5 also mentioned that managerial buy-in with regard to systems is often problematic and can hinder research activities. The RGL stated that trying to convince managers to set up systems is ‘an uphill battle’.

13. Additional storage issues

RGL6 stated that their group was concerned about not being able to find proper cloud-hosted storage, and that the size of their data contributes to the issue, resulting in the data being at risk.

14. Encryption issues

According to RGL6, having to encrypt sensitive data is a data rescue obstacle. Concerns about the possible degradation following encryption were also mentioned by the respondent.

15. Lack of data processes

One of the interviewed RGLs (RGL7) mentioned that there is a lack of formal data management processes in the institute, and that many of the current processes need improvement.

16. Data rescue skills, insight and experience required

Two of the RGLs detailed how data rescue requires certain skills, experience and insight into the data and the research discipline.

RGL3 stated that it is a challenge to find someone who understands the group’s data, and how to work on that platform. The RGL also stated that the group has a data specialist who is close to retirement but is also the only member of the group who knows how to handle and interpret the data. They are, however, in the process of training people accordingly.

RGL6 stated that it is vital for a data rescuer to select the correct format and be able to put the data into a format usable by others, and that the format should not necessarily be a PDF.

The respondent further stated the following:

> ‘Scanning it into a PDF and labelling the PDF is one thing. But there needs to be some metadata to it or there needs to be some description what is in the PDF in order to make it useful for somebody else, otherwise how would they know. Of course, if somebody else wants to use it at a later stage, they would not want it in a PDF format. They would want it in editable raw format.

> It needs to be curated into a dataset. At this stage it's just a transcript on an interview; it's not really useful in its current format. Well, I think it's useful in the sense that it's interesting, and there is interesting information in it. But useful in the sense of somebody being able to use it, it needs to be in a different format.'

17. Visibility/accessibility of data management experts

RGL2's comments regarding difficulty in obtaining data management assistance and advice suggest that there is a need for data expertise in the institute to be more visible. RGL2 mentioned the importance of being able to contact a data management expert when creating a data management plan.

18. Data rescuer to understand context of the data

RGL6 mentioned the importance of remembering or understanding the information held in the datasets. The respondent also remarked that a data rescuer should understand the type of information required for the data to make sense to future users of the data.

RGL1 agreed with the above and stated that it is more important to have a grasp of the context of the data, than it is to have subject knowledge or scientific expertise.

19. Subject knowledge is required

One of the respondents (RGL6) stated that adequate subject knowledge and expertise is an essential part of data rescue. The RGL further stated that it is important to understand the field of research and to know 'what is relevant to whom'.

20. Difficulty in locating the data and making sense of it

According to RGL3, difficulties in tracing/locating data, and making sense of it, constitute a data rescue obstacle. The respondent mentioned that much of their data are old, and that it is unlikely that all the data have been labelled and recorded accurately.

21. Dissatisfaction with current document management system (Vibe)

One of the interviewed RGLs (RGL3) expressed concerns about the institute's current DMS, and regarded the problems experienced with the system as a data rescue challenge.

> 'I know last year they also looked at backing up some data and trying to figure out how we can actually archive it safely on the [*organisational platform*]. They were actually talking about putting it on Vibe if I remember correctly. But Vibe isn't the

> platform for such huge quantities of data, and I think they are having difficulty at the moment, so …'

22. Data quantity/volume

Three RGLs (RGL3, RGL6 and RGL8) stated that the group's data quantities (and dataset size) are data rescue challenges. RGL3 captured the sentiment by stating that their group's data specialist had studied their data and had informed the RGL that archiving the data was impossible due to the scope of the data.

23. Data rescue and data management is up to the individual

According to two of the interviewed RGLs, the fact that data rescue is currently up to the individual is a challenge. RGL7 described the scenario as one where each researcher has a personal 'pet project' and mentioned that group members had their own 'workaround' regarding data storage practices.

RGL4 described this challenging issue succinctly:

> 'Some people do it, some don't. It's up to that person.'

24. Research culture

Two of the interviewed RGLs stated that research culture, and hanging on to conventional methods and options, can be a data rescue obstacle. RGL2 mentioned that researchers still prefer laboratory notebooks in hard copy format, finding them more convenient and portable than more secure digital versions. The same respondent also remarked that raw data, not collected in a digital format, would remain in its original format, even after being processed.

RGL8's comment agreed with the view above, as the respondent mentioned that a change in research culture is required for data rescue to become an established research activity.

25. Increase in project resources required

RGL8 mentioned that data rescue would only become viable once more resources per project were made available as the status quo did not allow for adequate data curation resources.

26. Automatic metadata logging equipment is required

One of the interviewed RGLs discussed software and equipment that could make data management easier, thereby lessening the risk of data loss. RGL5 stated that having equipment capable of automatically logging metadata would be an ideal scenario.

5.4.10.2 Summary

The table below provides a summary of data rescue challenges experienced by study respondents when rescuing data, or as envisaged during a hypothetical data rescue project.

Table 5.9: Data rescue obstacles

	RESEARCH GROUP LEADERS							
DATA RESCUE OBSTACLES/CHALLENGES	**RGL1**	**RGL2**	**RGL3**	**RGL4**	**RGL5**	**RGL6**	**RGL7**	**RGL8**
Cost		X		X			X	X
Data quantity/volume			X			X		X
Time required to rescue data			X	X		X		
Data rescuer to understand context of the data	X					X		
Research culture		X						X
Data rescue skills, insight and experience required			X			X		
Finding a suitable platform/having institutional infrastructure			X				X	
Data rescue and RDM is up to the individual				X			X	
Limitations of [*organisation*] data deposit platform						X	X	
Manpower				X		X		
Equipment/hardware/software required						X		X
Not knowing what has already been digitised/rescued	X							
Not knowing whether data rescue is worth it	X							
Visibility/accessibility of RDM experts		X						
Displeasure with current DMS			X					
Difficulty in locating the data, making sense of it			X					
Problematic ICT assistance				X				
Automatic metadata logging equipment is required					X			
Systems are not easy, fast, safe					X			
Additional storage issues						X		
Essential to have subject knowledge						X		
Encryption issues						X		
Data rescue involves a great deal of manual work						X		
Lack of data rescue processes							X	
Limitations of current configuration system							X	
Increase in project resources required								X

A total of 26 obstacles have been identified, with respondents mentioning between two and 11 obstacles each. The most common data rescue challenges were found to be cost, data quantity/volume, time required to rescue data, and the fact that a data rescue venture would require skills, insight and experience.

The next section contains a discussion of the findings into data rescue obstacles, and is followed by a section stating the implications emanating from these findings.

5.4.10.3 Discussion

As revealed via interview findings and portrayed in Table 5.9, RGLs have collectively mentioned a total of 26 data rescue challenges or obstacles. The myriad of stated challenges provides an insightful glimpse into the problems envisaged by researchers when asked to imagine a data rescue scenario, or when asked to list challenges experienced during past rescue efforts.

While the number and range of listed challenges are worrying, the existence of challenges is not unexpected. Previous research into the data management practices at the institute had revealed lack of adherence to data management best practices, an absence of vital data management platforms or tools, and data management procedures not yet in place (Patterton, 2014; Patterton, 2016; Patterton, Bothma & Van Deventer, 2018). In addition, it is noticeable that many of the stated data rescue challenges show similarity with factors leading to data being at risk, such as the high incidence of ICT-related factors, lack of processes, and problems with the current and available institutional systems. These risk factors were mentioned and discussed in Section 5.4.7: Data at risk: Factors.

Commenting on each of the factors is an essential part of this discussions chapter and is done accordingly below. The commenting will be followed by a listing of broader categories, as many of the listed factors share the same root cause or origin.

1. Data rescue involves a great deal of work

As can be seen from the initial Data Rescue Workflow Model (see Section 3:5: Data Rescue Workflow Model), and the various rescue models described in Chapter 3, data rescue involves numerous steps, activities and skills. It is not a same-day activity, and ideally not something to be completed on the side while continuing with research activities. Data rescue should preferably be viewed as a separate research project and given the same levels of planning, dedication, and adherence to best practices. It is precisely this effort required that plays a major part in data rescue not being considered.

Various respondents have commented on workload, and when considering the institute's various restructuring initiatives, downsizing efforts, and frequent freezing of non-vital positions, the work required to perform data rescue is a major obstacle to its attainment.

2. Data rescue is time-intensive

Many of the activities forming part of a data rescue project require a great deal of time; often the time required for rescue activities may not be known at the start of a project. Activities such as locating all the data belonging to a given dataset, locating its documentation, locating metadata, assigning metadata, and determining whether duplicates of the data exist, are some of the tasks that can end up requiring more time than was anticipated. With study respondents already mentioning the work pressures of their groups' researchers, it comes as no surprise that the time constraints accompanying rescue activities would be seen as an obstacle by researchers.

3. Data rescue entails added pressure on manpower

This obstacle is similar to the two previous obstacles and is connected to researchers being under considerable pressure (at the institute) due to meeting contractual deadlines, meeting their own key performance targets, and dealing with downsizing, restructuring and frequent freezing of non-key positions.

4. Data rescue is a costly activity

Apart from data rescue placing a premium on manpower and available researcher time, it is also an activity that often requires additional funding, or obtaining and using equipment not held by the research group. Scanners, digital cameras, camera lenses, camera stands, software readers, secure storage, repository fees and archival materials are all items often required for data rescue activities. Should a research group not have these items already in stock, not have the funds to purchase them, or be unable to borrow them from elsewhere in the institute, data rescue will need to be aborted due to lack of required funds or physical resources.

5. Data rescue impacts on resources

Adding to the above points: for data rescue projects to succeed, an increase in current available or allocated manpower, time, funding, equipment, servers, platforms, assistance and training is required.

6. Data rescue requires specific skills, insight and experience

A successful data rescue project ideally requires people with the following collective range of skills:

- disciplinary/subject knowledge,
- an understanding of the context of the data,
- ability to operate scanner and digital camera,
- familiarity with archive practices,
- ability to create suitable metadata for the data at risk,
- ability to draft and design a project plan and proposal,
- ability to implement and manage a project,
- can determine where assistance, collaboration and external input are required,
- an understanding of data management best practices and requirements (e.g., the importance of drafting a DMP),
- ability to evaluate and select suitable repositories,
- an understanding of the importance of data sharing and a DOI (or other handle),
- an understanding of the basic data rescue steps, and
- previous data rescue experience will be an advantage, though not an absolute requirement.

The listed skills and attributes are some of the major skills required from data rescue team members. It is unlikely that all team members will possess all skills; in reality some rescuers will be subject specialists while others will have repository knowledge, and others more skilled in data management practices.

Nonetheless, for data rescue to be successful, a range of skilled team members is required, and the lack of available skills or experience can be viewed as a data rescue challenge.

7. It is vital that data rescuers understand the context of the data

The importance of this aspect when selecting data rescue team members cannot be understated and shows the need for metadata creation and data documentation at the time of the original data generations. In addition, this stated challenge also shows the need for proper handover in ensuring recently collected data are understood and treated accordingly.

8. Subject knowledge is required when rescuing data

The need for subject knowledge as a data rescue obstacle is a major problem, as it implies that researchers do need to be part of a data rescue team. As stated earlier, work pressure and manpower availability concerns mean that this obstacle is one that will affect the implementation of many data rescue projects.

9. Difficulty in determining whether data rescue is worthwhile

While data assessment is vital, such an assessment tool, specifically drafted for this institute, is currently not available to researchers. Uncertainty about the value of data, and whether the use of limited resources will be worthwhile, is a problematic issue.

10. Difficulty in determining whether data have already been rescued

Poor recordkeeping, documentation and handover have resulted in researchers not able to tell whether data have already been rescued. Additionally, not knowing whether older paper-based data have already been digitised is another related data rescue challenge.

11. Difficulty in locating data and making sense of it

Older data not used after project completion are often difficult to locate, as it might be hidden in storerooms, cupboards or laboratories. In some instances, it may even still be stored on a computer that is no longer in use. These obstacles, emanating from poor handover, and lack of data management best practices, add to the workload required from data rescuers.

Additionally, locating the data is often just a first rescue step, and can be followed by difficulty in making sense of the data and contextualising it. The same root causes stated in the previous paragraph can be applied here.

12. Data rescue involves huge data volumes, quantity and scope

The advent of big data, coupled with decades of data that had not been stored or managed securely, can lead to a scenario where researchers are too overwhelmed to envisage a successful data rescue initiative. The feeling of 'one step forward, two steps back' can be applied here, with researchers unconvinced that the rescue of some data at risk is a beneficial exercise.

13. Problematic ICT assistance can affect data rescue

Several respondents have stated that ICT assistance has at times been less than ideal. Statements allude to ICT staff unable to consider 'out of the box' thinking, not being aware of suitable alternatives, and recommending inferior systems and platforms. While these aspects may well hinder the progress of any attempted data rescue effort, it also leads to distrust in colleagues and a consideration of tools and options not recommended by institutional management.

14. Lack of visibility of data managers and data rescue experts

With the recent appointment of an institutional data librarian, data management links on the institutional blog, and the institutional data management procedure approved, it is expected that imminent institution-wide data management training sessions will increase awareness of data-skilled staff.

15. Encryption issues surface when considering/attempting data rescue

At least two respondents have mentioned that while a recommended data deposit tool was being considered, the encryption and anonymisation requirements, applicable to their data, have resulted in dissatisfaction with the tool.

The sensitive nature of data held by the stated respondents requires that it may not be visible to external parties. The recommended data deposit tool has the additional requirement of encryption and anonymisation, thereby casting doubt on the tool's privacy options.

Respondents also expressed doubt regarding their eventual use of the data deposit tool, even though the tool's administrators were housed on the institute's premises and had a long history of involvement with institutional data-related activities.

16. Lack of data rescue processes in the institute

Data rescue, involving many activities forming part of best data management practices, is largely absent at the selected institute, and this aspect correlates highly with the institute's data management procedure only approved after the study's data collection activities were completed. The expected positive changes in this regard, considering the approval of the data management procedure during July 2021, will be discussed in the Recommendations chapter. While the procedure was still in its final review stage at the time of this study, it was unlikely that researchers would have been exposed to a data rescue process, its involved steps, and best practices contained within.

17. Data management and data rescue is up to the individual

Currently, data management at the selected institute, and by implication, data rescue, is up to the individual. With the institute's data management procedure approved only after the study's data collection activities had been completed, the resultant implications entailed no data rescue processes, and no relevant standards and guidelines. Data rescue efforts and the understanding thereof currently depend entirely on the decisions and insight of individual research groups, or even smaller units (i.e., individuals).

18. Current research culture may not include data rescue or even data management

With data management being a new venture at this institute, thereby implying decades of non-adherence to data management best practices, the research culture had, with time, become one of not managing data well. Despite this researcher having access to a limited number of qualitative discussions, data rescue, as a sub-activity of data management, appeared to be regarded as a new or unknown practice. It is anticipated that this perspective necessitates the need for training, awareness sessions, and assistance over time before data rescue will be accepted by the research sphere. In lieu of generalising about the current research culture in all research groups, further investigation into the matter is required.

19. Apparent dissatisfaction with the current document management system

Apparent dissatisfaction with the current document management system, described as not being user-friendly and slow, is a data rescue obstacle. Despite it being the official and recommended institutional DMS, it does not meet the needs of many researchers. Such a situation can result in users considering other platforms, data rescue projects not making use of the recommended system, and lack of confidence in other or future ICT decisions and recommendations.

20. Data rescue would benefit from automatic metadata logging equipment

This issue was only suggested as an obstacle by one of the respondents, yet its inclusion as a potential challenge is linked to the problematic issue of metadata. As reported in previous research into the institute's data management practices, metadata creation is an activity often neglected by researchers (Patterton, 2014; Patterton, 2016; Patterton, Bothma & Van Deventer, 2018). It is therefore understandable that a machine capable of logging automatic metadata would lighten the research workload.

21. Storage issues surface when considering/attempting data rescue

Problems related to the storage of data was an obstacle mentioned by several respondents. Displeasure and distrust were detected during interviews with respondents regarding storage; many believed the current institutional storage options were insufficient, not secure, and not userunfriendly. Although most of these concerns were related to data in electronic format, worries about the availability and affordability of suitable storage medium for unique and valuable physical data samples were also detected.

22. Proposed data deposit platform has limitations

The recommended data deposit platform has the advantages of its administrators being in close physical proximity to researchers (same campus), and admittedly has a history of close involvement with many institutional data infrastructure and data management activities. This same organisation also hosts a national data management planning tool and is in the process of creating a platform able to mine DOIs for submitted data.

Despite these seeming advantages, the data deposit tool was mentioned by respondents as not meeting their requirements, still being in a test phase, and putting demands such as anonymisation and encryption onto researchers. It was also not clear whether and when the data deposit tool would be able to assign a DOI to the data deposited.

Lack of satisfaction with this tool, despite its seeming advantages, will inevitably lead to researchers making use of other data deposit platforms. The matter requires further investigation.

23. Current institutional systems appear to not be easy, fast or safe

Respondents expressed doubts regarding some of the recommended institutional systems, and stated that the platforms were not easy to use, had displayed slow performance speeds, and might not be safe to use. The implications of researcher dissatisfaction with institutional systems will invariably lead to institutional distrust, and avoidance of the systems where possible. This may need further investigation.

24. Current configuration system not meeting demands

Some research groups make exclusive use of a configuration system for managing their projects, outputs, records and data. Originally, these heavily invested systems were thought to also perform as a DMP tool and perform several data management-related functions. Nonetheless, the system was recently stated to not meet data management requirements. By implication, such dissatisfaction can lead to data being mismanaged, or the research group having to invest in an improved system.

25. Data rescue requires suitable platforms and institutional infrastructure

Ideally, institutes and entities expecting or promoting data management and data rescue should, at the very least, be able to provide researchers with the required infrastructure and platforms. Examples of such platforms include a reliable document management system providing secure and sufficient storage, backup options with similar characteristics, a data management planning tool, and an institutional data repository. At the time of writing, respondents were not satisfied with the institutional platforms and infrastructure within their reach. This matter requires further investigation.

26. Data rescue requires specialised equipment, hardware and software

Data rescue projects will often require equipment such as scanners, digital cameras, archive boxes and labels, specialised readers for outdated data, and other data conversion tools. Comments made by respondents reveal the perception that the required and necessary equipment and tools are not readily available. This obstacle underlines the need for collaboration, sharing of institutional items, and careful reconsideration of readers and hardware regarded as 'outdated'. Outsourcing of certain data rescue activities, as reported by one of the RGLs (see Section 5.4.9.10: Outsourcing the rescue of data at risk to an external party), can also be considered.

A good example of the need for outdated equipment was the recent request by the Swedish National Archives for obsolete equipment and playback devices, required for work with obsolete media and in preserving such media for the long term (Open Preservation Foundation, 2021). Often, the need for such outdated equipment, written off and disposed of decades ago, is only realised when attempting to rescue media that can only be read via discarded equipment.

5.4.10.4 General overview of data rescue obstacles

As was the case with factors putting data at risk (see Section 5.4.7: Data at risk: Factors), data rescue problems and challenges listed by respondents show that the multitude of challenges can be clustered into a few broader categories. In addition, some of the obstacles can be placed into more than one broader category, as overlap is evident. The overlapping of categories also reveals that the root causes of data rescue challenges could be more alike than dissimilar. The categories chosen for clustering data rescue factors closely mirror the broader categories used in the section on Data at risk: Factors (see Section 5.4.7). The broader categories are detailed below.

Institute-related challenges

Obstacles in this category emanate from institutional policies, procedures and guidelines. It could also be the result of the institute not having the relevant policies, procedures and guidelines in place. Examples of such resultant challenges and obstacles are problematic ICT assistance, lack of data-related infrastructure, and lack of data rescue processes.

Infrastructure and ICT-related challenges especially are cause for concern, as the accessibility and security of modern data depend on infrastructure, ICT-related systems and ICT support. With modern data being mostly in electronic format, saved on institutional servers, backed up to similar devices, and relying on systems to provide the required access requirements and restrictions, these challenges are issues that cannot be ignored. Put simply, the status quo is not beneficial to working with either historic or modern data.

Equipment-related challenges

These are obstacles emanating from the equipment available to the research group. One could also state that this category is therefore linked to the previous heading, as institutional policies often determine the equipment to be purchased, permissible, or recommended. Examples of such obstacles include the stated lack of applicable equipment, systems being slow, systems not being user-friendly, and the configuration system not meeting the group's envisaged needs. In addition, the desire for equipment to log metadata automatically, and storage issues mentioned as troublesome, also fall into this category.

Resource-related challenges

While linked to 'institute-related challenges' and even 'equipment-related challenges', this category refers to a research group's lack of means to perform data rescue successfully. Included in this category are statements related to lack of manpower, funding, time, skills, and data understanding should rescue of data be required or considered. Ability, means and skills to deal with big data (either in volumes held, or rate of generation) also fall under this heading.

Challenges related to subpar data management practices

This category can also be described as 'behaviour-related' challenges, or even the institute's research culture. Examples of challenges falling into this category include data management and data rescue being up to the individual, inability to assess data value, and not knowing whether paper-based data have already been digitised.

With infrastructure and ICT-related challenges featuring strongly under institutional challenges, behavioural obstacles feature similarly when considering data management practices. Some of the behavioural obstacles can be mitigated through applicable training: an understanding of data management best practices, and insight into data assessment steps are crucial requirements.

On the other hand, a preponderance of behavioural factors exist due to the current research environment, research culture, or economic climate. Factors contributing to subpar practices may include manpower constraints, funding difficulties, or contractual obligations.

Scholarly comparison

Investigating data rescue challenges and obstacles did not form part of this study's literature review. Moreover, it was not a common feature of data rescue studies, as the published outputs were mostly the result of successful data rescue efforts. By implication, such ventures would not be reporting on insurmountable data rescue obstacles resulting in aborted rescue projects.

Obtaining details of data rescue problems facing this institute's respondents was vital to determine issues that can be addressed in future. It would also indicate areas which might benefit from guidelines accompanying the revised Data Rescue Workflow Model.

5.4.10.5 Implications of identified data rescue challenges

The following listed implications, resulting from findings into data rescue obstacles at the selected institute, have been identified:

- Data rescue activities and data rescue projects will not be ready for implementation at the selected research institute until many crucial data rescue challenges have been addressed.
- Difficulty in getting a data rescue off the ground is regarded as a likely scenario.
- The absence of data rescue projects can be ascribed to a combination of obstacles or challenges, and not the result of one single factor. Similarly, addressing a single obstacle only (e.g., implementing data rescue awareness/training, or ensuring a reliable document management system), will not result in a sudden update of data rescue activities.
- The uptake of non-approved systems and tools, due to dissatisfaction with the current options, is a troubling yet likely scenario.

Recommendations regarding the obstacles and challenges experienced or stated are proposed in Chapter 6, which comprises the study recommendations. Section 6.4.5 contains steps put forward to deal with the stated data rescue challenges.

5.4.11 Summary of interview findings

This segment of the results chapter discussed the results of one-on-one interviews conducted with selected RGLs of the investigated research institute. Eight semi-structured, in-depth interviews were conducted making use of Skype as the virtual interview platform. The intention behind the use of the virtual one-on-one interviews included establishing the nature of the data at risk, data rescue activities performed, and data rescues challenges experienced.

With all interviewed parties already found to have data at risk within their research group (via web-based questionnaire responses), the one-on-one interviews revealed additional details about the data at risk. It was established that all formats feature in the institute's data at risk, that data at risk were not time-specific, and that the data at risk were housed in a range of locations. Factors resulting in data being at risk, or contributing towards data loss or damage, were also established. A large number of risk factors were identified. Risk factors can be grouped into distinct categories entailing institutional, ICT, equipment, or researcher factors.

Interviewed participants also discussed the value of their data, and their data sharing practices. All RGLs stated that their data have long-term value; furthermore, parties interested in the data varied, but were most commonly indicated to be funders and fellow researchers. Not all research groups share data, and even fewer groups make use of a data repository for data sharing and curation.

Results of the interviews also provided brief details about the data management activities performed by RGLs, with data backups being the most commonly performed task. No evidence of institute-wide uptake of best practices with regard to data management could be established.

Experience of data rescue was minimal, with only one group providing evidence of a rescue venture involving the identification of data at risk, digitisation of the data, and adding the converted data with its metadata to a discipline-specific repository. It was interesting to note that most of this rescue project's activities had been outsourced to external parties.

Respondents provided details of a large number of challenges experienced, or hypothetically experienced, when planning data rescue. Rescue obstacles involved issues related to institutional factors, equipment problems, lack of resources, and subpar data management practices. Data rescue challenges experienced by most researchers relate to the cost of data rescue, data quantity/volume, time required to rescue data, and the fact that a data rescue venture would require skills, insight and experience.

During the interviews, participants were informed about the initial Data Rescue Workflow Model created by this researcher, and were invited to participate in the next data collection stage, which entailed a review of the model by interviewed researchers, and providing feedback directed towards the improvement of the model.

Results of this section contributed towards several of the study's research objectives, with information gathered addressing the objectives concerned with data rescue practices, trends, needs and challenges.

The next section contains the findings of the feedback provided by respondents after reviewing the initial Data Rescue Workflow Model.

5.5 Results: Feedback on initial Data Rescue Workflow Model

This section contains feedback supplied by RGLs after studying the initial model. Detailed information on the initial model is found in three previous sections, namely Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics, Section 3.8: Initial Data Rescue Workflow Model: Summary, and Section 3.9: Stages of initial Data Rescue Workflow Model. This stage of data collection (i.e., feedback on the model) formed the third of four data collection phases of the study.

5.5.1 Introduction

Interviewed RGLs (see Section 5.4: Results: One-on-one interviews) were informed about the initial Data Rescue Workflow Model during the conclusion of the interviews and asked whether they would be willing to participate in the next stage of the study. The interviewer supplied a brief description of the next stage and the type of feedback required. None of the interviewed RGLs expressed refusal regarding participation in the next phase. Following the interview, links to the initial model and accompanying documents were emailed to interviewed RGLs. RGLs were requested to view the model and its accompanying documents, provide feedback, and offer suggestions for amendments or alternatives to the model.

Additionally, RGLs were reminded that participation was voluntary, that feedback could be simplistic or detailed, and that providing feedback only on selected portions of the model or documentation was also permissible. Feedback could also be supplied in the format of their choice, be it an email, or electronically in the comments box of each web-based data rescue document, or by means of any virtual interview platform.

Invited RGLs were given a full calendar month to supply feedback. A reminder email was sent to RGLs four weeks after the original email was sent, with the deadline for feedback extended by one month.

The layout of findings with regard to feedback received adheres closely to the layout and chronology of the initial model (see Section 3.7: Initial Data Rescue Workflow Model: Description and characteristics). As such, the feedback section of the chapter starts with feedback on the model summary, followed by feedback on each of the nine main model stages, and concludes with general comments not assigned to a specific stage of the model. The feedback is preceded by details regarding the sample and responses.

5.5.2 Findings

This section contains the results obtained via the feedback provided by RGLs after they had viewed the initial model. The findings section is followed by a discussion of the resultant feedback provided by RGLs.

5.5.2.1 Sample and responses

Eight RGLs were invited to participate in this data collection stage of the study, comprising feedback on the initial Data Rescue Workflow Model, four of whom supplied feedback. For ease of continuance, respondents are referred to by the same pseudonym used in the previous section. Respondents involved in the feedback section were RGL1, RGL4, RGL5 and RGL6. Respondents who provided feedback had not previously stated (during the web-based questionnaire or interview) to have a more urgent need for data rescue than RGLs who declined to provide feedback.

RGL1 made use of email and Skype to provide feedback, while RGL4, RGL5 and RGL6 used email as a feedback tool.

All supplied responses have been documented in the sections below. Should a section not contain responses from any of the RGLs, it is due to no feedback being provided.

5.5.2.2 Feedback: Data Rescue Workflow Model summary

RGL6 stated that the summary overview was 'nice and helpful'.

5.5.2.3 Feedback: Stage 1 – Project initiation (and accompanying guidelines)

RGL1 submitted detailed feedback comprising concerns and problematic issues applicable to this section. The respondent's remarks are listed below.

RGL1 stated that locating data is not as difficult when one has any of the following:

- context,
- diverse data,
- unique data, and
- corporate memory.

The respondent elaborated on the issue of data location by stating that researchers working in an era of data abundance experience difficulties in locating and making sense of data, especially when a search mechanism does not allow for efficient data retrieval.

RGL1 also mentioned several concerns regarding difficulties experienced when trying to locate specific emails, and expressed dissatisfaction with the search capabilities of the institutional email system. While emails, under certain conditions, can certainly be used as research data, the respondent's concerns are more applicable to the area of records management. The topic of email searchability is not relevant to the topic at hand, and therefore not elaborated on here.

RGL1 further mentioned the following issues as problems which might surface during the execution of Stage 1 of the model, and when attempting to locate data:

- Data can be associated with a different user account.
- Data are not visible to the person trying to locate the data (due to access or viewing restrictions).
- The original data collectors, or people involved with the project, might have resigned or retired.
- It often takes a long time to find a specific document among thousands of similar documents.
- Metadata addition and categorisation did not adhere to best practices, resulting in data being difficult to trace and use in later years.

- Data might have been destroyed by a subsequent researcher, even though the original data collector stored it securely. Its value was not realised by a subsequent user.
- The different modalities of a project's electronic data, be it emails, ZIP files, text files or webpages, can complicate the storage and resultant locating of such data. RGL1 stated that not being able to locate data effectively and efficiently can hamper a planned data rescue project.

It should be noted that much of the feedback regarding Stage 1 of the model is not feedback on the correctness and usability of the model. While the supplied feedback has merit, it cannot be viewed as feedback relevant to the improvement of the model.

5.5.2.4 Feedback: Stage 2 – Storage and preservation (and accompanying guidelines)

Remarks regarding this stage of the initial Data Rescue Workflow Model are listed below.

- RGL1 stated that arranging paper-based media simply by date or topic may at times be difficult.
- RGL1 also remarked that it is important to store data in such a manner that large volumes of data are not daunting to a researcher trying to process the data.
- RGL4 required clarity on the existence of a localised, standardised archive space (i.e., storage space).

As was the case with feedback regarding Stage 1, much of the feedback regarding Stage 2 of the model is not feedback on the model, but are comments related to the general difficulties around managing data.

5.5.2.5 Feedback: Stage 3 – Create electronic inventory (and accompanying guidelines)

No stage-specific feedback was supplied by respondents.

5.5.2.6 Feedback: Stage 4 – Paper media imaging (and accompanying guidelines)

RGL1 described how choice of conversion format can affect future data access, as there are 'a lot of data not in a ubiquitous format'. The RGL also described related practices within their research group, and how switching from one repository system to another had resulted in data loss. The gist of RGL1's response is that it is important to ensure that the chosen final format (after imaging or conversion) does not result in information loss, and that the selected format is able to stand the test of time.

RGL1 also provided feedback on the time-dependency of data formats and explained that tapes had been used to create backups, but that the tape drive had to be replaced due to failure. Unfortunately, the replacement drive also ended up malfunctioning. This scenario resulted in tapes not being accessible to the research group at all. A similar scenario, but linked to external hard drives, was described. The RGL stated that their hard drives only worked on extremely old USB interfaces, and that ensuring the accessibility of their electronic data had proven to be difficult.

As was the case with the previous stage's feedback, much of the feedback regarding Stage 4 of the model was not related to the model, but instead pertained to general difficulties around managing data.

5.5.2.7 Feedback: Stage 5 – Digitisation of data values (and accompanying guidelines)

RGL1's remarks applicable to media imaging, and the possible effects of format choice (see Section 5.5.7: Stage 4: Paper media imaging) are also relevant to this section.

5.5.2.8 Feedback: Stage 6 – Describing the data (and accompanying guidelines)

RGL1's feedback posted in a previous section (Section 5.5.2.3: Stage 1: Project initiation), regarding description of data, can also be applied to this section.

5.5.2.9 Feedback: Stage 7 – Make data discoverable (and accompanying guidelines)

A range of issues, applicable to data discoverability, were raised by respondents.

- RGL4 sought clarity regarding the storage location of data.
- RGL4 sought clarity regarding the existence of a localised archive space storage location of data.
- RGL4 sought clarity regarding the existence of a standardised archive space.
- RGL4 sought clarity regarding the accessibility of such a storage space to all [*institute*] staff.
- RGL4 asked whether the storage location would entail a cloud-based platform.
- RGL5 mentioned that there should be an exit point for confidential data; that data should not be discoverable outside the [*institute*].
- RGL1 remarked that their group had experienced a case where researchers had left the institute, and the group was unable to recover or restore the data from the group's repositories.
- RGL1 also emphasised the importance of this stage and mentioned that it is often necessary to describe data in higher detail than what is initially thought.
- RGL1 stated that data are often not described in sufficient detail, as contractual agreements do not state it as a requirement.
- RGL1 further commented on the difficulty in deciding which metadata and data documentation needed to be captured to ensure future use.
- RGL1 also described the difficulty in deciding at what point information is seen to be captured sufficiently.

- Lastly, RGL1 summarised the difficulty of this stage by remarking that choices to be made at this stage are not easy, as resources might not be available, or the importance of data description is underestimated.

As was the case with feedback regarding the previous stages, much of the feedback regarding Stage 7 of the model is not feedback on the model, but are comments related to the general difficulties around managing data.

5.5.2.10 Feedback: Stage 8 – Archive the data

The aspects listed below were stage-applicable issues raised by respondents.

- RGL4 sought clarity regarding the existence of a localised archive space.
- RGL4 sought clarity regarding the existence of a standardised archive space.
- RGL1's feedback regarding format obsolescence, mentioned earlier in the Stage 4 feedback segment, is also applicable in this section on data archiving and data migration.
- RGL1's feedback regarding formats being time-dependent, mentioned earlier in the Stage 4 feedback segment, is also applicable in this section on data archiving and data migration.

As was the case with feedback regarding the previous stages, the feedback regarding Stage 8 of the model is not feedback on the model, but are comments related to the general difficulties around managing data.

5.5.2.11 Feedback: Stage 9 – Project closure

No stage-specific feedback was supplied by respondents.

5.5.2.12 General feedback

The **positive comments** listed below were made regarding the initial model as a whole.

- 'The model is well formulated.' (RGL4)
- 'The model is detailed.' (RGL4)
- 'The model is beautiful and elegant.' (RGL1)
- 'The model is very clear.' (RGL1)
- 'The model is concise in terms of the process flow.' (RGL1)
- 'The flow chart covers most of what needs to be done.' (RGL5)
- 'The diagrams relating to each phase are clear and make the process easy to understand. All stage diagrams make sense to me.' (RGL6)
- 'No problems were seen with the process itself.' (RGL1)

The **concerns** listed below relate to the model in its totality.

- The model was described as long (RGL4).
- The model was described as complex (RGL4).
- It was stated that it might prove difficult to implement the model (RGL4).
- It was mentioned that researchers are under time constraints and have enormous workload (RGL4).
- One of the respondents stated that there is a lack of available data rescue resources (RGL4), while another mentioned that there is a lack of resources to convert data at risk to an electronic format that can be archived digitally (RGL5).
- The hope was expressed that data rescue be performed by the institute's library services, instead of SET staff battling with workload (RGL4).
- Concern was expressed that researchers might shy away from data rescue activities (RGL4).
- The model's underlying tasks and activities were described as scary and complex (RGL1).
- Data rescue problems are not necessarily linked to issues with technical 'stuff', such as formats (RGL1).
- RGL1 stated that some of the tasks imply significant efforts, costs and time.
- One of the respondents stated that it is possible that some of these tasks will fail, due to their complexity (RGL1).
- One of the respondents stated that data rescue would need to be funded (RGL1).
- One of the respondents stated that a data checklist that is ticked off might not necessarily indicate that data are no longer at risk, as lack of resources, insight or capabilities might hinder best practices (RGL1).
- It was mentioned that researchers might underestimate the importance of data, and not think it might be that useful in future (RGL1).
- One of the respondents stated that not all data can realistically be saved (RGL1).
- It was stated that the inability to apply best data rescue practices now could cost the group dearly in future (RGL1).

While the above concerns were plentiful, it must be pointed out that most of these concerns are not relevant to the validity of the model, and do not add anything to how the model could be revised/improved.

Suggestions for an **amended or improved Data Rescue Workflow Model** (seen as a whole) were also put forward, and are listed below.

- The model includes paper-based media only; RGL6 suggested that more formats be addressed.
- The model should clearly show how confidential data should be treated (RGL5).

Additionally, suggestions for extensions to the model were also forthcoming, and are listed below.

- RGL4 suggested that a few standard data rescue templates be formulated, as this would result in the data rescue process being easier to follow, and easier to complete.
- RGL6 suggested that the model should indicate which standards are applicable when storing the converted data.
- RGL6 stated that the model should indicate which standards are applicable when digitising data.
- It was unclear whether the entire model would be implemented, or only certain stages or activities (RGL4).
- RGL5 expressed concerns regarding implementation of the model, citing lack of resources. The respondent mentioned that they would be interested in participating in a future data rescue case study rolled out by this researcher.

The next section comprises a discussion of feedback received from RGLs.

5.5.3 Discussion

This section contains a discussion of feedback received from RGLs after they had reviewed the initial model. The discussion also indicates which suggestions will lead to an improvement of the initial model and will be used when creating a revised model. Discussion on feedback received will feature under the headings of concerns, improvements and extensions.

5.5.3.1 Concerns

The most crucial concern received refers to the fact that the model focused on the rescue of paper-based data only. With respondents indicating via the web-based questionnaire and the in-person interview that data at risk entailed various format (see Figure 2 and Table 4), the importance of this concern is evident.

Concern was also expressed regarding the treatment of confidential data: it was mentioned that there should be an exit point for confidential data, as this data should not be discoverable outside the [*institute*]. This issue is of particular concern in the steps involving data sharing to a repository, long-term preservation of data, and during project closure when promotion of rescued data will occur. Confidential data, and data forming part of contractual research, are usually subject to sharing restrictions, and it is therefore vital that the data sharing and data promotion are not seen as a mandatory data rescue activity.

Another concern raised, and one that will be implemented into the revised version of the model, related to lack of clarity about the number of stages and activities forming part of data rescue. It was not clear to RGLs if the whole model, or only certain stages and activities, would feature during data rescue.

Requests for standardised templates and additional guidelines, coupled with hesitancy regarding the implementation and uptake of the model, suggest that respondents are not yet confident in their understanding of data rescue. Adding to the hesitancy of RGLs were the stated concerns regarding the model's complexity and length.

Concerns regarding the manpower, time and effort required by researchers was a common theme, and was rooted in the available research complement of many research groups. These aspects include dealing with work pressure to meet research deadlines, prioritisation of certain key performance areas, the 'publish or perish' scourge, contractual obligations, and the quest for funders or clients. In addition, aspects such as institutional restructuring, the forming of new groups, disbanding of previous groups, and freezing of non-critical positions have left researchers with limited opportunity to consider new activities including data rescue. While these concerns are not linked to the model, they do indicate the need for a realistic and feasible view of data rescue, and the vital role that the non-SET component can play in contributing towards data rescue activities. These aspects are also addressed in the applicable sector of the Recommendations chapter (see Section 6.4.8, Section 6.4.9, Section 6.4.10, Section 6.4.11 and Section 6.4.13).

With the feedback stage taking place during the country's strict lockdown period, COVID-19 restrictions and the lockdown demands can also be viewed as a factor inhibiting acceptance of new roles and responsibilities. At the time of data collection, many scientists were still working from home, or under strict protocol (implying non-ideal work conditions) on work premises.

Another issue mentioned by respondents, relevant to the practice of data rescue and not the model per se, relates to the availability of the necessary resources (other than human resources) to conduct data rescue. This concern about the required equipment, infrastructure and tools was not entirely unexpected. Similar concerns were also expressed during the in-depth interviews with respondents, which preceded the distribution of the initial data rescue model. Respondents mentioned that their groups did not have the indicated data rescue equipment, such as scanners, digital cameras, archival boxes, or funds for long-term preservation.

The reasons listed above are all issues estimated to affect the views of respondents regarding the initial data rescue model, and also the new and additional roles required from researchers should the model, or data rescue activities be implemented.

5.5.3.2 Improvements

Taking into consideration the primary concern expressed (i.e., the model including paper-based data only), the main amendment to the model would be the inclusion of different data formats, and the description of the specific steps unique to different formats.

Another area of improvement pertains to confidential data, and including within the model the data rescue steps applicable to such data. Data rescue stages and activities involving confidential data include the stage linked to the sharing of data to a repository, long-term preservation in a data archive, and project closure when rescued data are usually promoted. The model would need to indicate that not all data are shareable, and that not all rescue projects and their linked data should be publicised when no longer at risk.

With concerns raised about the number of stages required to rescue data, it was considered vital that the revised model include two levels of data rescue. As a result, the revised model made provision for 'full rescue' and 'partial rescue'. These concepts, created by this researcher during this stage of the study, are explained in Table i: Clarification of key terms. It became clear that not all data rescue efforts would be able to implement all rescue stages and activities included in the initial model, and that factors such as lack of resources might make 'partial data rescue' a more viable option. The updated model would therefore indicate when 'full rescue' would be required or possible, and also the conditions more suited to 'partial rescue'.

The requests of respondents regarding guidelines and templates have necessitated the need for these appendices to form part of a data rescue workflow model. As a result, the revised model would include the following guidelines:

- guidance on data assessment,
- guidance on data rescue project planning,
- guidance on data management plans,
- guidance on storage of paper data,
- guidance on archives and SHEQ,
- guidance on storage of physical samples and specimens,
- guidance on digitisation of paper data,
- guidance on the use of metadata, and
- guidance on use of data repositories.

In addition, sample templates for the creation of data inventories would also feature as a workflow appendix.

Concerns about the model complexity showed the need for a simple and user-friendly data rescue model. Linked concerns about the length of the model or its complexity indicated that novice data rescuers should have access to a descriptive summary of the model. Such a summary should be visible prior to consulting the more detailed stage-specific graphics of the model. To meet this need of researchers, the revised model would include a compact summary containing enough detail to briefly describe each stage, without cluttering the image or overwhelming the data rescue novice. The model should also contain stage-specific graphics with more details and clarification to meet the needs of rescuers preferring separate graphics.

Additional crucial enhancements to the model would include the following:

- a data assessment activity, and its linked guidance,
- a data destruction activity, and
- the use of distinct colours and shapes to indicate roles and activity types.

A full list of differences between the initial model and the revised model can be found in Table 5.10: Comparison between initial model and revised model.

5.5.3.3 Extensions

While already mentioned in Section 5.5.3.2, the revised model should include the following added features:

- guidance documents and sample templates, and
- a model version indicating full rescue, and a model version indicating partial rescue.

In addition, the model should consist of a compact summary, an extended summary, and at least one image for each of the data rescue stages.

A full list of differences between the initial model and the revised model can be found in Table 5.10: Comparison between initial model and revised model.

The next section offers conclusions reached after examining the feedback findings, and includes a listing of resultant implications.

5.5.3.4 Conclusions and implications

Despite the varied feedback received from researchers regarding the initial model, it became evident that additional steps or initiatives were required to compensate for the feedback from a small sample. As the feedback received was limited and did not provide sufficient critical input to extensively revise the model, it would be necessary to embark on an additional data collection method, entailing one which would require examination of the initial model by an additional sample. Sample members were anticipated to provide additional critical feedback enabling revision of the initial model. The decision was made to host a mini focus group session, with research library experts from the selected institute being focus group participants. The mini focus group, its contribution and findings are detailed in Section 5.7 of this chapter.

In addition, the academic review emanating from discussions around the model also pointed out several shortcomings of the first model and made many suggestions for improvements. Suggested improvements were incorporated in the list of implications below and in the revised model (Section 5.6), as well as in the final recommended model (Section 6.2.1.7).

Implications emanating from the feedback

- The most important suggestion emanating from the feedback received proposes that the model cater for more than paper-based data. As revealed via the web-based questionnaire and in-person interview stages, many research group have data at risk that are not paper-based; these media should feature in a revised version of the initial model.
- With a large part of the research at the selected institute being contractual in nature, and data not readily shareable, it is vital that a revised model indicates the steps for data differing in openness. To add to this: researchers should be made aware that even though data rescue commonly entails the upload of data to a public repository, the nature of research at the selected institute means that data rescue will not necessarily include this step.
- Other suggestions refer to minor tweaking of the model only.
- Feedback received from RGLs, although helpful in some aspects of model revision, was at times unrelated to the model, or to research data.
- The feedback received regarding the initial model was not as detailed or critical as anticipated. It also did not meet the needs of this study regarding critical input, and recommendations for a workable data rescue model.
- However, after reviewing all feedback received, and combining it with relevant input obtained during other data collection stages, it became apparent that the initial model needed revision.
- It was clear that additional feedback would need to be obtained via other institutional parties.
- The decision was made to obtain feedback and input from institutional experts who were not directly involved with research activities. Such a selection of experts would provide a different yet informed perspective on data rescue, and a data rescue workflow model.
- The data collection method used, the group of experts, and the findings of the data collection method (mentioned in the previous bullet) are discussed, respectively, in Section 4.5.4: Mini focus group session, Section 4.7.3: Sample C: Institutional experts (not SET based), and Section 5.7: Results: Mini focus group session.

5.5.4 Summary of feedback section

This segment of the results chapter contained the feedback supplied by selected RGLs after they had studied the initial Data Rescue Workflow Model. This segment also featured comments made regarding activities and tools related to data rescue, and comments related to the model's underlying activities (e.g., comments regarding difficulty in locating data are related to the model's Stage 1: Project initiation).

In short, general concerns related to the model itself referred to the following aspects:

- the fact that the model focuses on paper-based data,
- the lack of information regarding rescue of confidential data,
- the complexity of the model or some of the steps, and
- the length of the model.

These concerns were directed at the information contained within the images featuring the separate rescue stages and its accompanying appendices.

The model's single page summary was favourably received.

Concerns raised are vital and expected: the model comprises several steps and many of the activities were new to researchers, or had never been performed by them. The concern was also linked to researchers who feared that they would be the sole participants in data rescue projects, that there would be no or minimal training, and that data rescue would be an institutional requirement.

Feedback also indicated that a data rescue model should provide guidance on the following topics:

- locating data prior to a data rescue project,
- assessing the data prior to initiation of a data rescue project,
- adding context to the data in the absence of original data creators,
- creation of metadata and data documentation,
- obtaining safe and secure storage space or storage locations,
- rescuing data that are confidential in nature,
- ensuring rescued data are accessible to more than one person,
- dealing with encryption requirements from storage providers, and
- dealing with future format obsolescence.

Apart from the above suggestions regarding guidance, it was also proposed that the model could benefit from the inclusion of standard templates, and an indication of standards applicable to the relevant data rescue stages or activities.

While not related to the model, feedback received from Sample B has indicated that researchers are:

- apprehensive regarding the additional workload a data rescue project might cause,
- concerned that data at risk might not be located, understood, or be rescuable,
- concerned about the lack of data rescue skills, and
- concerned about the cost of data rescue implementation.

Whilst being helpful in some aspects of model revision, much of the feedback received from RGLs was unrelated to the model or to research data. The feedback received regarding the initial model was not as detailed and critical as anticipated, and did not meet the needs of this study regarding critical input and recommendations for a workable data rescue model.

After reviewing all feedback received and combining it with relevant input obtained during other data collection stages, it was evident that the initial model needed revision. The revision of the model and the changes implemented are discussed in the next section.

5.6 Revising the initial Data Rescue Workflow Model

This section describes the reasons leading up to the revision of the initial model, and the changes made to the model.

5.6.1 Rationale of a revised Data Rescue Workflow Model

Two main factors contributed to the initial model deemed in need of revision and amendment:

- Feedback received from Sample B had necessitated several amendments to the initial model.
- In-depth discussions during the academic review stage brought about a realisation that major changes to the initial model were required. A period of nine months had passed between the date of initial model creation, and the date of analysis of the Sample B model feedback. This passage of time had included discussions with academic peers, leading to an increased understanding of data at risk, data rescue activities, and the steps which should be included in a data rescue workflow model.

The importance of revising the model prior to the next data collection stage (a mini focus group session with Sample C, see Section 4.5.4: Mini focus group session) was to ensure that Sample C would be able to comment on both the initial and the revised model.

Evaluating researcher feedback and considering the limitations of the model (based on insight gained) resulted in a number of features earmarked for inclusion in a revised version of the model. The features considered vital for inclusion in the revised model, missing from the initial model, are listed below.

- The model should include data at risk formats other than paper-based media.
- The model should also describe how to treat data that are not involved in all data rescue stages, e.g., how to treat certain paper-based data that will not be digitised but are at risk.
- The model should stipulate how to handle data that are found to be not fit for rescue, and that would no longer be stored and preserved.
- The model should describe how to treat data that are found to be fit for eventual rescue, but cannot be rescued immediately.
- The placement of the master data rescue inventory step should feature earlier in the model.
- There should be simplification of activities and terminology regarding the different data inventories.
- The model should contain a clear indication of roles and responsibilities.
- The model should ideally indicate the areas of involvement of the institutional section/parties involved with data rescue.
- The model should provide a legend explaining all shapes, colours and symbols.
- The model should provide guidance for the rescue of modern electronic data currently not in a preservation format.
- The ambiguity around the terms ‘imaging’ and ‘digitising’ should be addressed.
- The model should indicate where and how outsourcing of data rescue can be implemented.
- The model should describe the data rescue steps applicable to the rescue of confidential data.
- The shapes used in the diagrams should clearly indicate different activities, steps, tasks and outputs.
- The use of colour in the diagram to indicate distinct roles and responsibilities should be considered.

Furthermore, it was also considered important to implement the following changes:

- the renaming of certain stages, and the use of more suitable terminology for stage headings,
- replacing ambiguous terminology found elsewhere in the model with more suitable terms,
- combining data rescue steps (diagrammatically) where possible to minimise confusion and ‘shape’ overload, and
- creating a compact summary diagram, as well as an extended summary diagram.

Aspects mentioned in the two bulleted lists above were evaluated for inclusion in the revised model.

5.6.2 Characteristics of the revised Data Rescue Workflow Model

Differences between the initial and revised model are listed in the table below.

Table 5.10: Comparison between initial model and revised model

INITIAL DATA RESCUE WORKFLOW MODEL	REVISED DATA RESCUE WORKFLOW MODEL
Model consisted of a summarised image, and an image for each of the data rescue stages	Model consisted of a compact summary, an extended summary, and at least one image for each of the data rescue stages
Model consisted of nine stages	Model consisted of eight stages
Stages are: • Project Initiation • Storage and Preservation • Create Inventories • Imaging of Media • Digitisation of Media • Describing the Data • Making Data Discoverable • Archive the Data • Project Closure	Stages are: • Data Rescue Preparatory Stage • Data Rescue Planning • Data Storage and Preservation • Digitisation • Documenting the Data • Data Sharing • Long-Term Preservation • Project Closure
Model assumes data are paper based	Model includes different data formats
Model describes rescue of paper data	Model describes rescue of paper data, early digital data, modern digital data, and samples
Model describes full data rescue only	Model describes full and partial data rescue
Model does not indicate roles of institutional parties	Model indicates roles of institutional parties
Only one option per stage is described	Different options per stage are described
Creation of inventories is seen as a separate stage	Inventories form part of other stages
Data destruction is not addressed	Data destruction forms part of the model
Data assessment is understated	Data assessment features strongly
Data imaging and digitising of paper data values are separate stages	Data imaging and data digitising are included in the same stage
Colours are used to distinguish between activities and outputs	Colours are used to indicate responsibly parties
Model requires minimal shape explanation	Model requires explanation of range of shapes and colours
Model does not include outsourcing option	An outsourcing option forms part of the model

The revised model consisted of:

- a compact summary (Figure 5.5; also Section 5.6.3: Compact summary),
- an extended summary (Figures 5.8, 5.9, 5.10 and 5.11; also Section 5.6.5: Extended summary), and
- separate diagrams for each of the eight data rescue stages (Figure 5.12 through to Figure 5.27, also Section 5.6.6: Stage-specific sections of model).

The decision to draft two different summaries was because the summaries would have different objectives. The compact summary consisted of a single-page diagram and provided an overview of the main data rescue stages, without delving into roles, responsibilities, decisions, outputs, or rescue sub-activities. The extended summary contained more detail than the compact summary, and included indications of full and partial rescue, different data formats, and decisions made during the rescue process.

In addition, the model contained a separate figure for each of the eight main data rescue stages. As the model caters for full rescue as well as partial rescue, some of the stages featured a figure for the partial data rescue activities of the stage, and a figure displaying the full rescue activities of the stage. An example of such an occurrence is the data rescue planning stage, showing separate figures for full data rescue and partial data rescue.

Each of the model's sections is described in more detail in the remainder of this section.

5.6.3 Compact summary

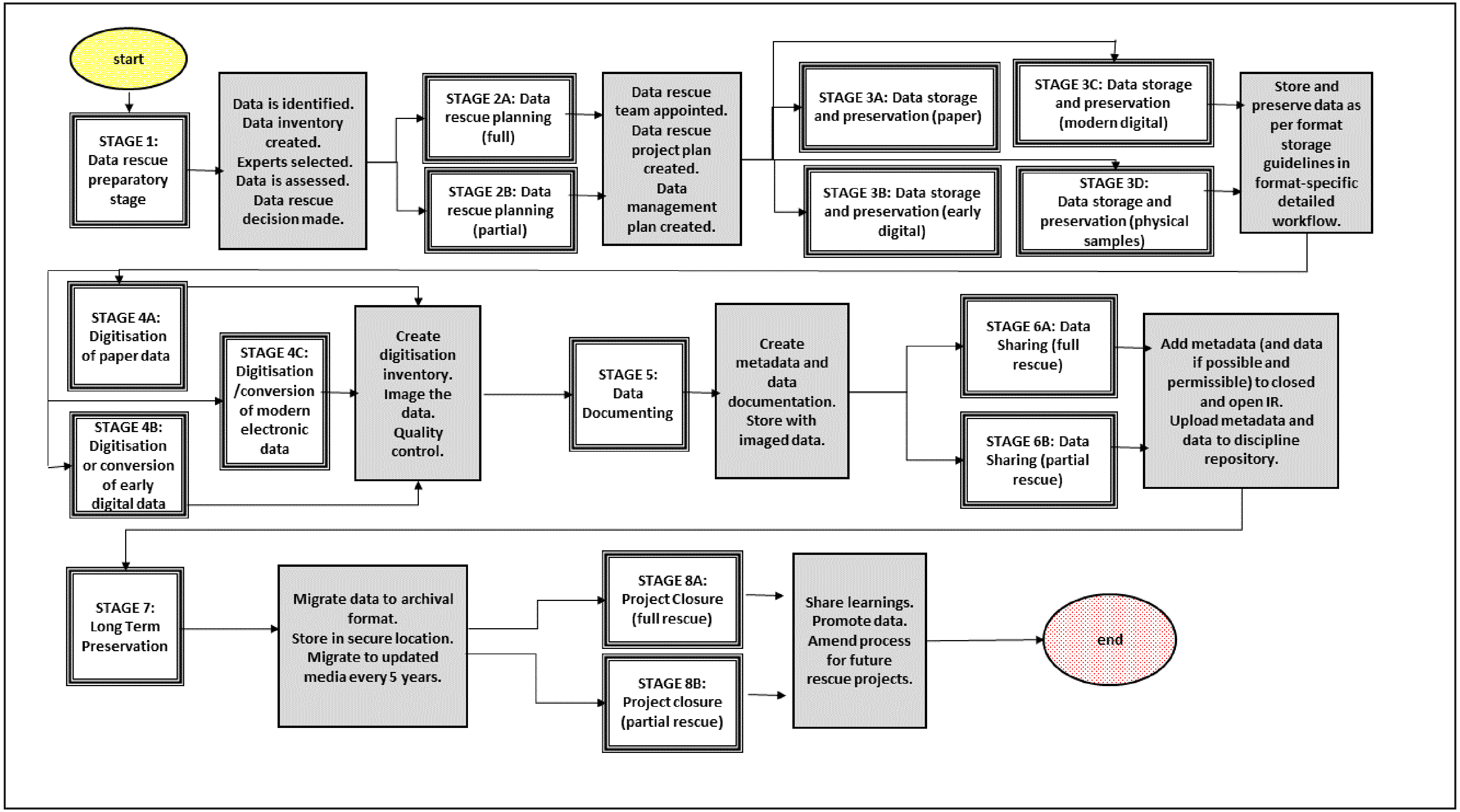
The compact summary portrays the main data rescue stages of the revised model, and the main stages forming part of a data rescue project. The compact summary can be viewed in Figure 5.5: Compact summary of revised Data Rescue Workflow Model.

While showing some similarities to the summary image of the initial model (see Section 3.8), the summaries are not identical. Major differences include the following:

- the compact summary forming part of the revised model displays eight stages, while the summary of the initial model shows nine stages,
- the compact summary contains brief description of the main activities forming part of the various stages, while the summary of the initial model only features the name of the stage,
- the compact summary indicates that two types of data rescue may be performed, namely full rescue and partial rescue, while the summary of the initial model assumes all rescue will include all steps of the model,
- the compact summary indicates that different data formats are involved, while the summary of the initial model does not stipulate the data formats included in the model, and
- the compact summary minimises the use of colour, adhering to a monochromatic look, while the summary of the initial model used colours in random fashion.

The compact summary is directed towards data rescue novices and provides users of the revised model with a one-page overview of the model, and its involved rescue stages. Being a very brief portrayal of the process, it does not supply details on roles and responsibilities, vital decisions to be made, templates to be used, or outputs to be created. The compact summary is not intended to portray workflow; instead, it is a diagram encompassing the main stages of the actual workflow model.

Once the compact summary has been viewed and understood by the novice data rescuer, the extended summary should be studied. The extended summary is discussed in Section 5.6.5, and follows the next section, containing the legend to the model.



start
STAGE 1: Data rescue preparatory stage
Data is identified. Data inventory created. Experts selected. Data is assessed. Data rescue decision made.
STAGE 2A: Data rescue planning (full)
STAGE 2B: Data rescue planning (partial)
Data rescue team appointed. Data rescue project plan created. Data management plan created.
STAGE 3A: Data storage and preservation (paper)
STAGE 3B: Data storage and preservation (early digital)
STAGE 3C: Data storage and preservation (modern digital)
STAGE 3D: Data storage and preservation (physical samples)
Store and preserve data as per format storage guidelines in format-specific detailed workflow.
STAGE 4A: Digitisation of paper data
STAGE 4B: Digitisation or conversion of early digital data
STAGE 4C: Digitisation /conversion of modern electronic data
Create digitisation inventory. Image the data. Quality control.
STAGE 5: Data Documenting
Create metadata and data documentation. Store with imaged data.
STAGE 6A: Data Sharing (full rescue)
STAGE 6B: Data Sharing (partial rescue)
Add metadata (and data if possible and permissible) to closed and open IR. Upload metadata and data to discipline repository.
STAGE 7: Long Term Preservation
Migrate data to archival format. Store in secure location. Migrate to updated media every 5 years.
STAGE 8A: Project Closure (full rescue)
STAGE 8B: Project closure (partial rescue)
Share learnings. Promote data. Amend process for future rescue projects.
end

Figure 5.5: Compact summary of revised Data Rescue Workflow Model

5.6.4 Legend to the revised Data Rescue Workflow Model

The two images included in this section comprise a legend to the extended summary, and a legend to the stage-specific stages.

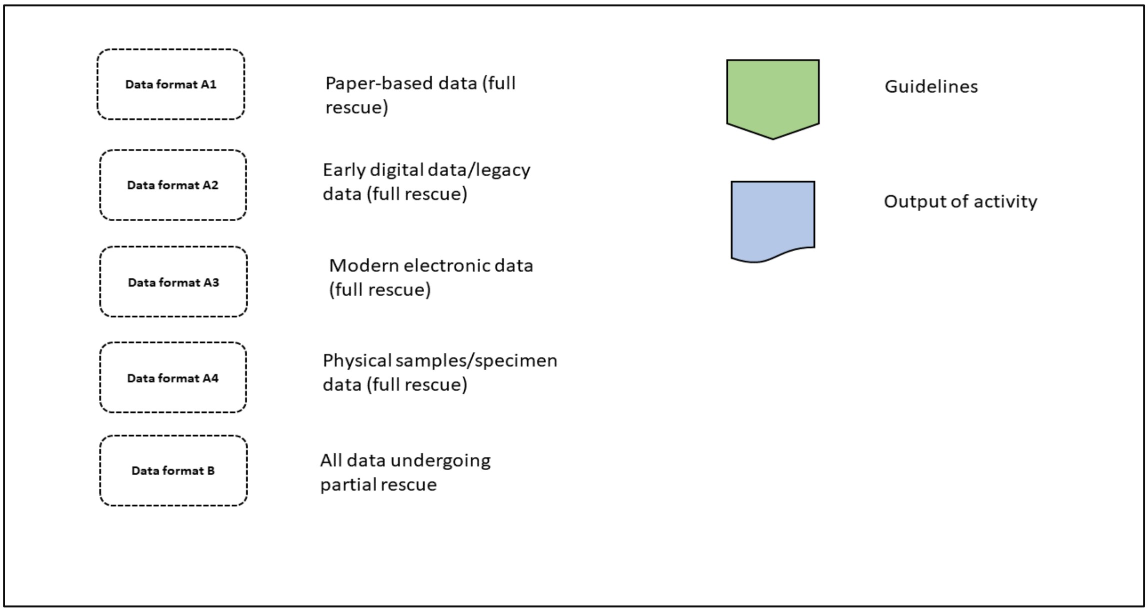

Data format A1
Paper-based data (full rescue)
Data format A2
Early digital data/legacy data (full rescue)
Data format A3
Modern electronic data (full rescue)
Data format A4
Physical samples/specimen data (full rescue)
Data format B
All data undergoing partial rescue
Guidelines
Output of activity

Figure 5.6: Legend of revised Data Rescue Workflow Model (a)

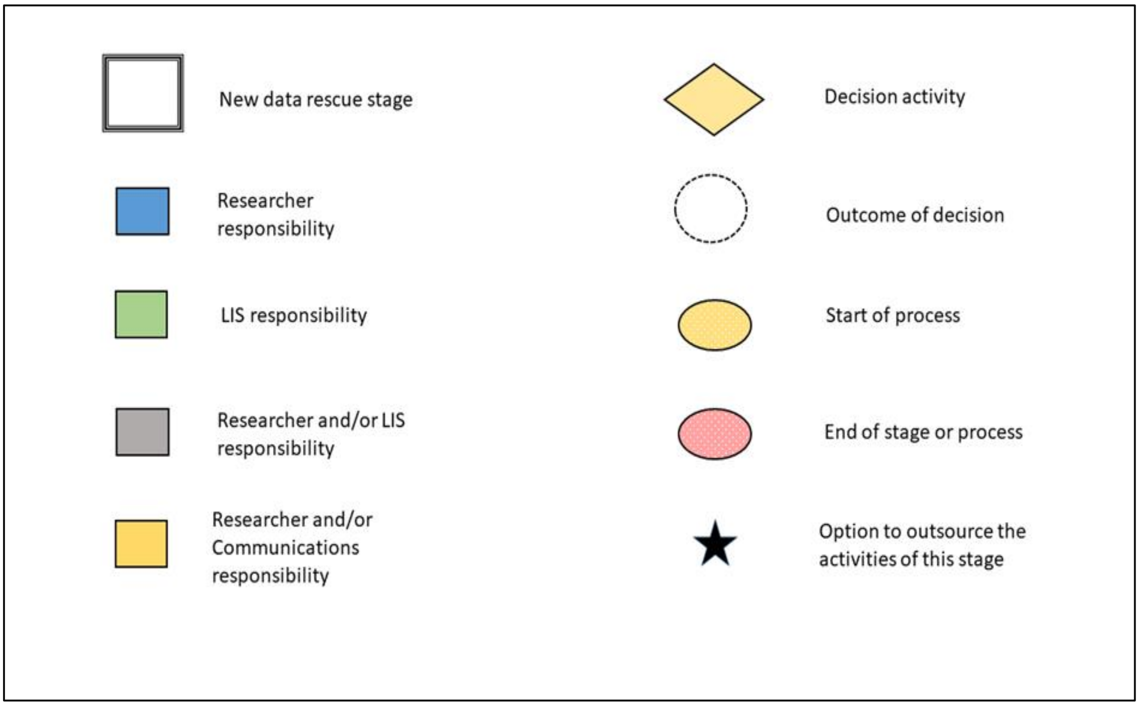

New data rescue stage
Researcher responsibility
LIS responsibility
Researcher and/or LIS responsibility
Researcher and/or Communications responsibility
Decision activity
Outcome of decision
Start of process
End of stage or process
Option to outsource the activities of this stage

Figure 5.7: Legend of revised Data Rescue Workflow Model (b)

Stage-specific stages are described in a subsequent section.

5.6.5 Extended summary

The extended summary (see Figures 5.8, 5.9, 5.10 and 5.11), forming part of the revised model, portrays the main stages, activities, decisions and outputs of the proposed data rescue workflow. The extended summary is also suited to hardcopy format and can be accommodated by an A3 size paper.

Unlike the compact summary, the extended summary does not have its equivalent in the initial model. It was created to provide more than brief information about the data rescue stages involved, data formats included, partial and full rescue option and steps, outputs to be created, and different rescue tasks related to different data formats.

For the purposes of this model, full data rescue can be described as the rescue of data involving all the data rescue workflow model stages. Partial data rescue, on the other hand, entails the rescue of data where only certain of the rescue steps included in the model can be performed. Reasons for opting to perform partial rescue can include lack of resources, or lack of metadata or data documentation. An example of partial data rescue is the following scenario:

- a collection of historic paper-based data being discovered,
- data were determined to have metadata and data documentation; alternatively, this can be added by subject experts,
- data were assessed and found to be valuable and of use to other researchers,
- there are insufficient resources and infrastructure for the data to be digitised,
- metadata and data documentation can be uploaded to a repository, and
- the paper-based data are stored and preserved according to best practices.

In the above scenario, only certain data rescue stages forming part of the revised model would be performed, and this data rescue venture, consisting of certain activities only, is referred to as partial data rescue.

It is likely that the extended model will suffice as guidance for many data rescue projects or teams. While the extended summary does give an indication of roles and responsibilities, such as showing the role of the research library services during the 'Data sharing' stage, it is not the main contribution of the extended summary. This summary is meant to serve as guidance for rescue projects where teams prefer that role indication be up to them, or where the homogenous nature of rescue teams does not lend itself to role allocation based on regular professional duties.

As the goal of the extended summary is not to indicate roles and responsibilities, data rescue teams making use of the model have the added option of accessing and viewing stage-specific images (see Sections 5.6.6.1 up to Section 5.6.6.8). The stage-specific parts of the model are described next.

5.6.6 Stage-specific sections of model

The stage-specific graphics provide more detail than was portrayed in the compact or extended summary of the model. Features forming part of the stage-specific portrayals are as follows:

- Stage-specific graphics are intended to indicate workflow; the summaries, by contrast, are not workflows but informative data rescue diagrams.
- Roles and responsibilities are colour-coded; Figure 5 displays and clarifies the colours used in the model.
- The roles and responsibilities of SET-based staff, research library professionals, and shared responsibility are included in the legend.
- Stage-specific graphics provide format-specific details. As such, there are format-specific schemas for each of the rescue stages, should the stage not be generic in nature. Examples of format-specific schemas include Stage 3A: Data storage and preservation (paper data), Stage 3B: Data storage and preservation (early digital data) and Stage 3C: Data storage and preservation (modern electronic data). Conversely, Stage 5: Documenting the data, is a generic stage, not format-specific, and does not have a separate graphic for the different formats.

Each of the eight data rescue stages is briefly discussed in Section 5.6.3.1 up to Section 5.6.3.8.

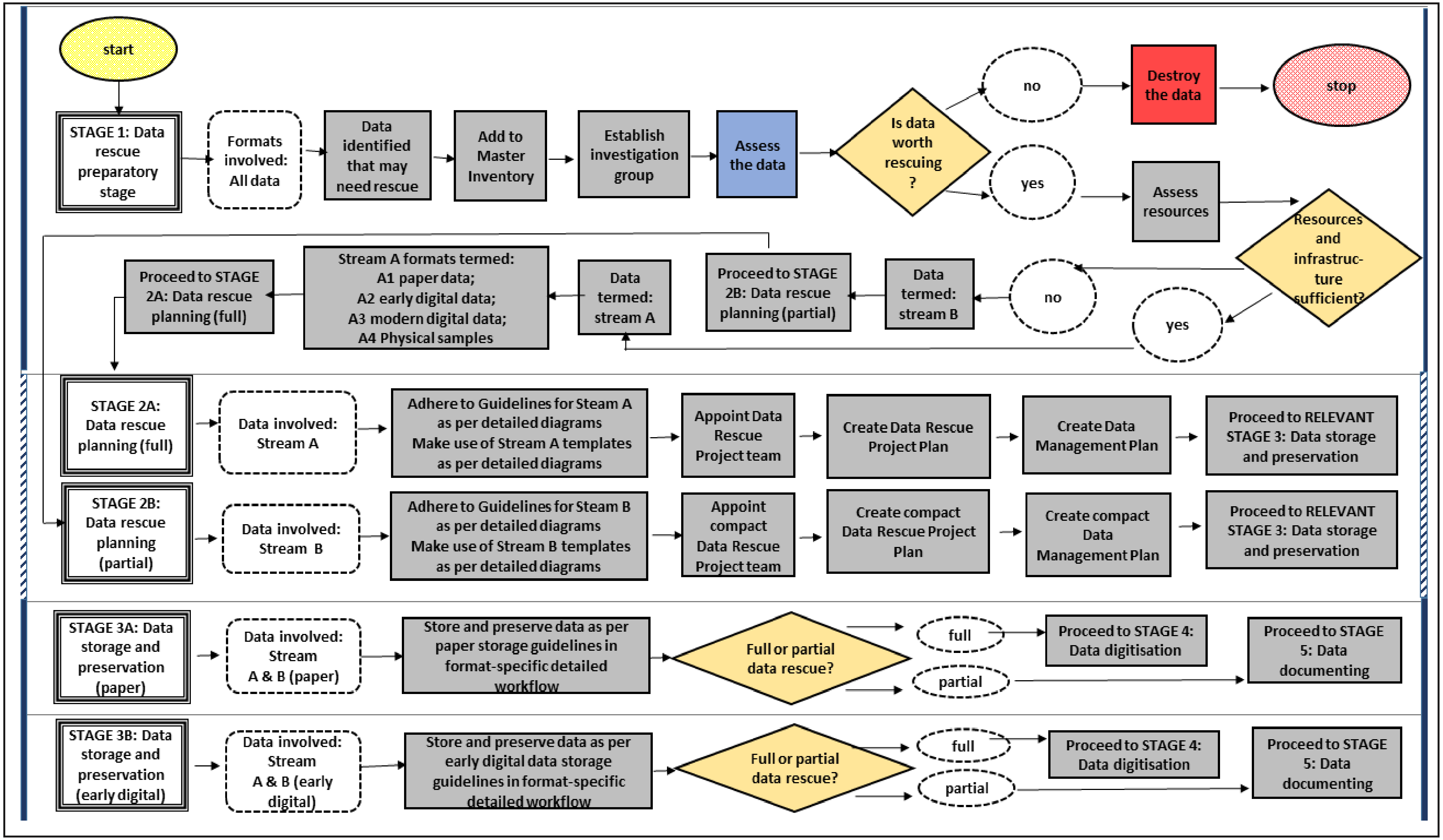

start
STAGE 1: Data rescue preparatory stage
Formats involved: All data
Data identified that may need rescue
Add to Master Inventory
Establish investigation group
Assess the data
Is data worth rescuing ?
no
Destroy the data
stop
yes
Assess resources
Resources and infrastruc-ture sufficient?
no
yes
Data termed: stream B
Proceed to STAGE 2B: Data rescue planning (partial)
Data termed: stream A
Stream A formats termed: A1 paper data; A2 early digital data; A3 modern digital data; A4 Physical samples
Proceed to STAGE 2A: Data rescue planning (full)
STAGE 2A: Data rescue planning (full)
Data involved: Stream A
Adhere to Guidelines for Steam A as per detailed diagrams Make use of Stream A templates as per detailed diagrams
Appoint Data Rescue Project team
Create Data Rescue Project Plan
Create Data Management Plan
Proceed to RELEVANT STAGE 3: Data storage and preservation
STAGE 2B: Data rescue planning (partial)
Data involved: Stream B
Adhere to Guidelines for Steam B as per detailed diagrams Make use of Stream B templates as per detailed diagrams
Appoint compact Data Rescue Project team
Create compact Data Rescue Project Plan
Create compact Data Management Plan
Proceed to RELEVANT STAGE 3: Data storage and preservation
STAGE 3A: Data storage and preservation (paper)
Data involved: Stream A & B (paper)
Store and preserve data as per paper storage guidelines in format-specific detailed workflow
Full or partial data rescue?
full
partial
Proceed to STAGE 4: Data digitisation
Proceed to STAGE 5: Data documenting
STAGE 3B: Data storage and preservation (early digital)
Data involved: Stream A & B (early digital)
Store and preserve data as per early digital data storage guidelines in format-specific detailed workflow
Full or partial data rescue?
full
partial
Proceed to STAGE 4: Data digitisation
Proceed to STAGE 5: Data documenting

Figure 5.8: Extended summary of Data Rescue Workflow Model (a)

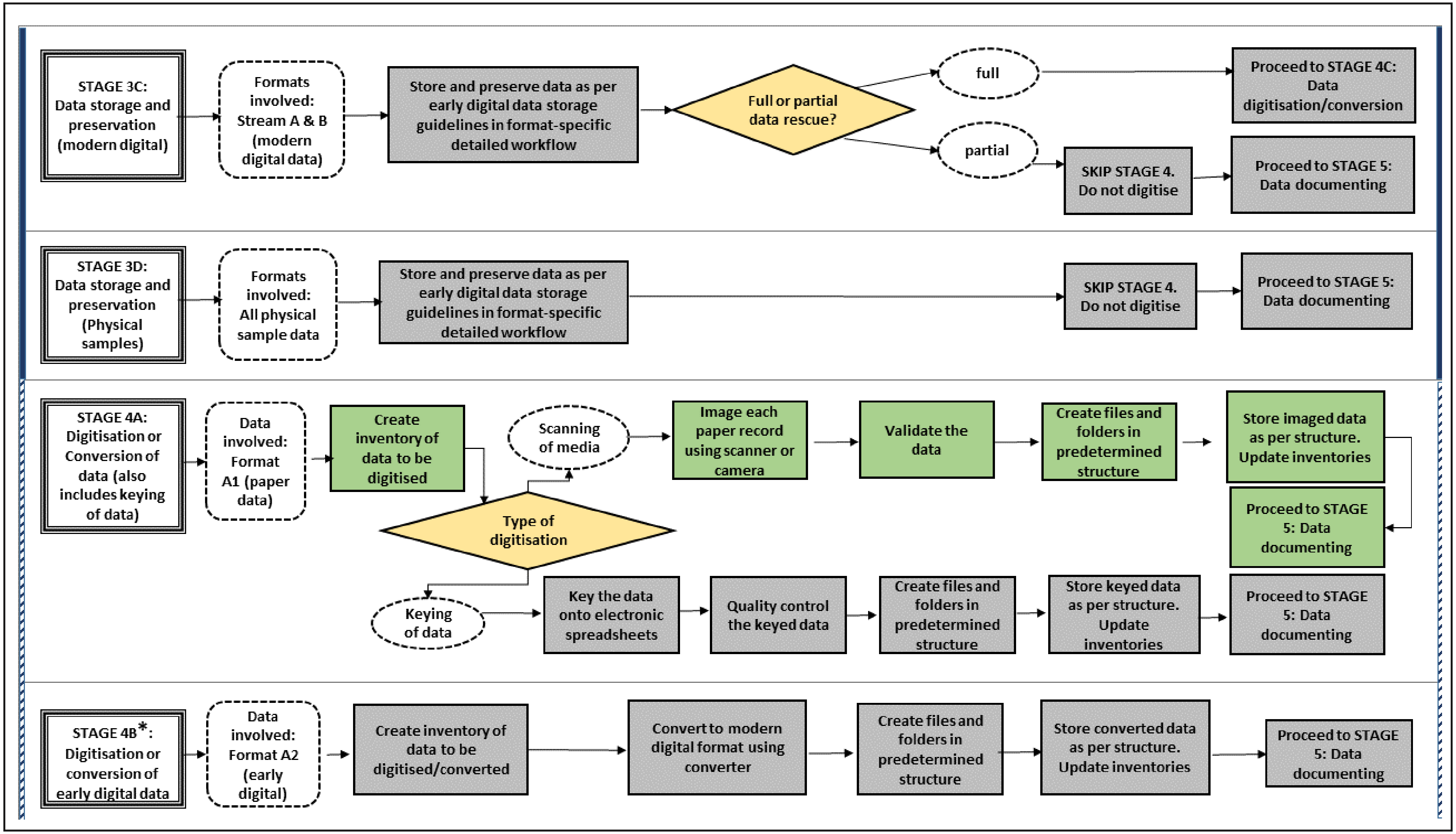

STAGE 3C: Data storage and preservation (modern digital)
Formats involved: Stream A & B (modern digital data)
Store and preserve data as per early digital data storage guidelines in format-specific detailed workflow
Full or partial data rescue?
full
partial
Proceed to STAGE 4C: Data digitisation/conversion
SKIP STAGE 4. Do not digitise
Proceed to STAGE 5: Data documenting
STAGE 3D: Data storage and preservation (Physical samples)
Formats involved: All physical sample data
Store and preserve data as per early digital data storage guidelines in format-specific detailed workflow
SKIP STAGE 4. Do not digitise
Proceed to STAGE 5: Data documenting
STAGE 4A: Digitisation or Conversion of data (also includes keying of data)
Data involved: Format A1 (paper data)
Create inventory of data to be digitised
Type of digitisation
Scanning of media
Image each paper record using scanner or camera
Validate the data
Create files and folders in predetermined structure
Store imaged data as per structure. Update inventories
Proceed to STAGE 5: Data documenting
Keying of data
Key the data onto electronic spreadsheets
Quality control the keyed data
Create files and folders in predetermined structure
Store keyed data as per structure. Update inventories
Proceed to STAGE 5: Data documenting
STAGE 4B*: Digitisation or conversion of early digital data
Data involved: Format A2 (early digital)
Create inventory of data to be digitised/converted
Convert to modern digital format using converter
Create files and folders in predetermined structure
Store converted data as per structure. Update inventories
Proceed to STAGE 5: Data documenting

Figure 5.9: Extended summary of revised Data Rescue Workflow Model (b)

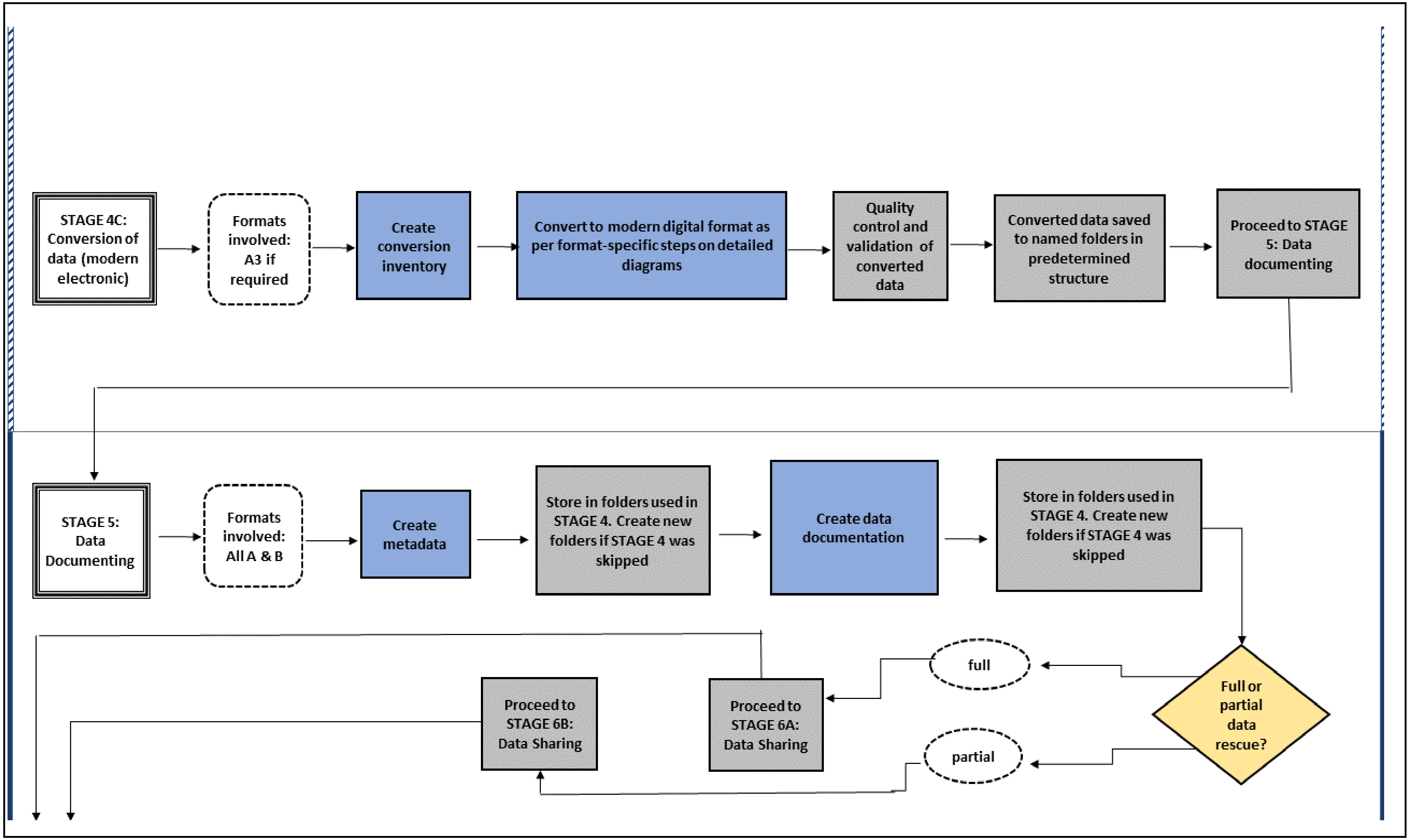

STAGE 4C: Conversion of data (modern electronic)
Formats involved: A3 if required
Create conversion inventory
Convert to modern digital format as per format-specific steps on detailed diagrams
Quality control and validation of converted data
Converted data saved to named folders in predetermined structure
Proceed to STAGE 5: Data documenting
STAGE 5: Data Documenting
Formats involved: All A & B
Create metadata
Store in folders used in STAGE 4. Create new folders if STAGE 4 was skipped
Create data documentation
Store in folders used in STAGE 4. Create new folders if STAGE 4 was skipped
Full or partial data rescue?
full
partial
Proceed to STAGE 6A: Data Sharing
Proceed to STAGE 6B: Data Sharing

Figure 5.10: Extended summary of revised Data Rescue Workflow Model (c)

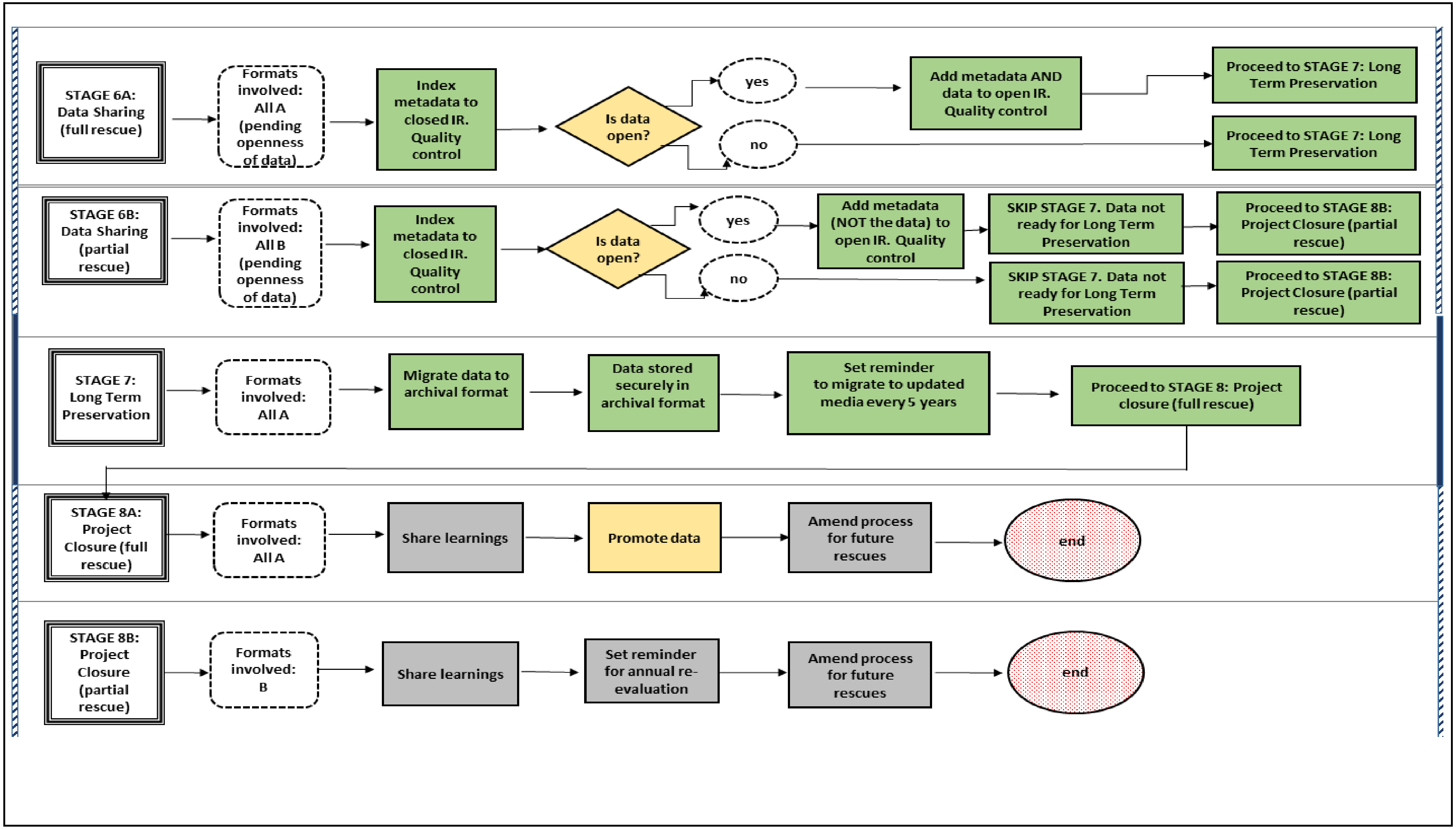

STAGE 6A: Data Sharing (full rescue)
Formats involved: All A (pending openness of data)
Index metadata to closed IR. Quality control
Is data open?
yes
no
Add metadata AND data to open IR. Quality control
Proceed to STAGE 7: Long Term Preservation
Proceed to STAGE 7: Long Term Preservation
STAGE 6B: Data Sharing (partial rescue)
Formats involved: All B (pending openness of data)
Index metadata to closed IR. Quality control
Is data open?
yes
no
Add metadata (NOT the data) to open IR. Quality control
SKIP STAGE 7. Data not ready for Long Term Preservation
SKIP STAGE 7. Data not ready for Long Term Preservation
Proceed to STAGE 8B: Project Closure (partial rescue)
Proceed to STAGE 8B: Project Closure (partial rescue)
STAGE 7: Long Term Preservation
Formats involved: All A
Migrate data to archival format
Data stored securely in archival format
Set reminder to migrate to updated media every 5 years
Proceed to STAGE 8: Project closure (full rescue)
STAGE 8A: Project Closure (full rescue)
Formats involved: All A
Share learnings
Promote data
Amend process for future rescues
end
STAGE 8B: Project Closure (partial rescue)
Formats involved: B
Share learnings
Set reminder for annual re-evaluation
Amend process for future rescues
end

Figure 5.11: Extended summary of revised Data Rescue Workflow Model (d)

5.6.6.1 Stage 1: Data Rescue Preparatory Stage

The revised model contains a stage describing the steps and activities taking place prior to the creation of a data rescue project plan. It therefore includes activities such as locating data, assessing data, determining whether the institute has sufficient resources to rescue the data, and implementing the destruction of data should the data not be found to have value. The stage also caters for the treatment of data should the data be found to have value, but not able to be rescued at the time of data assessment.

In the initial model, preparatory activities were not accentuated, and were included with activities such as project planning, project team selection, and creation of a data management plan. Stage 1: Data rescue preparatory stage of the revised model separates activities and makes a definite distinction between preparatory activities before the decision to rescue data is made, and the first steps of an actual data rescue project.

5.6.6.2 Stage 2: Data Rescue Planning

Data rescue planning and its accompanying activities, such as appointing a data rescue project team, drafting a project plan, and drafting a DMP, feature as a separate stage in the revised model. These activities all formed part of the first stage, titled Project initiation, in the initial model. The separation of activities was considered important, as this enabled the simplification of description of both stages, enabled the indication of responsible parties, and resulted in a clearer distinction between planning a rescue project and determining whether a rescue project is a feasible, realistic venture.

The revised model also distinguishes between data rescue planning activities during full rescue, and data rescue planning during a partial rescue project. While not vastly different, the data rescue project plan for partial rescue will omit the description of rescue activities, responsibilities and outcomes of those stages not forming part of the project.

The responsibility for the preparatory stage is mostly shared between the SET-based sector and the library and information services sector, with researchers expected to assess the value of data. Waste disposal or ICT services will play a role during data destruction, depending on the data format.

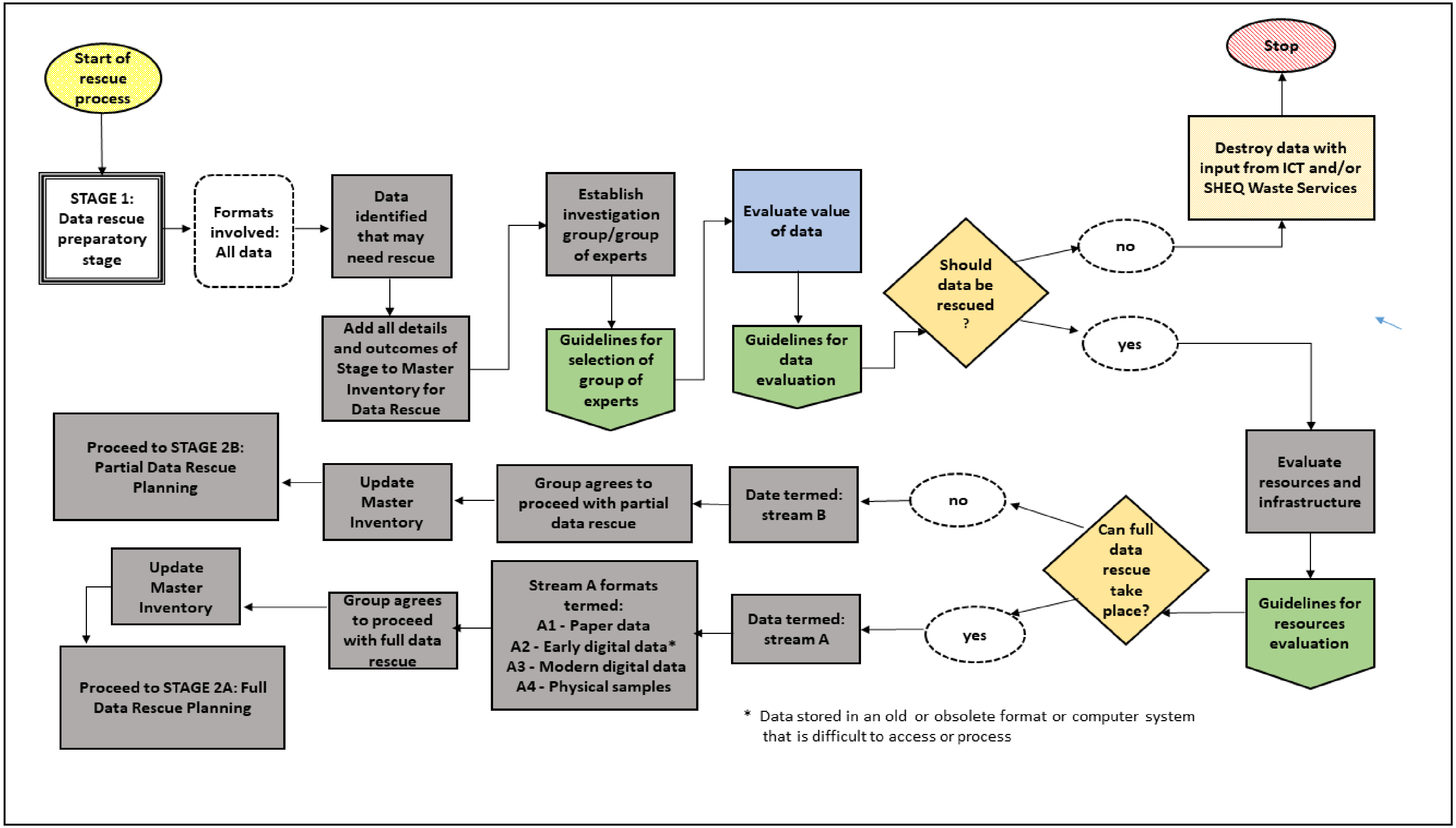

Start of rescue process
STAGE 1: Data rescue preparatory stage
Formats involved: All data
Data identified that may need rescue
Add all details and outcomes of Stage to Master Inventory for Data Rescue
Establish investigation group/group of experts
Guidelines for selection of group of experts
Evaluate value of data
Guidelines for data evaluation
Should data be rescued ?
no
Destroy data with input from ICT and/or SHEQ Waste Services
Stop
yes
Evaluate resources and infrastructure
Guidelines for resources evaluation
Can full data rescue take place?
no
Date termed: stream B
Group agrees to proceed with partial data rescue
Update Master Inventory
Proceed to STAGE 2B: Partial Data Rescue Planning
yes
Data termed: stream A
Stream A formats termed:
A1 - Paper data
A2 - Early digital data*
A3 - Modern digital data
A4 - Physical samples
Group agrees to proceed with full data rescue
Update Master Inventory
Proceed to STAGE 2A: Full Data Rescue Planning
* Data stored in an old or obsolete format or computer system that is difficult to access or process

Figure 5.12: Stage 1: Data rescue preparatory stage

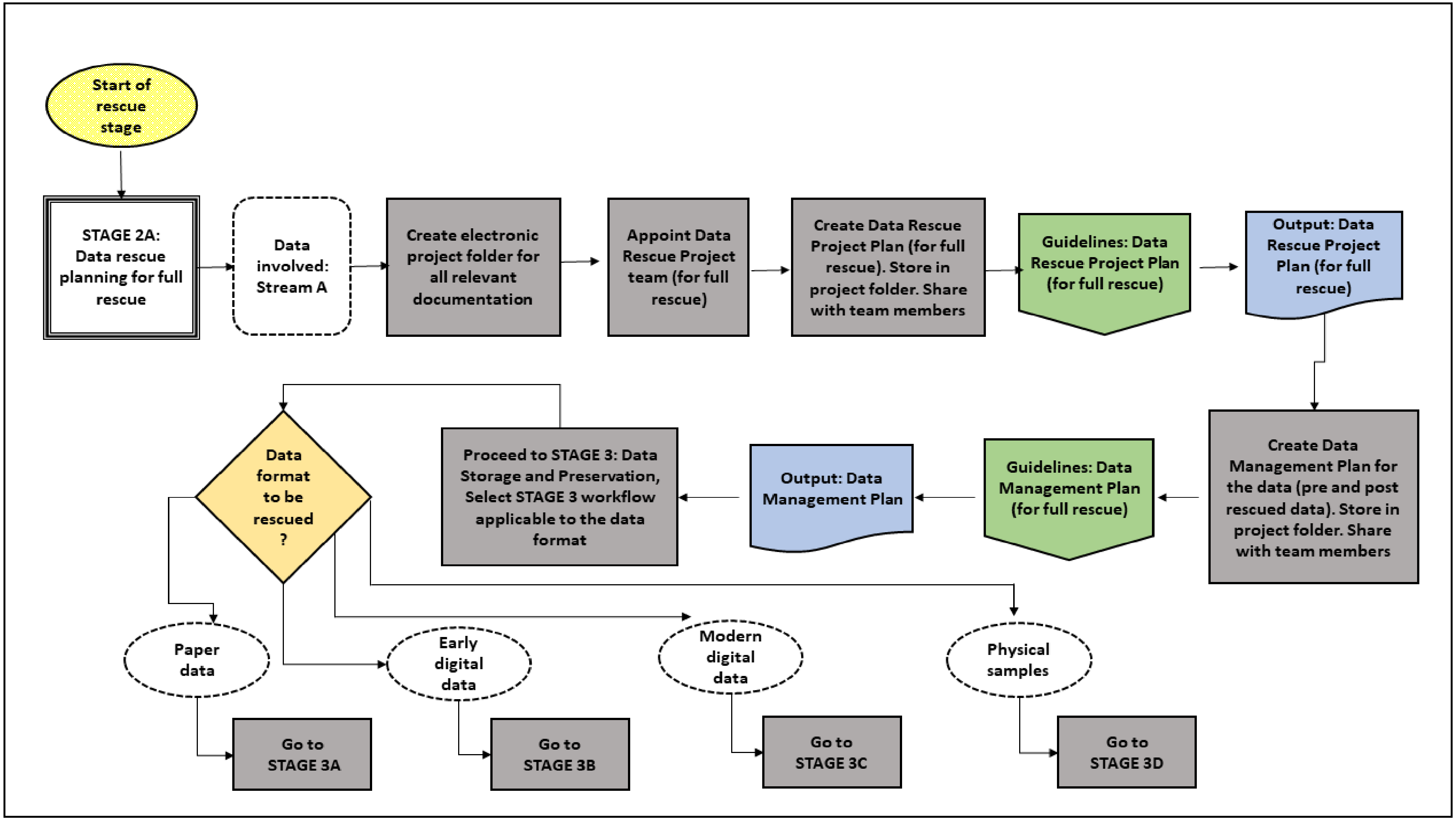

Start of rescue stage
STAGE 2A: Data rescue planning for full rescue
Data involved: Stream A
Create electronic project folder for all relevant documentation
Appoint Data Rescue Project team (for full rescue)
Create Data Rescue Project Plan (for full rescue). Store in project folder. Share with team members
Guidelines: Data Rescue Project Plan (for full rescue)
Output: Data Rescue Project Plan (for full rescue)
Create Data Management Plan for the data (pre and post rescued data). Store in project folder. Share with team members
Guidelines: Data Management Plan (for full rescue)
Output: Data Management Plan
Proceed to STAGE 3: Data Storage and Preservation, Select STAGE 3 workflow applicable to the data format
Data format to be rescued ?
Paper data
Early digital data
Modern digital data
Physical samples
Go to STAGE 3A
Go to STAGE 3B
Go to STAGE 3C
Go to STAGE 3D

Figure 5.13: Stage 2A: Data rescue planning (full rescue)

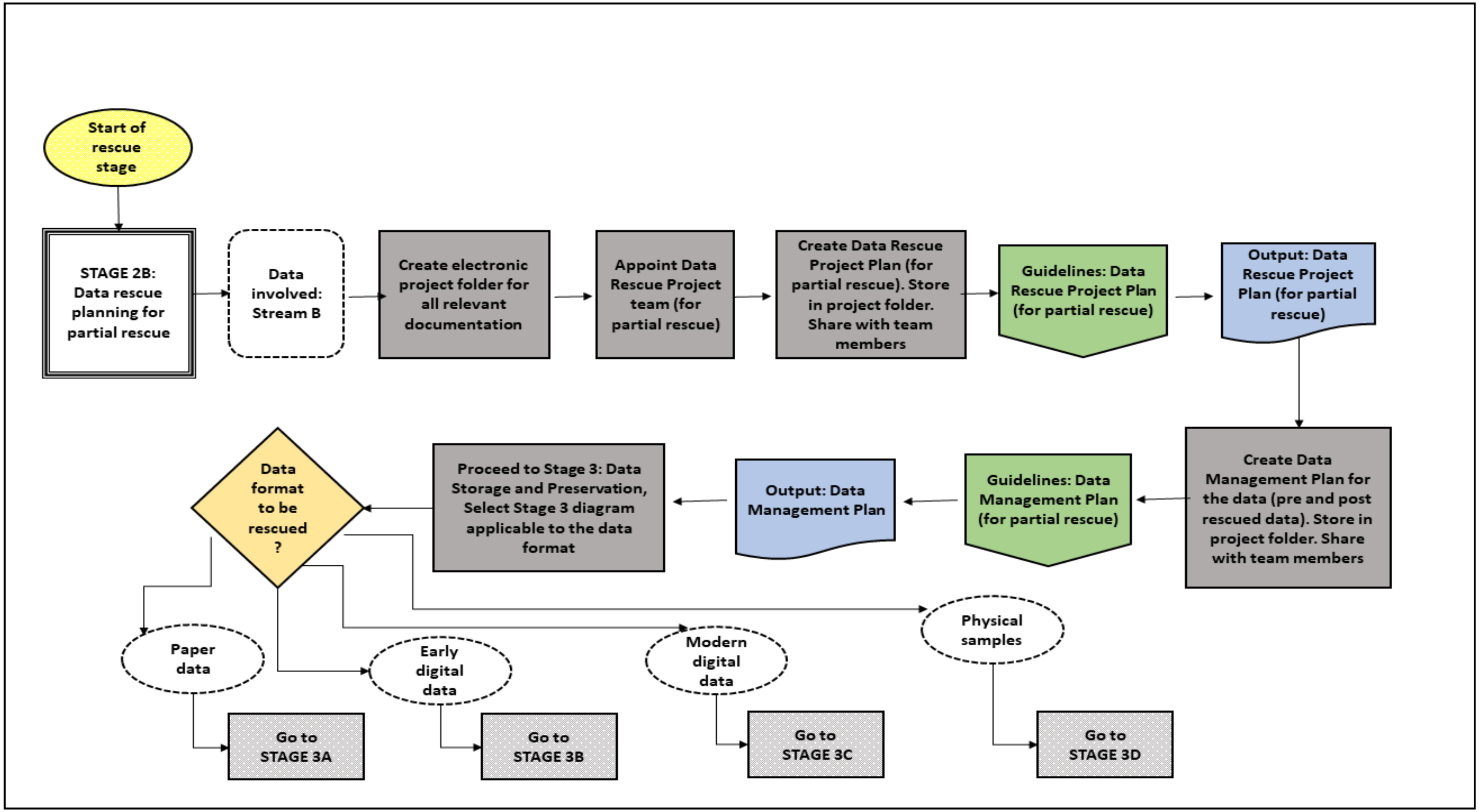

Start of rescue stage
STAGE 2B: Data rescue planning for partial rescue
Data involved: Stream B
Create electronic project folder for all relevant documentation
Appoint Data Rescue Project team (for partial rescue)
Create Data Rescue Project Plan (for partial rescue). Store in project folder. Share with team members
Guidelines: Data Rescue Project Plan (for partial rescue)
Output: Data Rescue Project Plan (for partial rescue)
Create Data Management Plan for the data (pre and post rescued data). Store in project folder. Share with team members
Guidelines: Data Management Plan (for partial rescue)
Output: Data Management Plan
Proceed to Stage 3: Data Storage and Preservation, Select Stage 3 diagram applicable to the data format
Data format to be rescued ?
Paper data
Early digital data
Modern digital data
Physical samples
Go to STAGE 3A
Go to STAGE 3B
Go to STAGE 3C
Go to STAGE 3D

Figure 5.14: Stage 2B: Data rescue planning (partial rescue)

5.6.6.3 Stage 3: Data Storage and Preservation

While the initial model only catered for the storage and preservation of paper-based media, the revised model includes other formats in this stage. The revised model contains the following format-specific graphics:

- Stage 3A: Storage and preservation of paper-based media
  - Stage 3A refers to the storage and preservation treatment of paper-based data and includes aspects such as using gloves when handling data, labelling of boxes and shelves, and ideal paper archive conditions.
  - Stage 3A describes the treatment of paper-based media that will undergo digitisation (i.e., full rescue), and paper-based media that will not be digitised (i.e., partial rescue).
- Stage 3B: Storage and preservation of early digital data
  - Stage 3B refers to the storage and preservation treatment of data in an early digital format. The fact that data are at risk entails adherence to best practices to ensure that it is secure and accessible until converted, or for the near future (if not being converted).
  - Stage 3B describes the treatment of early digital data that will be converted to a common, open, modern format (i.e., full rescue), and early digital data that will not be converted (i.e., partial rescue). Data destruction after digitisation is not mandatory.
- Stage 3C: Storage and preservation of modern electronic data
  - Stage 3C refers to the storage and preservation treatment of data in a modern electronic format. Modern electronic data are seen to be at risk when in a format not commonly used (e.g., requiring specialised software to be read) or when data are without metadata and data documentation. It can also be at risk when not stored securely, or when file and folder naming conventions do not adhere to best practices.
  - Storage and preservation practices involving modern electronic data being at risk entail transferring the data to a secure storage location, ensuring backups are done, and implementing a planned and documented file and folder naming convention and structure.
- Stage 3D: Storage and preservation of physical samples
  - Stage 3D refers to the storage and preservation of physical samples, or specimens. With physical samples data covering a heterogeneous range of

artefacts, elements and components, a generic data rescue model cannot cater for all physical samples data.

- The revised model contains a link to guidelines containing a list of specimen storage documentation for different disciplines and specimen types.
- The crucial message conveyed via the graphic for Stage 4D is that data rescuers should store the specimens according to best practices for the discipline and data type. Additional research and reading may be required from data rescuers when implementing Stage 3D.

Responsible parties for Stage 3 include researchers and/or research library staff (mainly archival personnel).

5.6.6.4 Stage 4: Digitisation

This stage of the revised model contains the following graphics:

Stage 4A: Digitisation of paper-based media

- The Stage 4A graphic entails the digitisation of paper-based media, involving either the scanning of paper via scanner or camera, or the keying of paper data onto electronic media.
- Data rescuers can access guidelines.

Stage 4B: Conversion of early digital media

- The Stage 4B graphic entails the conversion of early digital data to a common, open, modern digital format.
- Depending on the resources available, the conversion can be performed by members of the data rescue team, or be outsourced to trustworthy and experienced digitisation services.

Stage 4C: Conversion of modern electronic data

- The Stage 4C graphic entails the conversion of modern electronic data to a common, open, modern digital format.
- This stage is applicable in instances where the data have been generated using specialised equipment, or when uncommon software is required to access the data. Data sharing and depositing data in a data repository require data to be in a common format.

This stage only involves data undergoing full rescue.

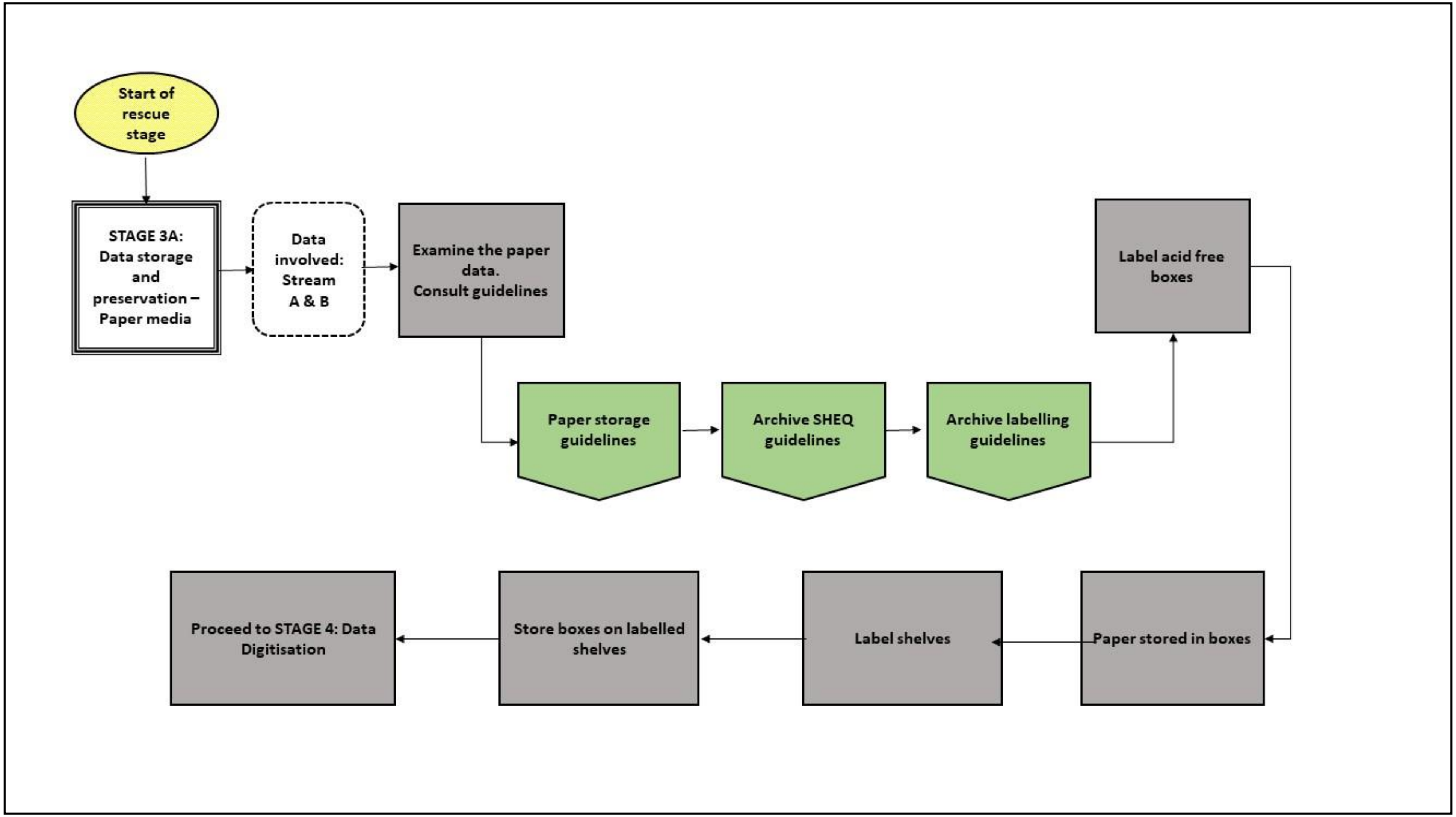

Start of rescue stage
STAGE 3A: Data storage and preservation – Paper media
Data involved: Stream A & B
Examine the paper data. Consult guidelines
Paper storage guidelines
Archive SHEQ guidelines
Archive labelling guidelines
Label acid free boxes
Paper stored in boxes
Label shelves
Store boxes on labelled shelves
Proceed to STAGE 4: Data Digitisation

Figure 5.15: Stage 3A: Data storage and preservation (paper media)

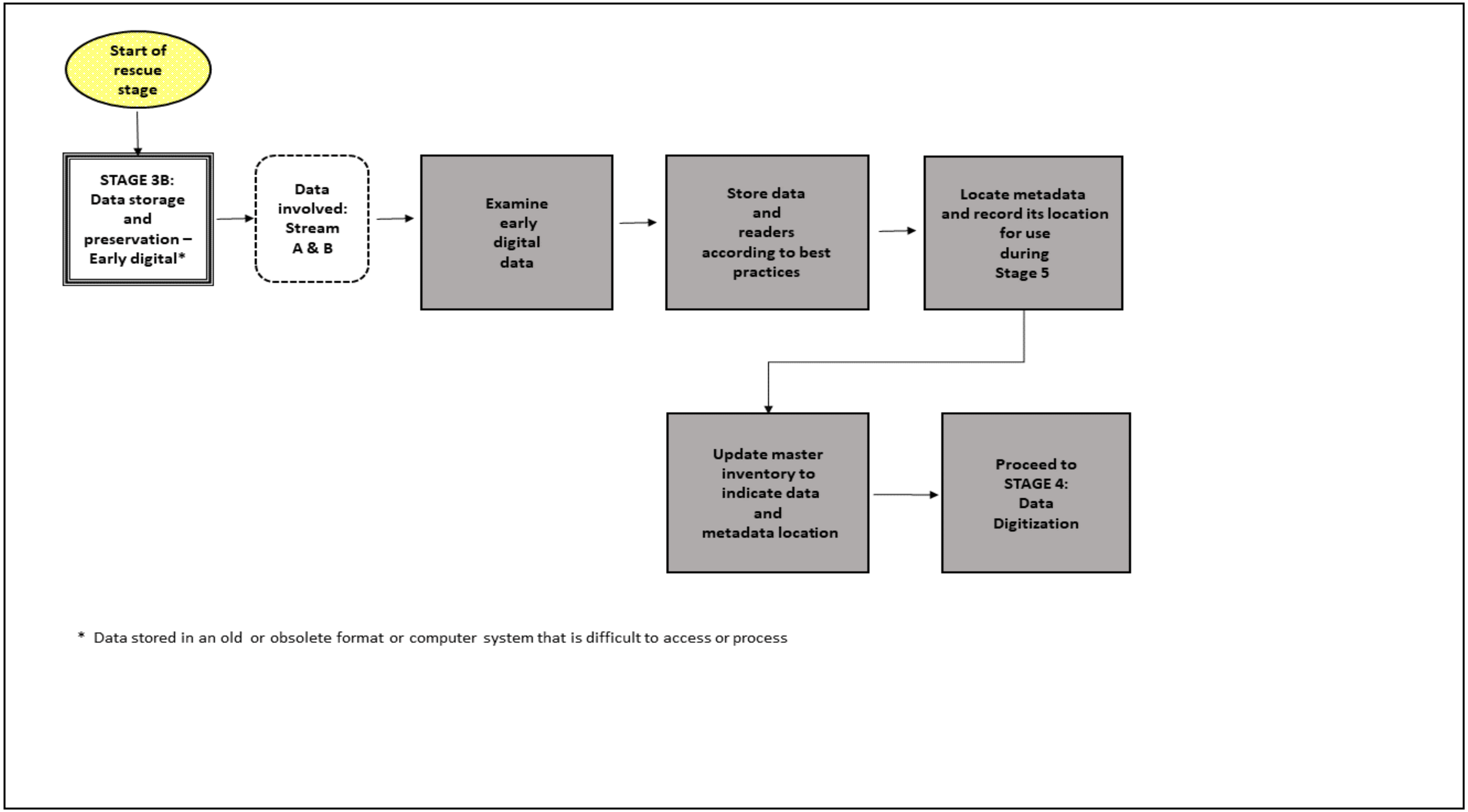

Start of rescue stage
STAGE 3B: Data storage and preservation – Early digital*
Data involved: Stream A & B
Examine early digital data
Store data and readers according to best practices
Locate metadata and record its location for use during Stage 5
Update master inventory to indicate data and metadata location
Proceed to STAGE 4: Data Digitization
* Data stored in an old or obsolete format or computer system that is difficult to access or process

Figure 5.16: Stage 3B: Data storage and preservation (early digital media)

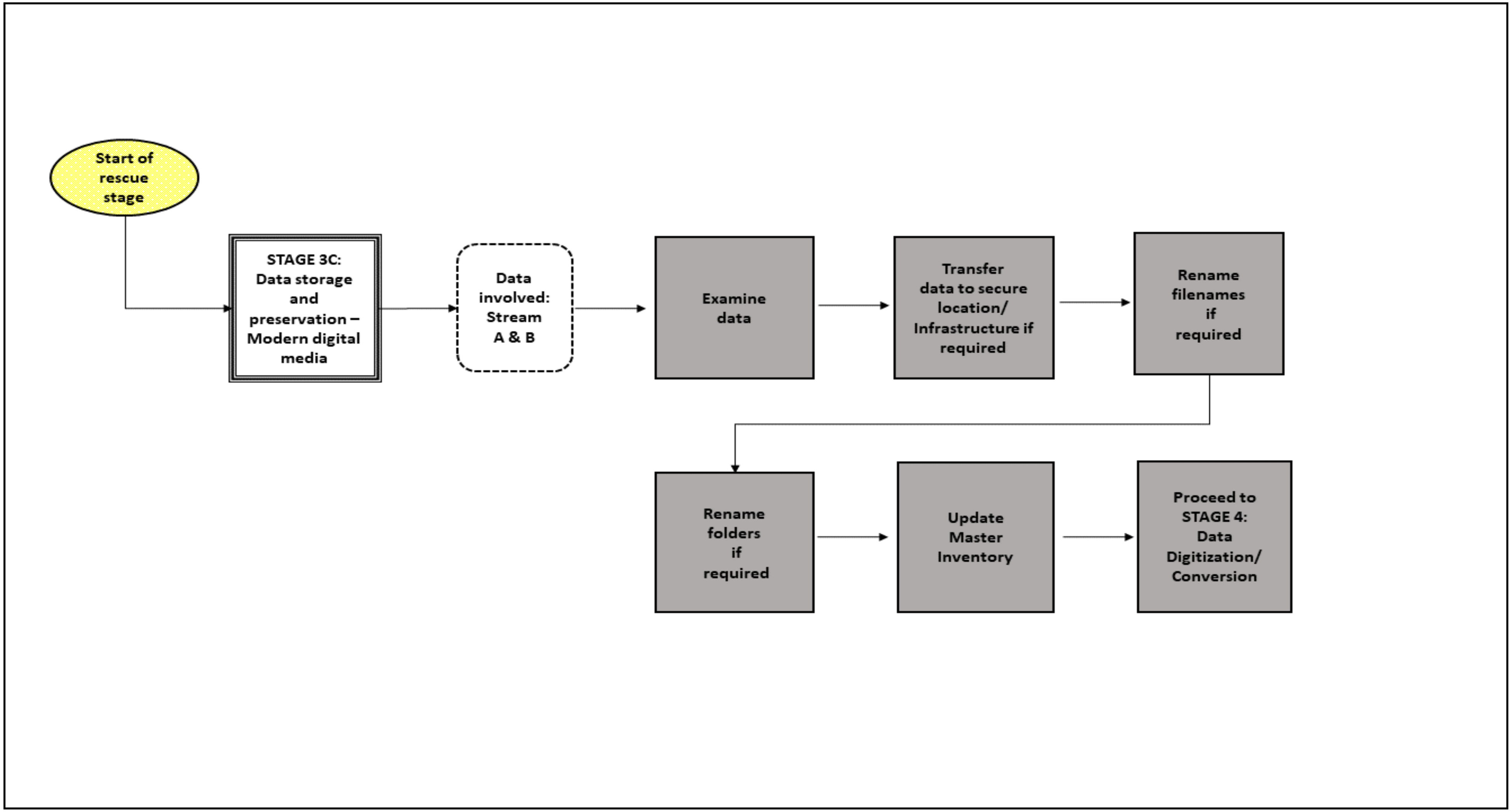

Start of rescue stage
STAGE 3C: Data storage and preservation – Modern digital media
Data involved: Stream A & B
Examine data
Transfer data to secure location/ Infrastructure if required
Rename filenames if required
Rename folders if required
Update Master Inventory
Proceed to STAGE 4: Data Digitization/ Conversion

Figure 5.17: Stage 3C: Data storage and preservation (modern digital media)

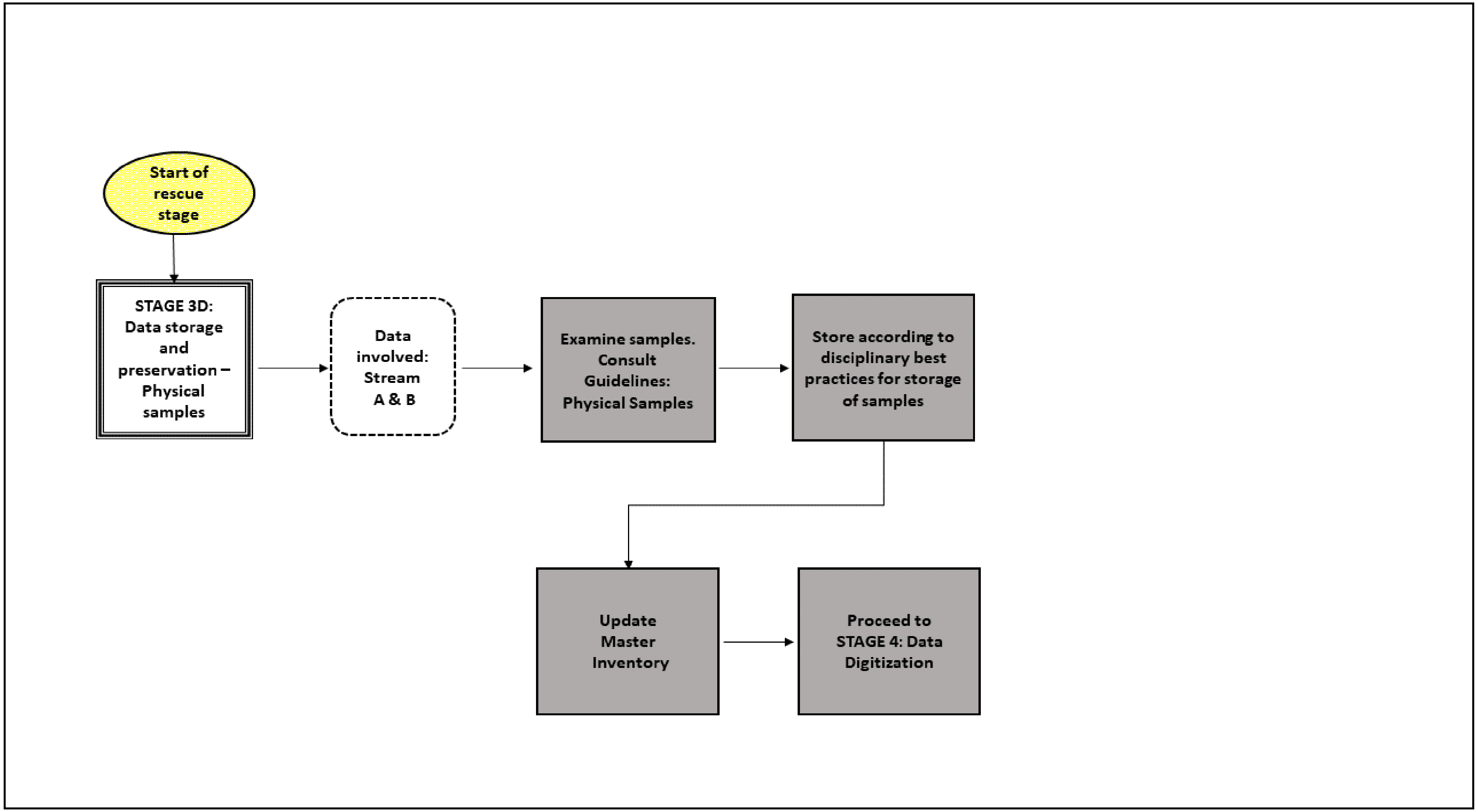

Start of rescue stage
STAGE 3D: Data storage and preservation – Physical samples
Data involved: Stream A & B
Examine samples. Consult Guidelines: Physical Samples
Store according to disciplinary best practices for storage of samples
Update Master Inventory
Proceed to STAGE 4: Data Digitization

Figure 5.18: Stage3D: Data storage and preservation (physical samples)

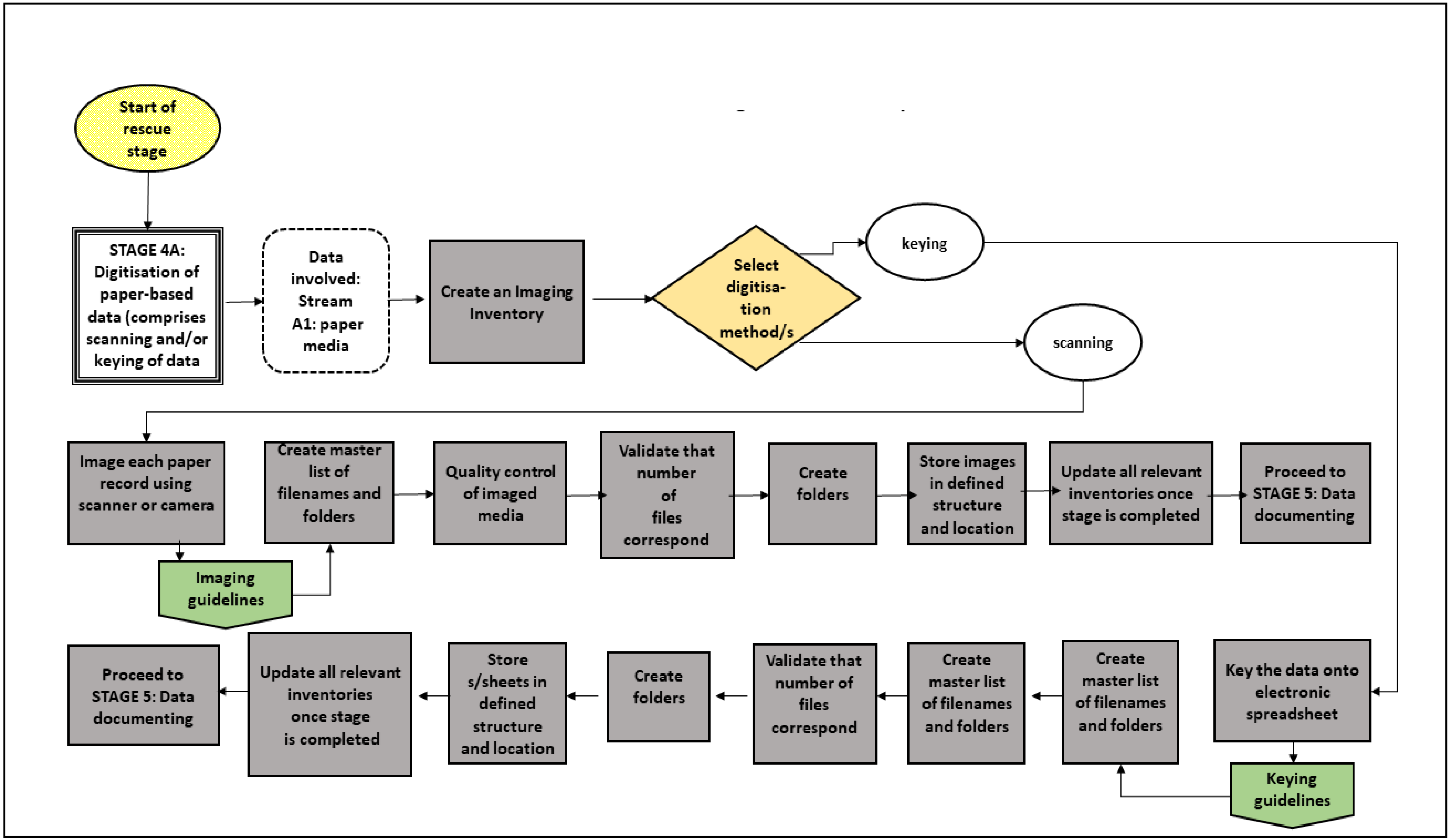

Start of rescue stage
STAGE 4A: Digitisation of paper-based data (comprises scanning and/or keying of data
Data involved: Stream A1: paper media
Create an Imaging Inventory
Select digitisa-tion method/s
keying
scanning
Image each paper record using scanner or camera
Imaging guidelines
Create master list of filenames and folders
Quality control of imaged media
Validate that number of files correspond
Create folders
Store images in defined structure and location
Update all relevant inventories once stage is completed
Proceed to STAGE 5: Data documenting
Key the data onto electronic spreadsheet
Keying guidelines
Create master list of filenames and folders
Create master list of filenames and folders
Validate that number of files correspond
Create folders
Store s/sheets in defined structure and location
Update all relevant inventories once stage is completed
Proceed to STAGE 5: Data documenting

Figure 5.19: Stage 4A: Digitisation of paper-based data

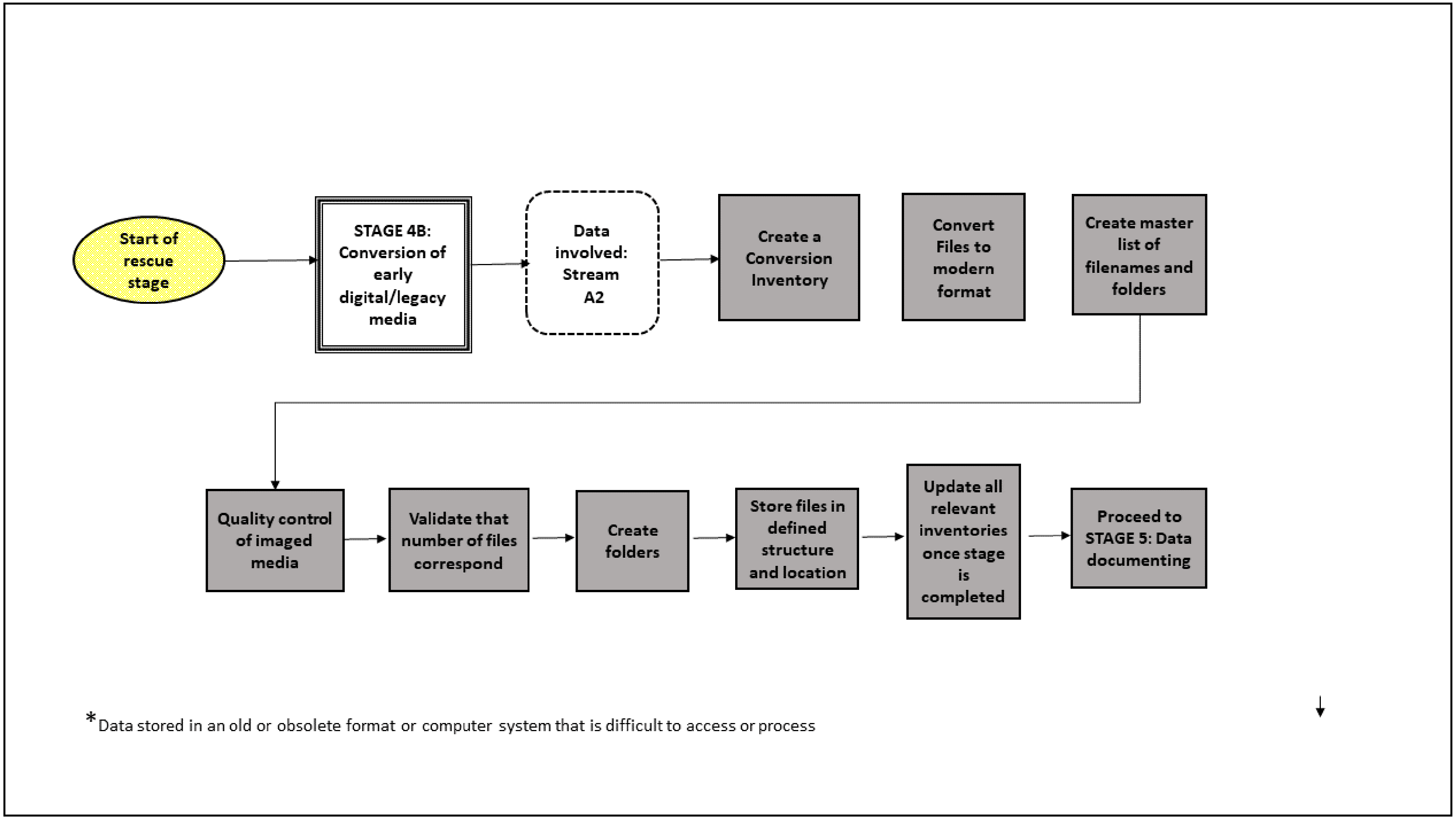

Start of rescue stage
STAGE 4B: Conversion of early digital/legacy media
Data involved: Stream A2
Create a Conversion Inventory
Convert Files to modern format
Create master list of filenames and folders
Quality control of imaged media
Validate that number of files correspond
Create folders
Store files in defined structure and location
Update all relevant inventories once stage is completed
Proceed to STAGE 5: Data documenting
*Data stored in an old or obsolete format or computer system that is difficult to access or process

Figure 5.20: Stage 4B: Conversion of early digital media

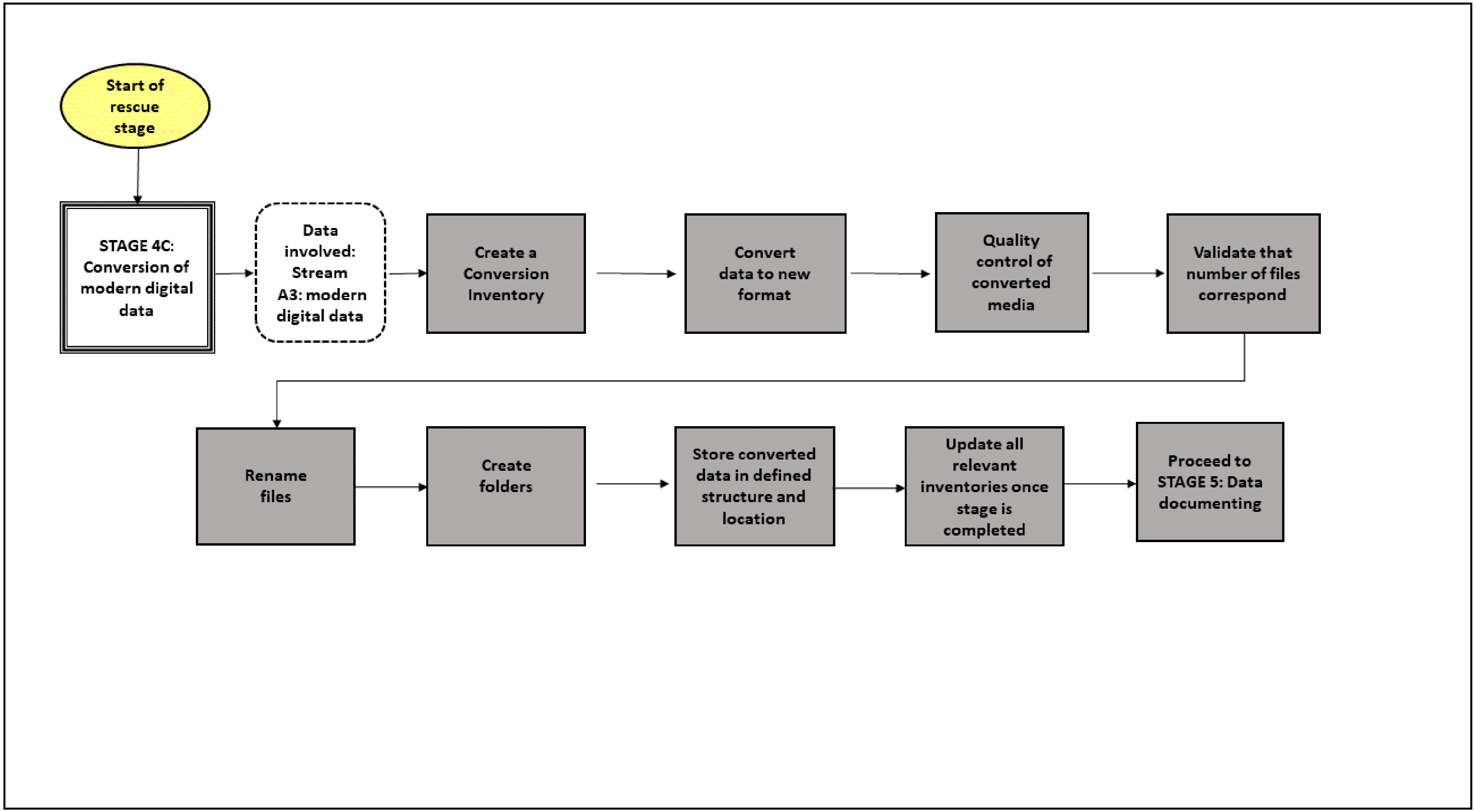

Start of rescue stage
STAGE 4C: Conversion of modern digital data
Data involved: Stream A3: modern digital data
Create a Conversion Inventory
Convert data to new format
Quality control of converted media
Validate that number of files correspond
Rename files
Create folders
Store converted data in defined structure and location
Update all relevant inventories once stage is completed
Proceed to STAGE 5: Data documenting

Figure 5.21: Stage 4C: Conversion of modern digital data

This stage does not involve physical samples data.

Responsibility of this stage depends on the skills of team members, and available resources. The stage can involve researchers, research library staff, or an external organisation.

5.6.6.5 Stage 5: Documenting the Data

This stage consists of describing the data by adding metadata and data documentation to the data, either by creating such documents or adding metadata and data documentation that had already been available. The steps contained within this stage show minor difference when the initial model and the revised model are compared.

Data rescuers can view a document containing guidelines for the creation of metadata and are also able to use a generic Dublin Core metadata template.

Responsible parties for documenting of data include researchers, who are likely to be assisted by research library personnel who would provide guidance on metadata standards and crucial metadata fields.

5.6.6.6 Stage 6: Data Sharing

Stage 6 involves all rescued data, irrespective of type of rescue (partial and full rescue) and data format. Stage 6 portrays the typical sharing activities forming part of a data rescue project. The graphic titled Stage 6A portrays data sharing of data undergoing full rescue, while the graphic titled Stage 6B shows data sharing for partially rescued data.

Stage 6 of the revised model portrays the steps to be taken for data sharing at the selected institute. Institute-specific data sharing steps portrayed in the graphic entail the following:

- adding the metadata of all rescued data to the institutional closed repository,
- adding the metadata of non-confidential data to the institutional open repository (this step can also include dataset upload under certain conditions), and
- uploading the dataset and metadata of non-confidential data to a discipline-specific repository.

Assuring that a DOI is assigned to a non-confidential dataset (via upload to a discipline-specific repository) is a recommended step, as it enables the tracking of data sharing and enables data citation.

Responsible parties for this stage will mostly involve the library and information services sector assisted in part by SET-based staff.

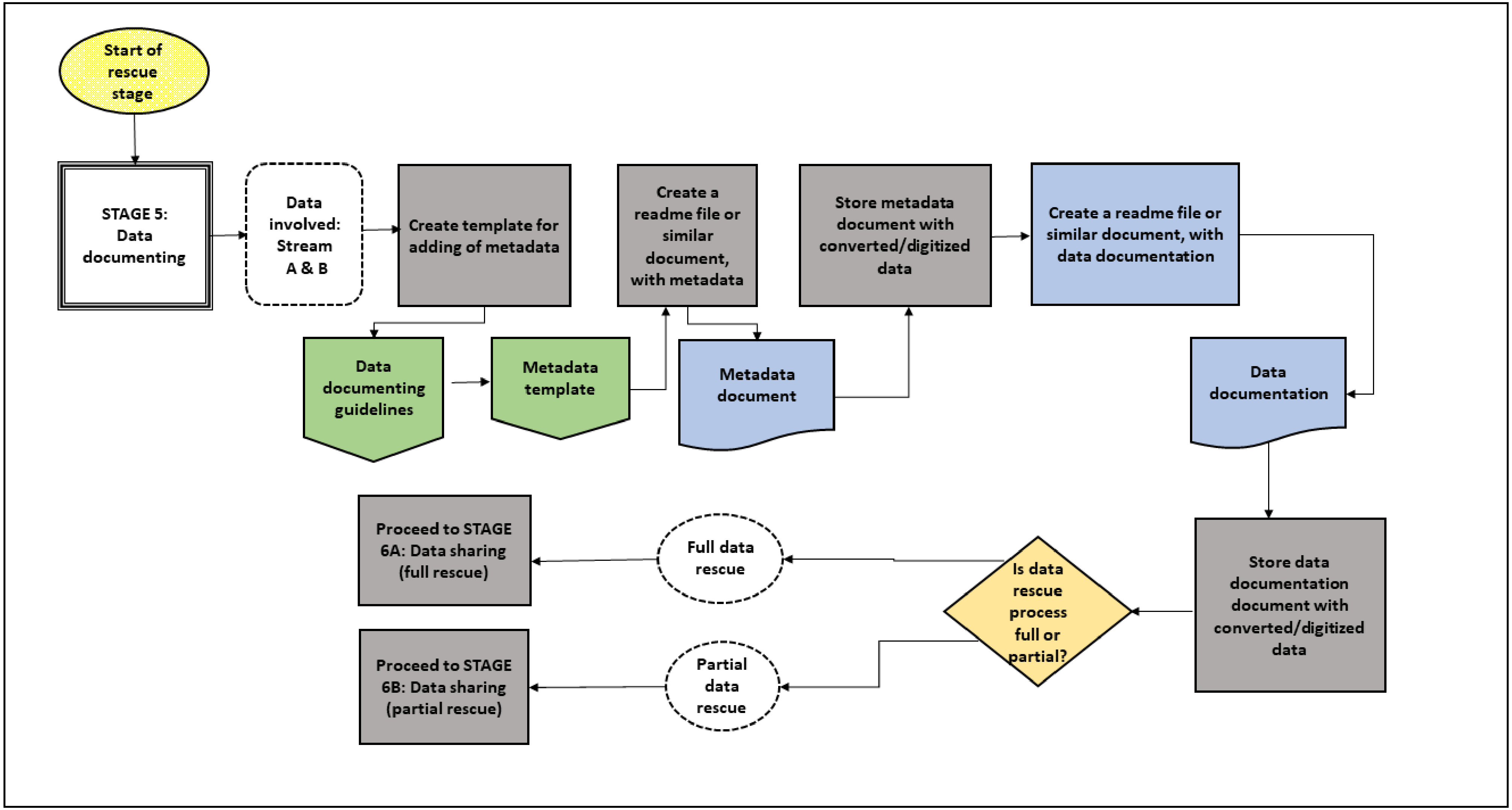

Start of rescue stage
STAGE 5: Data documenting
Data involved: Stream A & B
Create template for adding of metadata
Create a readme file or similar document, with metadata
Store metadata document with converted/digitized data
Create a readme file or similar document, with data documentation
Data documenting guidelines
Metadata template
Metadata document
Data documentation
Proceed to STAGE 6A: Data sharing (full rescue)
Full data rescue
Is data rescue process full or partial?
Store data documentation document with converted/digitized data
Proceed to STAGE 6B: Data sharing (partial rescue)
Partial data rescue

Figure 5.22: Stage 5: Data documenting

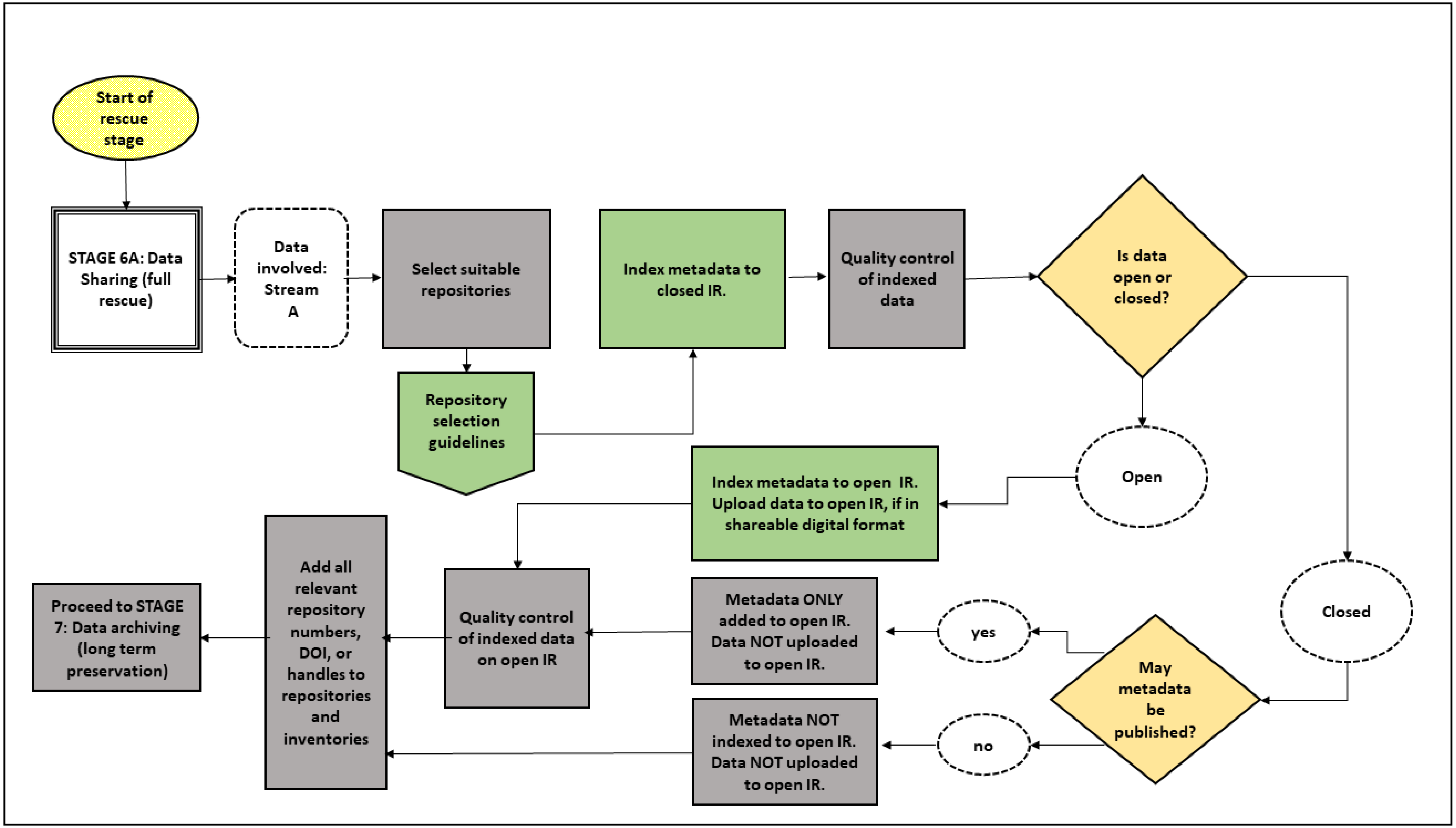

Start of rescue stage
STAGE 6A: Data Sharing (full rescue)
Data involved: Stream A
Select suitable repositories
Repository selection guidelines
Index metadata to closed IR.
Quality control of indexed data
Is data open or closed?
Open
Closed
Index metadata to open IR. Upload data to open IR, if in shareable digital format
May metadata be published?
yes
no
Metadata ONLY added to open IR. Data NOT uploaded to open IR.
Metadata NOT indexed to open IR. Data NOT uploaded to open IR.
Quality control of indexed data on open IR
Add all relevant repository numbers, DOI, or handles to repositories and inventories
Proceed to STAGE 7: Data archiving (long term preservation)

Figure 5.23: Stage 6A: Data sharing (full rescue)

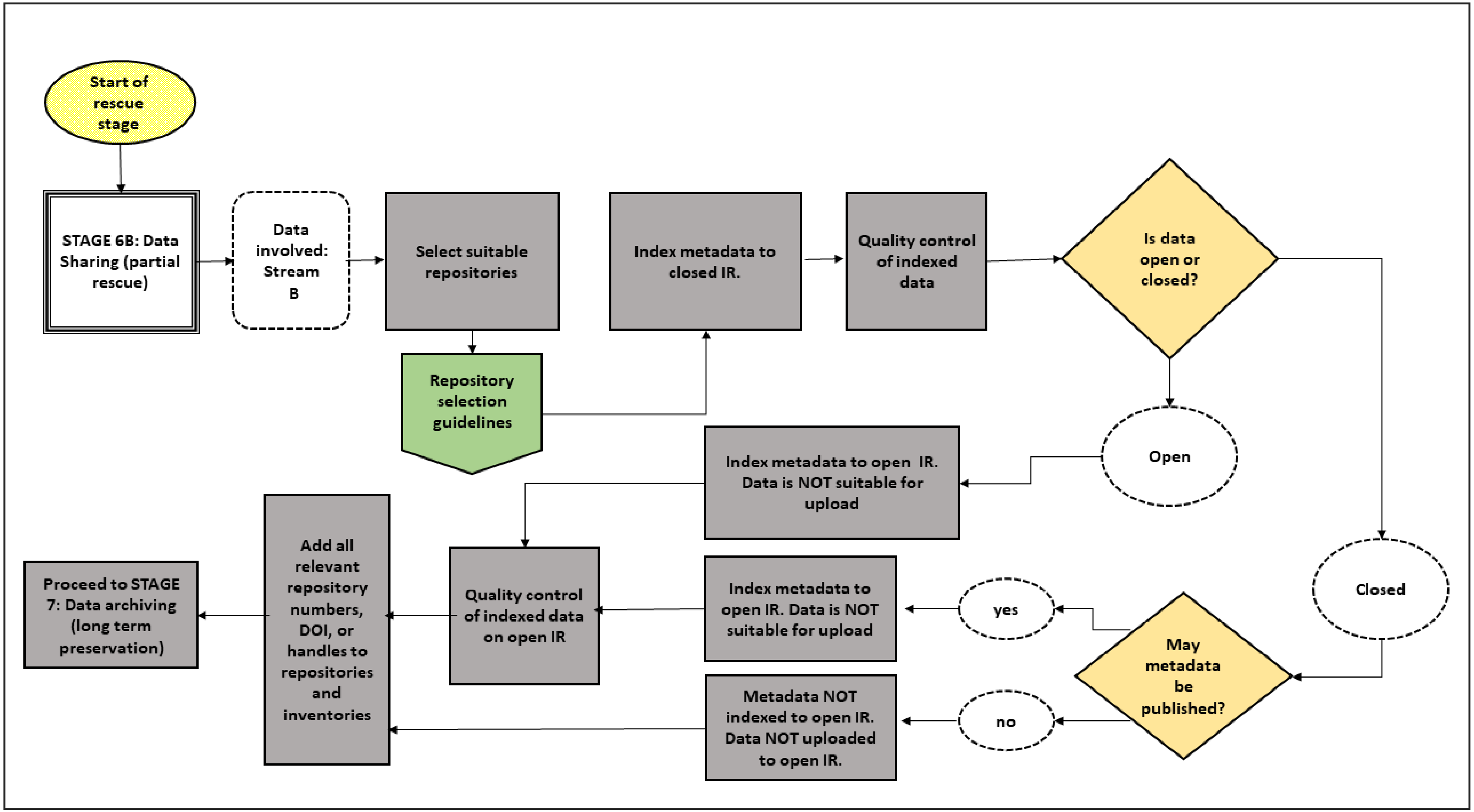

Start of rescue stage
STAGE 6B: Data Sharing (partial rescue)
Data involved: Stream B
Select suitable repositories
Repository selection guidelines
Index metadata to closed IR.
Quality control of indexed data
Is data open or closed?
Open
Closed
Index metadata to open IR. Data is NOT suitable for upload
May metadata be published?
yes
no
Index metadata to open IR. Data is NOT suitable for upload
Metadata NOT indexed to open IR. Data NOT uploaded to open IR.
Quality control of indexed data on open IR
Add all relevant repository numbers, DOI, or handles to repositories and inventories
Proceed to STAGE 7: Data archiving (long term preservation)

Figure 5.24: Stage 6B: Data sharing (partial rescue)

5.6.6.7 Stage 7: Long-Term Preservation

Once converted or digitised/keyed, all electronic data will undergo the same long-term preservation steps.

Long-term preservation can occur simultaneously with conversion, or later.

Many of the tasks listed on the diagram are archival tasks and not part of the data rescue workflow per se. If the data rescue team does not include the archivist as part of the rescue team, then it should be ensured that the archive used does data migration and performs regular checking of data access.

Responsibility for tasks is shared between research library staff (including archival staff) and the SET-based component of the data rescue team. ICT assistance and involvement may feature in certain projects, depending on the preservation format selected or required.

5.6.6.8 Stage 8: Project Closure

The last stage of the revised model is similar to the last stage of the initial model, also titled Project Closure.

The biggest difference between the two stages is that the former caters for full rescue and partial rescue, while the latter assumes that all data rescue projects will implement similar project closure activities. With partial rescue, part of the final stage of the revised model describes the annual review of partially rescued data. Determining whether the reasons for not implementing full rescue (e.g., lack of infrastructure and resources) are still relevant, and acting accordingly is an important part of the final stage.

Responsibility for tasks is shared between the SET-based component of the data rescue team, the institute’s communications division, and applicable research library professionals.

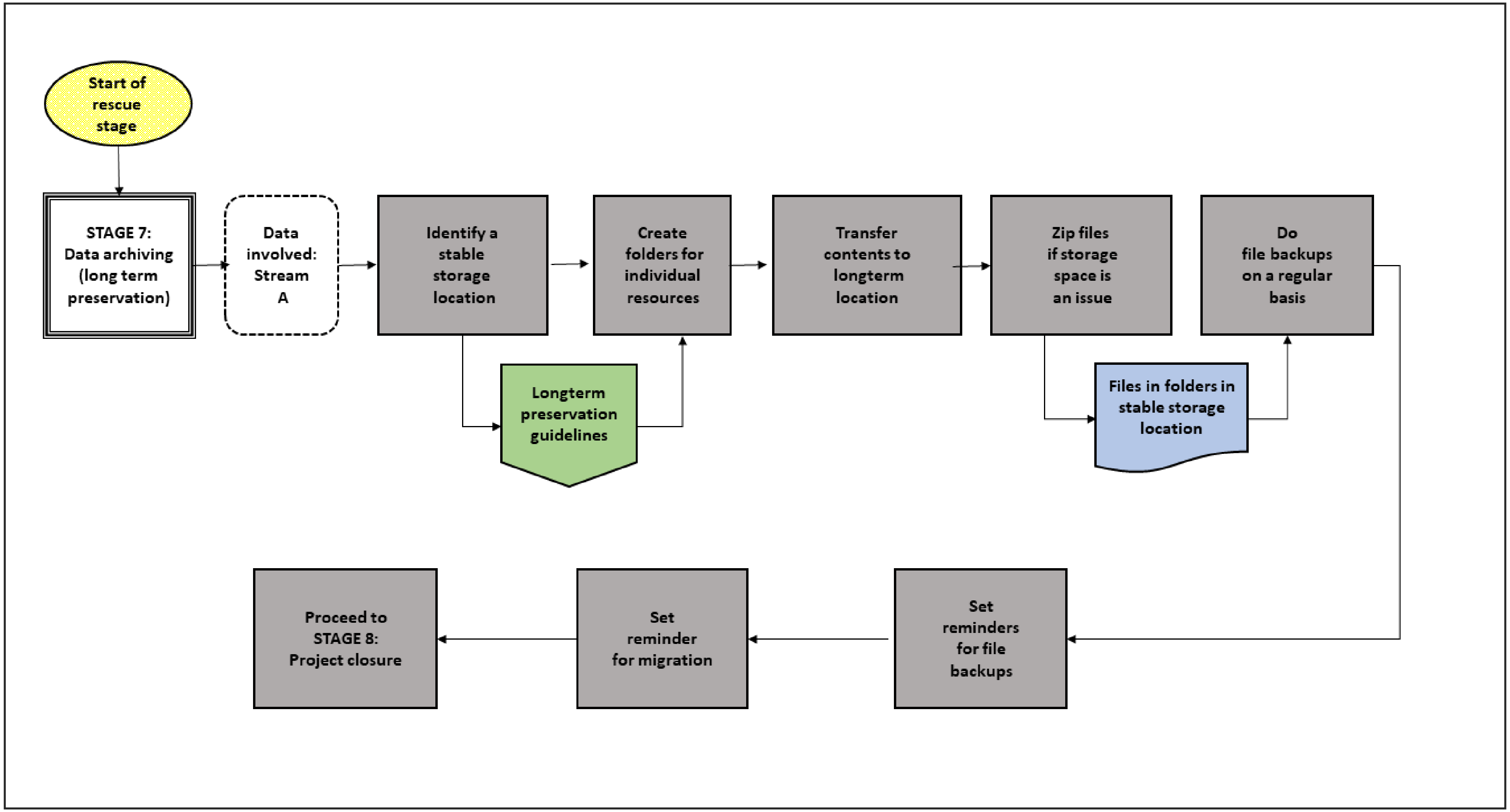

Start of rescue stage
STAGE 7: Data archiving (long term preservation)
Data involved: Stream A
Identify a stable storage location
Create folders for individual resources
Transfer contents to longterm location
Zip files if storage space is an issue
Do file backups on a regular basis
Longterm preservation guidelines
Files in folders in stable storage location
Set reminders for file backups
Set reminder for migration
Proceed to STAGE 8: Project closure

Figure 5.25: Stage 7: Data archiving

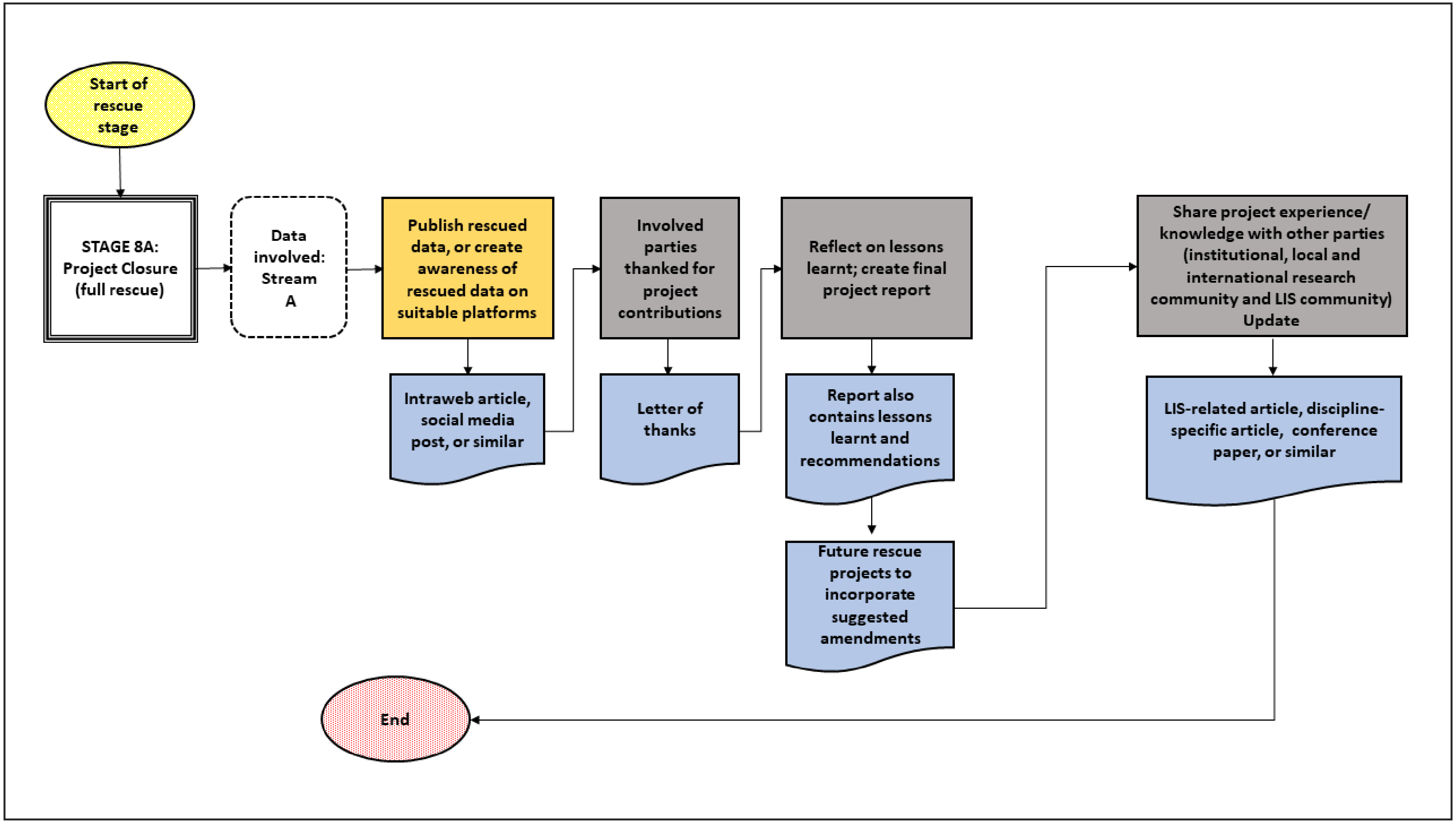

Start of rescue stage
STAGE 8A: Project Closure (full rescue)
Data involved: Stream A
Publish rescued data, or create awareness of rescued data on suitable platforms
Intraweb article, social media post, or similar
Involved parties thanked for project contributions
Letter of thanks
Reflect on lessons learnt; create final project report
Report also contains lessons learnt and recommendations
Future rescue projects to incorporate suggested amendments
Share project experience/ knowledge with other parties (institutional, local and international research community and LIS community) Update
LIS-related article, discipline-specific article, conference paper, or similar
End

Figure 5.26: Stage 8A: Project closure (full rescue)

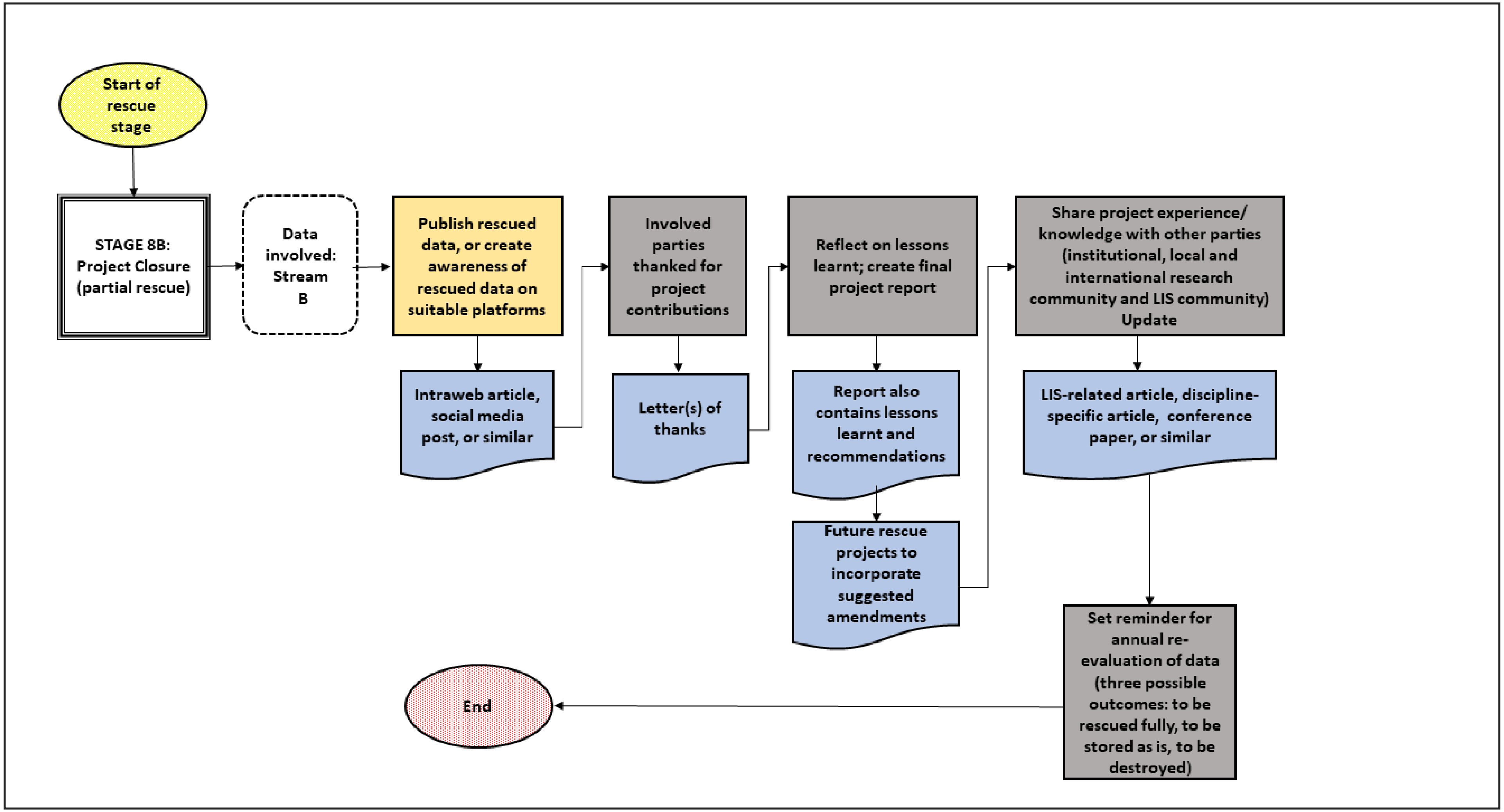

Start of rescue stage
STAGE 8B: Project Closure (partial rescue)
Data involved: Stream B
Publish rescued data, or create awareness of rescued data on suitable platforms
Intraweb article, social media post, or similar
Involved parties thanked for project contributions
Letter(s) of thanks
Reflect on lessons learnt; create final project report
Report also contains lessons learnt and recommendations
Future rescue projects to incorporate suggested amendments
Share project experience/ knowledge with other parties (institutional, local and international research community and LIS community) Update
LIS-related article, discipline-specific article, conference paper, or similar
Set reminder for annual re-evaluation of data (three possible outcomes: to be rescued fully, to be stored as is, to be destroyed)
End

Figure 5.27: Stage 8B: Project closure (partial rescue)

5.6.7 Summary of revision of model

This section described the reasons for amending the initial model and describes the newly created revised model. The rationale for drafting a revised model is for two main reasons:

- unsatisfactory feedback received from Sample A after reviewing the initial model, and
- insight gained during the period between date of creation of the initial model, and analysing the feedback obtained from Sample B. A period of approximately nine months had passed since drafting the initial model and the completion of the feedback stage described in Section 5.5 of this study. This period, combined with exposure to researcher viewpoints and subsequent data rescue documentation provided the researcher with additional data rescue knowledge, and an understanding of features and steps required from a data rescue workflow model.

The revised model differs from the initial model in various respects; the differences are tabulated in Table 5.10. Major differences are as follows:

- The initial model entailed paper-based media rescue while the revised model includes various data formats.
- The initial model assumed that all data rescue stages were applicable to all data being rescued. The revised model differentiates between partial rescue and full rescue, and acknowledges that not all data will be involved in all the stages.
- The initial model had nine main rescue stages; the revised model has eight stages. Some stages have also undergone a name change.
- The initial model consists of a summary and stage-specific stages; the revised model consists of a compact summary, an extended summary, and stage-specific diagrams.

The revised model incorporated many of the suggestions put forward during the feedback session; suggestions can be viewed in Section 5.5: Results: Feedback on Data Rescue Workflow Model.

Both the initial and the revised model were discussed during the next data collection stage. The mini focus group session, involving three institutional research library experts (i.e., not SET-based staff), provided additional and novel feedback regarding the components of a usable data rescue workflow model that would include the roles and responsibilities of the research library.

The next section of this chapter contains the findings and discussion of the mini focus group session involving research library experts.

5.7 Results: Mini focus group session

5.7.1 Introduction

A mini focus group session was held with institutional experts who were not employed in an SET-based role, but whose specific experience would have led to insight into data rescue activities, archival procedures, researcher behaviour and research library roles. It was expected that these purposively selected experts would be able to provide insightful and novel recommendations regarding the initial model, and the revised model.

5.7.2 Sample and responses

A sample of three experts were approached and invited to be part of a mini focus group session. The purpose of the mini focus group session, comprising a discussion of the two models, was explained to invited parties. All three experts accepted the invitation and participated enthusiastically in the focus group session.

The respective focus group participants possessed the following characteristics:

- LIS Expert 1: more than three decades of library and information services experience, with institutional workflows involving library, SET staff and ICT responsibilities, institutional repository experience, and project management being particular areas of expertise.
- LIS Expert 2: three decades of experience in both disciplinary research as a scientist, and library and information services expertise in user training, publishing advice, and information scientist activities.
- LIS Expert 3: close on a decade's experience in records management, archival roles, activities and procedures, plus the drafting and editing of institutional workflows.

5.7.3 Findings

Feedback received during the focus group session include recommendations based on extensive experience with institutional workflows, insights into the behavioural patterns of the selected institute's SET-based section, and knowledge of research library activities. In addition to these recommendations, participants also put forward personal preferences regarding an ideal model.

Findings related to feedback supplied during the focus group session and related to the separate data rescue models are provided below. The first version of the model received little feedback, while the revised version, though positively received, produced elaborate and novel feedback and recommendations. As a result, chapter feedback related to the latter version of the model comprises separate sections, such as the model's appearance, feedback related to the summary image, feedback related to the individual stages, and institute-specific feedback.

5.7.3.1 Initial Data Rescue Workflow Model

- No distinct feedback categories are included here as the feedback was minimal.
- LIS Expert 2 welcomed the look of the Summary image of the model's initial version:

> *'I like the clear, clean first draft summary.'*

5.7.3.2 Revised Data Rescue Workflow Model

This section includes focus group feedback pertaining to the revised model.

General layout and appearance (including use of shapes, colours and connectors)

Several observations regarding the general look of the model were made, with recommendations and suggestions also put forward by participants. Feedback included the following:

- Participants stated that the arrows on some diagrams were confusing and did not make sense at times.
- Participants stated that some of the arrows seem to be going in the wrong direction.
- Participants stated that it might be prudent to check whether the included shapes had been used correctly, and that their usage was in line with their standardised meaning.
- Participants stated that the model often made use of two different streams joining a single block, and that this should never be done when creating a workflow model. This researcher was reminded that even when the resultant block includes the same activity for both streams, two streams should not join up in a single activity block.
- This researcher was reminded that when creating a new flowchart row, it is good practice to take a short route with the connector. It was stated that the connector should not go all the way back to the opposite margin.

General content (including terminology, details and available information)

General feedback regarding the terminology, descriptions and information contained within the revised model is listed below. Feedback is applicable to both summaries, and the individual stage-specific diagrams.

- LIS Expert 3 mentioned that the diagrams needed additional explanatory details on the diagram itself. Conversely, a different participant preferred viewing the compact summary, and being verbally guided through the summary.
- This researcher was asked whether the created images were intended to form a flowchart (indicating workflow activities), or whether it is more closely related to a diagram or model.
- Participants stated that the model should make use of verbs, rather than nouns, when describing a workflow activity. For example, LIS Expert 2 suggested using the words 'Archive the data', instead of 'Data archiving'.

  > *'"Describing data" is a better term [sic] than "data documentation".'*

- LIS Expert 2 stated that the stage-specific diagrams were overwhelming, due to the multitude or blocks included in each stage:

  > *'There are too many blocks.'*

- The idea was put forward that the model somehow needs a way to link back to the original data assessment decisions.
- LIS Expert 1 suggested that the model use a symbol to indicate that the activity, system or term is applicable to the selected institute only:

  > *'Perhaps add an asterisk when something is applicable to a CSIR scenario.'*

- Participants suggested that the model, or guidelines accompanying the model, should clearly define what is meant by the terms, phrases and concepts forming part of the model. Examples of words or phrases requiring clarification include:
  - 'data at risk',
  - 'data rescue',
  - 'partial data rescue',
  - 'full data rescue',
  - 'data documentation',
  - 'metadata',
  - 'data inventory',
  - other inventories listed in the model, and
  - 'data storage and preservation'.

- Adding to the above, participants suggested that there should be a link to a data rescue glossary.
- In addition, participants suggested that there should be a linked list of data rescue acronyms and abbreviations.
- Participants also felt it was crucial to stress that the stage titled as 'Data storage and preservation' refers to the pre-rescued data. The reader should be aware that the term, as used in the model, does not refer to the storage and preservation of data in its rescued, modern digital format.
- Participants mentioned that they were unsure what the various main stages entailed and stated that a concise description or clarification of each stage should ideally be provided.

Compact summary

Recommendations regarding the compact summary of the revised model (see Section 5.6.3: Compact summary) are provided below.

- The idea of a single-page simplified summary was lauded, with participants stating that such a shortened version of the model was required for inexperienced users of the model, or newcomers to the concept of data rescue.
- Participants recommended that the simplified summary only contain the main data rescue stages, with reference to sub-activities or processes only included to clarify what the stage entailed.
- Participants suggested that the shapes used in the summary be changed to distinguish them from the shapes used in the stage-specific images.
- Participants recommended that the summary image not be described as a summary of the workflow, or a summarised workflow. Instead, it should be referred to as a diagram, as it is merely a figure containing the main data rescue stages. The compact summary was stated to not meet the requirements of a 'workflow' portrayal.

Extended summary, and stage-specific images

Feedback regarding the extended summary and stage-specific images was enthusiastic, illuminating and abundant. Recommendations and critique regarding the different rescue stages are listed below.

- Stage 1 (Data rescue preparatory stage)
    - Participants suggested that the guidelines accompanying the concept of 'partial rescue' also include data that do not have metadata or data documentation.

  - Adding to the above point was the recommendation that the importance of metadata and data documentation be emphasised in the guidelines accompanying the data assessment activity.
  - Participants recommended that a checklist be created to assist the role-players who would be assessing the data during this stage.
  - LIS Expert 3 mentioned that it is important to indicate who will be appointing the data assessment team:

    'Who will appoint the data assessment team?'

  - LIS Expert 1 mentioned that it is important that data assessment involve researchers, as they have best insight into data value:

    'Data assessment should be done by researchers.'

  - Participants mentioned that it is important to indicate where the pre-rescue data will be stored.
  - Participants suggested that the responsibility (i.e., the responsible division) for managing and maintaining the 'Master Inventory' be indicated within the model.
- Stage 2 (Data rescue planning)
  - LIS Expert 3 mentioned that it is important to provide details regarding the team member selection process:

    'How will the team members be selected?'

- Stage 5 (Data documenting)
  - Participants stated that the phrase 'Describing data' is a better option than 'Data documenting'. The participants stated that 'data documenting' is a term not easily understood and may mean different things to different role-players.
- Stage 6 (Data sharing)
  - Participants recommended that the model should rather describe a generic option, and not be institute specific.
  - Participants stated that the option of supplying a link to the data be included in this stage.
  - Participants stated that the option of interested parties contacting the data owner/author, and not uploading data to a repository, be included in this stage.
  - Participants mentioned that the institute's library and information services unit should be the main party (or even the only party) involved with indexing of data.

  - Participants stated that the use of an open access institutional repository does not automatically mean that the FAIR data principles[43] are adhered to. One of the participants suggested that the model include steps to ensure that the concept of FAIR data is addressed.
- Stage 7 (Long-term preservation):
  - The migration of data to an archival format should be done earlier. Currently, it is only done at the end of the rescue project, before the 'Project Closure' stage.
  - LIS Expert 3 stated the following:

    'The archival stage should include the activity of regular monitoring.'

  - The participant further clarified that it is important to test on a regular basis whether all archival formats can be accessed and viewed (i.e., annually test the data).
  - LIS Expert 3 also declared that it is important to state in the guidelines when data should be zipped.
- Stage 8 (Project closure)
  - Participants stated that the promotion of data would depend on the type of data, and data ownership. It was recommended that this stage include guidelines describing how different data stakeholders influence the promotion of rescued data.
  - Participants suggested that the graphics be simplified, as the current portrayal seemed to indicate duplication of certain activities.

Institute-specific critique and observations

The following observations were made by focus group participants:

- It is realistic to expect that most data rescue at the selected institute will be partial in nature.
- It is realistic to expect that researchers at the selected institute will only be involved with data rescue should the data in question be of benefit to them.
- LIS Expert 1 stated that it is realistic to expect this institute's researchers to only be really involved in stage 1, and metadata creation:

  'Researchers probably will only be really involved in Stage 1, and metadata creation. The remainder of activities will be done by CSIRIS.'

[43] https://www.go-fair.org/fair-principles/

- Adding to the above: it is realistic to expect the institute's research library services to perform the remainder of the data rescue activities.
- The model should stipulate whether the library and information services division of the selected institute will accept boxes filled with old data being brought to the library.

All recommendations, suggestions and concerns will be evaluated for inclusion in an updated version of the model. It is important that changes to the revised model (resulting in the final model) result in an improved model, and not merely be included for the sake of having a model different to the previous versions.

5.7.4 Discussion

The previous section clearly shows that the suggestions for changes to the revised model were insightful, diverse and abundant. While the focus group session displayed general approval of the revised model, it was not accepted as a final or usable model. The nature and volume of feedback received during the focus group session paved the way for additional amendments to the revised model. When combining the feedback received with insight gained over time, it was evident that the revised model be edited to accommodate a range of aspects put forward.

The following themes, emanating from the received feedback, were evaluated as being crucial to a user-friendly and usable model:

- Model appearance:
    - certain connector points should be redrawn to display consistency in direction,
    - despite suggestions from participants regarding connector direction, it was decided that all diagrams and workflows be read from left to right, wherever possible,
    - ensure that connectors should contain one arrow point only,
    - ensure that two connectors should not join a single shape, and
    - apply consistency in use of shapes, which should agree with standardised practices (where possible).
- Terminology:
    - model name to be decided on and used consistently; the current descriptions include the words and phrases: flowchart, model, data rescue model, and data rescue workflow model interchangeably,
    - a data rescue glossary should be one of the available guidelines,
    - stage names to ideally involve a verb, rather than a noun (where possible), and
    - stage names to be self-explanatory, where possible.

- Data assessment
  - the model should provide guidelines or a checklist.
- Data rescue team
  - the model should provide guidelines on data rescue team selection.
- Confidentiality of data
  - clarify and describe how confidential data will be treated through the various stages, and
  - of particular importance: data sharing and promotion, and how it relates to confidential data.
- Use of inventories:
  - model to indicate whether the master inventory is to be updated regularly, or recreated in subsequent projects,
  - in general, the use of inventories should be explained better and simplified,
  - consider the option of a project inventory for each project,
  - indicate where inventories should be stored securely, and
  - indicate responsible party for inventories storage and management.
- Metadata:
  - consult previous projects of the group or unit, prior to the creation of a metadata template,
  - a standardised metadata template is not considered feasible should an institute incorporate a range of research disciplines, and
  - consider whether data without metadata can be rescued.
- Data sharing
  - the model should include references to FAIR data, and how repository selection can assist in ensuring data are FAIR, and
  - description of generic options is crucial, as not all institutes make use of the options described in the initial and revised versions of the model.
- Archiving of data:
  - it is crucial to make use of the term 'preservation format' in this graphic,
  - state whether all files are zipped, or whether the zip activity is in parallel with activities not requiring zipping of files,
  - it is crucial that the archives and repositories selected show adherence to best practices, thereby ensuring that activities such as data migration be executed by the repository without it forming part of the data rescue workflow,

  - regular monitoring of archived data should be included in this stage, and
  - the model should indicate that converting data to a preservation format is not necessarily done at the end of a rescue project.
- Project closure:
  - slight amendments to this stage graphic should be made to indicate parallel activities, such as publishing and marketing, and
  - accommodate the requirements of different stakeholders, or different data confidentiality types.

Themes listed above have been deemed valuable for inclusion in an updated version of the revised model. The updated model will be presented and described in the next chapter.

5.7.5 Summary of mini focus group session

This section of the results chapter reported on the findings obtained via a mini focus group session which included this researcher and three purposively selected institutional research library experts (i.e., not SET staff) based at the involved institute.

It was anticipated that the feedback provided during the mini focus group session would produce valuable input regarding both data rescue models and put forward novel suggestions regarding a usable data rescue workflow model. It would assist in exposing this researcher to a different data rescue perspective than had been possible during the previous data collection stages.

Participants preferred the revised model to the initial model, but collected input suggested that the revised model would also benefit from modifications and redrafting. Discussions around the revised model brought about proposed changes regarding several aspects of the model. Categories of critique involved the model's look and appearance, and the terminology used and clarity of content. Suggestions and recommendations regarding an improved model included references to the compact summary, the extended summary, and the stage-specific diagrams of the revised model.

Suggestions related to overall changes to the revised model included supplying links to a data rescue glossary, replacing ambiguous terminology with more suitable terms, making use of verbs instead of nouns (where possible), and ensuring the underlying instructions and meaning regarding connectors and arrows are non-ambiguous.

Suggestions for amendments for stage-specific images were mostly directed at the data planning stage, the data sharing stage, and the data archiving stage. Discussions around the first stage (Stage 1: Preparatory stage) resulted in many suggestions, with clarification regarding roles and responsibilities, data assessment, assessment team selection, and the storage of pre-rescue data required by participants. It is suspected that the multitude of suggestions and recommendations emanating after viewing of Stage 1 are due to the following:

- the stage involving a varied range of activities,
- the stage potentially requiring several guidelines as assistance,
- the stage involving different outcomes (i.e., data destruction, data rescue to proceed, partial or full rescue decided on, or data rescue placed on hold), and
- the stage potentially involving participants from a range of institutional sectors, i.e., research library, research sector, waste division, and ICT.

All suggestions and recommendations were evaluated before being implemented in a next version of the model. Incorrect information, or proposed changes not leading to an improved model, were either discarded or only implemented under certain diagrammatical conditions. An example of such a suggestion includes the proposed change regarding arrow direction, and the recommendations that connectors are not taken all the way to the opposite margin. The uniformity of arrow flow direction from left to right is regarded to be crucial to the consistency of the model.

Following an evaluation of critique received, the revised model was updated. The final version of the model is described in Chapter 6: Recommendations.

5.8 Summary of results

This chapter represented the core findings of the study which were obtained via the various methods applied to collect and analyse information. The four methods used in this study comprised a web-based questionnaire, virtual one-on-one interviews, feedback provided after examining a first version of a Data Rescue Workflow Model, and a mini focus group session. The sample involved with the web-based questionnaire comprised RGLs at the selected institute, while the sample for the one-on-one interviews comprised RGLs who had indicated to have data at risk, or to have performed data rescue activities. In contrast with the mentioned SET-based participants, the mini focus group session involved research library experts, also based at the selected research institute.

The amount of data collected via the various data collection stages was relatively small, and comprised 22 fully-completed short web-based questionnaires, eight virtual in-person interviews and a focus group with three people. All participants are employed at the same research institute. Even though the numbers are small, the data collected and the resultant findings provide sufficient information pertaining to the topics that were investigated on site: the nature of data at risk, factors causing data to be at risk, the data rescue activities performed and the obstacles faced when contemplating or performing data rescue. While RGL feedback on the data rescue model lacked depth and critical input, the feedback is not necessarily linked to the low number of researchers. It was continually observed that the lack of data rescue experience played a big part in the model feedback being of a lower critical quality than anticipated.

A second aspect linked to the low number of participants is linked to the study being a case study. The decision to focus on a specific research institute only, and not involve several institutes (i.e., survey research), was discussed in Section 4.4.4: Rationale for case study use. This study aimed at gathering in-depth information about the specific case, and did not focus on collecting data from a diverse range of research institutes in different stages of RDM implementation and using dissimilar systems. The number of participants who participated in the focus group is small but they are all experts in their own right and was regarded as sufficient for a mini focus group, and for representation of the institute's small research library. Participants were purposefully selected because of their experience; also suitably experienced; enlarging the group would have resulted in attendance by staff with little or no experience of data, archiving, workflow or research practices.

The main findings of the study revealed that data at risk is prevalent within the selected institute, and that a range of factors contribute to data being at risk of loss or damage. Data at risk were found to include all data formats and be present in a range of research disciplines. In addition, older data as well as recently collected data run the risk of being at risk.

Data rescue is not a common activity at the selected institute, with a small percentage of SET-based respondents indicating to having performed any sort of data rescue activity. An even smaller percentage of SET-based respondents were found to have drawn up or be in possession of data rescue documentation. Researchers have also indicated that multiple obstacles and challenges exist when performing or envisaging data rescue activities. All interviewed parties identified several actual or hypothetical data rescue obstacles and challenges.

Only one of the interviewed parties was found to have participated in a data rescue project, entailing identification of data at risk, converting it to a modern digital format, uploading the converted data and its accompanying data documentation to a discipline repository, and ensuring the data are preserved in the long term. An interesting caveat to this rescue venture is that a large part of the project involved the input of external digitisation services.

An initial Data Rescue Workflow Model, created by this researcher, was reviewed by the same group of researchers who had participated in the virtual one-on-one interviews. Participants supplied limited feedback on the model after reviewing the model and its various graphics, guidelines and templates. Feedback was required to assist with the drafting of a usable model to be used for handling data at risk. Feedback received, although useful in certain aspects, was mostly facile in nature and could not serve as sole contributor to a revised version of the initial model. With the passage of time that had occurred after the creation of the initial model, additional insight was also acquired via the academic review process, and by combining this with the feedback received, a revised model was drafted.

Due to the lack of in-depth feedback received from researchers after they had reviewed the initial model, it was crucial to involve more experts in the process of model review. The decision was made to implement a mini focus group session consisting of three research library experts from the institute, and thereby obtain additional feedback and suggestions on the model. Both the initial and revised model were discussed and critiqued by the three research library experts during the session. The outcome of the mini focus group session was a substantial collection of insightful and useful recommendations regarding a workable and user-friendly workflow model for data rescue and was further enhanced by feedback from the academic review. Feedback provided also served to provide a unique perspective on data rescue activities, as previous data collection stages had only involved SET-based experts. Feedback received from research library experts was evaluated, and a large component of feedback ideas incorporated when drafting the final model.

The next section comprises the conceptual reflection of key concepts emanating from this study.

5.9 Conceptual reflection

With data collection, data analysis and the largest part of the discussion of findings completed, a next activity comprises a reflection on the study’s key concepts. This reflection entails a critical summary of some of the findings and learnings, with the researcher’s own perspective also applied to the theoretical literature and factual findings.

The reflection section focuses on three key concepts emanating from the study; the concepts are (i) data at risk, (ii) data rescue and (iii) the involvement of the library and information services in data rescue activities. The conceptual reflection strategy used in this section adheres in part to the ‘What? So what? Now what?’ model of reflective thinking, often implemented after a critical incident that has taken place and one requires the extraction of learning from the event (University of Edinburgh, 2020). This researcher will also be asking the ‘What if’ question, thereby providing a slant on the ‘So what’ stage of reflection, as this slight difference in nuance enables the consideration of alternatives, exclusions and ambiguities linked to the key concepts.

The reflection section also refers to the updated conceptual framework of the study; the framework is displayed in Figure 5.28: Updated conceptual framework of study. When compared with the earlier version of the framework (see Figure 2.1), the updated framework includes additional non-causal links to variables including the data rescue publications involved in the study's content analysis, the various and renamed versions of the data rescue model, the anticipated data rescue activities of the library and information services sector, the required data rescue training, and the effect of data rescue involvement on the library and information services identity.

5.9.1 Data at risk

The findings of this study have revealed the preponderance of data at risk, with such data existing in all geographic locations, present in diverse research disciplines and displaying a great range in data formats. In addition, the age of such data varied between being a few decades old to data collected several centuries ago. The non-causal link between data at risk and the variables stated is also portrayed in Figure 5.28: Updated conceptual framework of study. As a result of realising the prevalence of data at risk, this researcher arrived at the following questions regarding this key concept at hand:

- What data are included when stating that data are at risk? What data are excluded when referring to the concept of data at risk?
- Is the current definition of data at risk the most suitable one?
- Are data at risk merely the result of sub-par data management practices?
- Besides rescuing data, what will limit the prevalence of data at risk, and are current data collections not being subjected to the same fate?

Different definitions are used to describe the data involved in rescue studies, and four of these are provided below as examples. The four definitions were selected as they cumulatively represent the viewpoints of a data rescue interest group, two authors of a book chapter on data at risk and two prolific authors in the areas of data at risk and data rescue.

- CODATA refers to such data as the 'many sets of scientific data which are not in modern electronic formats and whose information is therefore not accessible to the research that needs it' (2013). This description refers to and assumes that modern electronic data are accessible, that rescued data are in a modern electronic format, and that all data at risk can be converted to a modern electronic format.

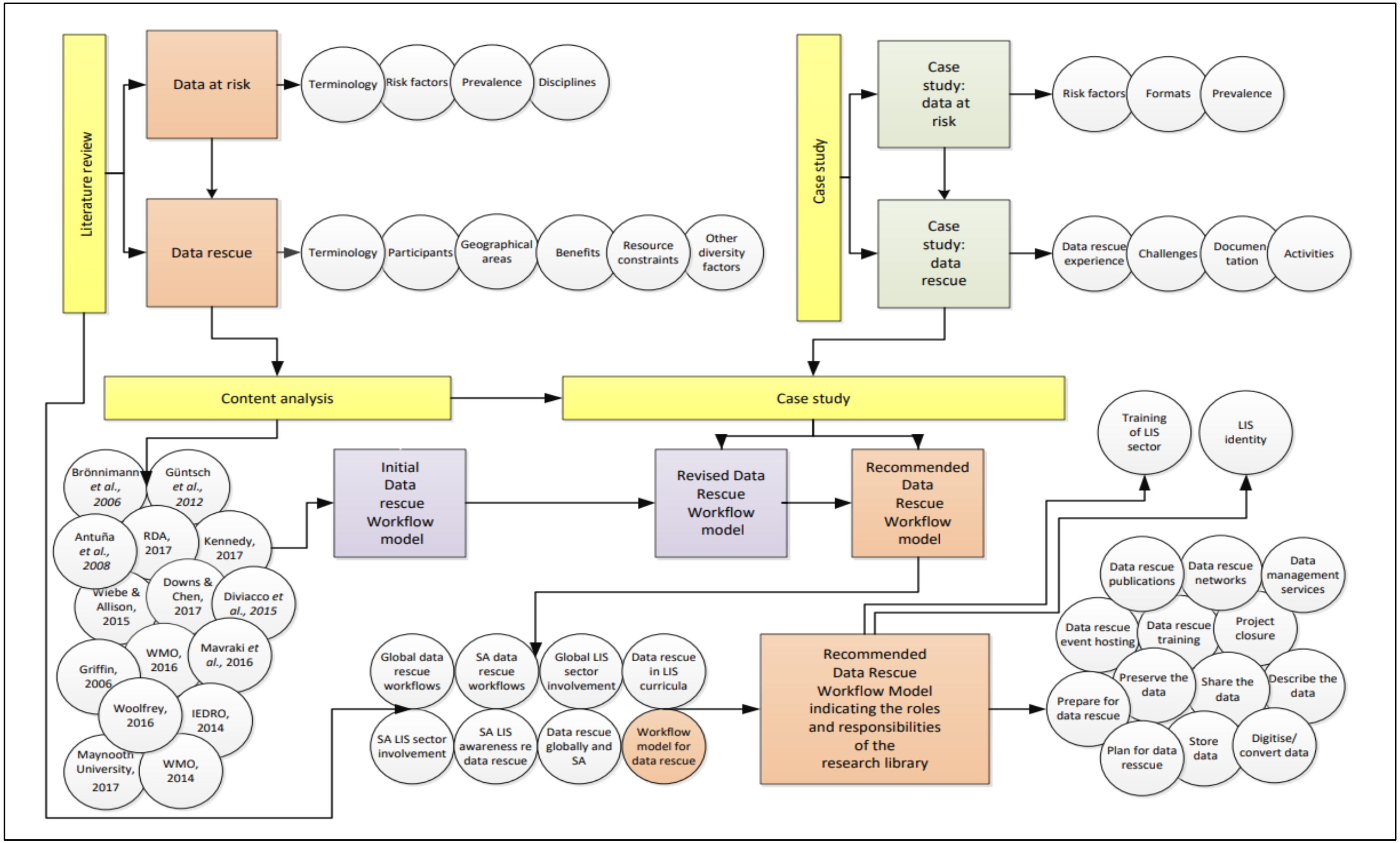


Figure 5.28: Updated conceptual framework of study

- Downs and Chen (2017) state that the data are at risk of loss. This description refers to and assumes that loss is the primary and only reason why data are at risk, and in need of rescue.
- Murillo adds a different slant to the concept by referring to such data as being 'endangered' (2014). This description is novel in that it considers that a range of threatening variables may affect the longevity of data.
- Griffin (2005) includes the words 'lost' and 'endangered' in an early publication and also subsequently states that 'data at risk' is a blanket descriptor of the non-electronic (mostly pre-electronic) data which are subject to a multitude of hazards of various types and origins (2015). While these descriptions are more comprehensive than the preceding definitions, they do not consider the risk-related outcomes of data in a modern format.

A critical look at the definitions used in the cited examples has resulted in recommendations regarding awareness of data at risk (see Section 6.4.1: Create awareness around data at risk). Suggestions comprise the following: descriptions of data at risk should also refer to physical samples and specimens, data being in a modern format does not ensure that the data are not at risk, and the definitions emphasising loss as the only risk-related outcome is not sufficient. To clarify: a hypothetical collection of moon dust specimens (i.e., samples data), currently stored inappropriately and without metadata is as much at risk as are hypothetical moon data (i.e., images) currently housed on older format magnetic tape where the tape reader is in need of repairs. In the same vein: modern data stored on a portable hard drive only should be seen as equally at risk-prone as older paper-based data housed in moth-ridden boxes in a dusty backroom. It is important that the idea of 'loss' linked to data at risk should not be seen to only refer to the quality of no longer being available, but to also include conditions such as being partially damaged, or no data context being available. For this reason, this researcher maintains that valuable modern digital data and samples, suitable for sharing, may also run the risk of being endangered, and has therefore included these formats in the study's recommended Data Rescue Workflow Model (see Section 6.2.1.7).

With the realisation that references to data at risk should be clearer and possibly wider, a follow-up issue relates to the topic of risk, and the nature of risks referred to. Are modern data, stored appropriately, backed up securely, accompanied by excellent metadata and data documentation but sensitive in nature, deemed to be 'at risk'? Such data certainly does carry the inherent risk of revealing information to parties outside of the agreed-on audience. A second example: does the inadvertent non-destruction of sensitive medical data, contrary to what was stated in the informed consent agreement, render the data to be 'at risk'? There is clearly an element of risk present, as the risk of unethical data access or sharing has now been intensified.

Should the answer to the two scenarios above be 'yes', then this would imply that all data are linked to an aspect of risk. This is certainly an opinion to be considered, and one closely linked to a comment made to this researcher after explaining the gist of her studies to the interested party, i.e.: '*But all data are at risk!*' Modern digital data are indeed at risk of being hacked; Johar mentions malware, malicious mobile apps, physical security threats and insecure networks as common methods of gaining access to data (2017).

Should all data then potentially be at risk, it becomes necessary to convey with more clarity the type of data that would benefit from the steps stipulated in this study's data rescue workflow model. Merely referring to data at risk could potentially convey the misleading idea that the hypothetical non-anonymised HIV studies data, stored securely and yet at risk due to an insecure network, are to benefit from the rescue model. Author Services (2023) listed several types of data that should not be shared, and one could rightly state that there are grave risks linked to the incorrect handling of modern personal data that are not anonymised, modern data that are not owned by the researcher, commercially sensitive data, modern data posing a security risk, *sub judice* data and data linked to threatened species. However, these types of data will not benefit from the data rescue workflow model, and do not fall into the same category as data at risk that need to undergo the stages included in the model. Such data are only included in a data rescue workflow model should the data need to be shared within the research group (or only with parties having permissible access), or when the ethical sharing of data, under certain restrictions, is permissible. This researcher wishes to emphasise that this study's data rescue model, and future adaptations thereof, should not take the place of a data management model aimed at ensuring ethical sharing of data, or a model stipulating how sensitive data should be managed and preserved. The rescue model discussed in this study deals with the treatment of data of value to the research community, or under rare circumstances, sensitive data to be accessed in future by the research group or researcher only.

A clear distinction should be made between data running the risk of unethical research behaviour and accidental exposure, and of data deemed to be at risk due to its outdated format, lack of metadata, poor storage conditions, or deterioration of the medium. While the former type is linked to best research practices and adhering to ethical guidelines, the latter are data falling under the definition used in this study. Indeed, study participants were informed about the type of data referred to by this researcher during the data collection phases, with the following definition supplied (see Appendix 2): 'For the purposes of this study, data at risk is defined as research data in any format at risk due to:

- deterioration of the media (e.g., paper/microfilm damage, pest damage, war/unrest, loss of inventories), or

- catastrophic loss (e.g., only one set of records, vulnerability to fire/floods/disasters, obsolete storage media, obsolete file format, archival destruction, loss of human knowledge related to use of the dataset).'

A second issue emanating from reflecting on data at risk pertains to the number and range of factors, found in documented outputs and stated by interview participants, causing data to be at risk. The combined list of factors is extensive and reveals that a range of factors ultimately play a role in affecting access to and use of data. While the implementation of a data rescue project is an effective way of ensuring the data is accessible in future, merely relying on data rescue activities is not the only way to address the prevalence of data at risk. Implementing data rescue clearly has benefits for the specific datasets being rescued, but does not ensure that the causes of data being at risk are addressed, and does not result in the implementation of steps minimising the occurrence of data at risk in future.

The introduction of research data management (RDM) in the research sphere and the accompanying DMP forming part of data management best practices is an excellent way of minimising the prevalence of data at risk. Adherence to an approved DMP ensures that metadata accompany data, that applicable data will be preserved in the long term and that the access conditions surrounding the data are considered and met. However, RDM will not influence, affect or change the activities linked to data generated prior to the introduction of RDM at an institute. Aspects such as loss of human skills (Levitus, 2012; WMO, 2014; Wyborn *et al*., 2015), changing priorities (Arrouays *et al.,* 2017; Gaudin, 2017; McGovern, 2017) and funding and policy (Gaudin, 2017) are examples of potential continual research upheavals that may affect the handling and storage of data. At the selected research institute, participants have eluded to the value of data not being recognised, researchers being unaware that data are at risk, and the poor handover activities when researchers leave the research group or institute. When these practices form part of an institute's research culture and are no longer frowned upon, mitigating steps might be required to ensure that the proliferation of data at risk does not continue. Recommended steps to deal with the many factors putting data at risk include awareness training regarding data at risk and risk factors, formalised handover steps linked to data-related matters when a researcher retires or resigns, consideration of a dedicated institutional data repository, and formalised steps regarding data, data readers and data skills whenever a research group is terminated, or institutional refocus activities are implemented. Chapter 6 expands on recommendations pertaining to data at risk.

Ultimately, data at risk are as much a part of the research environment as laboratories, grants, experiments and scientific outputs. The fact that the research library and researchers are discussing data at risk is an indication that divisions working with data have taken note of the data management errors of the past and are investigating ways to ensure that history does not repeat itself.

5.9.2 Data rescue

What is data rescue? A basic web search will most likely convey to the general internet user that data rescue refers to the recovery of data, usually by means of specialised software supplied by a commercial entity. To a researcher, a first encounter with the term might bring about visions of an expert from the institute's ICT division executing a series of activities in an effort to recover said researcher's data stored on the crashed laptop's hard drive. In casual conversations with colleagues and friends, the aforementioned description – the recovery of modern digital data that have crashed, or been hacked, or are somehow inaccessible – was determined by this researcher to be the perspective often held by other parties. Within the context of this study, data rescue refers to a slightly different issue: the process of securing data at risk of being lost due to deterioration or simple obsolescence of the storage media, natural hazards, theft or vicious destruction, and ensuring that data can be easily accessed and used. While scholarly information about data rescue can be found via use of scholarly search engines, locating grey literature regarding the topic is a more complicated task. Additionally, the descriptions of data rescue, as seen in the context of this study, also reveal an aspect regarded as worrying (and thought-limiting) by this author. Many studies define the process as one targeting data that 'had been recorded manually and mechanically on paper and filmstrips' (IEDRO, 2014), or 'data, from both film and tape' (Gallaher, 2015) or 'consolidating the paper records' (ACRE, 2019). Refreshingly, the singular instances of reporting on the rescue of samples (Stanley *et al.,* 2020) as well as metadata (Hills, 2015) show that the rescue of data at risk is wider than the rescue of only paper-based data and early digital data.

Based on the aforementioned, two issues immediately come to mind, and each is questioned in turn. With the use of the term 'data rescue' in the mainstream media automatically being associated with the recovery of compromised modern digital data, is the use of the term 'data rescue' still correct, and is the term relevant to the activities referred to in this study? Will the activity of data rescue ever gain a foothold, even in the research environment, when the use of the term immediately directs the attention to ICT services and the retrieval of inaccessible institutional electronic records? It would appear that this is an issue worthy of consideration, and the idea that the term 'data rescue' could in future be replaced with 'data conservation' (Hills, 2019) or even a novel, innovative term is a likely scenario.

The second issue is linked to the discussion involving formats, presented in Section 5.9.1: Data at risk. With many data rescue studies reporting on the rescue of early digital data, what would be the reason for the lower incidence of documented rescue linked to samples at risk and even modern digital data at risk? Possible explanations for this might well be that modern electronic data benefit from recently-implemented RDM practices, that samples are generally less susceptible to deterioration, or that the projects dealing with the rescue of samples are not publishing their data rescue exploits. Whatever the reason, the hope is expressed that details of these projects are also made available and that the information contained within will form part of the current arsenal of data rescue literature.

Despite the prevalence of paper-based and early digital data in published data rescue outputs and the scarcity of rescue information linked to other data formats, this researcher has included the rescue of samples and modern digital data in the study's model. The quest for wider inclusion in data rescue projects may even be taken one step further: within the data rescue sphere, should data-related rescue activities not include and encompass more than only data? With different data formats already included in the rescue model, consideration should also be given to the rescue of data-related facets such as the rescue of data readers, the preservation of skills and knowledge linked to the data at risk, and the preservation of data rescue skills. When taking all of these into account, it becomes evident that in the bigger scheme of data rescue that proper handover, job-shadowing, succession planning, creation of asset inventories and comprehensive data rescue reporting belong in the data rescue sphere.

Data rescue has been described as the process of identifying threatened data and capturing the data in digital format for long-term preservation and reuse (DataFirst, 2022), as the recovery and reuse of scientific data that were not initially accessible for research (Elsevier, 2016) and an activity that assures that future generations of scientists and other data users have access to all the information necessary (WMO, 2016). These cited descriptions convey the idea that the sharing of data is a crucial data rescue activity, and it is worth noting here that all data rescue studies examined by this researcher included the sharing of rescued data. One could go so far as to say that the main driving force behind all rescue ventures is the intention to make data available to future researchers. The section of this study explaining the rationale behind data rescue also lends credence to this fact (see Section 2.5: Rationale for data rescue); outcomes such as extending a subject's knowledge base, providing impetus for future research projects and extending the coverage in data repositories are all linked to the sharing of rescued data with a wider audience.

A crucial part of any data rescue venture, and one which cannot be illustrated with ease via a diagram or model, is the role of teamwork, collaboration and cooperation when rescuing data. Even in this study, the inclusion, illustration and description of the 'team' aspect is limited, as the model is generic in nature and the team composition and importance are bound to vary between projects. Nevertheless, the discussion around data rescue teams is an important one, even in a study that focuses on the input of a specific sector of the selected research institute. It needs to be realised that successful data rescue will not be achieved when only making use of the research library, or when implementing a rescue project and not involving sectors outside of a research group. Data rescue is a collaborative effort, and while a project might be driven or managed by a single institutional sector, the inputs and skills of many sectors form part of a fruitful rescue venture.

The fact that a successful data rescue project is based on the inputs of different participants, with various skills, is one demonstrated in many studies. The WMO, in providing guidance pertaining to hydrological data rescue, states that a key factor in the success of their past data rescue projects was the inclusion of an appropriate mix of technical skills (2014). The skills included managerial skills and responsibilities, corporate, scientific and technical understanding, and dedication to hard work with attention to detail. Another example of collaborative data rescue efforts is the description of US federal data rescue events as involving librarians, scientists, technologists, and other open data advocates to build a broad and resilient coalition (Allen, Stewart & Wright, 2017). Another interesting collaborative effort is the one involving the European Space Agency (ESA) and the Vatican Library and is a striking example of successful cooperation between two organisations from different research disciplines (European Space Agency, 2018). This effort comprised the preservation, management and exploitation of archived Vatican documents and records by using the ESA's 'FITS' format to ensure that future generations will have access to the digitised religious books.

The recovery of historic moon data on old IBM Mark V data tapes was thought to be an impossible vision, as a reader for the tapes could not be located. Through cooperative efforts involving scientists, NASA, Sydney University, a data recovery company called SpectrumData and the Australian Computer Museum Society, a compatible IBM tape drive reader was eventually traced and the data recovered (Modine, 2008). Lastly, a request placed on the internet and on social media sites by the Swedish National Archives in 2019, asking whether there were institutes in possession of obsolete equipment and willing to donate the equipment resulted in productive responses, thereby enabling access of records held by the archives (Open Research Foundation, 2021).

These examples illustrate that data rescue should not take place in a vacuum. The expectation that a research group can independently rescue their own data, or that a sector of a research library can execute data rescue unaided is neither ideal nor likely feasible. Assembling a data rescue team at the start of a project is an important preparatory step, and it is vital that a rescue team contains a combination of expertise linked to the research discipline, the formats in question, digitisation, and data management tasks such as the creation of a DMP, repository selection and metadata standards and creation. Through gaining insight into the collaborative aspect of data rescue, this researcher is now also cognizant of the fact that asking RGLs about the group's past data rescue projects was an unrealistic and misinformed step.

An aspect linked to data rescue collaboration between the research library and researchers relates to buy-in: the research library should be prepared for varying levels of willingness when approaching researchers and requesting participation in a data rescue project. It is anticipated that willingness could range between eager participation to outright refusal. With this study's interview findings showing that researchers are concerned about the effort and time required to rescue data, that manpower is limited, that they might not have rescue skills, and that data location is often difficult, this researcher is aware of potential backlash during the initial phases of future data rescue projects.

Additional aspects which could potentially influence researchers' eagerness to be involved with the rescue of their data include issues linked to resource constraints, indicated in a study by Rappert and Bezuidenhout (2016). With findings revealing that researchers have negative data sharing perceptions, show low levels of data sharing, convey reluctance to share data and experience difficulties in gathering data generated by students, the outlook for agreeing to participate in a crucial data rescue activity – sharing the data – is rather bleak. Add to this the related findings regarding heavy workload and it is likely that embarking on a research project is bound to be met with bouts of reluctance.

Data rescue entails a collaborative effort in which endangered data, in any format and of value to future research, are subjected to a series of steps including sharing the data and ensuring data are available for future use. As discussed in Section 2.5, rescuing data is a way of extending the knowledge base of a subject, it increases scientific accuracy and provides impetus for future research projects. Implementing a library-driven data rescue project could be met with enthusiasm from researchers excited about sharing their data with a global audience, or with discontent and reluctance from resource-constrained scientists who regard data as personal property and might have experienced negative data sharing outcomes. Obtaining enthusiastic buy-in after deciding to embark on a data rescue project within a research institute might prove to be a skill just as vital as the data rescue skills linked to digitisation, metadata creation and repository selection. Add to this equation the expected involvement of the institute's ICT sector and Communications department, and the planned data rescue project becomes a balancing act of persuasion, pleading and promises of beneficial outcomes. The sobering statement, 'we will not be able to save all data' (Griffin, 2015) was made by the 2015

Chair of the International Astronomical Union's Working Group on the Preservation and Digitization of Photographic Plates and the Chair of the 2015 CODATA Task Group on Data At Risk, and even though her words refer to the voluminous nature of data at risk, it might also be applied to the potential reluctance of researchers to have their data rescued and shared.

The next segment entails reflecting on the recommended and future involvement of the library and information services during data rescue.

5.9.3 Involvement of library and information services in data rescue

A key objective of this study was to indicate the roles and responsibilities of the research library within data rescue. In determining the sector’s involvement, it also became necessary to contemplate the implications of promoting the assimilation of data rescue by the LIS sector. It is imperative that the potential factors affecting the discipline’s adoption by the LIS sector, the potential LIS training requirements and the potential effect of data rescue inclusion on the LIS identity be envisaged and considered. This section reflects on the recommended adoption of data rescue by the library and information services sector mainly via a retrospective look at factors linked to the implementation of RDM by the same sector, roughly a decade ago.

The research library has not always been involved with data-related activities and it would be incorrect to view the LIS sector’s current association with RDM as an established one. Compared with ‘traditional’ library services such as cataloguing, loaning of material, literature searches and collection management the inclusion of research data management in its service streams is a new addition. The period 2012–2014 was a time of upheaval in many library circles, with the concept of RDM being introduced to the LIS sector, RDM training materials and training opportunities surfacing, and new data-related LIS positions being created. This period coincided with the implementation of the new mandate by the US National Science Foundation (NSF) during January 2011, requiring researchers to include a DMP with their proposals for funding (Harvard Medical School, Research Data Management, 2018). As a result of this step, researchers, new to the DMP concept, were approaching libraries, as default providers of research services, for assistance.

By 2012 the Liber Working Group On E-Science/Research Data Management had already published an online report and guide containing 10 recommendations for libraries to get started with research data management (2012). As expected, the primary recommendation pertained to DMPs and the provision of support linked to the topic. Additional recommendations included the development of metadata standards and provision of metadata services, participation in institutional research data policy development, providing services for data storage, discovery and permanent access and promotion of research data citation. Even back then, the idea of the LIS sector becoming part and parcel of data-related services, with the involvement of the library estimated to include far more than involvement in DMP guidance, was starting to take hold.

A year or two later and the advent of scholarly publications related to libraries and their involvement in RDM had started. Examples of this trend include the 2013 article by Cox and Pinfield that looked at current activities and future priorities of libraries involved with research data management, and the article by Antell *et al*., investigating the participation of science librarians in data management (2014). In a similar vein was Burnett’s 2013 article posing questions about the role of the librarian in RDM. The use of non-standardised terminology and overlapping of RDM with similar library tasks are evident when reading the article of Creamer *et al*., titled ‘An Assessment of Needed Competencies to Promote the Data Curation and Management Librarianship of Health Sciences and Science and Technology Librarians in New England’ (2012).

It had taken less than three years since the NSF DMP mandate implementation for RDM to be associated with libraries. As a result, the prevalence of RDM training geared towards this sector had also increased. Online publications such as the *Data Management for Libraries: A LITA Guide* (American Library Association, ALA Store, 2014) and online training courses such as the University of Edinburgh’s MANTRA (Rice & MacDonald, 2013) and the Digital Curation Centre’s three-hour-long RDM training course for librarians (Jones, Guy & Pickton, 2013) provided evidence of the strong emerging link between RDM and libraries.

Even though the research library linked to this study’s selected institute did not experience RDM interest from researchers during this era (i.e., 2011–2014), the institute’s library had begun to show signs of RDM inclusion since 2013. In 2013, and several years prior to the appointment of a data librarian or the drafting of an institutional data management policy, an institutional research library representative delivered a presentation on novice data managers at the 5th African Conference for Digital Scholarship and Curation (Patterton, 2013). A year later, the same research LIS professional had created an animated video titled ‘An introduction to the basics of Research Data’ and shared it via the YouTube platform (Patterton, 2014). Slowly and surely, and prior to the establishment of research data services at the selected research library, RDM was being included as a part of the portfolio of institutional library services.

A decade later and the RDM–library link is prominent at several South African institutes: the University of Cape Town libraries has RDM as part of its digital library services (University of Cape Town, UCT Libraries, 2023), the University of Pretoria has a Research Data Management page on the Department of Library Services platform (University of Pretoria, Department of Library Services, 2022), and UNISA's library offers Research Data Management services as part of research support (UNISA, Library, 2022).

Based on the information provided above and the proposed inclusion of data rescue in the research library repertoire, the questions may well be asked: What development path is predicted for the data rescue–research library association? Is the proposed involvement a bridge too far or is it the start of a seamless integration of a new service stream into the library sphere? This researcher considers several issues to count in favour of the proposed union between data rescue and the research library, and four conducive factors are listed in this section.

1. **Current involvement in related activities:** Many libraries are already participating in activities showing great similarity to activities forming part of a generic data rescue project. The documented involvement and activities are described in Section 2.6.4.2, and based on literature and mini focus group feedback the adoption of data rescue by the research library is a logical next step.
2. **Involvement and knowledge of RDM:** An aspect linked to the previous paragraph relates to the fact that the LIS sector has successfully adopted RDM, with many research libraries providing RDM training, employing data-experienced LIS professionals, hosting institutional data repositories and being involved with the quality control of metadata and data uploaded to repositories. The overlap of RDM with data rescue is a factor that can only enhance the research library adoption of this new service stream.
3. **Adaptation to change:** The LIS sector's history regarding change and adaptation is admirable and a source of motivation. Major changes during the last five decades include the introduction of CDs, the introduction of computers, the internet, electronic books, electronic communication, electronic scanning, electronic library systems, digitisation and the option of remote work. New positions have also been created, and the selected research library is a fitting example: 2019 saw the introduction of three new positions at the library, namely, a systems librarian, a data librarian and a digitisation clerk.

   Drummond is of the opinion that change is important and that librarians need to confront change head-on in order to evolve and grow as information professionals (2016). Drummond further stated that the profession must accept uncertainty in the workforce and that working with change is more beneficial to development than continuing to work the way the sector always has.

However, the ability to adapt to change does not imply that the process of change takes place in the absence of concerns and uneasiness. In line with this, Antell's 2014 survey of science librarians explored the RDM roles and responsibilities, both new and traditional, that science librarians had believed were necessary to meet the demands of RDM (2014). The results revealed feelings of both uncertainty and optimism, with participants uncertain about the roles of librarians, libraries, other campus entities and the skills that would be required. Despite the uncertainty optimism was also detected, with participants excited about applying current skills to this emerging field of librarianship. Additional challenges emanating from Burnett's study comprised finding the time for RDM, having to learn new skills, establishing credibility with researchers and learning to deal with unpublished work and raw data, as opposed to published materials (2013). A study by Cox and Pinfield refers to the RDM challenges associated with skills gaps, resourcing and cultural change (2014), while related findings were reported by Perrier, Blondal and Macdonald (2018), who stated that libraries were experiencing uncertainty around RDM roles and relationships.

The examples above illustrate that the ability of the library and information sector to deal with change is an advantageous trait; when confronted with change, libraries show realism in being apprehensive about new skills required and services to be provided, but also portray fortitude and enthusiasm, enabling the successful integration of a new service stream. When reflecting on the developments during the past decade, the almost effortless entrenchment of RDM within libraries is testament to the sector's resilience and adaptation to change. Based on this factor only, it is likely that the research library's uptake of data rescue tasks will follow a similar trajectory.

4. **Adapt or die:** The final factor predicted by this researcher to enhance the adoption of data rescue by the LIS services sector is the harsh reality that research library disinterest could result in disastrous outcomes for the library. As stated by Andrikopoulou, Rowley and Walton, libraries should grasp new opportunities and add to their portfolio as a means of ensuring their own future (2022). The authors further declared that the contribution of libraries should be valuable and visible and that failing to deliver may result in their roles being taken over, aspects of roles being downgraded and status and salaries being affected (2022). The emergence of data rescue is therefore not only an opportunity to be grasped, but one that could ensure

the future of this service support sector in times of uncertainty, restructuring and downsizing.

The four factors mentioned above paint a positive picture of the anticipated adoption of data rescue by the research library. With the study proposing that libraries be involved with data rescue, and the research library being a plausible candidate for such a role, it is worthwhile considering the potential implications of this additional service stream for the identity and character of the research library and the research library professional.

The opinion of Klein and Lenart regarding the identity of librarians is thought-provoking; the authors claim that there 'never was, and never ought to be a fixed essentialist librarian identity' (2020). According to them, the librarian identity has always been fluent, dynamic and responsive to user needs. Explanatory details regarding this stance refer to the exclusion and homogeneity in the workplace, high attrition rates of minority librarians, exploitation and alienation of an underrepresented workforce as well as stereotyping. The LIS identity, in adapting to the demands of a dynamic profession, is therefore described as complex and fluid.

Tucker mentions the importance of lifelong learning, of acquiring new knowledge, of including LIS in more environments and of career shifts later in life (2021). The author refers to the concept of an 'information architect', and associates the term 'evolving' with the librarian's skillset, mindset, and professional identity.

Hays and Studebaker discussed the librarian trait of being 'open-minded' and how this can lead to the adoption and assumption of a teacher identity, for example (2019). The study by Fraser-Arnott links to the aforementioned studies by claiming that a shift in library and librarian identities occurs as libraries and library workers evolve and explore new practices (2021), while Andrikopoulou, Rowley and Walton state that involvement in and leadership of RDM university practices has the potential to re-shape the library's role, image and identity (2022).

Two pertinent questions, linked to librarian identity, are posed by this researcher when considering the evolving nature of the LIS identity and data rescue adoption:

- How will research library professionals, responsible for many data rescue activities, view themselves? Will it be as data curators, data librarians, data rescuers, data conservationists, or even a novel term? This question emanates from a similar question asked by Andrikopoulou, Rowley and Walton, when discussing professional identity and allegiance of research data managers (2022).

- How will the research library deal with the hypothetical potential overlap between data rescue services and similar services in an institute, and will similar skills sets affect allegiance over time? As was the case with the previous question, this question emerged after viewing a study by Andrikopoulou, Rowley and Walton regarding data management, library identity and institutional sectors (2022).

It is far too soon to answer the questions above and to predict the effect of data rescue adoption on the LIS identity. What can be stated with certainty is that the sector's history has shown clear evidence of the identity being in flux through continually evolving, adapting to new user needs, undergoing job title changes and dealing with the introduction of new tools, equipment, systems and services.

With the identity of the librarian showing affinity towards change and transformation, the perception might be created that implementation steps will entail smooth sailing on the LIS-linked ocean of acceptance and serenity. Such a stance is not necessarily correct; data rescue assimilation has significant implications for many issues and the aspects linked to training is a topic worthy of consideration.

The inclusion of data rescue into the research library's service stream is bound to have implications for the training of the sector, the tertiary LIS curriculum, and the training of researchers. As a reflective first step it is worthwhile taking note of aspects linked to RDM training during the past decade, such as training requirements stated by RDM library staff, RDM skills and attributes required and the range of available training options and platforms.

As early as 2013 the statement was made by Burnett that the LIS community should consider whether 2013 was the time for the subject of research data management to move into the educational mainstream and to include RDM in the curriculum of newly developing postgraduate LIS courses (2013). The prediction proved to be correct; data management is now an established postgraduate LIS subject or module and forms part of the degree contents of many LIS qualifications. The module on data management, forming part of the honours degree in Information Science at the University of Pretoria, is a fitting example (University of Pretoria, Information Science, 2022).

The implementation of RDM as an LIS service has been accompanied by studies investigating the training requirements of library professionals working with data, or even studies providing insight into the skills and traits ideally displayed by librarians working with data. Examples of these outputs include the study involving Ghanaian university libraries, reporting that technical skills and competency sets should be improved via training and other capacity building programmes (Frederick & Run, 2019), and the study by Tang and Hu, which stated that institutional commitment to resources and training opportunities is crucial if RDM services are to grow (2019). RDM training requirements have also been shown to be strongly linked to the need for cooperation and sharing of knowledge. The need for stronger collaboration with faculties or researchers to develop more discipline-based curriculums for RDM and more application-based approaches for teaching RDM (Xu, 2022), and stating the importance of the development of a global community of practice where data librarians work together, exchange information, help one another to grow, and strive to advance RDM practice around the world (Tang & Hu, 2019) are testament to this fact.

Federer's study into ideal data librarian qualities has revealed the existence of two types of data librarians: data generalists, who provide data services across a variety of fields, and subject specialists, who provide more specialised services to a specific discipline (2018). Her findings also suggest that data librarians provide a broad range of services to researchers and therefore need a variety of skills and expertise. In addition, RDM workers considered a broad range of skills and knowledge to be important to their work, especially 'soft skills' and personal characteristics such as communication skills and the ability to develop relationships with researchers. Traditional library skills such as cataloguing and collection development were considered less important (Federer, 2018).

RDM librarians are currently able to make use of various training options geared towards their profession: the RDM course run by the Library Juice Academy (2023) and the established online MANTRA course managed by the University of Edinburgh (EDINA and Data Library, University of Edinburgh, 2013) immediately come to mind. The Research Data Management Librarian Academy is another valuable RDM training platform, and is described as a free online professional development programme for librarians, information professionals, and other professionals who work in a research-intensive environment throughout the world (Research Data Management Librarian Academy, 2017). Another helpful contribution to the field of RDM training is the publication by Barbrow, Brush and Goldman containing links to data rescue training materials, or resources that may prove beneficial to novice data rescuers (2017).

Lessons learnt via an examination of the RDM training path and applicable to the concept of data rescue training for research library professionals are listed below.

- It is not too early to include data rescue as part of postgraduate RDM modules at universities and other tertiary education establishments.
- It is important that ideal and mandatory data rescue skills and traits be identified, and that data rescue training options and materials consider and include the identified aspects. It might

also be beneficial to distinguish between a research library professional who is a generalists data rescuer, and one who holds subject expertise.

- Adding to the point above: training options should cater for the novice data rescuer and the experienced rescuer.
- Online training options, be it in the form of an academy, a course or tutorials, should be created and widely advertised.
- Data rescue groups and collaboration, exchange of ideas, and sharing of learned knowledge and experiences are bound to be required by data rescue practitioners.
- The creation of an annual updated compilation of data rescue training options, materials and community groups is strongly suggested.
- Ideally, training should be provided not only by presenters with theoretical knowledge of data rescue, but also by representatives from both the research sector and the research library who are able to impart knowledge gained through practical data rescue experience.

Based on the above, several training recommendations are put forward in Section 6.4.12: Amend and adapt the LIS curriculum to include data rescue.

It is not possible to predict with certainty what the future holds for the proposed LIS sector–data rescue association, but reflecting on the past RDM–LIS issues has provided glimpses into the integration of a new service stream into the research library's portfolio of services. As mentioned in the section on the LIS sector's ability to adapt to change, the uptake of RDM by the LIS sector was not free from obstacles and uncertainties around the implementation of a new service. Despite challenges, the discipline of RDM is now an established and flourishing part of the library and information services sector of many institutes, globally and in South Africa. With RDM showing many similarities with data rescue, including activities performed, skills required, system familiarity, data understanding and suitable candidates for inclusion, the assimilation of data rescue into the research library environment has the potential to show similar success under the right conditions and support mechanisms. Promotion of data rescue should ideally be accompanied by an understanding of the feedback, events and outcomes that had surfaced during the recent LIS adoption of RDM.

Rambo provides helpful advice regarding new LIS services by stating that a new library service can only be successful when researchers are able to see the value in what they are requested to do, that the library should ensure that there are sufficient resources (primarily in the form of people) and ICT expertise to dedicate to a new effort, and that libraries should demonstrate a degree of flexibility (2015). By implication, successful data rescue implementation requires that the benefits of data rescue are known to researchers, that ICT are informed about their roles and are in agreement with their inclusion and able to provide the requested services, and that the library has dedicated and skilled employees forming part of the research library data rescue team. The library should be adaptable when it comes to issues around visits to research groups, deadlines imposed upon the research sector, expectations regarding researcher involvement in data rescue, and data sharing restrictions concerns expressed by researchers.

The next chapter entails establishing whether the study's research questions, stipulated in Section 1.4, have been answered. The chapter will also present and discuss the recommended Data Rescue Workflow Model. Recommendations with regard to the implementation of data rescue at a research institute, the limiting of data-at-risk factors, and addressing data rescue challenges, are also presented. Additional recommendations regarding the training of parties involved in data rescue, and including the topic in LIS curricula, are set forth.

CHAPTER 6: RECOMMENDATIONS

6.1 Introduction

The previous chapter portrayed and discussed the results of the study obtained via the different data collection methods. These methods comprised a web-based questionnaire completed by the selected institute's research group leaders (RGLs), virtual-one-on-one interviews held with selected RGLs, feedback supplied by interviewed RGLs, and findings emanating from a mini focus group session held with selected research library experts. Findings obtained via empirical methods revealed details on the selected institute's data at risk, data rescue experiences, and data rescue challenges encountered. It also provided details regarding requirements for a data rescue workflow model, and information on the required adaptations and modifications pertaining to the initial Data Rescue Workflow Model to suit institutional data rescue needs.

In this closing chapter, the researcher presents the study's main research question and research sub-questions and discusses how these questions have been addressed. The section containing research questions is followed by a summary of the study's main findings, after which conclusions regarding the findings are presented.

Based on the study findings and considering the main research questions of the study, several recommendations have been put forward. Ideas for future research emanating from this study are also stipulated.

This chapter is ended by a study conclusion.

6.2 Research questions

This section addresses the study's sub-questions, the main findings obtained when answering the sub-questions, and the resulting implications before moving over to the main research question. The research questions, used to guide and centre the research, are addressed in reverse order. Reverse order was selected, as answering each of the sub-questions successfully would in turn lead to and enable the answering of the main question. Conversely, not addressing or not being able to answer any of the sub-questions would lead to difficulty in answering the study's main research question.

A practical example of the relationship between the study's main research question and sub-questions is as follows: through identifying current data rescue workflows and guidelines, gathering information about library and information services participation in data rescue, and formalising acquired data into a data rescue workflow model, this researcher was in a favourable position to describe how the roles and responsibilities of the research library can be included in a data rescue workflow.

6.2.1 Research sub-questions, findings and implications

This study endeavoured to answer eight research sub-questions, the main findings of which are briefly stated in this section. This section also refers to relevant previous sections in the text for elaborative details.

6.2.1.1 The current data rescue frameworks/workflows

The findings related to this research question were detailed fully in Chapter 3: Literature and the creation of a data rescue workflow model, and to a lesser extent also addressed in Chapter 2, comprising the literature review. Details supplied below contain a summarised account of documented data rescue frameworks and workflows.

Overview: Data rescue workflows, models and processes were examined to ascertain typical, generic, or tried-and-tested data rescue steps. Workflows and frameworks were also scrutinised to establish whether discipline-specific workflows and rescue activities differ from discipline-agnostic steps, and to investigate the possible description of participants and their role and responsibilities (see Section 3.3, Section 3.4 and Section 3.6). The perusal of documented workflows and frameworks enabled the gathering of information and insight leading to the creation of a data rescue workflow model.

Purpose of documented workflow publications: The various data rescue workflows and frameworks selected, and elaborated on in Section 3.3, differed in complexity, scope, data formats rescued, and research disciplines involved. Each of these added towards an overall understanding of the data rescue discipline. This understanding in turn enabled the drafting of an initial workflow indicating the main steps and tasks forming part of a data rescue venture (see Section 3.6 and Section 3.7).

Terminology: This section makes use of the terms ‘framework’ and ‘workflow’, but documented literature has also referred to the following terms when describing data rescue steps: ‘guidelines’, ‘steps’, ‘practical strategies’, ‘roadmap’, ‘cookbook’ and ‘model’ (see Section 3.2: Introduction to data rescue workflows, models and processes analysed). All relevant data rescue steps found in documented literature, irrespective of terminology, have been consulted when reviewing and analysing the data rescue concept referred to in this research question.

'Data rescue' vs 'data refuge': It is crucial to differentiate between the activity conventionally referred to as 'data rescue' by international data rescue entities and used in scholarly articles, and the term 'data refuge' used when describing the provision of a haven for US federally funded data. Elaborative details are provided in Section 2.2: 'Data rescue' and related terminology.

The focus of this study is the conventional data rescue workflow, and not the workflow forming part of 'data refuge'.

Complexity: Documented literature describing data rescue workflows range between simplistic and detailed. Simple workflows often mentioned the broad and main stages of data rescue only, while detailed and intricate data rescue models often required more than only basic knowledge of certain research disciplines. Examples of this varying feature can be found in Section 3.2: Introduction to data rescue workflows, models and processes analysed.

Discipline-specific versus generic: Documented data rescue workflows tend to fall into either a generic category or a discipline-specific category. Examples of these differing characteristics, and the different discipline-relevant data rescue frameworks, consulted when creating this study's initial Data Rescue Workflow Model, are provided in Section 3.3 and Section 3.4.

The assumption that discipline-specific workflows are bound to be complex in nature have proven to be unfounded. Many of these workflows are valuable for use in a discipline-agnostic environment as well, with the data rescuer able to discard or adapt the activities found to be inapplicable to the discipline of data at hand.

Publication type: As is the case with publications on data rescue, documented data rescue workflows and frameworks were found in a wide range of published outputs. This researcher is of the opinion that the choice of publication type is frequently linked to the objective of the publication. More details are provided in Section 2.6.4.2, under the heading 'Data rescue publications'.

Commonalities of conventional data rescue workflows: Whilst published sources examined most often featured data belonging to the environmental sciences, data rescue also involved disciplines outside of the environmental sciences sphere. Scrutinising published data rescue projects linked to different disciplines has provided insight into the commonalities found between data rescue projects, and an understanding of typical steps forming part of most rescue projects. Activities commonly found in data rescue projects have been discussed in Section 3.6: Content analysis and the creation of a Data Rescue Workflow Model.

Recommended data rescue models to use as guidance: Essential, recommended documented workflows and manuals have been identified and were discussed in Chapter 3, Section 3.3. The chapter entailed a review of published literature containing data rescue workflows, models and guidance.

Unique workflow features and problematic areas: An investigation of documented data rescue workflows also brought to light unique concepts and activities mentioned in documented sources. These concepts and data rescue tasks were identified as a value-adding feature and were incorporated into the study's workflow model where feasible and relevant.

Examples of these prominent and valuable data rescue additions include managing the data rescue venture as a project, thereby indicating the mandatory inclusion of a project plan and its accompanying features. Another uncommon activity come across, yet regarded to be a crucial part of any data rescue project, is the creation of a data management plan (DMP). Details are provided in Section 3.4: Summary of workflows.

Role-players: The participation of researchers, data curators, LIS professionals, archivists, web specialists, software engineers, database specialists, equipment and data format experts and technicians (unique to the specific study), expert volunteers (unique to the study discipline or equipment), and citizen scientists have been mentioned in the collection of data rescue documentation consulted (see Section 2.6.4). The crucial role of collaboration has been shown in various data rescue projects; more detail is provided in Section 2.6.4.2 under the heading 'Data rescue collaboration'.

While the participation of different role-players during data rescue have been stated in publications, publications seldom make mention of the specific data rescue roles and responsibilities of experts, professions or positions. In other words, data rescue models tend to not limit the execution of specific data rescue activities to definite sectors, such as the research sector, the library and information services sector or the ICT sector.

Summary: This section addressed the research question concerned with current data rescue workflows/models. Data rescue workflows/models were discussed in detail in Chapter 3, Section 3.3 and Section 3.4. This section of the recommendations chapter provides an overview of current data rescue workflows/models found in published literature. Aspects such as terminology, complexity, role-players, and the issue of discipline-specific versus generic workflows formed part of the overview.

6.2.1.2 The current SA workflows in data rescue vs international best practices

Little documented evidence was found of data rescue workflows implemented in the South African context. The low prevalence of published South African workflows is not unexpected, given the lack of pervasiveness of documented data rescue projects in SA. This summarised overview refers to the workflows of three South African data rescue projects.

The brief details available regarding the climatological data rescue projects (detailed in Section 2.6.2.2: Data rescue in South Africa) portray similarities with several of the international paper data rescue workflows described in this study (see Section 3.3.7: Data rescue workflow: Recovery of 'dark' data (zooplankton data), Section 3.3.9: Data rescue workflow: IEDRO (climate and environmental data), and Section 3.3.12: Data rescue workflow: World Meteorological Organization (climate data) as comparison). As only a basic outline of this local rescue project could be sourced, an in-depth comparison with other workflows is not possible.

The workflow steps forming part of the South African sociological data rescue project (see Section 3.3.11: Data Rescue Workflow: DataFirst) resemble the steps of many of the recommended paper rescue models described in Section 3.3: Data rescue workflows, models and processes. Detailed workflow steps concerning this project were stated to form part of an article not yet published, making current in-depth comparisons with other workflows not a viable option.

As with the two other South African data rescue ventures, little detail is available concerning the rescue of audio and video data at risk at the selected research institute. The available information regarding the broad rescue steps (see Section 5.4.9.1, under the heading 'Outsourcing the rescue of data at risk to an external party') shows the rescue project to adhere to the generic data rescue steps of locating data at risk, digitisation to a modern format, and uploading the new format data to a disciplinary repository. With older format audio and video tapes not featuring in the workflows analysed in Section 3.3, additional comparisons are not achievable.

Summary: While the published number of South African data rescue workflows is small, and finer details of the data rescue projects not yet available, their broad descriptions of the workflow steps agree with best practices/guidelines internationally. In-depth information about the activities forming part of each stage is required before reaching a more definitive conclusion regarding similarities and differences between local data rescue workflows and best practices workflows steps.

6.2.1.3 The current documented state of library and information services involvement with data rescue – global community

As discussed in Section 2.6.4.2, containing details regarding the participation and involvement of the LIS sector during data rescue activities, evidence of library and information services involvement was found in documented outputs. In the context of this study, the concept of library and information services, and the LIS sector, involves libraries (academic, research, special, and in certain instances also public libraries), physical archives, librarians, information custodians, information specialists, data managers, data curators, archivists, indexers, repository managers, and technicians or assistants assisting the LIS professionals. The concepts of library and information services as well as 'LIS' investigated and discussed therefore include the physical buildings, the infrastructure and services provided, and the involvement and contributions of LIS sector professionals as well as LIS sector semi-professionals. Library and information services involvement and participation in data rescue activities, as found in published sources, are summarised below.

Identification of four levels of LIS involvement in data rescue: A presentation identifying the four possible levels of LIS sector involvement in the rescue of modern digital data was published by an LIS sector professional. The four levels of LIS involvement can be summarised as raising data rescue awareness (including organising workshops or events, and administering surveys), web archiving of modern digital data, rescuing data required by a community, and harvesting 'uncrawlable' document(s) through a data rescue event. Details are provided in Section 2.6.4.2, under the heading 'Levels of data rescue involvement'.

The presentation describing the four levels is viewed as a valuable point of departure. Additional and/or clarifying tasks illustrating library and information services activities with regard to data rescue are briefly mentioned under the headings in the remainder of this section.

General involvement (brief non-specific mentions): Cursory reporting on LIS sector involvement in data rescue has been reported over the years, and in different publication types by different entities. International data rescue organisations, university library websites, international meteorological organisations, popular news companies, and researchers are all examples of sectors that have briefly mentioned the participation of the LIS sector in data rescue projects. More information is provided in Section 2.6.4.2: LIS sector (including archivists).

Data storage: A review of relevant literature has shown that libraries, including academic libraries, research libraries and national libraries have traditionally been used as secure storage locations for paper-based media. Other legacy formats, such as punch cards or magnetic tapes, were also found to be housed in libraries and archives forming part of libraries. Historically, librarians and information scientists have also been responsible for the management and preservation of artefacts.

Documented sources have also referred to the fact that in more recent times, early digital formats (e.g., floppy discs and stiffy discs) were commonly sent to researchers' affiliated libraries after project completion. National libraries were mentioned to contain valuable datasets.

More details about the involvement of the library and information services sector in data storage can be found in Section 2.6.4.2, under the headings 'Data storage, curation, and long-term preservation', 'Locating data' and 'Data inventories'.

Data curation and preservation: Information custodians have traditionally engaged in the curation of research data via traditional library activities as well as special collection practices. More recently, these practices are increasingly being seen as an important future area of engagement, with many research librarians and data managers becoming involved in the curation of data, and this involvement forming part of documented literature. This topic is discussed in Section 2.6.4.2, under the heading 'Data storage, curation, and long-term preservation'.

Catalogues and inventories: Documented outputs often refer to libraries and library groups when recommending sources from which to download inventories of valuable historical data assets (see Section 2.6.4.2, under headings titled 'Data inventories' and 'Participation in data rescue projects', respectively).

Inventory development: At least one published output described the involvement of a library and information services group in the creation of an inventory required prior to rescuing data. Details of an instance of a country-specific library and information services group developing an inventory of survey files in need of rescue were traced in the group's blog. Details regarding the involvement of the library and information services sector during inventory development, or environmental scanning, are provided in Section 2.6.4.2, under the heading 'Environmental scan of data rescue initiatives'.

Data sharing: Investigated literature reported on the rescue of data documentation, and how this activity enabled the eventual sharing of rescued data on an academic data portal. Details are provided in Section 2.6.4.2, under the heading 'Data sharing'.

Skills, knowledge and experience: Several sources refer to the skills set, knowledge and experience of LIS professionals, and how these attributes can add value to data rescue efforts. Examples of the beneficial contributions of the library and information services include the ability to locate data, ability to track experts who would clarify data, data management skills of library and information services data curators, repository knowledge, and an understanding of document indexing, metadata creation and metadata standards. More details are provided in Section 2.6.4.2, under the headings 'Locating data', 'Data assessment' and 'Data sharing'.

Consortia, working groups, interest groups, networks: As documented in the study's literature review chapter (see Section 2.6.4.2, Data rescue groups), there is evidence of LIS professionals forming groups or participating in groups for the purpose of enhancing data rescue collaboration, knowledge and services. Section 2.6.4.2 also provides details of a consortium formed by research libraries to move the 'Data Refuge' concept to a more sustainable footing.

Conferences and other meetings: Examples of libraries and library and information services groups being involved in data rescue conferences, organising conferences centred on data rescue, and participation in virtual meetings to discuss topics related to data rescue were also traced. Details are provided in Section 2.6.4.2, under the heading 'Host data rescue event'.

Planned and hosted data rescue/refuge events: Documented sources have revealed that libraries have been instrumental in the organisation of data rescue events, with more details provided in Section 2.6.4.2, under the heading 'Host data rescue event'.

Extended involvement in data rescue projects: Evidence of library and information services involvement in data rescue projects was found, with the rescue of historical administrative data by library and information services professionals, the rescue of health data via recovery of required missing data documentation, and a data rescue project managed by the US National Agricultural Library being examples of LIS-sector involvement (see Section 2.6.4.2).

Data rescue publications: Several published instances were found of library and information services groups, involved in data rescue, who had drafted and published data rescue-related publications. These examples are listed and detailed in Section 2.6.4.2, under the heading 'Data rescue publications'.

Participation in a data rescue survey: At least one study involved information professionals as participants in a survey related to data at risk. The intention of the survey was to gather information about the valuable library and information services perspective on the data at risk predicament and preservation challenges. Additional details are supplied in Section 2.6.4.2, under the heading 'Data at risk survey participation'.

Unique LIS perspective on data rescue: Obtaining an understanding of the unique LIS perspective of data at risk and data rescue was regarded as crucial for data preservation planning. Library and information services professionals' multidisciplinary knowledge, their curation skills, and experience as information professionals have been cited as contributing to the profession's unique and valuable perspective on data rescue. Details regarding this topic are provided in Section 2.6.4.2, under the heading 'Unique data at risk perspective'.

LIS training and education: Documented instances of references to data rescue training implemented and recommended for LIS students have been found (see Section 2.6.4.2, LIS training). A prime example pertains to the findings obtained after involving library and information services staff in a survey on data at risk encountered, and data rescue practices executed. The survey findings led to the authors concluding that the results had implications for the training of information custodians. In South Africa, an instance of graduate students being exposed to the practice of data rescue is already in existence; details are provided in Section 2.6.4.2 under the heading 'LIS training'. The topic is further broached as a recommendation in Section 6.4.12: Amend and adapt the LIS curriculum to include data rescue.

Data assessment/Reuse potential: Evidence was found of literature sources mentioning that experts within the library and information services sector are regarded as vital contributors during data assessment and when the reuse potential of data needs to be determined. Details regarding library and information services involvement during data assessment are found in Section 2.6.4.2, under the heading 'Data assessment'.

Summary: Documented sources have mentioned or described the involvement of libraries and the library and information services sector workforce, including librarians, information custodians, data curators and archivists, in a diverse number of data rescue activities and concepts. Documented involvement ranged from general and brief mentions to descriptions of participation in a single rescue activity, to involvement throughout an entire rescue project.

Published documentation reporting on the library and information services sector and data rescue has also reported on data rescue aspects indirectly related to data rescue, such as the data at risk perspectives of data curators, data rescue publications emanating from the library and information services sector, and the data rescue training of LIS students.

The inference can be made that the library and information services workforce has the potential to play a role in most activities forming part of a data rescue project. In addition, library and information services professionals as well as semi-professional staff can be involved in data rescue projects.

6.2.1.4 The current documented state of data rescue awareness – South African library and information services community

A review of documented sources found neither any dedicated nor regular library and information services involvement/participation in data rescue activities and projects in SA. This researcher wishes to emphasise that not being able to find documented involvement does not equate with no library and information services awareness within the South African library and information services community.

While completing this study, this researcher also presented an hour-long lecture on data at risk and data rescue to a university postgraduate LIS class (see Section 2.6.4.2, under 'LIS training'). The issue is discussed as a recommendation in Section 6.4.12.

6.2.1.5 The current documented state of data rescue involvement – South African library and information services community

The current state of library and information services involvement in South African data rescue projects, as detected via documented sources, is minimal. Apart from the involvement of three different university libraries during a documented data rescue involving apartheid era datasets, no published evidence of library and information services involvement during data rescue could be ascertained. More details are found in Section 3.3.11: Data Rescue Workflow: DataFirst (sociological data). The data rescue project was also publicised in LIS-related media, including articles in an online university library news magazine and a popular international online library magazine (see Section 2.6.2.2: Data rescue in South Africa).

6.2.1.6 The current documented state of data rescue globally and in South Africa

It is evident from the study's literature review reporting on data rescue, and the empirical findings of the study involving participants from the selected institute, that the practice referred to as 'data rescue' displays a range of features and elements. To answer the research question above, several data rescue subheadings are listed below, with brief study findings and referenced sections forming part of this section. It should be noted that the concept of 'data at risk', being a precursor to 'data rescue', is also included in several of the subheadings.

It is important to mention that an absolute and archetypal data rescue project does not exist, and that it is unwise to assign certain fixed categorical attributes to data rescue projects. Data rescue characteristics listed and expanded on below provide a summary of common results emanating from the study's literature review, and the findings obtained via the empirical data collection methods.

Data at risk (overview): The prevalence of data at risk was found in all corners of the earth, and involved a range of disciplines, data formats and data collection periods (see Section 2.6.1, Section 2.6.2 and Section 2.6.3). Due to the role played by environmental data in the creation of reliable and valid climatic models and in providing glimpses into environmental changes over time, most documented instances of data rescue involve climate data. In addition to the prevalence of climatic data rescue, several subject areas forming part of the sphere of environmental sciences commonly feature in published sources. Other disciplines involved in data rescue include the study areas of astronomy data, the health sciences data, and sociological studies (see Section 2.6.1).

Regarding the South African situation, documented sources refer to the existence of environmental data from the 19th century, with Section 2.6.2.2: Data rescue in South Africa providing details. More recent South African data at risk emanating from research in the sociological sciences and humanities have also been documented; details are found in Section 3.3.11: Data Rescue Workflow: DataFirst (sociological data), and Section 5.4.6.6, under the heading ‘Data value and data sharing: RGL6’.

At the selected multidisciplinary research institute, most respondents, performing research in a varied range of subject areas, declared having data at risk in their research groups, as reported in Section 5.3.4: Prevalence of data at risk.

Data at risk formats: A review of published rescue outputs revealed a wide range of data formats involved in rescue projects. Paper-based media rescue has involved ship logs, diaries, maps, photographs, lighthouse data, observatory records and historical newspapers. Examples of non-paper legacy formats rescued include various magnetic tape formats, microfiche, microfilm, early digital media, photographic plates and punch cards. Physical specimens have also been the focus of data rescue projects. The US ‘Data Refuge’ movement during 2016–2019 involved the rescue of federally funded data in modern digital formats. Section 2.2.3, titled ‘Data Refuge’ reports on these aspects in more detail.

In SA, paper-based historic data were mentioned in all documented data rescue efforts (see Section 2.6.2.2 and Section 3.3.11). Data collected from the selected research institute showed that data at risk included paper data, samples, magnetic tape data, early digital data, modern digital data, audio data, video data and modern digital data (see Sections 5.3.5: Formats of data at risk). The single instance of data rescue at the selected institute involved audio data in an early digital format (see Section 5.4.9.1, under the heading ‘Outsourcing the rescue of data at risk to an external party’).

Data at risk period: Documented data rescue sources investigated demonstrated the date ranges of data collection of data involved in rescue projects. As such, historic data, early digital data, and even modern digital data featured in documented data rescue efforts. Several data rescue publications referred to the rescue of meteorological data from the 17th and 18th centuries, while the rescue of modern and digital data was proof that the other end of the date spectrum is also included in data rescue activities (see Section 2.6.3: Data sources and formats).

In SA, the documented data rescue projects have shown that data involved in rescue projects originated from as early as meteorological data collected during 1829, right up to humanities data generated in the late 1990s (see Section 2.6.2.2, and Section 5.4.6.6, under the heading 'Data value and data sharing: RGL6').

Data at risk at the selected research institute included older data such as data from the 1970s on floppy disks, paper data collected during the 1980s, and older audio data collected during the 1990s. Conversely, results of the online questionnaire distributed at the selected institute revealed that modern data and recent data are frequently at risk. More details are provided in Section 5.3.5, Section 5.3.6 and Section 5.4.5.

Factors leading to data being at risk: Both the literature review of the study and the in-depth interviews held with respondents from the selected institute revealed a range of factors putting data at risk of loss or damage. Overall, risk factors as found globally (see Section 2.3) and those listed by researchers at the selected institute (see Section 5.4.7) showed great similarity. The unique nature of the selected institute and its available resources and policies resulted in study participants revealing several risk factors not mentioned in global literature. Risk factors mentioned by study participants provided a glimpse into the link between data at risk and the available institutional infrastructure, systems, policies and support services.

Scope of projects: Published data rescue projects ranged between huge undertakings and smaller data rescue ventures. Two good examples of the former include the Old Weather project involving millions of weather, ocean and sea-ice observations recorded by mariners and scientists over the past 150 years, and the DRAW project involving thousands of person-hours of work (see Section2.6.4.6: Citizen scientists). An example of a smaller rescue project is the rescue of a collection of fish audio data, as described in Section 2.5.11.

Data rescue entities: Many groups are involved with data rescue, and their involvements range from the provision of training, right through to project implementation and the donation of the necessary basic data rescue equipment. Details of such groups are provided in Section 2.6.5: Data rescue entities and interest groups, and Section 2.6.4.2, under the heading 'Data rescue groups'.

Evidence of the involvement and support provided by a global data rescue organisation for a South African rescue project was found. Section 2.6.2.2, titled 'Data rescue in South Africa' provides more details.

Interest Groups: While smaller national and even regional data rescue interest groups exist, evidence of extensive data rescue involvement by global interest groups featuring members from across the globe was found (see Section 2.6.5, Data rescue entities and interest groups).

In SA there are as yet no groups solely focusing on the rescue of data. Collaborations that have taken place during the few documented projects (see Section 2.6.2.2: Data rescue in South Africa) are temporary in nature.

Involved parties: As mentioned earlier, it is not possible to define a 'typical' data rescue project, mention the 'typical' stakeholders, or describe the 'typical' roles and responsibilities of various data rescue participants. Documented literature has mentioned the following data rescue participants: researchers, data curators, archivists, historians, web designers, coders, volunteers and citizen scientists. In addition, global rescue organisations and entities are frequently involved in environmental data rescue projects. More details about involved parties are provided in Section 2.6.4: Data rescue participants and contributors.

In SA there is currently limited published documentation pertaining to data rescue participants. It is expected that more elaborate details regarding the sociological data rescue project's involved parties are to be published in a planned scholarly article authored by the project manager (see Section 2.6.7: Other data rescue outputs). With regard to the series of South African historic meteorological data rescue projects, it was found that participating parties included meteorological researchers, meteorological students, and collaborations with a meteorological office based abroad (see Section 2.6.2.2: Data rescue in South Africa).

At the selected institute, with data rescue projects being minimal (only one instance reported), the involved parties were described as institutional researchers, external digitisers, other experts in the same humanities-related subject area, and external discipline repository staff (see Section 5.4.9.1, under the heading 'Outsourcing the rescue of data at risk to an external party').

Manuals and guidance: This study's literature review revealed a range of freely available published data rescue manuals and guidelines (see Section 2.6.6, Data rescue manuals and guidance). Many of these manuals were discipline-specific, with the most prominent and often-recommended publications focused on climate data rescue, hydrological data rescue and marine species data rescue.

Another highly regarded data rescue guidance publication and described by the WMO as an example of 'best practices' is a poster detailing a data rescue project in the classroom setup. The data rescue steps and processes included in several published guidelines and manuals are discussed in more detail in Chapter 3: Literature and the creation of a data rescue workflow model.

No evidence of manuals for the rescue of South African data at risk was found. It is presumed that the meteorological data rescue activities taking place in SA would have received guidance from the global data rescue initiative supporting the local rescue activities (see Section 2.6.2.2: Data rescue in South Africa).

No data rescue manuals or guidance are currently in place at the selected institute. The single instance of a complete data rescue project had enlisted the assistance and guidance of an external data digitisation entity and other discipline-specific experts (see Section 5.4.9.1, under the heading 'Outsourcing the rescue of data at risk to an external party').

Other published data rescue outputs: Documented data rescue outputs have revealed that data rescue reporting, publishing and sharing are done using a wide range of publications. The reporting or published relaying of data rescue ventures has been found in scholarly journals, popular magazines, conference papers, blog postings, webpages of data rescue entities, SlideShare items, book chapters, reports, library webpages, university webpages and annual reports. YouTube videos detailing the efforts of global data rescue entities, and introductory videos to data rescue and data archaeology, have also been traced (see Section 2.6.7 for details).

Documented outputs of South African environmental data rescue projects were traced in annual reports or progress reports of supporting international data rescue initiatives, supplemented by mentions on the initiative's website (see Section 2.6.2.2: Data rescue in South Africa). Details of the South African sociological data rescue project could be found in online university news articles, university library webpages, and on the webpages of the data centre managing the project. A future scholarly article pertaining to the latter project is being finalised, while a presentation on the data rescue venture was also forwarded upon request (see Section 2.6.7: Other data rescue output).

Workflows: As discussed in detail in Section 3.3: Data rescue workflows, models and processes, several documented outputs contain data rescue workflows or frameworks. Data rescue workflows are also addressed in another sub-question; details can be viewed in Section 6.2.1.1 and Section 6.2.1.2.

Meetings and related ventures: Section 2.6.8 provides details about documented data rescue meetings and related ventures. Several data rescue workshops have been held during the last decade, and organised by, or taking place under the auspices of a range of involved data rescue entities. Workshops were held in different countries and on different continents. Evidence of at least one data rescue workshop per year was found.

No documented evidence of workshops, conferences or similar events related to data rescue in SA could be traced. It is, however, assumed that unpublished meetings forming part of data rescue projects have taken place.

No evidence of data rescue-related meetings at the selected research institute could be ascertained via the study's empirical data collection methods.

Terminology: The terminology around data rescue is discussed in Section 2.2. Most published outputs still use the term 'data rescue' when referring to the activity of converting data at risk to a modern electronic format and uploading the converted data with its data documentation to a repository for wider use. A smaller number of publications have used the term 'data curation' when describing similar activities. It is worth mentioning that a major international data rescue interest group recently made the decision to change the name of the group to refer to 'data conservation' instead of 'data rescue'. Details are provided in Section 3.3.13: Data rescue workflow: RDA (discipline-agnostic).

In SA, the term 'data rescue' is the only documented phrase found.

Data rescue and politics: Following soon after the inauguration of the 45th president of the USA, concerned scientists and citizens took it upon themselves to organise data rescue events, directed at rescuing federally funded environmental research data. These scientists and citizens, apprehensive about the new administration's support of environmental sciences research, and the future accessibility of environmental research data, participated in data rescue events entailing the scraping and archiving of environmental data to secure archives. The movement was also referred to as 'Data Refuge', and involved thousands of volunteering scientists, web experts, librarians, archivists, academics, coders and other volunteering parties. Details are provided in Section 2.2.3: Data Refuge.

Data rescue and citizen science: Many prominent data rescue projects make use of volunteers as citizen scientists, and examples of such projects are discussed in more detail in Section 2.6.4.6. These projects typically require human input, as the data transcription tasks are impossible for computers due to diverse and idiosyncratic handwriting (see Section 2.6.4.6: Citizen scientists).

No evidence of involvement of citizen scientists in South African data rescue projects could be established. The same finding held true for data rescue at the selected research institute.

Data rescue challenges: With published data rescue projects being successful in nature, the prevalence or nature of obstacles and challenges did not readily feature in data rescue publications. It was, however, a topic forming part of the empirical phase of the study, with results elaborating on the numerous obstacles and challenges mentioned by study participants. These challenges had their roots in one or more of four broad categories, namely, institutional factors, equipment problems, lack of resources, and subpar data management practices (see Section 5.4.10).

The mitigation of such data rescue obstacles is also discussed later in this chapter and takes the shape of a recommendation emanating from this study (see Section 6.4.5: Address and mitigate data rescue challenges).

Data rescue overview: As is the case with data at risk, an investigation of relevant literature revealed that data rescue shows a global presence, with data rescue projects completed in a vast number of countries and across several continents. Documented data rescue projects most commonly feature the rescue of historic paper-based environmental data, but several other subject areas were also mentioned in data rescue outputs. Examples of such research areas include data emanating from lunar expeditions, stratospheric measurements, physics, and religious studies. Documented data rescue projects in SA have involved the environmental sciences, sociological sciences, and the humanities.

With a range of data rescue publications scrutinised, involving a review of the workflows of the projects, their complexity, parties included, disciplines involved and scale of rescue, it is fair to say that the concept of a 'typical' rescue project does not exist. At most, there is evidence of commonalities (refer to Section 3.6, Section 3.7.1 and Section 3.7.2). It is important to be aware of the difference in nuance between 'conventional' data rescue and the concept of 'data refuge', which is also commonly referred to in the US media as 'data rescue'. Both concepts are focused on ensuring that data at risk are available for future use, with conventional data rescue involving all formats, disciplines and origins of data, and data refuge concerned with US federally funded environmental data, in modern electronic format, at risk of being removed by the US government.

The differentiation described above does not mean that all conventional 'data rescue' projects deal with historic data only. Data rescue projects can also target data in a modern electronic format; the recent name change of the RDA Data Rescue Interest Group to the RDA Data Conservation Interest Group, to convey the group's interest in ensuring modern data are accessible and not at risk, is testament to this school of thought (see Section 2.2.2).

In the South African data rescue environment, published accounts of data rescue projects are rare, with this researcher able to trace only three documented South African rescue entities. Of these three, two of the identified rescue entities had been involved in more than one rescue project, with one of the entities participating in sociological data rescue, and the other participating in meteorological data rescue. The third entity, identified while interviewing RGLs at the selected institute, had been involved in a single rescue project only. Adding to this is the obvious conclusion that data rescue at the selected institute is an uncommon activity.

6.2.1.7 Formalising the theory and practice in a workflow model for a data rescue

The research question posted above is addressed in the following manner:

- This section commences with a brief introduction to the recommended Data Rescue Workflow Model, and the factors leading to the creation of the current and recommended version of the model. Factors touched on include the suggestions and recommendations put forward via study participant feedback (discussed in Section 5.7.3), academic review (discussed in Section 5.7.4) and vital data rescue activities identified during the content analysis of various data rescue models (Section 3.6: Content analysis and the creation of a Data Rescue Workflow Model).
- A table showing the main differences between the three models used during this research follows the concise summary; the table contains details of the initial model, the revised model and the final recommended model.
- A series of three different diagrammatical portrayals of the recommended model follow the comparison table, and each portrayal contains a relevant description and explanation. The three different diagrammatical portrayals form the study's Data Rescue Workflow Model, and represent three different layers of the same model.
- The series of diagrams comprise the following layers:
    - a single-page diagram summarising the recommended model, and its brief accompanying description,
    - a four-page compact summary of the recommended model, and its accompanying description, and
    - detailed description and portrayal of the individual stages of the recommended model.

The next section contains an introduction to the recommended model.

Key features of the recommended Data Rescue Workflow Model: The recommended model is the third and final model created during this study. It follows on from the initial model (described in Chapter 3: Section 3.7, Section 3.8 and Section 3.9) and the revised model (described in Chapter 5: Section 5.6.3, Section 5.6.5 and Section 5.6.6). The model here is an amended and improved version of the previous models, as a series of feedback and academic review activities had necessitated changes to the model. The changes made, and differences between the initial model and amended model are included in Table 5.10, while the comparison between the study's three models (initial, amended and recommended) is shown in Table 6.1.

The recommended model consists of three layers, displayed graphically, namely:

- a data rescue summary diagram,
- a data rescue compact workflow, and
- detailed stage-specific workflows.

While the graphics differ in appearance, shapes used and data rescue details provided, they are all part of the same recommended data rescue process. The main difference between the different graphics pertains to the level of data rescue details provided, with the summary diagram portraying the main rescue stages and tasks, the compact workflow providing more details, and the stage-specific workflows portraying detailed information on the tasks performed, outputs delivered, decisions made, and available guidance of each data rescue stage.

The recommended data rescue process, as portrayed via the study's Data Rescue Workflow Model is as follows:

- The workflow model is a generic data rescue model and is not discipline-specific.
- Diagrams and flowcharts illustrate the recommended workflow.
- The recommended workflow consists of eight main stages, namely:
    1. preparing for data rescue,
    2. planning for data rescue,
    3. storing and preserving the data,
    4. digitising or converting the data,
    5. describing the data,
    6. sharing the data,
    7. preserving the data (long term), and
    8. project closure.

- The model provides for two distinct types of data rescue, namely:
  1. full data rescue, and
  2. partial data rescue.
- The terms ‘full data rescue’ and ‘partial data rescue’ were constructed by this researcher to differentiate between data involved in all proposed data rescue stages, and data involved in certain stages only.
- Full data rescue refers to a rescue process including all eight rescue steps included in the study’s recommended model.
- Partial data rescue refers to a rescue process only including some of the rescue steps included in the study’s recommended model. For the most part, partial data rescue will refer to a rescue process where data are not digitised to a modern electronic format, not uploaded to a data repository, and not preserved for the long term. All other rescue steps, including the secure storage of data at risk, and upload of relevant metadata and data documentation to a data repository, form part of partial data rescue.
- The model provides data rescue steps for four different data formats comprising:
  1. paper-based data,
  2. data in an early digital format,
  3. modern digital data, and
  4. samples or physical specimens.
- While the model does not limit the execution of its included tasks to certain professions or institutional positions, the section dealing with stage-specific descriptions makes recommendations regarding ideal incumbents for certain rescue tasks.

The next section contains a tabular comparison between the initial model (see Section 3.7, Section 3.8 and Section 3.9), the revised model (see Section 5.6) and the recommended model.

Comparison between the three different models featuring in this study: Table 6.1 provides a comparison between the three models used during this research, namely:

- the initial Data Rescue Workflow Model,
- the revised Data Rescue Workflow Model, and
- the recommended Data Rescue Workflow Model.

Table 6.1: Comparison between the three data rescue workflow models used in this study

MODEL FEATURES	INITIAL MODEL	REVISED MODEL	RECOMMENDED MODEL
Composition	Model consists of a summarised image, and an image for each of the data rescue stages	Model consists of a compact summary, an extended summary, and at least one image for each of the data rescue stages	Model consists of a compact summary, an extended summary, and at least one image for each of the data rescue stages
Number of stages	Model consists of nine stages	Model consists of eight stages	Model consists of eight stages
Names of stages	Stages are: • Project Initiation • Storage and Preservation • Create Inventories • Imaging of Media • Digitisation of Media • Describing the Data • Making Data Discoverable • Archive the Data • Project Closure	Stages are: • Data Rescue Preparatory Stage • Data Rescue Planning • Data Storage and Preservation • Digitisation • Documenting the Data • Data Sharing • Long-Term Preservation • Project Closure	Stages are: • Data Rescue Preparatory Stage • Planning for Data Rescue • Storing and Preserving the Data • Digitising/Converting the Data • Describing the Data • Sharing the Data • Preserving the Data (for long term) • Project Closure
Names of stages	Stage names provided but not displaying consistency	Stage names provided but not displaying consistency	Stage names to consistently include a verb
Names of stages	Stage names reported to be ambiguous	Stage names reported to be ambiguous	Stage names are self-explanatory
Data formats	Model assumes data are paper based Model describes rescue of paper data	Model includes different data formats Model describes rescue of paper data, early digital data, modern digital data, and samples	Model includes different data formats Model describes rescue of paper data, early digital data, modern digital data, and samples
Type of rescue	Model describes full data rescue only	Model describes full and partial data rescue	Model describes full and partial data rescue
Roles	Model does not indicate roles of institutional parties	Model indicates roles of institutional parties	Model indicates roles of institutional parties
Options per stage	Only one option per stage is described	Different options per stage are described	Different options per stage are described
Inventories	Creation of inventories is seen as separate stage	Creation of inventories forms part of other stages	Creation of inventories forms part of other stages

MODEL FEATURES	INITIAL MODEL	REVISED MODEL	RECOMMENDED MODEL
Data destruction	Data destruction is not addressed	Data destruction forms part of the model	Data destruction forms part of the model
Data assessment	Data assessment is understated	Data assessment features strongly	Data assessment features strongly
Data imaging and data digitising	Data imaging, and digitising (keying) of paper data values are separate stages	Data imaging and data digitising are described in one stage	Data imaging and data digitising are described in one stage
Use of colours	Colours are used to distinguish between activities and outputs	Colours are used to indicate responsibly parties	Colours are used to indicate responsibly parties
Shapes used	Model requires minimal shape explanation	Model requires explanation of range of shapes and colours	Model requires explanation of range of shapes and colours
Outsourcing	Model does not include outsourcing option	An outsourcing option forms part of the model	An outsourcing option forms part of the model
Guidelines	Provided	Provided	Provided
Directional flow	Workflow model shows instances of reading from right to left AND left to right	Workflow model shows instances of reading from right to left AND left to right	Diagram and workflow model read left to right consistently
One-page summary	Resembles a summarised workflow Simplistic Lacks vital information present in subsequent models' summaries	Resembles a summarised workflow Described as a 'compact summary' Shapes are similar to shapes of four-page summary or detailed stages	Is a diagrammatical representation, NOT a workflow Described as a 'diagram' Shapes do not mimic the shapes used during the four-page summary or detailed stages
Four-page summary	Does not feature in initial model	Created using PowerPoint Cluttered appearance; too much detail for a compact summary Has instances of several arrows joining one task/block	Created using Visio Simplified without sacrificing crucial details No instances of more than one arrow joining a single process/task block
Detailed stages (explanatory details)	Stage-specific images do not contain a summary	Stage-specific images do not contain a summary	All stage-specific images consistently contain a summary of stage activities
Detailed stages (roles and responsibilities)	Involvement of other parties not indicated	Images show potential involvement of different sectors, comprising mainly SET base and library and information services sector	Library and information services sector roles indicated in Section 6.2.2, with potential involvement of other sectors also stated Sector involvement not indicated in current section (Section 6.2.1.7)

MODEL FEATURES	INITIAL MODEL	REVISED MODEL	RECOMMENDED MODEL
Detailed stages (shapes, symbols, colours)	Sufficient differentiation between shapes Contents of detailed stages are more cause for concern than their appearance	Separate shape is used to indicate guidance Some stage images contain too much detail and are overwhelming to novices Outputs forming part of stage are placed between two processes/task activities	Asterisk was used to indicate that guidance is available for stage/task Stage images were simplified without sacrificing crucial details Outputs were placed above or below process shape/task activity, not between tasks
Stage 1	Creation of 'Data rescue progress document' not in model No guidance available on selecting team of experts No guidance available on establishing institutional rescue feasibility	Creation of 'Data rescue progress document' not in model No guidance available on selecting team of experts No guidance available on establishing institutional rescue feasibility	Creation of 'Data rescue progress document' forms part of stage Option to consult: 'Data Assessment' guidelines Option to consult: 'Selecting team of experts' guidelines Option to consult: 'Assessing institutional rescue feasibility'
Stage 2	'Progress document' not in model	'Progress document' not in model	Progress document to be updated
Stage 3	'Progress document' not in model	'Progress document' not in model	Progress document to be updated
Stage 4	'Progress document' not in model	'Progress document' not in model	Progress document to be updated
Stage 5	'Progress document' not in model	'Progress document' not in model	Progress document to be updated
Stage 6	'Progress document' not in model	'Progress document' not in model	Progress document to be updated
Stage 7	Certain activities (which were archival responsibilities) assigned to data rescuer 'Data to be in preservation format' not mentioned in stage Regular monitoring was not a mentioned activity 'Data rescue progress document' not in model	Certain activities (which were archival responsibilities) assigned to data rescuer 'Data to be in preservation format' not mentioned in stage Regular monitoring was not a mentioned activity 'Data rescue progress document' not in model	Certain preservation tasks (i.e., archive responsibilities) not included in model Stage includes: 'Ensure data are in preservation format' Regular monitoring included in stage Updating 'Data rescue progress document' is part of stage
Stage 8	Marketing/creating awareness and publishing are included as separate activities	Marketing/creating awareness and publishing are included as separate activities	Marketing/creating awareness and publishing are parallel activities

Data Rescue Summary Diagram: This single-page diagram, portrayed in Figure 6.1, is a synopsis of the main data rescue stages and activities of the recommended model. It provides an outline of the proposed data rescue process, with its main objective being the portrayal of data rescue to rescue novices within a single page. As such, the image includes the main stages and crucial activities, but excludes aspects such as:

- roles and responsibilities,
- references to guidance documents,
- references to all outputs created during the various stages, and
- relevant decisions.

These excluded features (roles, responsibilities, guidance documents, outputs and relevant decisions) are discussed in a subsequent section of the chapter.

When compared to the previous single-page summary portrayed in Figure 5.5, Section 5.6.3, the updated image displays several differences. The major differences, also listed briefly in Table 6.1, are described below.

- The updated image features the term ‘diagram’ in the title while the previous summarised image, as discussed in Chapter 5, was described as a ‘compact summary’ only. The name change was requested during the study’s mini focus group session, as participants believed that although a summary was useful and even vital, the image failed to meet the requirements of a ‘workflow’. It was consequently recommended that the image be renamed.
- The diagram makes use of unique shapes as far as possible and does not attempt to emulate the workflow shapes of the more detailed images. As a result, this summary diagram features the main stages identified by a square or rectangle, and the encompassing tasks captured in a shape encircling the square or rectangle.
- As indicated in Table 6.1, the updated data rescue summary diagram also features the updated stage names, with verbs included in stage names where feasible. In addition, name changes were anticipated to form a clearer picture of the relevant stage and are regarded as being less ambiguous that the previous stage names.

As the objective of the summary diagram is to provide data rescue novices, or inexperienced users of the model with a single page overview of the recommended data rescue workflow, it was considered vital that the diagram convey the following information regarding data rescue:

- It is important to include within the first stage an activity involving a decision as to whether data rescue should proceed, or whether data rescue is not a viable option.

- It is important that the data rescue project undergoes a process of project planning, and that planning also involves the appointment of a project team and the creation of a data management plan.
- It is important that the pre-rescued data be stored securely prior to it being rescued.
- It is important that data at risk be accompanied by metadata and data documentation.
- Data should ideally be converted to a common, open, modern digital format, if feasible.
- Data should ideally be shared, provided there are no accompanying privacy or confidentiality restrictions.
- It is crucial that converted data in common, open, modern digital format be preserved in the long term.
- With the rescue of data being viewed as a project, it is important that proper project closure steps be executed.

Based on the aspects listed above, the summary's main stages and crucial tasks, mirrored in the detailed data rescue compact workflow, and the comprehensive stage-specific workflows, are as follows:

- **Stage 1: Preparatory Stage**
  - This stage involves preparatory activities performed to evaluate the data and institutional resources and to determine whether data rescue should proceed. The stage also involves the destruction of data that will not be rescued.
  - Data that are potentially at risk and could potentially be rescued are identified.
  - A data inventory of the identified data is created.
  - A team of experts, tasked with determining the value of the data, is selected.
  - The team of experts assess the data and decide whether the data are worthy of rescue. Following that, institutional data rescue resources are evaluated.
  - Should data be deemed valuable, and rescue resources are available, the rescue process will continue, and the process proceeds to Stage 2.
  - This stage also includes a decision made on whether full data rescue or partial data rescue will be pursued.
  - Should data be deemed valuable, but rescue resources not be sufficient or available, the data are stored securely in its current format.
  - Data not worthy of rescue are destroyed, and the process is terminated.

  - Stage 1 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 2: Planning of data rescue**
  - This stage involves the planning activities performed after a decision has been made to proceed with the rescue of data.
  - A first planning step involves the appointment of a data rescue project team.
  - The stage requires the drafting of a data rescue project plan.
  - The stage requires the drafting of a data management plan applicable to the data to be rescued.
  - Stage 2 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 3: Storing and preserving data**
  - This stage involves the secure storage and preservation of the un-rescued data.
  - The pre-rescued data are stored and preserved as per the format-specific guidelines forming part of the detailed stage workflows.
  - Should data be undergoing full rescue, the process proceeds to Stage 4, entailing the digitisation or conversion of data.
  - Should data be undergoing partial rescue, the process proceeds to Stage 5, entailing the creation of metadata and data documentation.
  - Stage 3 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 4: Digitising or converting data**
  - This stage involves the digitisation of data to a common, open, modern electronic format, or the conversion of data in an older electronic format or proprietary modern format to a common, open, modern electronic format.
  - The first task to be completed during this stage entails the creation of a digitisation inventory, or a conversion inventory.
  - Following the creation of an inventory, the data are imaged, keyed or converted to a common, open, modern electronic format.
  - The scanned, keyed or converted data then undergo a process of validation and quality control.

  - The imaged, keyed or converted data are stored in pre-defined file and folder structures.
  - Stage 4 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 5: Describing the data**
  - This stage involves describing the data by creating metadata and data documentation linked to the data, thereby ensuring that the data can be understood and used by future users.
  - The creation of metadata and data documentation form part of full data rescue and partial data rescue, and are the main tasks executed during this stage.
  - The metadata and data documentation are stored in the same folder as the digitised/converted data.
  - Should data form part of partial rescue (Stage 4 was not involved, and data not converted), the data documentation and metadata are stored in an electronic folder and the details recorded in the project plan.
  - Both full and partial rescue data will be involved in the next stage (Stage 6), entailing the sharing of data.
  - Stage 5 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 6: Sharing the data**
  - This stage involves sharing the digitised or converted data and its metadata and data documentation. The stage also involves the sharing of metadata and data documentation of data that had not been digitised or converted. The sharing of data and its data documentation and metadata involves the upload of data, data documentation and metadata to data repositories.
  - For full rescue: digitised/converted data, metadata and data documentation are uploaded to identified data repositories.
  - For partial rescue: metadata and data documentation are uploaded to identified data repositories.
  - The full data rescue process proceeds to Stage 7, which involves the long-term preservation of digitised/converted data.

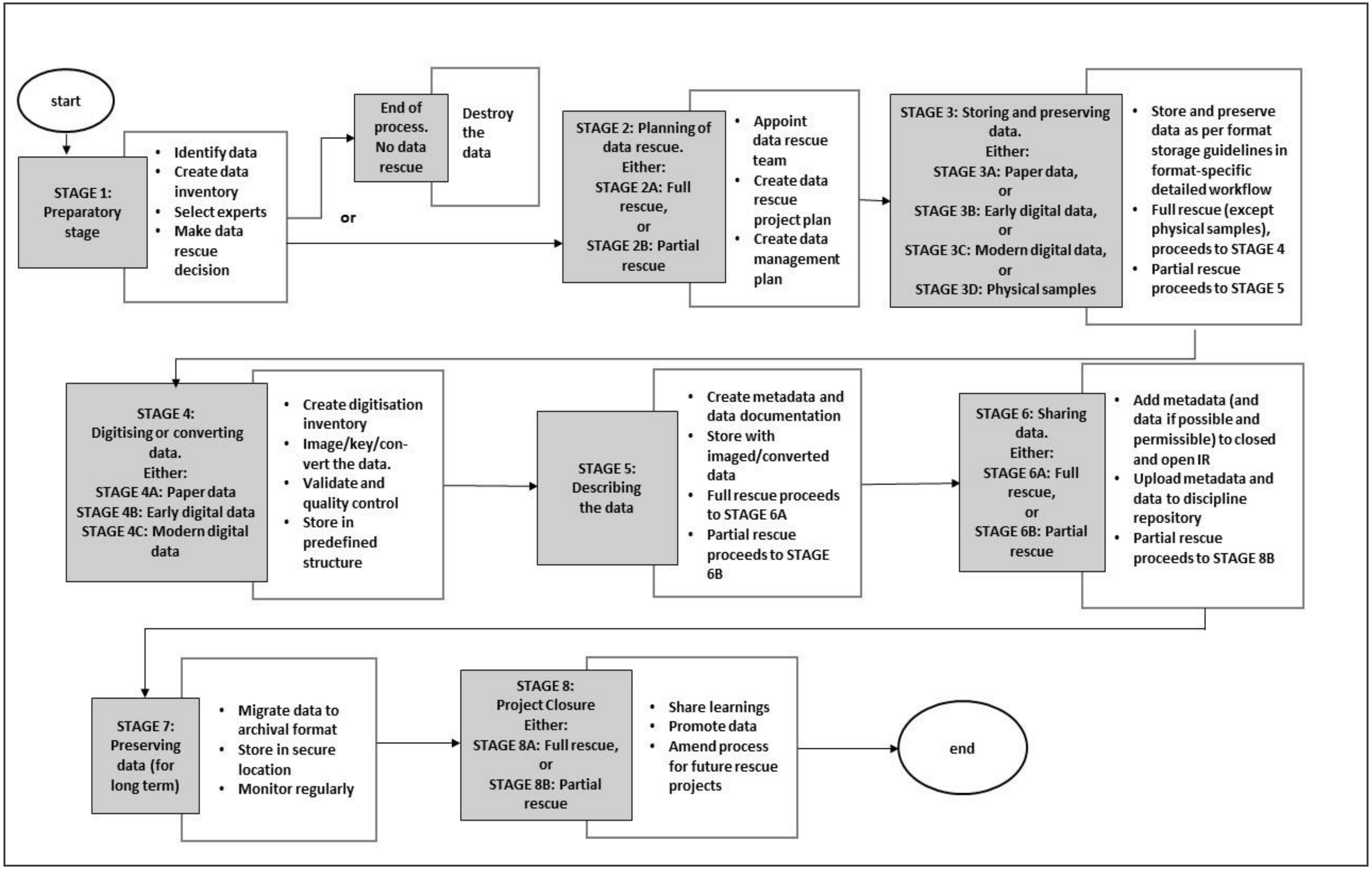

start
STAGE 1: Preparatory stage
• Identify data
• Create data inventory
• Select experts
• Make data rescue decision
or
End of process. No data rescue
Destroy the data
STAGE 2: Planning of data rescue. Either: STAGE 2A: Full rescue, or STAGE 2B: Partial rescue
• Appoint data rescue team
• Create data rescue project plan
• Create data management plan
STAGE 3: Storing and preserving data. Either: STAGE 3A: Paper data, or STAGE 3B: Early digital data, or STAGE 3C: Modern digital data, or STAGE 3D: Physical samples
• Store and preserve data as per format storage guidelines in format-specific detailed workflow
• Full rescue (except physical samples), proceeds to STAGE 4
• Partial rescue proceeds to STAGE 5
STAGE 4: Digitising or converting data. Either: STAGE 4A: Paper data STAGE 4B: Early digital data STAGE 4C: Modern digital data
• Create digitisation inventory
• Image/key/con-vert the data.
• Validate and quality control
• Store in predefined structure
STAGE 5: Describing the data
• Create metadata and data documentation
• Store with imaged/converted data
• Full rescue proceeds to STAGE 6A
• Partial rescue proceeds to STAGE 6B
STAGE 6: Sharing data. Either: STAGE 6A: Full rescue, or STAGE 6B: Partial rescue
• Add metadata (and data if possible and permissible) to closed and open IR
• Upload metadata and data to discipline repository
• Partial rescue proceeds to STAGE 8B
STAGE 7: Preserving data (for long term)
• Migrate data to archival format
• Store in secure location
• Monitor regularly
STAGE 8: Project Closure Either: STAGE 8A: Full rescue, or STAGE 8B: Partial rescue
• Share learnings
• Promote data
• Amend process for future rescue projects
end

Figure 6.1: Data rescue summary diagram

  - The partial data rescue process proceeds to Stage 8, which entails activities forming part of project closure.
  - Stage 6 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 7: Preserving the data (for long term)**
  - This stage involves ensuring the data are in a preservation format and thereby preserved in the long term.
  - A stable and secure storage location should be identified. It is crucial that the identified storage location ensures the regular migration of the data to an updated archival format.
  - Data, in a long-term preservation format, are stored in the identified location.
  - Regular monitoring of the rescued data is required.
  - Stage 7 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

- **Stage 8: Project closure**
  - This stage involves the final activities of the rescue process and entails ensuring that all loose ends are tied up.
  - The final data rescue project report is completed and shared with relevant stakeholders.
  - The rescue project details, learnings and experiences are shared; this can take the form of articles, presentations and social media posts.
  - The rescued data should be promoted within the relevant research community.
  - If required, the rescue process and workflow will be amended for future projects.
  - Stage 8 steps are described in more detail in the subsequent sections explaining the compact four-page workflow and the stage-specific workflows.

To recap: the summary diagram provides a single-page introduction to, and overview of the proposed data rescue process as portrayed via the more extensive data rescue compact workflow, and the detailed individual stages. This overview will be of use to novice data rescuers and persons planning to make use of the model when rescuing data. While not sufficient in terms of data rescue guidance, the summary diagram indicates the main rescue stages, the chronology of stages, and the main features of each stage.

Data Rescue Compact Workflow: The data rescue compact workflow is the second layer of the proposed Data Rescue Workflow Model and comprises a four-page summary of the data rescue process. The four-page process summary consists of four images, which are portrayed via Figure 6.2, Figure 6.3, Figure 6.4 and Figure 6.5.

The data rescue compact workflow has the following features:

- The workflow is portrayed via a flowchart; flowcharts are discussed in Section 3.5.2.
- The workflow comprises eight main stages; the stages are identical to the eight main stages listed for the summary diagram.
- As the workflow provides more data rescue details than the previous section's summary diagram, stages also provide the following information:
    - data formats involved,
    - whether the stage involves full rescue or partial rescue, and Stream A data and/or Stream B data,
    - the important decisions forming part of the stage, and
    - more details on tasks forming part of each stage (when compared to the summary diagram).
- The following shapes are used in the workflow flowchart:

⬭	Start of rescue process
⬭	End of rescue process
▢	Data formats involved
■	Data rescue task
◇	Decision
○	Decision outcome

The remainder of this section portrays the flowcharts comprising the data rescue compact workflow. The workflow, covering four pages, is accompanied by explanations of each stage illustrated in the flowcharts.

The data rescue compact workflow is described and clarified as follows:

- **Stage 1: Preparing for data rescue (part of Figure 6.2)**
    - This stage involves preparatory activities performed to evaluate the data and institutional resources to determine whether data rescue should proceed. The stage also involves the destruction of data that will not be rescued.
    - All data formats are involved in this stage.
    - A first activity entails identifying data that may need to be rescued.
    - Following data identification, the available details of the data are added to a master data inventory.
    - The creation of a data inventory is followed by the appointment of an investigation group who will be tasked with assessing and evaluating the data.
    - Data assessment can have two outcomes, namely data deemed worthy of rescue, and data not deemed as such.
    - Should data be seen not to have rescue value, data destruction will take place and the rescue process is regarded as terminated.
    - Should data be seen to have rescue value, a next step, comprising the evaluation of institutional data rescue resources, takes place.
    - Should resources be deemed to be insufficient, a process of partial data rescue will commence. Data forming part of partial rescue are referred to as 'Stream B' during the workflow. The partial rescue process then proceeds to Stage 2, where the directions for Stream B data are to be adhered to.
    - Should resources be deemed to be sufficient, a process of full data rescue will commence. Data forming part of full rescue are referred to as 'Stream A' during the workflow. Stream A data are further termed as A1 data (paper), A2 data (early digital data, A3 data (modern digital data), and A4 data (physical samples). The full rescue process then proceeds to Stage 2, where the directions for Stream A data are to be adhered to.
    - Stage steps applicable to Stage 1 are described in more detail in the stage-specific workflow portrayed in Figure 6.6.

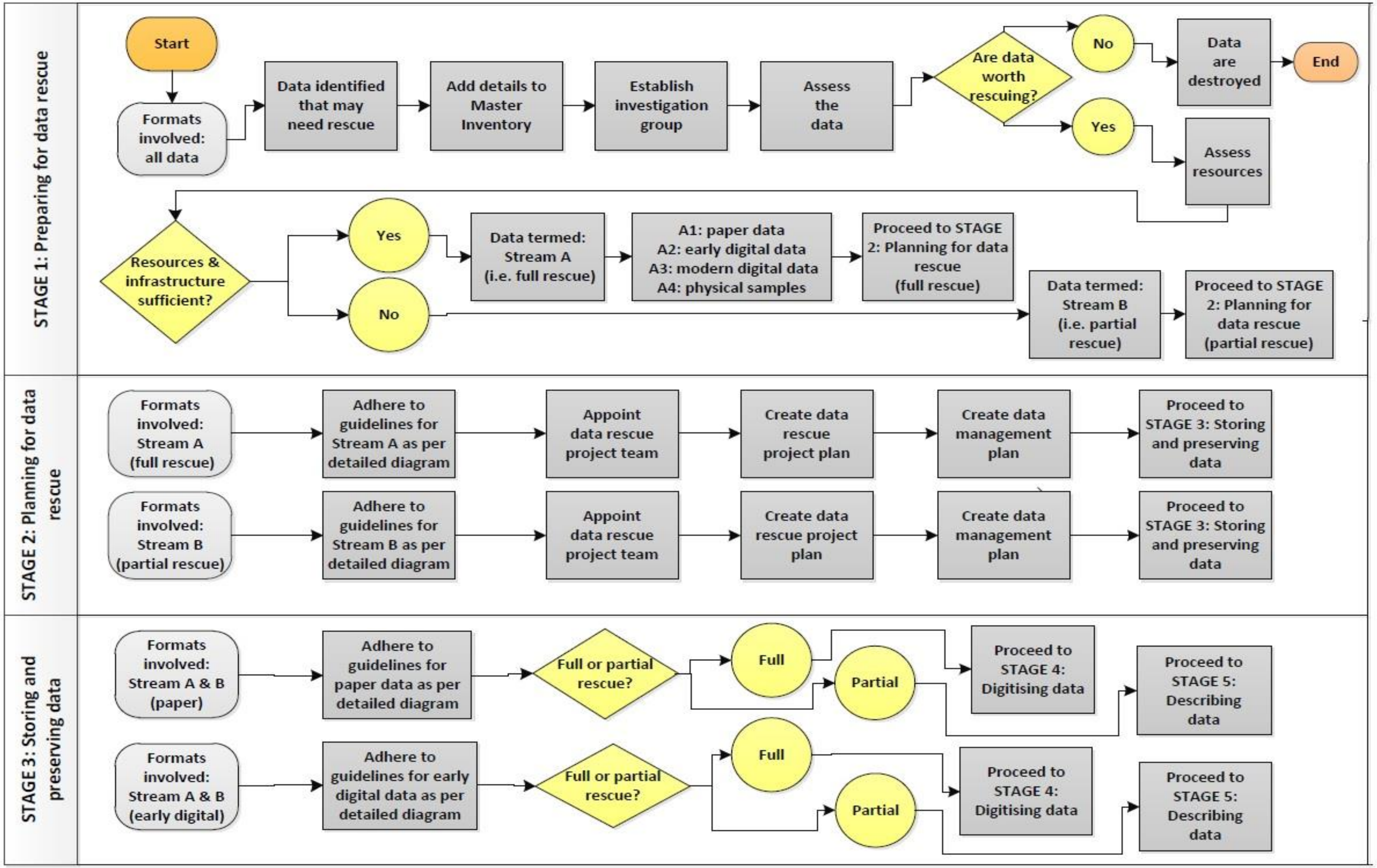

STAGE 1: Preparing for data rescue
Start
Formats involved: all data
Data identified that may need rescue
Add details to Master Inventory
Establish investigation group
Assess the data
Are data worth rescuing?
No
Data are destroyed
End
Yes
Assess resources
Resources & infrastructure sufficient?
Yes
Data termed: Stream A (i.e. full rescue)
A1: paper data
A2: early digital data
A3: modern digital data
A4: physical samples
Proceed to STAGE 2: Planning for data rescue (full rescue)
No
Data termed: Stream B (i.e. partial rescue)
Proceed to STAGE 2: Planning for data rescue (partial rescue)
STAGE 2: Planning for data rescue
Formats involved: Stream A (full rescue)
Adhere to guidelines for Stream A as per detailed diagram
Appoint data rescue project team
Create data rescue project plan
Create data management plan
Proceed to STAGE 3: Storing and preserving data
Formats involved: Stream B (partial rescue)
Adhere to guidelines for Stream B as per detailed diagram
Appoint data rescue project team
Create data rescue project plan
Create data management plan
Proceed to STAGE 3: Storing and preserving data
STAGE 3: Storing and preserving data
Formats involved: Stream A & B (paper)
Adhere to guidelines for paper data as per detailed diagram
Full or partial rescue?
Full
Partial
Proceed to STAGE 4: Digitising data
Proceed to STAGE 5: Describing data
Formats involved: Stream A & B (early digital)
Adhere to guidelines for early digital data as per detailed diagram
Full or partial rescue?
Full
Partial
Proceed to STAGE 4: Digitising data
Proceed to STAGE 5: Describing data

Figure 6.2: Data Rescue Compact Workflow (Stages 1–3)

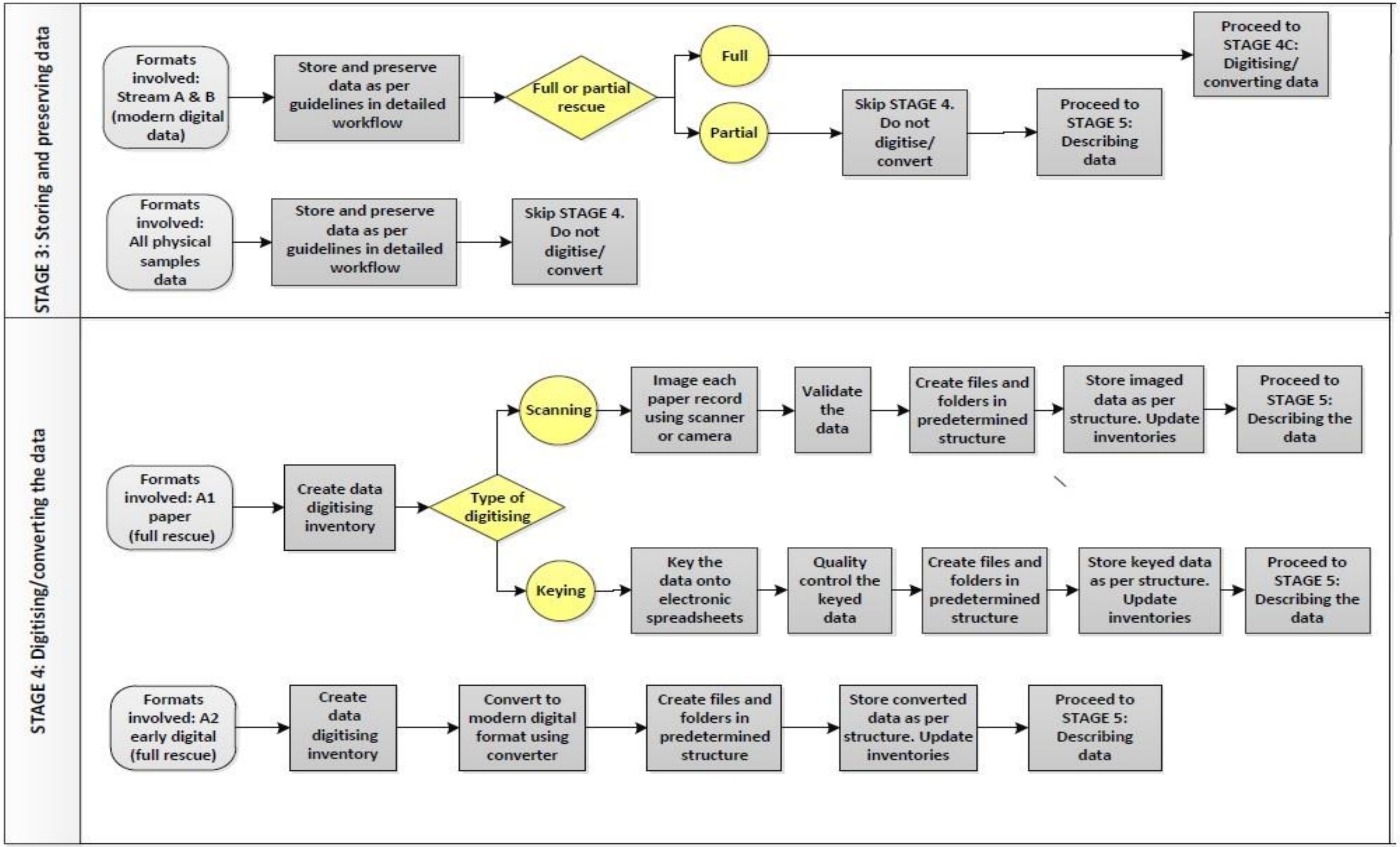

STAGE 3: Storing and preserving data
Formats involved: Stream A & B (modern digital data)
Store and preserve data as per guidelines in detailed workflow
Full or partial rescue
Full
Proceed to STAGE 4C: Digitising/ converting data
Partial
Skip STAGE 4. Do not digitise/ convert
Proceed to STAGE 5: Describing data
Formats involved: All physical samples data
Store and preserve data as per guidelines in detailed workflow
Skip STAGE 4. Do not digitise/ convert
STAGE 4: Digitising/converting the data
Formats involved: A1 paper (full rescue)
Create data digitising inventory
Type of digitising
Scanning
Image each paper record using scanner or camera
Validate the data
Create files and folders in predetermined structure
Store imaged data as per structure. Update inventories
Proceed to STAGE 5: Describing the data
Keying
Key the data onto electronic spreadsheets
Quality control the keyed data
Create files and folders in predetermined structure
Store keyed data as per structure. Update inventories
Proceed to STAGE 5: Describing the data
Formats involved: A2 early digital (full rescue)
Create data digitising inventory
Convert to modern digital format using converter
Create files and folders in predetermined structure
Store converted data as per structure. Update inventories
Proceed to STAGE 5: Describing data

Figure 6.3: Data Rescue Compact Workflow (Stage 3–4)

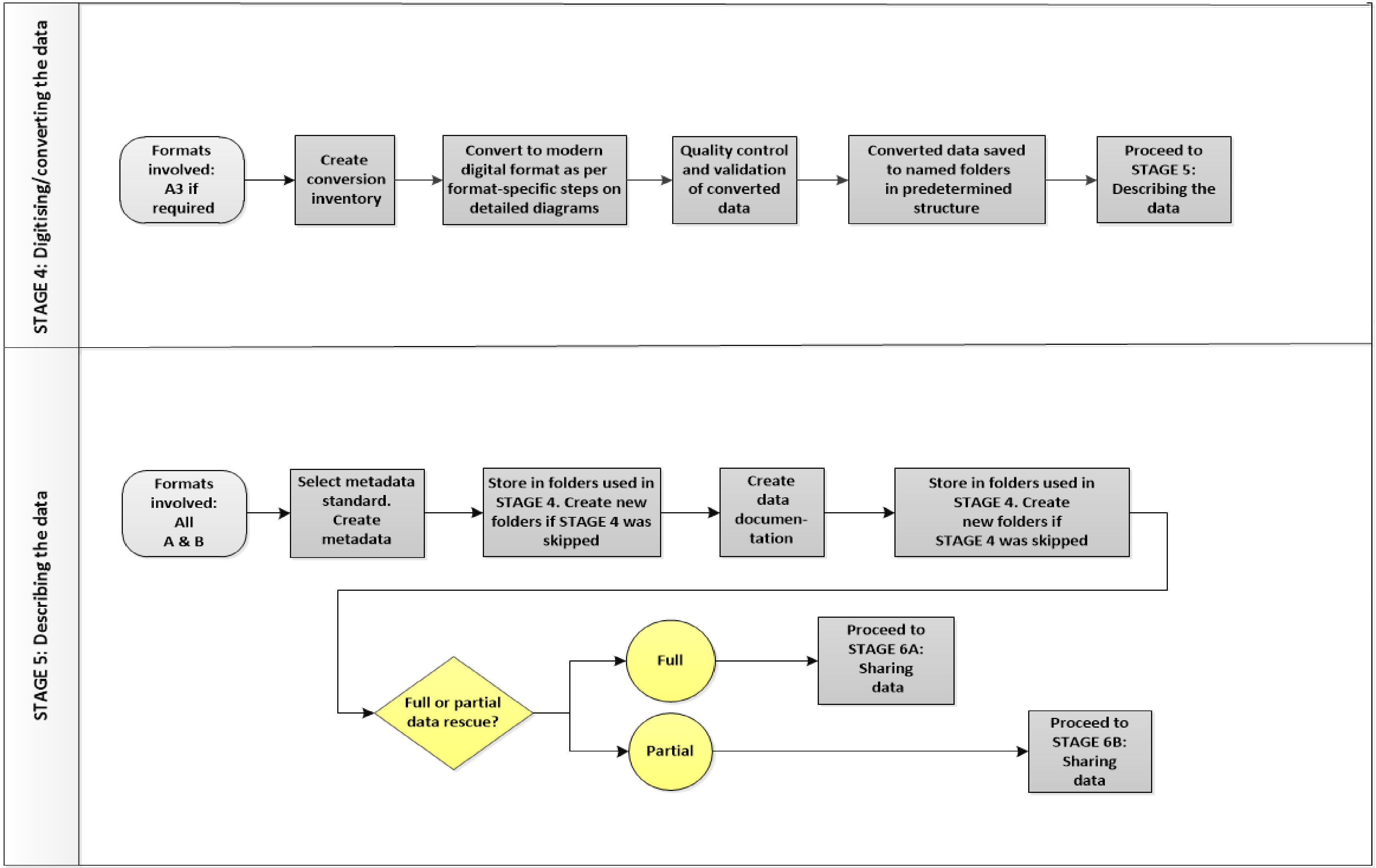

STAGE 4: Digitising/converting the data
Formats involved: A3 if required
Create conversion inventory
Convert to modern digital format as per format-specific steps on detailed diagrams
Quality control and validation of converted data
Converted data saved to named folders in predetermined structure
Proceed to STAGE 5: Describing the data
STAGE 5: Describing the data
Formats involved: All A & B
Select metadata standard. Create metadata
Store in folders used in STAGE 4. Create new folders if STAGE 4 was skipped
Create data documen-tation
Store in folders used in STAGE 4. Create new folders if STAGE 4 was skipped
Full or partial data rescue?
Full
Partial
Proceed to STAGE 6A: Sharing data
Proceed to STAGE 6B: Sharing data

Figure 6.4: Data Rescue Compact Workflow (Stages 4–5)

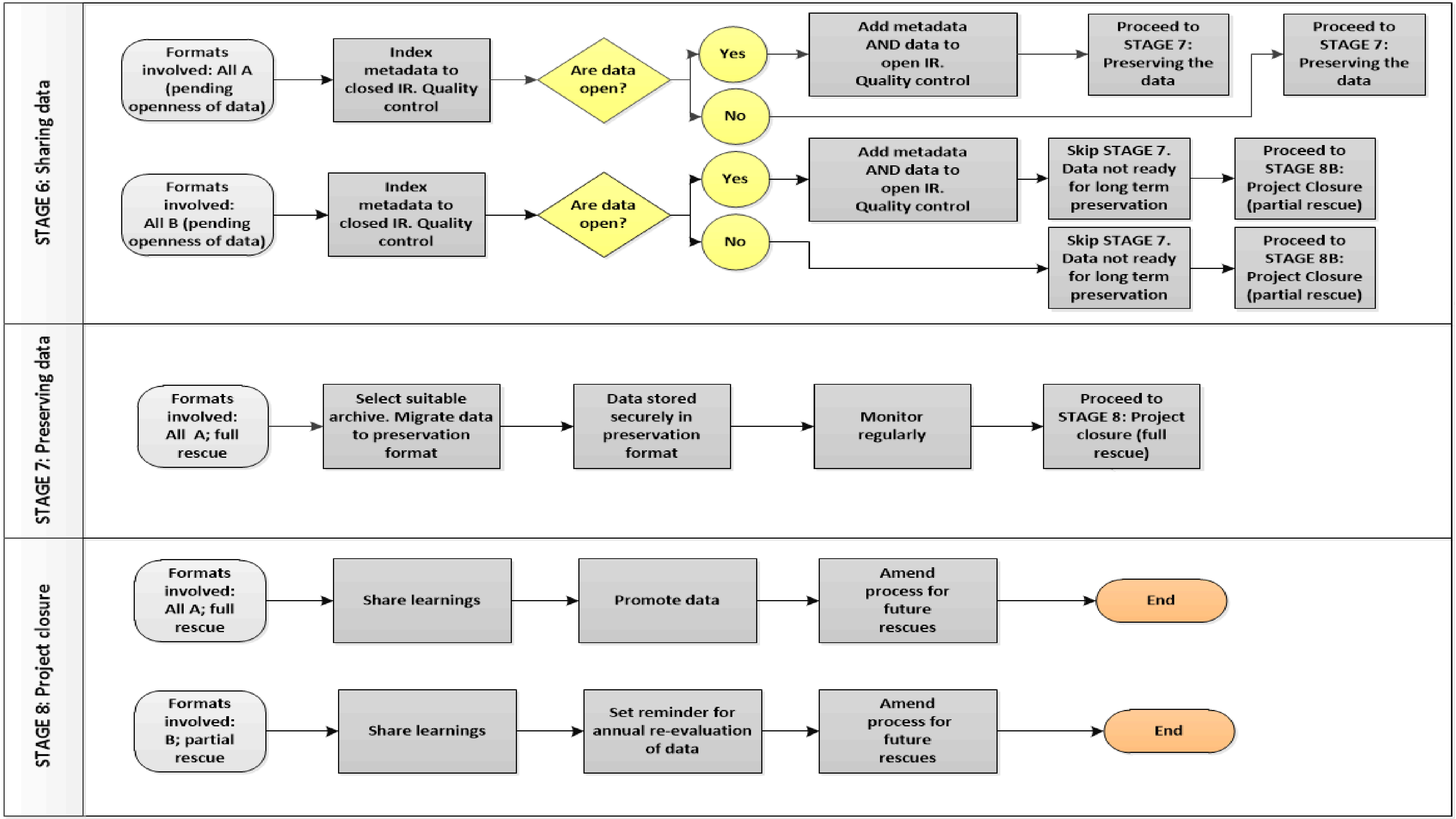

STAGE 6: Sharing data
Formats involved: All A (pending openness of data)
Index metadata to closed IR. Quality control
Are data open?
Yes
Add metadata AND data to open IR. Quality control
Proceed to STAGE 7: Preserving the data
No
Proceed to STAGE 7: Preserving the data
Formats involved: All B (pending openness of data)
Index metadata to closed IR. Quality control
Are data open?
Yes
Add metadata AND data to open IR. Quality control
Skip STAGE 7. Data not ready for long term preservation
Proceed to STAGE 8B: Project Closure (partial rescue)
No
Skip STAGE 7. Data not ready for long term preservation
Proceed to STAGE 8B: Project Closure (partial rescue)
STAGE 7: Preserving data
Formats involved: All A; full rescue
Select suitable archive. Migrate data to preservation format
Data stored securely in preservation format
Monitor regularly
Proceed to STAGE 8: Project closure (full rescue)
STAGE 8: Project closure
Formats involved: All A; full rescue
Share learnings
Promote data
Amend process for future rescues
End
Formats involved: B; partial rescue
Share learnings
Set reminder for annual re-evaluation of data
Amend process for future rescues
End

Figure 6.5: Data Rescue Compact Workflow (Stages 6–8)

- **Stage 2: Planning for data rescue (Figure 6.2)**
    - This stage involves the planning activities performed after a decision has been made to proceed with the rescue of data.
    - Both full data rescue and partial data rescue (i.e., Stream A data and Stream B data) are involved in this stage.
    - Both data streams undergo the following activities:
        - the selection/appointment of a data rescue project team,
        - the drafting of a data rescue project, and
        - the drafting of a data management plan.
    - Both the full rescue process and the partial rescue process then proceed to Stage 3, where the directions for the applicable data format are to be adhered to.
    - Stage steps applicable to Stage 2 are described in more detail in the stage-specific workflow portrayed in Figure 6.7 (for full data rescue) and Figure 6.8 (for partial data rescue).

- **Stage 3: Storing and preserving the data (Figure 6.2 and Figure 6.3)**
    - This stage involves the secure storage and preservation of the data prior to it being rescued.
    - Both full data rescue and partial data rescue (i.e., Stream A data and Stream B data) are involved in this stage. All data formats are involved in this stage.
    - Data of both data streams are stored and preserved according to the format-specific guidelines forming part of the stage-specific workflows.
    - Following completion of the stage, the next format-specific stages are indicated below.
        - Paper-based data forming part of full rescue proceed to Stage 4 (data will be digitised).
        - Early digital data forming part of full rescue proceed to Stage 4 (data will be converted).
        - Modern digital data forming part of full rescue proceed to Stage 4 (data will be converted).
        - Physical samples data forming part of full rescue completion of this stage proceed to Stage 5 (data conversion is not done).

    - Paper-based data forming part of partial rescue proceed to Stage 5 (data will not be digitised).
    - Early digital data forming part of partial rescue proceed to Stage 5 (data will not be converted).
    - Modern digital data forming part of partial rescue proceed to Stage 5 (data will not be converted).
  - Stage steps applicable to Stage 3 are described in more detail in the stage-specific workflow portrayed in Figure 6.9 (for paper-based data), Figure 6.10 (for early digital data), Figure 6.11 (for modern digital data) and Figure 6.12 (for physical samples data).

- **Stage 4: Digitising or converting the data (Figure 6.3 and Figure 6.4)**
  - This stage involves the digitisation of data to a modern electronic format, or the conversion of data in an older electronic format or proprietary modern format to a common, open, modern electronic format.
  - Digitisation of paper-based data entails the following steps:
    - A digitising inventory is created.
    - Digitising of paper data can entail imaging (data scanned or digitally photographed) or data keyed onto electronic spreadsheets.
    - For scanning of data:
      - image each page using scanner or camera,
      - validate the imaged data,
      - create files and folders according to predetermined structure,
      - store the imaged data in files and folders,
      - update the master inventory and the digitisation inventory, and
      - proceed to Stage 5.
    - For keying of data:
      - key data onto electronic spreadsheets,
      - quality control the keyed data,
      - create files and folders according to predetermined structure,
      - store the keyed data in files and folders,
      - update the master inventory and the digitisation inventory, and
      - proceed to Stage 5.
  - Conversion of early digital data entails the following steps:

        - A conversion inventory is created.
        - The data are converted to a common, open access, modern digital format.
        - Files and folders are created according to a predetermined structure.
        - The converted data are stored in the files and folders.
        - The master inventory and the conversion inventory are updated.
        - The rescue process proceeds to Stage 5.
    - Conversion of modern proprietary data in digital format entails the following steps:
        - A digitisation inventory is created.
        - The data are converted to a common, open, modern electronic format.
        - Files and folders are created according to a predetermined structure.
        - The converted data are stored in the files and folders.
        - The master inventory and the conversion inventory are updated.
        - The rescue process proceeds to Stage 5.
    - Stage steps applicable to Stage 4 are described in more detail in the stage-specific workflow portrayed in Figure 6.13 (for paper-based data), Figure 6.14 (for early digital data) and Figure 6.15 (for modern digital data).

- **Stage 5: Describing the data (part of Figure 6.4)**
    - This stage involves describing the data by creating metadata and data documentation linked to the data, thereby ensuring that the data can be understood and used by future users.
    - Both full data rescue and partial data rescue (i.e., Stream A data and Stream B data) are involved in this stage. All data formats are involved in this stage.
    - A first step entails selecting the appropriate metadata standard to be used. Alternatively, adherence to the metadata requirements of the repository to be used for data sharing (see Stage 6) will suffice.
    - Metadata are created and saved in an electronic format.
    - Data documentation are created and stored in an electronic format.
    - The metadata and data documentation of digitised/electronic data are stored with the data in the predefined structure mentioned in Stage 4.
    - The metadata and data documentation of non-digitised/non-electronic data are stored in a secure folder as described in the data management plan.
    - Full data rescue will proceed to Stage 6A.
    - Partial data rescue will proceed to Stage 6B.

  - Stage steps applicable to Stage 5 are described in more detail in the stage-specific workflow portrayed in Figure 6.16.

- **Stage 6: Sharing the data (part of Figure 6.5)**
  - This stage involves sharing the digitised or converted data and its metadata and data documentation. It also involves the sharing of metadata and data documentation of data that had not been digitised or converted. The sharing of data and its data documentation and metadata involves the upload of data, data documentation and metadata to data repositories.
  - Both full data rescue and partial data rescue form part of this stage, but the involved activities of each are not identical.
  - Full data rescue steps to be performed during this stage are as follows: selecting suitable data repositories, uploading data with metadata and data documentation to the repositories, and adding the DOI to relevant documentation.
  - Partial data rescue steps to be performed during this stage are as follows: selecting suitable data repositories, uploading metadata to the repositories, and adding the DOI to relevant documentation.
  - Stage steps applicable to Stage 6 are described in more detail in the stage-specific workflow portrayed in Figure 6.17 (for full data rescue) and Figure 6.18 (for partial data rescue).

- **Stage 7: Preserving the data for the long term (part of Figure 6.5)**
  - This stage involves ensuring the data are in a preservation format and thereby preserved in the long term.
  - This stage only involves electronic data forming part of full rescue. Partial rescue does not entail long-term preservation of data.
  - A suitable platform (digital archive) for the secure long-term preservation of the data is identified.
  - Ensure that data are in a format accepted by the archive.
  - Activities such as regular migration of the archived data are archival duties and are not performed by the rescue team.
  - Notwithstanding the previous bullet: access to the archived data should be monitored on a regular basis by a designated rescue team member.

    - The data rescue process moves on to Stage 8.
    - Stage steps applicable to Stage 7 are described in more detail in the stage-specific workflow portrayed in Figure 6.19.

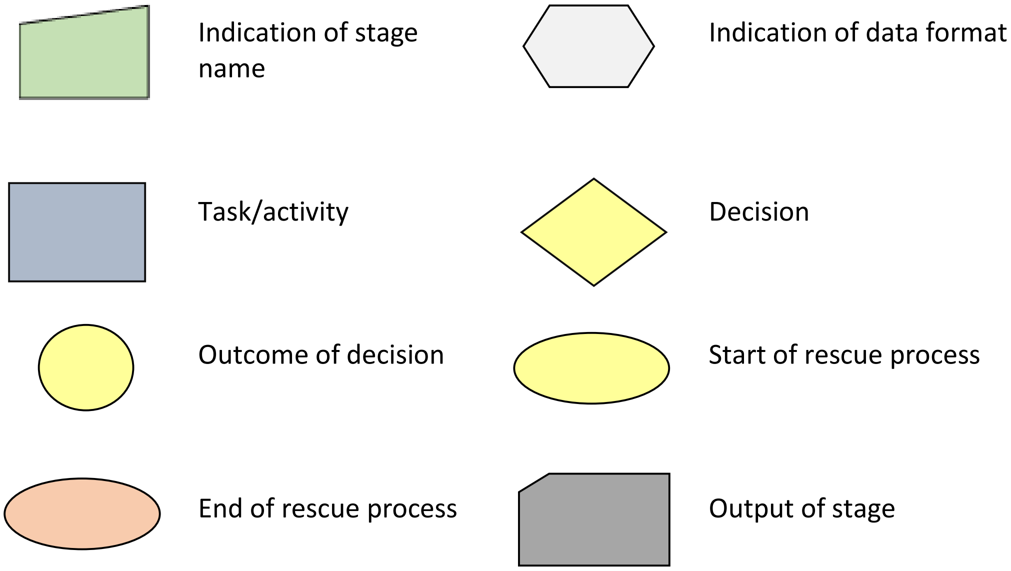
- **Stage 8: Project closure (part of Figure 6.5)**
    - This stage involves the final activities of the rescue process and entails ensuring that all loose ends are tied up.
    - Both full data rescue and partial data rescue form part of this stage. All data formats are involved in this stage.
    - Full data rescue involves the following activities:
        - learnings obtained via the rescue project should be shared with others,
        - the rescued data should be promoted, and
        - if required, the rescue process will be amended for future rescue projects.
    - Partial data rescue involves the following activities:
        - learnings obtained via the rescue project should be shared with others,
        - a reminder is to be set to re-evaluate institutional data rescue resources annually, and
        - if required, the rescue process will be amended for future rescue projects.
    - Stage steps are described in more detail in the stage-specific workflow portrayed in Figure 6.20 (for full data rescue) and Figure 6.21 (for partial data rescue).
    - The completion of this stage signals the end of the last stage of the rescue process.

The next section contains depictions and discussions of the detailed stage-specific data rescue workflows.

Detailed stages of the recommended Data Rescue Workflow Model: The detailed stages of the recommended model form the last section of the model. Although the activities contained within the stage-specific workflows are identical to the activities in the summary diagram and the compact workflow, this section of the model portrays activities in more detail, and include features such as:

- workflows specific to data formats,
- links to stage-specific or task-specific guidance documents, and
- indication of outputs linked to a stage or activity.

The stage-specific workflows are portrayed via flowcharts. The shapes used in the stage-specific flowcharts, together with a description of each shape, are indicated below.



The remainder of this section features the depiction of the separate stage-specific workflows, with each workflow accompanied by a description and clarification.

Stage 1: Preparatory stage

- This stage is depicted in Figure 6.6, and expands on the information provided in Figure 6.2, and its linked description.
- This stage involves preparatory activities performed to evaluate the data and institutional resources to determine whether data rescue should proceed. The stage also involves the destruction of data that will not be rescued.
- It is important to note that the implementation of a data rescue project does not necessarily coincide with the gathering of new data. Data rescue is concerned with the rescue (storage, sharing, preservation, promotion) of data that were identified as being at risk and with ensuring that the data are converted to a preservation format, described, and made available to the public. Data rescue (or the outcome of data rescue) only features in conventional new research projects when the researcher discovers that the data that were to be collected during the project, are already in existence, have been rescued, and can now be used during the new project. Adding the data to another dataset or two to form a longitudinal study is another possibility.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow, with additional clarifying details below.
- Guidance documents for this stage are as follows:

  - a document titled 'Guidance on data assessment' (see Appendix 11: Guidance on data assessment), and
  - templates for the creation of data inventories (see Appendix 17: Sample templates for the creation of data inventories).
- Recommended outputs to be created during this stage include:
  - a master data inventory,
  - a selected group of experts,
  - destroyed data if applicable,
  - a decision regarding the type of rescue that will be pursued (full or partial), and
  - data termed as Stream A or Stream B.

Stage 2A: Planning for data rescue (full rescue)

- The stage is depicted in Figure 6.7 and expands on the information provided in Figure 6.2 and its linked description.
- This stage involves the planning activities performed after a decision has been made to proceed with the full rescue of data.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (for full rescue), with additional clarifying details below.
- Guidance documents for this stage are as follows:
  - a document titled 'Guidance on data rescue project planning' (see Appendix 12), and
  - a document titled 'Guidance on data management plans' (see Appendix 13).
- Recommended outputs to be created during this stage include:
  - a project folder,
  - a project plan also containing a data rescue progress document, and
  - a data management plan.

Stage 2B: Planning for data rescue (partial rescue)

- The stage is depicted in Figure 6.8 and expands on the information provided in Figure 6.2 and its linked description.
- This stage involves the planning activities performed after a decision has been made to proceed with the partial rescue of data.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (partial rescue), with additional clarifying details below.
- Guidance documents for this stage are as follows:

    - a document titled ‘Guidance on data rescue project planning’ (see Appendix 12), and
    - a document titled ‘Guidance on data management plans’ (see Appendix 13).
- Recommended outputs to be created during this stage include:
    - a project folder,
    - a project plan also containing a data rescue progress document, and a data management plan.

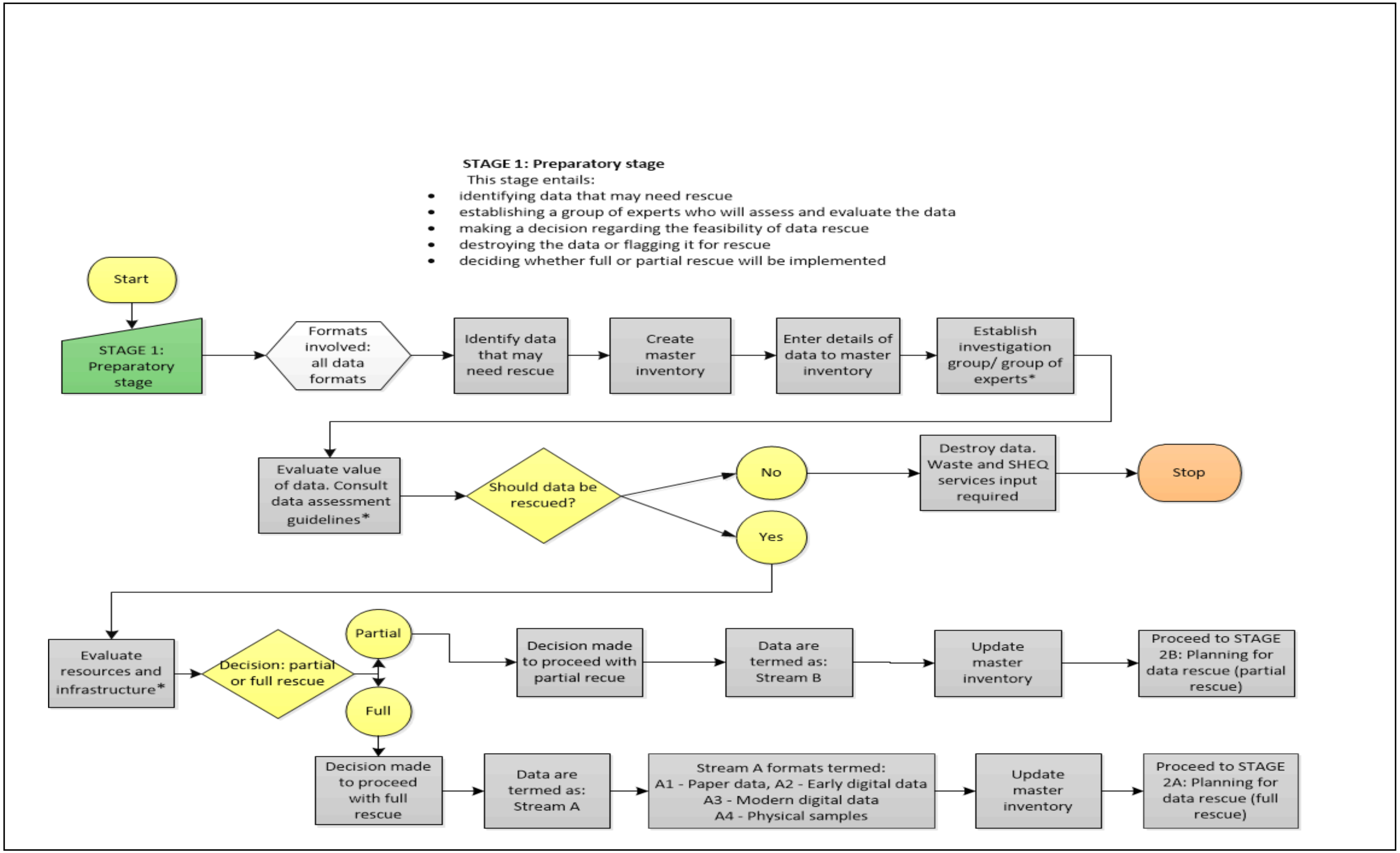

STAGE 1: Preparatory stage
This stage entails:
• identifying data that may need rescue
• establishing a group of experts who will assess and evaluate the data
• making a decision regarding the feasibility of data rescue
• destroying the data or flagging it for rescue
• deciding whether full or partial rescue will be implemented
Start
STAGE 1: Preparatory stage
Formats involved: all data formats
Identify data that may need rescue
Create master inventory
Enter details of data to master inventory
Establish investigation group/ group of experts*
Evaluate value of data. Consult data assessment guidelines*
Should data be rescued?
No
Yes
Destroy data. Waste and SHEQ services input required
Stop
Evaluate resources and infrastructure*
Decision: partial or full rescue
Partial
Full
Decision made to proceed with partial recue
Data are termed as: Stream B
Update master inventory
Proceed to STAGE 2B: Planning for data rescue (partial rescue)
Decision made to proceed with full rescue
Data are termed as: Stream A
Stream A formats termed: A1 - Paper data, A2 - Early digital data A3 - Modern digital data A4 - Physical samples
Update master inventory
Proceed to STAGE 2A: Planning for data rescue (full rescue)

Figure 6.6: Stage 1: Data rescue preparatory stage

Stage 2A: Planning for data rescue (full rescue)

This stage entails the following:

- appointing a team for the data rescue project
- creating a project plan for the data rescue project
- drafting a data management plan for the data rescue project

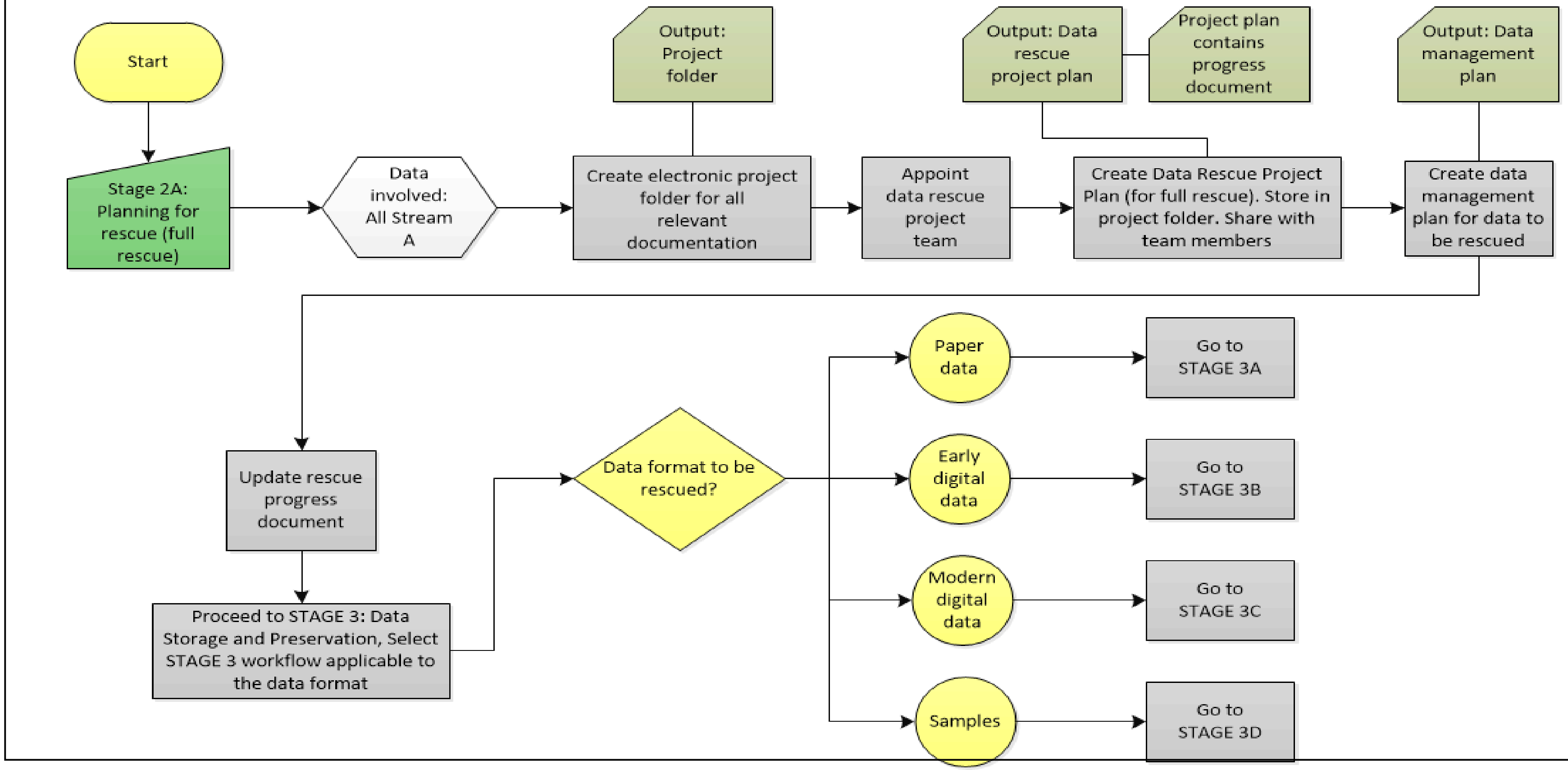

Start
Stage 2A: Planning for rescue (full rescue)
Data involved: All Stream A
Output: Project folder
Create electronic project folder for all relevant documentation
Appoint data rescue project team
Output: Data rescue project plan
Project plan contains progress document
Create Data Rescue Project Plan (for full rescue). Store in project folder. Share with team members
Output: Data management plan
Create data management plan for data to be rescued
Update rescue progress document
Proceed to STAGE 3: Data Storage and Preservation, Select STAGE 3 workflow applicable to the data format
Data format to be rescued?
Paper data
Early digital data
Modern digital data
Samples
Go to STAGE 3A
Go to STAGE 3B
Go to STAGE 3C
Go to STAGE 3D

Figure 6.7: Stage 2A: Planning for data rescue (full rescue)

Stage 2B: Planning for data rescue (partial rescue)

This stage entails the following:

- creating a folder for the data rescue project
- appointing a team for the data rescue project
- creating a project plan for the data rescue project
- drafting a data management plan for the data rescue project

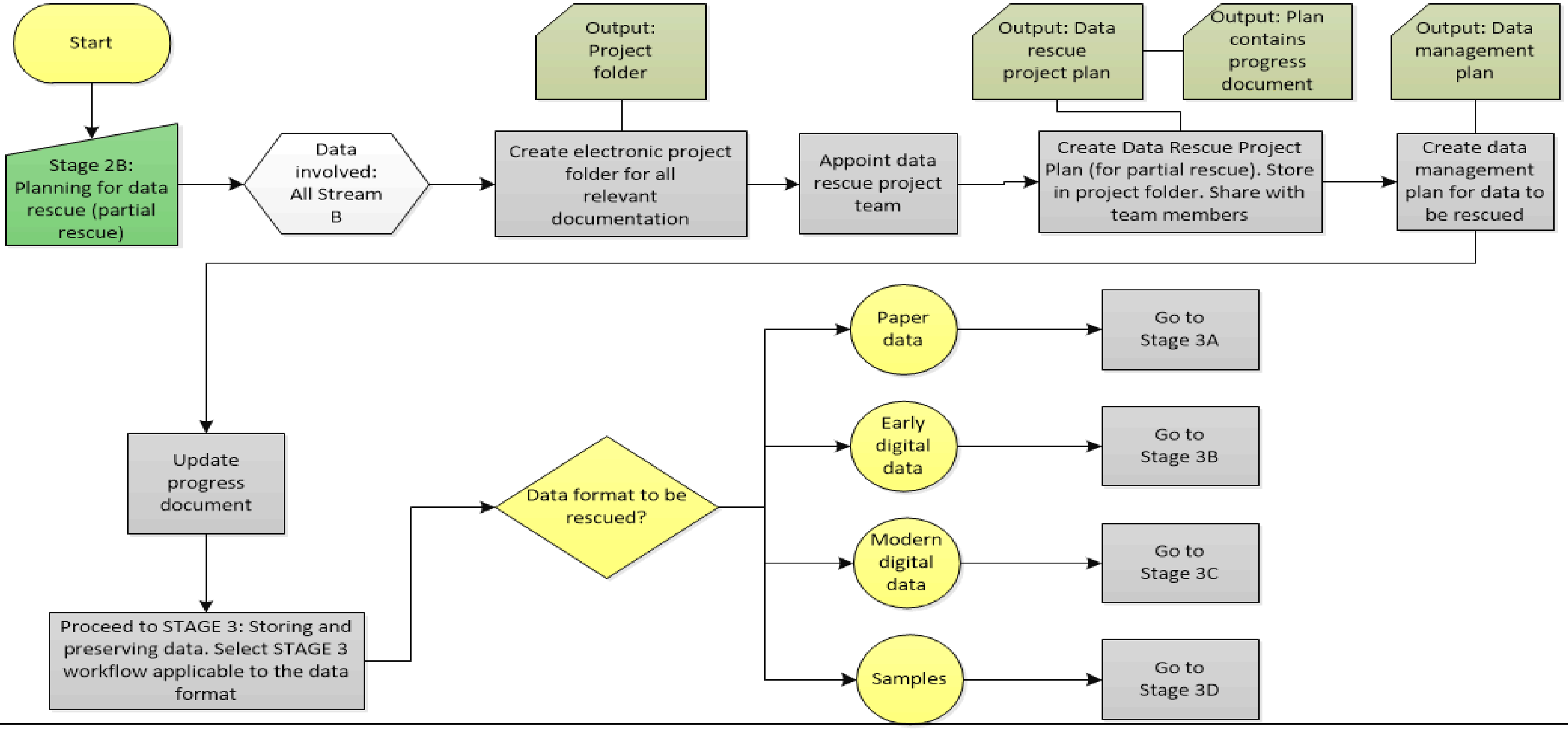

Start
Stage 2B: Planning for data rescue (partial rescue)
Data involved: All Stream B
Output: Project folder
Create electronic project folder for all relevant documentation
Appoint data rescue project team
Output: Data rescue project plan
Output: Plan contains progress document
Create Data Rescue Project Plan (for partial rescue). Store in project folder. Share with team members
Output: Data management plan
Create data management plan for data to be rescued
Update progress document
Proceed to STAGE 3: Storing and preserving data. Select STAGE 3 workflow applicable to the data format
Data format to be rescued?
Paper data
Early digital data
Modern digital data
Samples
Go to Stage 3A
Go to Stage 3B
Go to Stage 3C
Go to Stage 3D

Figure 6.8: Stage 2B: Planning for data rescue (partial rescue)

Stage 3A: Storing and preserving paper-based data

- The stage is depicted in Figure 6.9 and expands on the information provided in Figure 6.2 and its linked description.
- This stage involves the secure storage and preservation of paper data prior to it being rescued.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (paper data), with additional clarifying details below.
- Guidance documents for this stage are as follows:
    - a document titled 'Guidance on storage of paper data' (see Appendix 14), and
    - a document titled 'Guidance on archives and SHEQ' (see Appendix 15).
- Recommended outputs to be created during this stage include:
    - labelled boxes,
    - labelled shelves,
    - paper-based data placed inside boxes on shelves, and
    - updated data rescue progress document.

Stage 3B: Storing and preserving early digital media

- The stage is depicted in Figure 6.10 and expands on the information provided in Figure 6.2 and its linked description.
- This stage involves the secure storage and preservation of early digital data prior to it being rescued.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (early digital data), with additional clarifying details below.
- Recommended outputs to be created during this stage include:
    - early digital data stored as per best practices for the data format, and
    - updated rescue progress document.

Stage 3C: Storing and preserving modern digital data

- The stage is depicted in Figure 6.11 and expands on the information provided in Figure 6.3 and its linked description.
- This stage involves the secure storage and preservation of modern digital data prior to it being rescued.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (modern digital data), with additional clarifying details below.
- Recommended outputs to be created during this stage include:

  - file and folder structure for the unconverted data created as per institutional records management guidelines,
  - data stored in structure, and
  - updated rescue progress document.

Stage 3D: Storing and preserving samples

- The stage is depicted in Figure 6.12 and expands on the information provided in Figure 6.2 and its linked description.
- This stage involves the secure storage and preservation of physical samples data prior to it being rescued.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (physical samples data), with additional clarifying details below.
- Guidance document for this stage is as follows:
  - a document titled ‘Guidance on storage of physical samples and specimens’ (see Appendix 16).
- Recommended outputs to be created during this stage include:
  - physical samples data or specimens stored securely, and
  - an updated rescue progress document.

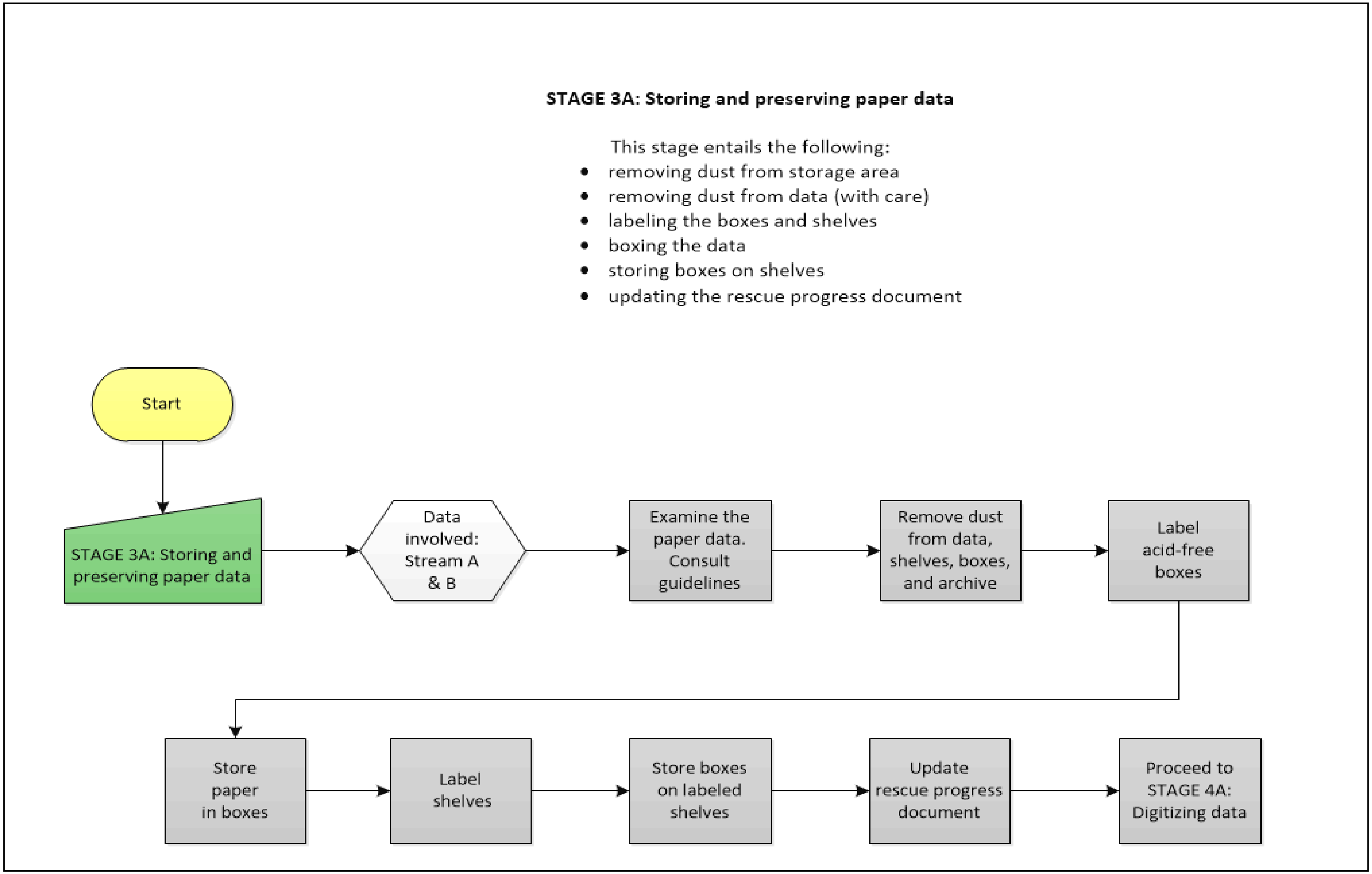

STAGE 3A: Storing and preserving paper data
This stage entails the following:
• removing dust from storage area
• removing dust from data (with care)
• labeling the boxes and shelves
• boxing the data
• storing boxes on shelves
• updating the rescue progress document
Start
STAGE 3A: Storing and preserving paper data
Data involved: Stream A & B
Examine the paper data. Consult guidelines
Remove dust from data, shelves, boxes, and archive
Label acid-free boxes
Store paper in boxes
Label shelves
Store boxes on labeled shelves
Update rescue progress document
Proceed to STAGE 4A: Digitizing data

Figure 6.9: Stage 3A: Storing and preserving paper data

STAGE 3B: Storing and preserving early digital media

This stage contains the following activities:

- examining the early digital data
- storing the data according to best practice for the format
- locating the metadata linked to the data
- updating the rescue progress document accordingly

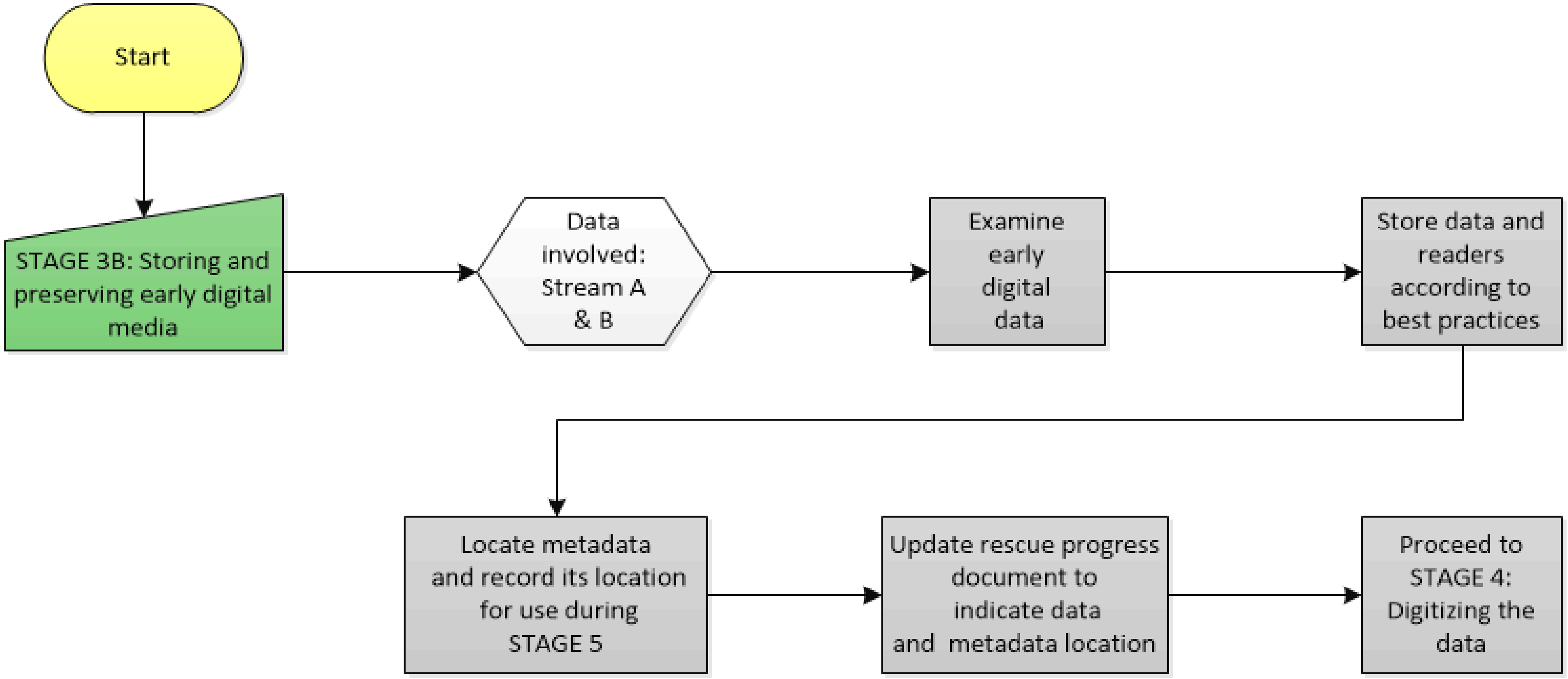

Start
STAGE 3B: Storing and preserving early digital media
Data involved: Stream A & B
Examine early digital data
Store data and readers according to best practices
Locate metadata and record its location for use during STAGE 5
Update rescue progress document to indicate data and metadata location
Proceed to STAGE 4: Digitizing the data

Figure 6.10: Stage 3B: Storing and preserving early digital media

STAGE 3C: Storing and preserving modern digital data

This stage entails the following:

- examining the data in modern digital format
- transferring the data to a secure location if required
- renaming filenames and folders if required
- updating the rescue progress document

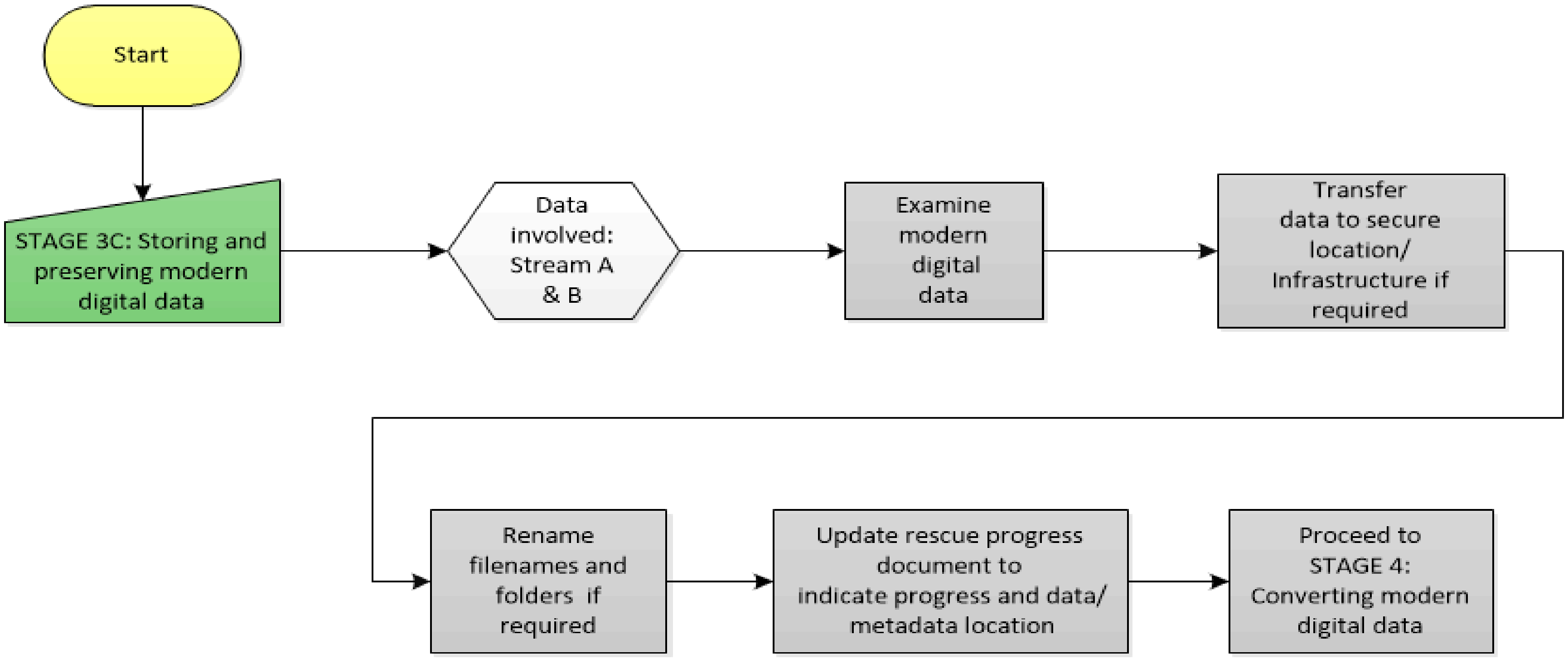
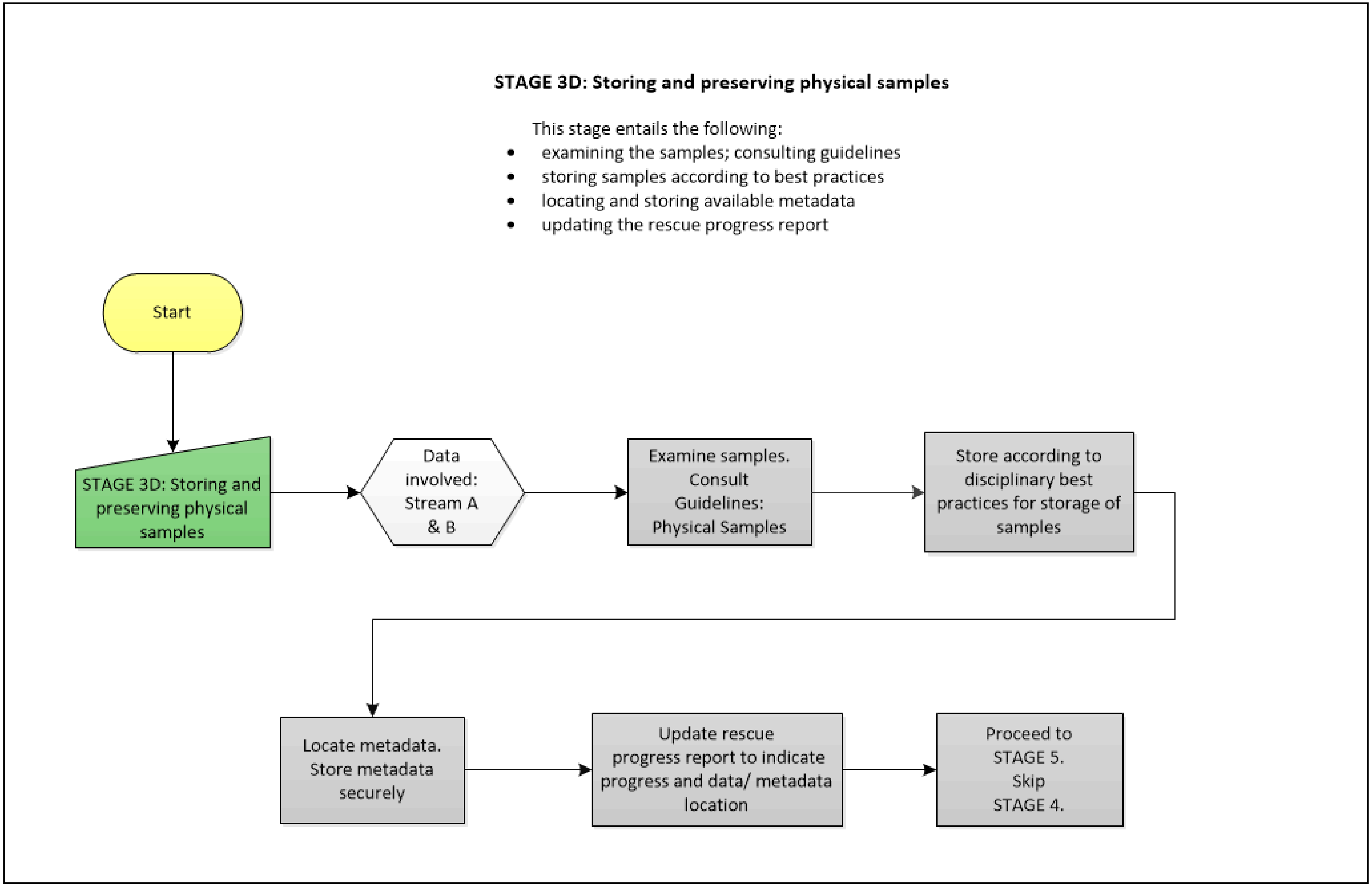

Start
STAGE 3C: Storing and preserving modern digital data
Data involved: Stream A & B
Examine modern digital data
Transfer data to secure location/ Infrastructure if required
Rename filenames and folders if required
Update rescue progress document to indicate progress and data/ metadata location
Proceed to STAGE 4: Converting modern digital data

Figure 6.11: Stage 3C: Storing and preserving modern digital data



STAGE 3D: Storing and preserving physical samples

This stage entails the following:

- examining the samples; consulting guidelines
- storing samples according to best practices
- locating and storing available metadata
- updating the rescue progress report

Start
STAGE 3D: Storing and preserving physical samples
Data involved: Stream A & B
Examine samples. Consult Guidelines: Physical Samples
Store according to disciplinary best practices for storage of samples
Locate metadata. Store metadata securely
Update rescue progress report to indicate progress and data/ metadata location
Proceed to STAGE 5. Skip STAGE 4.

Figure 6.12: Stage 3D: Storing and preserving samples

Stage 4A: Digitising paper-based data

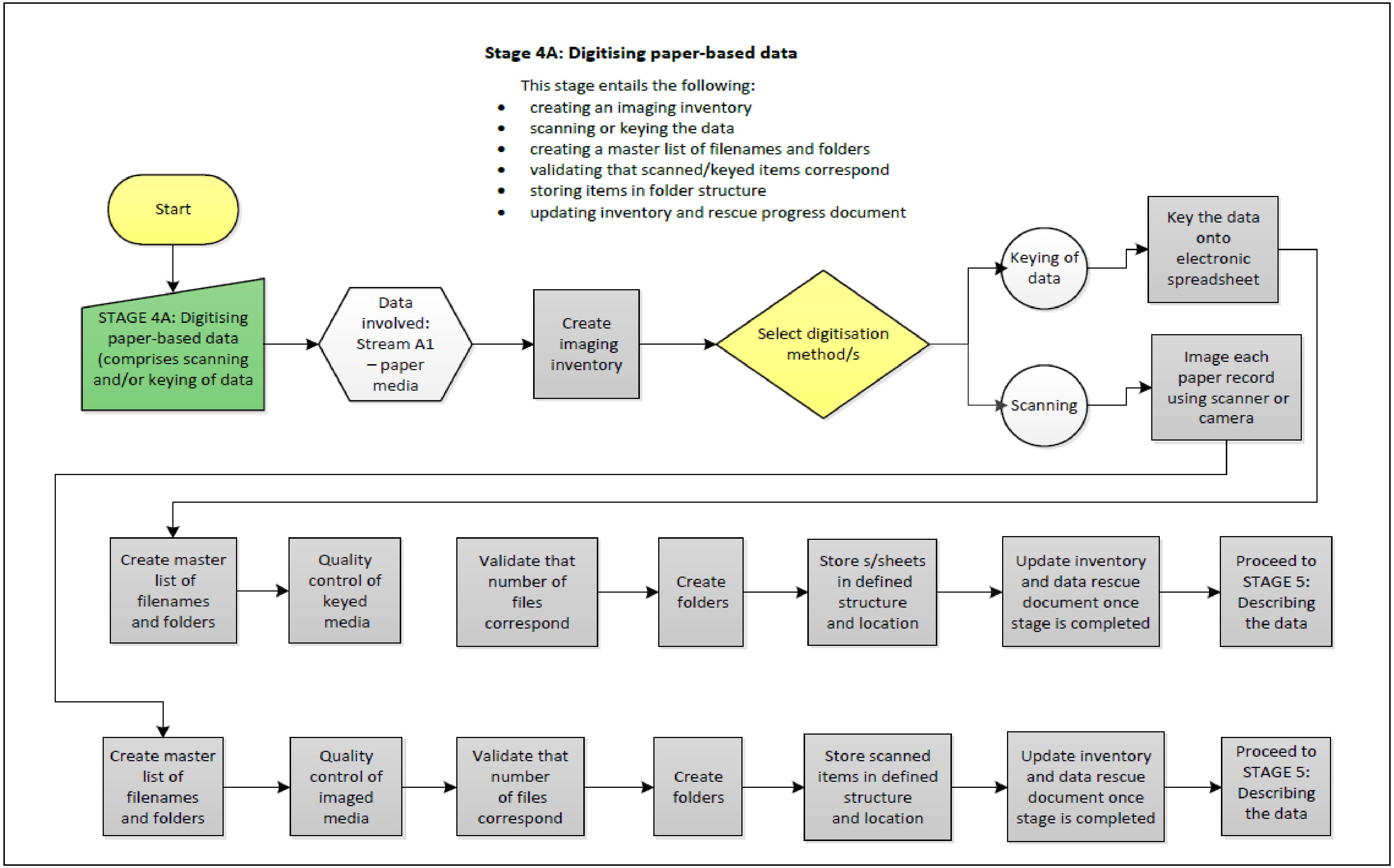
- The stage is depicted in Figure 6.13 and expands on the information provided in Figure 6.3.
- This stage involves the digitisation of paper data to a common, open, modern digital format.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (paper data), with additional clarifying details below.
- Guidance document for this stage is as follows:
  - a document titled 'Guidance on digitisation of paper data' (see Appendix 18).
- Recommended outputs to be created during this stage include:
  - predetermined file and folder structure created,
  - data digitised to a common, open, modern digital format, and
  - updated rescue progress document.

Stage 4B: Converting early digital data

- The stage is depicted in Figure 6.14 and expands on the information provided in Figure 6.3.
- This stage involves the conversion of early digital data to a common, open, modern digital format.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (early digital data), with additional clarifying details below.
- Recommended outputs to be created during this stage include:
  - predetermined file and folder structure created,
  - data converted to a common, open, modern digital format, and
  - updated rescue progress document.

Stage 4C: Converting modern digital data

- The stage is depicted in Figure 6.15 and expands on the information provided in Figure 6.4.
- This stage involves the conversion of modern proprietary data to a common, open, modern digital format.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (modern digital data), with additional clarifying details below.
- Recommended outputs to be created during this stage include:
  - predetermined file and folder structure created,
  - data converted to a common, open, modern digital format, and
  - updated rescue progress document.



Stage 4A: Digitising paper-based data
This stage entails the following:
• creating an imaging inventory
• scanning or keying the data
• creating a master list of filenames and folders
• validating that scanned/keyed items correspond
• storing items in folder structure
• updating inventory and rescue progress document
Start
STAGE 4A: Digitising paper-based data (comprises scanning and/or keying of data
Data involved: Stream A1 – paper media
Create imaging inventory
Select digitisation method/s
Keying of data
Key the data onto electronic spreadsheet
Scanning
Image each paper record using scanner or camera
Create master list of filenames and folders
Quality control of keyed media
Validate that number of files correspond
Create folders
Store s/sheets in defined structure and location
Update inventory and data rescue document once stage is completed
Proceed to STAGE 5: Describing the data
Create master list of filenames and folders
Quality control of imaged media
Validate that number of files correspond
Create folders
Store scanned items in defined structure and location
Update inventory and data rescue document once stage is completed
Proceed to STAGE 5: Describing the data

Figure 6.13: Stage 4A: Digitising paper-based data

STAGE 4B: Converting early digital data

This stage entails the following:

- creating a conversion inventory
- converting files to a modern format
- quality control and validation of converted media
- creating folders; storing files in predefined structure
- updating all relevant documents

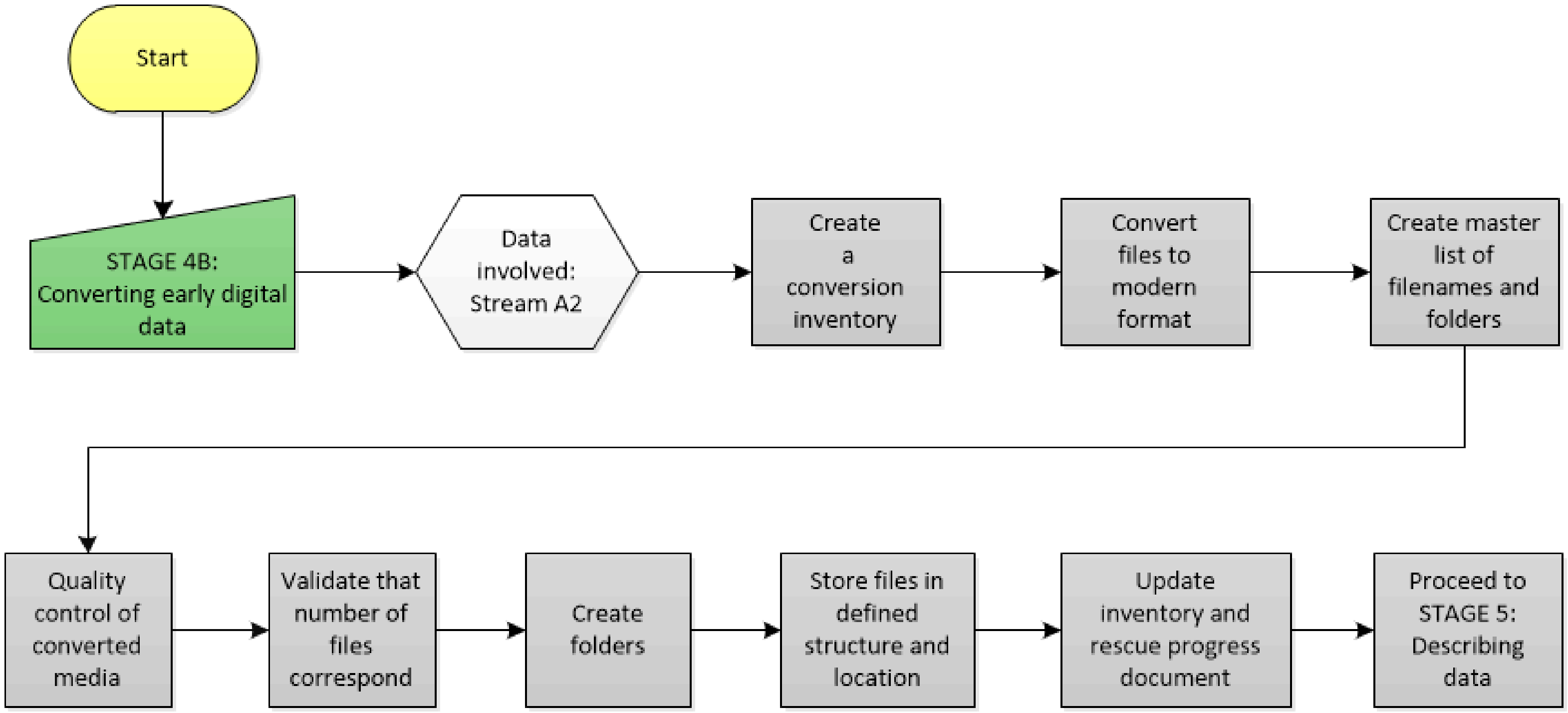

Start
STAGE 4B: Converting early digital data
Data involved: Stream A2
Create a conversion inventory
Convert files to modern format
Create master list of filenames and folders
Quality control of converted media
Validate that number of files correspond
Create folders
Store files in defined structure and location
Update inventory and rescue progress document
Proceed to STAGE 5: Describing data

Figure 6.14: Stage 4B: Converting early digital data

STAGE 4C: Converting modern digital data

This stage entails the following:

- creating a conversion inventory
- creating file and folder structure
- converting data to the required format
- quality controlling and validating the data
- renaming files if required
- storing converted data in structure
- updating relevant documents

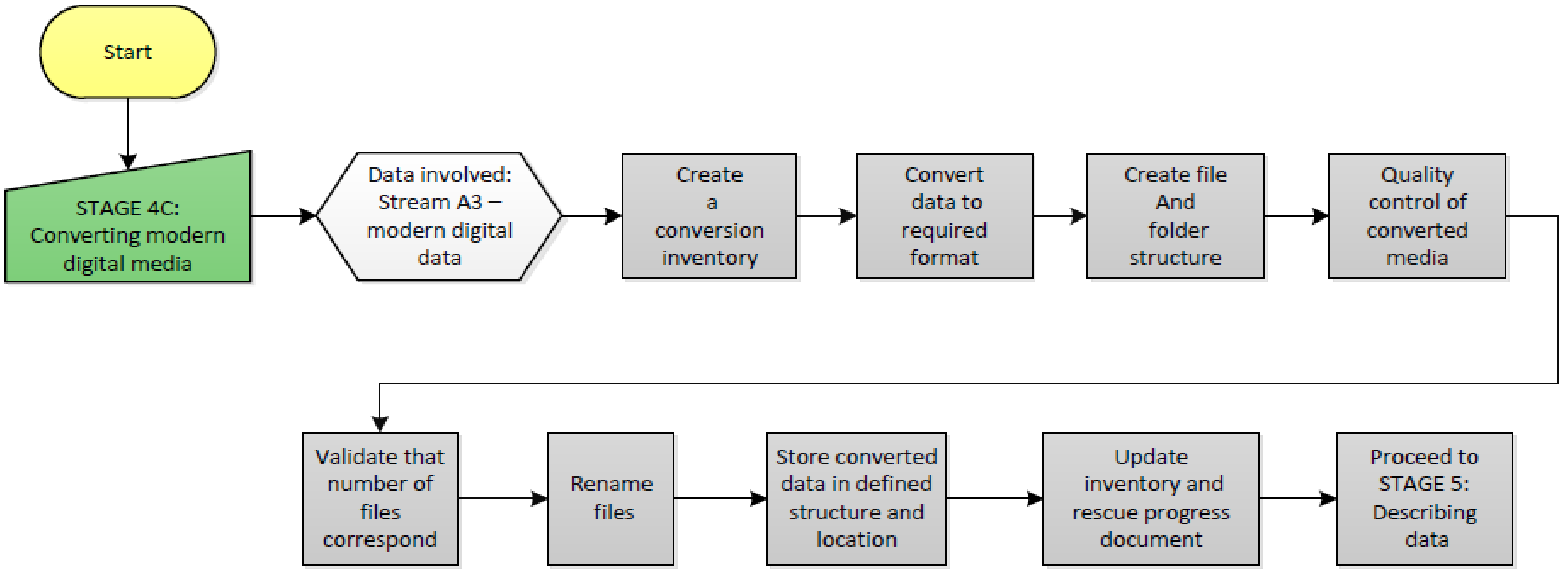

Start
STAGE 4C: Converting modern digital media
Data involved: Stream A3 – modern digital data
Create a conversion inventory
Convert data to required format
Create file And folder structure
Quality control of converted media
Validate that number of files correspond
Rename files
Store converted data in defined structure and location
Update inventory and rescue progress document
Proceed to STAGE 5: Describing data

Figure 6.15: Stage 4C: Converting modern digital data

Stage 5: Describing the data

- This stage is depicted in Figure 6.16 and expands on the information provided in Figure 6.4 and its linked description.
- This stage involves describing the data by creating metadata and data documentation linked to the data, thereby ensuring that the data can be understood and used by future users.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow, with additional clarifying details below.
- Data documentation can be described as documentation accompanying the data enabling a secondary user to make sense of and reuse the data.
- The stage requires the creation of a document containing data documentation, and can include aspects such as
  - purpose of data collection,
  - data collection procedures,
  - structure and organisation of data files,
  - sampling,
  - time and timing of data collection,
  - data validation and quality assurance, and
  - type of data manipulation performed.
- In addition, metadata will also be created. The metadata standard selected will depend on many factors, such as the specific data discipline, and the metadata required by the selected repositories (see Stage 6).
- It is suggested that a metadata template be created and that the required fields be completed by a person familiar with the research discipline.
- Both the metadata document and data documentation will be stored in the predetermined file structure also containing the converted data.
- Guidance document for this stage is as follows:
  - a document titled ‘Guidance on the use of metadata’ (see Appendix 19).
- Recommended outputs to be created during this stage include:
  - a metadata template with the required fields,
  - a completed metadata document,
  - a document containing data documentation, and
  - updated rescue progress document.
- The process then proceeds to the next stage, entailing the sharing of data.

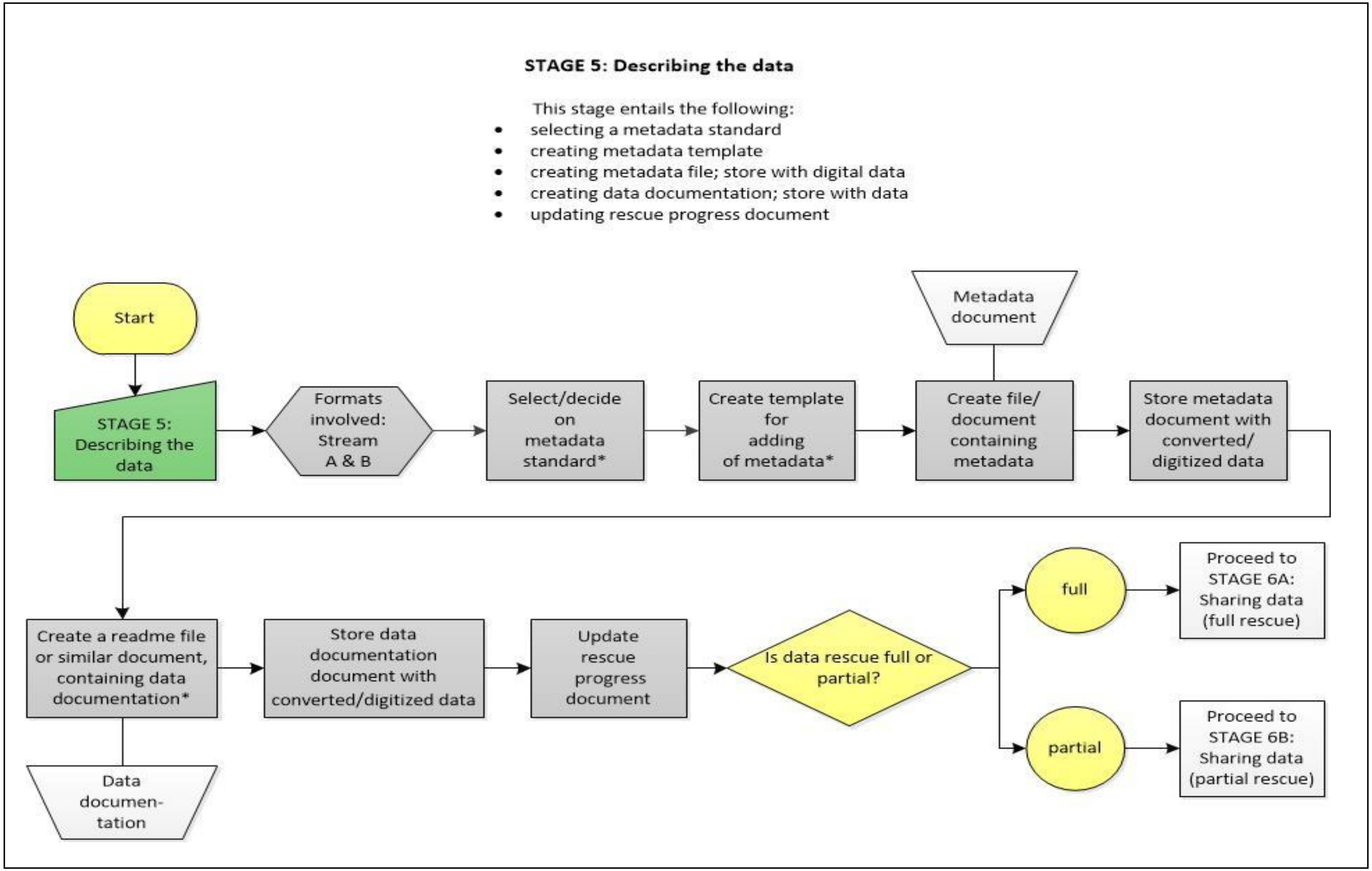

STAGE 5: Describing the data
This stage entails the following:
• selecting a metadata standard
• creating metadata template
• creating metadata file; store with digital data
• creating data documentation; store with data
• updating rescue progress document
Start
STAGE 5: Describing the data
Formats involved: Stream A & B
Select/decide on metadata standard*
Create template for adding of metadata*
Metadata document
Create file/ document containing metadata
Store metadata document with converted/ digitized data
Create a readme file or similar document, containing data documentation*
Data documen-tation
Store data documentation document with converted/digitized data
Update rescue progress document
Is data rescue full or partial?
full
partial
Proceed to STAGE 6A: Sharing data (full rescue)
Proceed to STAGE 6B: Sharing data (partial rescue)

Figure 6.16: Stage 5: Describing the data

Stage 6A: Sharing the data (full rescue)

- The stage is depicted in Figure 6.17 and expands on the information provided in Figure 6.5 and its linked description.
- This stage involves sharing the digitised or converted data and its metadata and data documentation. The sharing of data and its data documentation and metadata involves the upload of data, data documentation and metadata to data repositories.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow, with additional clarifying details below.
- Guidance document for this stage is as follows:
    - a document titled ‘Guidance on use of data repositories’ (see Appendix 20).
- Recommended outputs to be created during this stage include:
    - the uploaded data in selected data repositories, accompanied by guiding data documentation and required metadata, and
    - updated rescue progress document.
- The process then proceeds to the next stage entailing the long-term preservation of data.

Stage 6B: Sharing the data (partial rescue)

- The stage is depicted in Figure 6.18 and expands on the information provided in Figure 6.5 and its linked description.
- This stage involves the sharing of metadata data that had not been digitised or converted. The sharing of metadata involves the upload of the metadata to selected data repositories.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (partial rescue), with additional clarifying details below.
- Guidance document for this stage is as follows:
    - a document titled ‘Guidance on use of data repositories’ (see Appendix 20).
- Recommended outputs to be created during this stage include:
    - the uploaded metadata to selected repositories, and
    - updated data rescue progress document.
- The process then proceeds to Stage 8B, entailing the workflow steps forming part of data rescue project closure for partially rescued data.

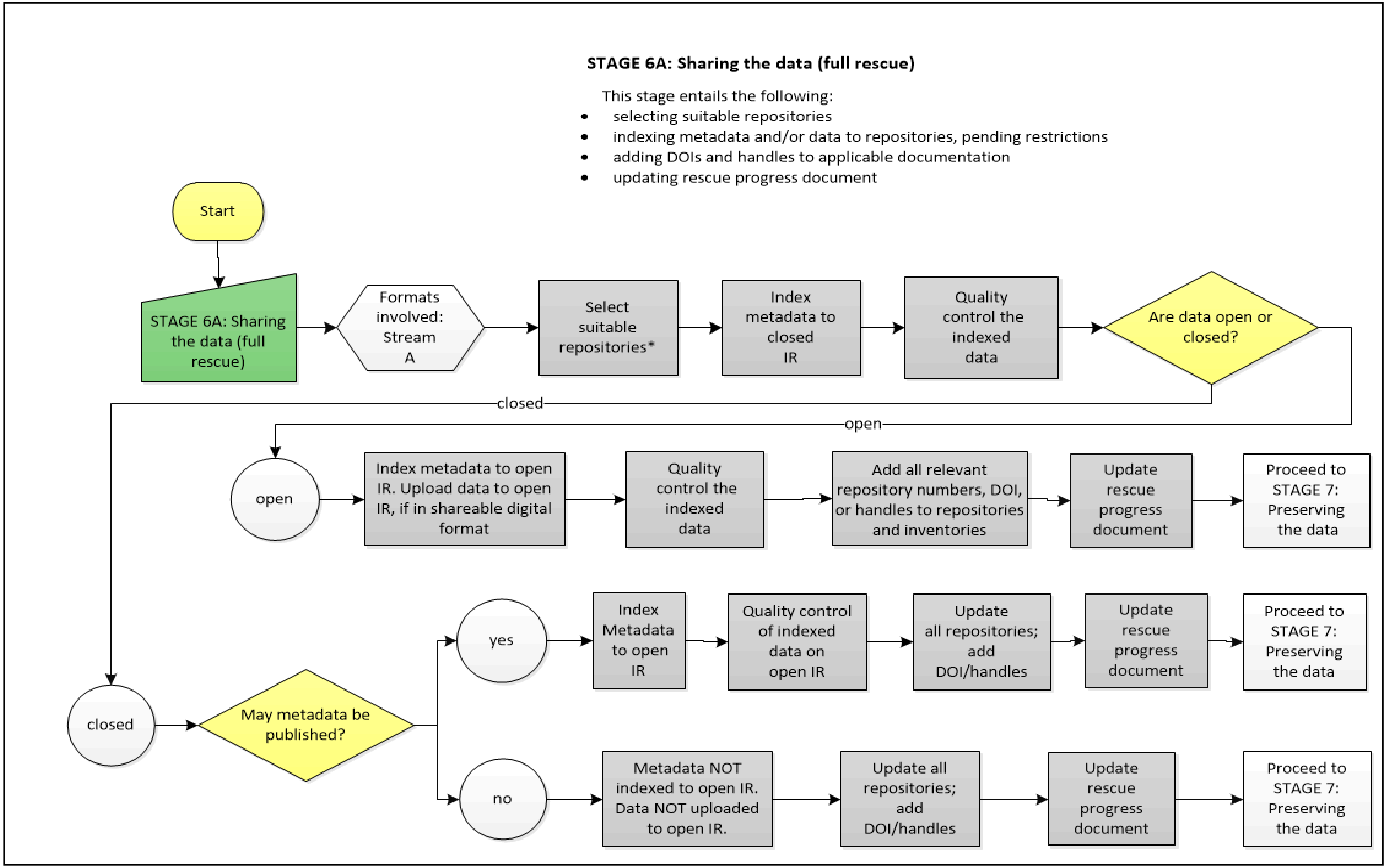
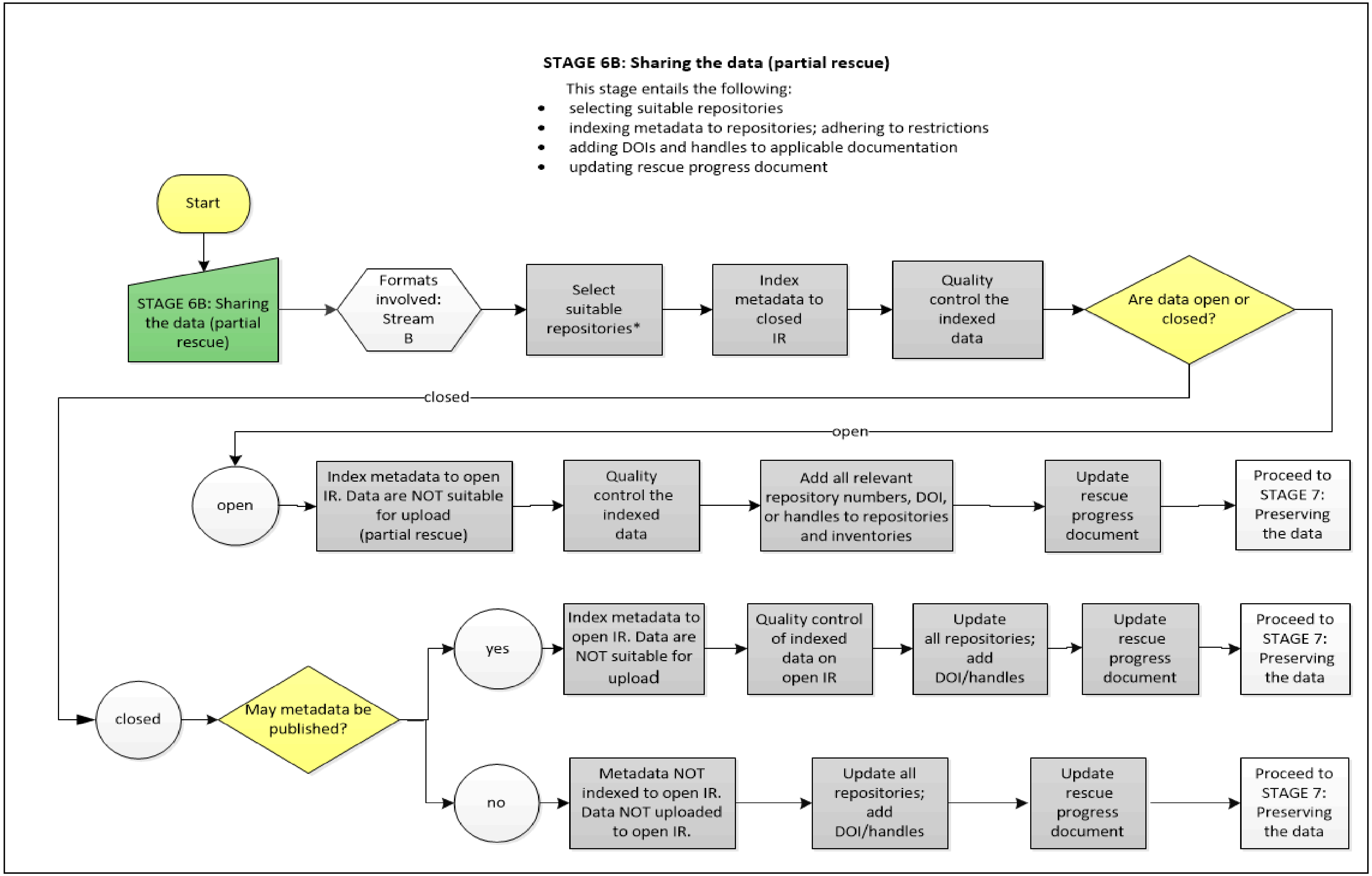

STAGE 6A: Sharing the data (full rescue)
This stage entails the following:
• selecting suitable repositories
• indexing metadata and/or data to repositories, pending restrictions
• adding DOIs and handles to applicable documentation
• updating rescue progress document
Start
STAGE 6A: Sharing the data (full rescue)
Formats involved: Stream A
Select suitable repositories*
Index metadata to closed IR
Quality control the indexed data
Are data open or closed?
closed
open
open
Index metadata to open IR. Upload data to open IR, if in shareable digital format
Quality control the indexed data
Add all relevant repository numbers, DOI, or handles to repositories and inventories
Update rescue progress document
Proceed to STAGE 7: Preserving the data
closed
May metadata be published?
yes
Index Metadata to open IR
Quality control of indexed data on open IR
Update all repositories; add DOI/handles
Update rescue progress document
Proceed to STAGE 7: Preserving the data
no
Metadata NOT indexed to open IR. Data NOT uploaded to open IR.
Update all repositories; add DOI/handles
Update rescue progress document
Proceed to STAGE 7: Preserving the data

Figure 6.17: Stage 6A: Sharing the data (full rescue)



STAGE 6B: Sharing the data (partial rescue)

This stage entails the following:

- selecting suitable repositories
- indexing metadata to repositories; adhering to restrictions
- adding DOIs and handles to applicable documentation
- updating rescue progress document

Start
STAGE 6B: Sharing the data (partial rescue)
Formats involved: Stream B
Select suitable repositories*
Index metadata to closed IR
Quality control the indexed data
Are data open or closed?
closed
open
open
Index metadata to open IR. Data are NOT suitable for upload (partial rescue)
Quality control the indexed data
Add all relevant repository numbers, DOI, or handles to repositories and inventories
Update rescue progress document
Proceed to STAGE 7: Preserving the data
closed
May metadata be published?
yes
Index metadata to open IR. Data are NOT suitable for upload
Quality control of indexed data on open IR
Update all repositories; add DOI/handles
Update rescue progress document
Proceed to STAGE 7: Preserving the data
no
Metadata NOT indexed to open IR. Data NOT uploaded to open IR.
Update all repositories; add DOI/handles
Update rescue progress document
Proceed to STAGE 7: Preserving the data

Figure 6.18: Stage 6B: Sharing the data (partial rescue)

Stage 7: Preserving the data (for the long term)

- The stage is depicted in Figure 6.19 and expands on the information provided in Figure 6.5 and its linked description.
- This stage involves ensuring the data are in a preservation format and thereby preserved in the long term.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow, with additional clarifying details below.
- Recommended outputs to be created during this stage include:
  - the data (in preservation format) uploaded to a secure long-term storage location, and
  - updated data rescue progress document.
- The process then proceeds to Stage 8A, entailing the workflow steps forming part of data rescue project closure for fully rescued data.

Stage 8A: Project closure (full rescue)

- The stage is depicted in Figure 6.20 and expands on the information provided in Figure 6.5 and its linked description.
- This stage involves the final activities of the rescue process and entails ensuring that all loose ends are tied up.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (full rescue), with additional details clarified below.
- Awareness for full rescued data must be created. Activities such as institutional intranet articles, scholarly articles in dedicated data journals or digital curation journals, conference presentations, and even data rescue workshops form part of awareness creation.
- It is important that all parties who formed part of the rescue project be thanked via a letter of thanks.
- The final project report must be completed and will be distributed to all members. In addition, institutional awareness of the published report is vital.
- The final report should include an indication of the learnings accompanying the rescue project, and rescue areas where amendments for future projects should be made.
- Recommended outputs to be created during this stage include:
  - an article/presentation related to the rescued data,
  - an article/presentation detailing the rescue project,

STAGE 7: Preserving the data

This stage entails the following:

- identifying a stable long term storage location
- creating folders for resources
- transfering the data to the stable location
- monitoring data regularly
- updating rescue progress document

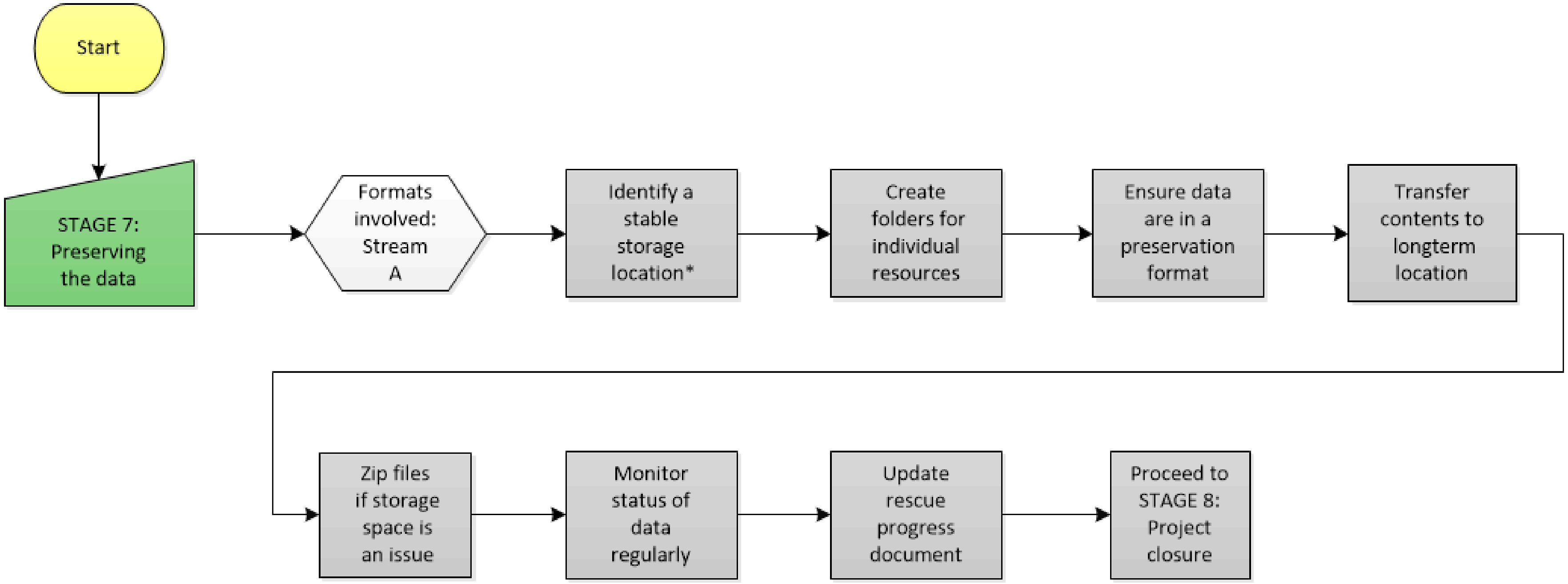

Start
STAGE 7: Preserving the data
Formats involved: Stream A
Identify a stable storage location*
Create folders for individual resources
Ensure data are in a preservation format
Transfer contents to longterm location
Zip files if storage space is an issue
Monitor status of data regularly
Update rescue progress document
Proceed to STAGE 8: Project closure

Figure 6.19: Stage 7: Preserving the data (for the long term)

  - a letter of thanks to project participants,
  - the final completed data rescue project report, and
  - updated data rescue progress report.
- The data rescue project is thereby completed. Certain activities are ongoing activities, such as regular monitoring of archived data, and recording of data citation metrics.

Stage 8B: Project closure (partial rescue)

- The stage is depicted in Figure 6.21 and expands on the information provided in Figure 6.5 and its linked description.
- This stage involves the final activities of the rescue process and entails ensuring that all loose ends are tied up.
- The description of this stage mirrors the description of the stage as provided for the data rescue compact workflow (partial rescue), with additional clarifying details below.
- Awareness for the partial rescue project should be created. Activities such as institutional intranet articles, scholarly articles in dedicated data journals or digital curation journals, conference presentations, and even data rescue workshops form part of awareness creation.
- Awareness creation of the partial rescue project can also include requests for collaboration to fully rescue the partially rescued data.
- Recommended outputs to be created during this stage include:
  - an article/presentation detailing the rescue project,
  - a letter of thanks to project participants, and
  - the final completed data rescue project report.
- It is important that all parties who formed part of the rescue project be thanked via a letter of thanks.
- The final project report must be completed and will be distributed to all members. In addition, institutional awareness of the published report is vital.
- An activity forming part of the drafting of the final report includes indicating the learnings gained during the project and stating areas where amendments to future rescue projects should be made.
- The data rescue project is thereby completed. Ongoing activities related to the project include an annual evaluation of institutional resources to determine whether the remaining data rescue activities can be performed.

STAGE 8A: Project closure (full rescue)

This stage entails the following:

- creating awareness of, and marketing the rescued data and the rescue project (via article, paper, presentation)
- thanking all involved parties
- creating and distributing the final project report

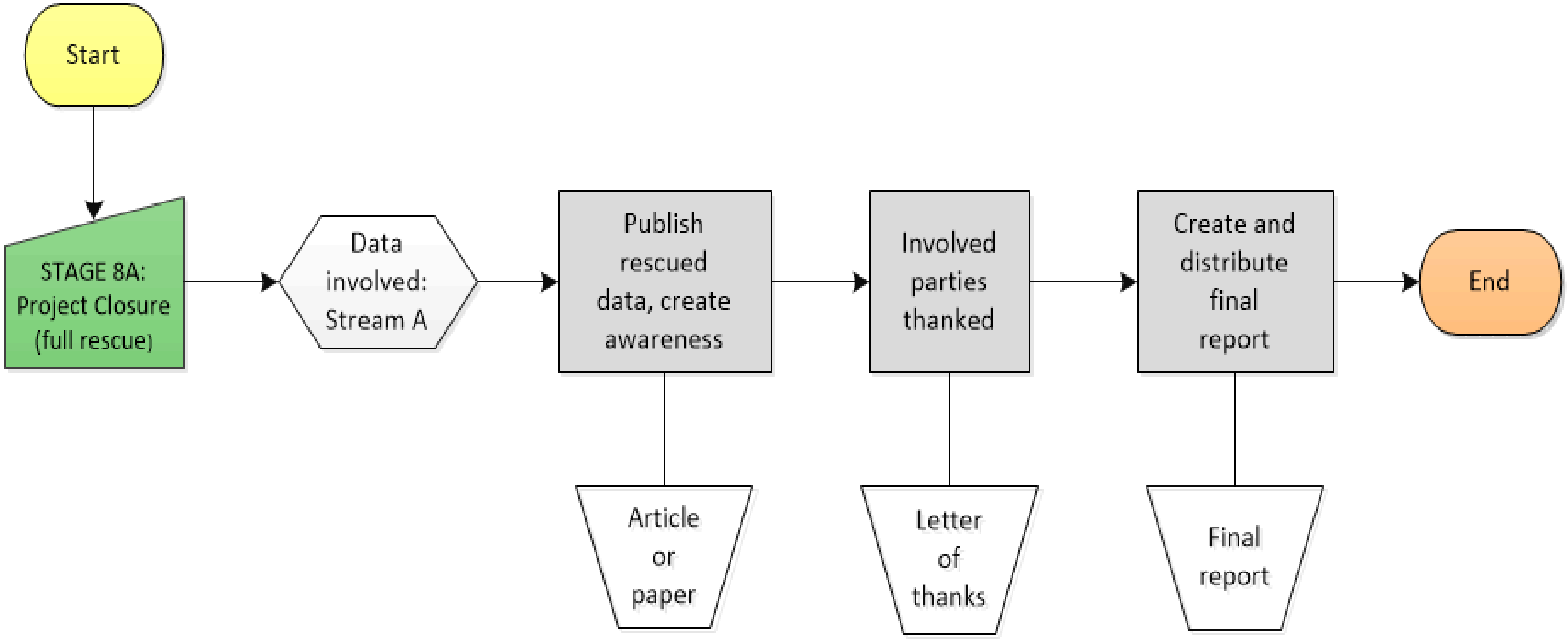

Start
STAGE 8A: Project Closure (full rescue)
Data involved: Stream A
Publish rescued data, create awareness
Involved parties thanked
Create and distribute final report
End
Article or paper
Letter of thanks
Final report

Figure 6.20: Stage 8A: Project closure (full rescue)

STAGE 8B: Project closure (partial rescue)

This stage entails the following:

- creating awareness of, and marketing the partially-rescued data and the rescue project (via article, paper, presentation)
- thanking all involved parties
- creating and distributing the final project report
- setting a reminder for annual re-evaluation of data

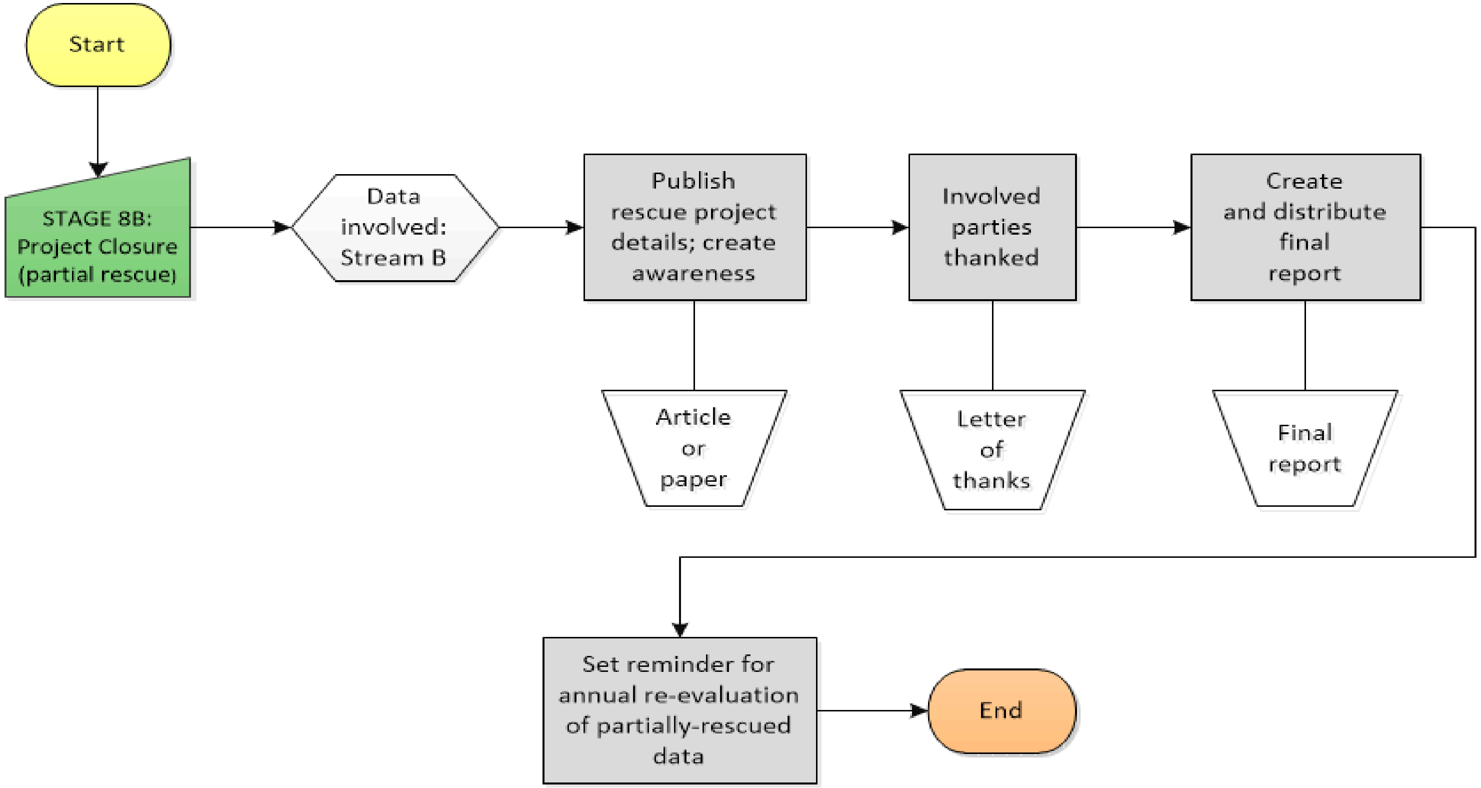

Start
STAGE 8B: Project Closure (partial rescue)
Data involved: Stream B
Publish rescue project details; create awareness
Article or paper
Involved parties thanked
Letter of thanks
Create and distribute final report
Final report
Set reminder for annual re-evaluation of partially-rescued data
End

Figure 6.21: Stage 8B: Project closure (partial rescue)

Summary: This section provided an answer to the research question concerned with the ways in which theory and practice could be formalised in a model for a data rescue workflow. In answering this question, this researcher has presented a model referred to as a 'recommended Data Rescue Workflow Model' consisting of three main layers. These layers, comprising a data rescue summary diagram, a data rescue compact workflow, and detailed stages of the recommended Data Rescue Workflow Model, have been described and clarified within this section.

6.2.1.8 Inclusion of data rescue topics in the LIS curricula

Several suggestions regarding this topic can be made; the topic is also addressed in Section 6.4.12: Amend and adapt the LIS curriculum to include data rescue.

With the crucial role of library and information services experts already indicated in the recommended model, a next step requires that library and information services staff be trained to be able to execute the required data rescue tasks. While some of the tasks will show similarities with tasks already being performed by library and information services professionals (e.g., uploading data to a repository, ensuring metadata are complete), many of the roles indicated in the data rescue workflow entail the library and information services professional understanding the entire data rescue process. In addition, certain activities will require either on-the-job training (such as discussed in Section 6.4.8), or training received via postgraduate studies or a course at a tertiary institute.

The following suggestions are made regarding the inclusion of data rescue topics in the LIS curriculum:

- It is suggested that the topics of data at risk and data rescue form part of the LIS curriculum.
- It is further suggested that the data rescue training be done at a level where students would have already dealt with research data or have been exposed to a module related to research data management.
- In the South African context, this would take place during honours- and masters-level LIS studies.
- It is not advisable to implement data rescue as an LIS topic earlier in the curriculum, as experience with and an understanding of research data is crucial.
- Adding on to the above points: it is suggested that students should already have come across and experienced concepts such as file formats, secure data storage, data licensing, metadata and data documentation, and data repositories before attending lectures on data at risk and data rescue.

- It is suggested that learners in data rescue ideally have practical data experience, and not only having attended lectures on data management. In the South African context, the inclusion of a mini-dissertation at honours level ensures practical data experience.
- During this study, this researcher presented lectures on data at risk and data rescue on three separate occasions. More details are provided in Section 6.4.12 (under the heading 'Amend and adapt the LIS curriculum to include data rescue'), and form part of the recommendations section of this chapter.
- At the very least, it is estimated that the lectures contain the following:
    - Lecture 1: Introduction to data at risk: data at risk examples, defining data rescue, importance of data rescue, data rescue projects and success stories.
    - Lecture 2: Explanation of the Data Rescue Workflow Model, including an explanation of guidelines, templates, and data rescue outcomes.

With virtual presentations being the norm during the COVID-19 pandemic, and the range of technical problems frequently affecting virtual live presentations, it might prove prudent for presenters to ensure that different versions of the presentation are available to students. Apart from a live presentation, alternative data rescue teaching tools could include a presentation uploaded to SlideShare, a slideshow presentation with audio uploaded to YouTube, or a presentation with handout notes (PDF format) being available to students.

A final suggestion pertains to the administering of a test to evaluate learners' grasp of the topic. Recurring and identical gaps in data rescue insights among most learners will also indicate that the topic presentation requires modification and adaptation.

The next section of this chapter addresses the study's main research question.

6.2.2 Main research question, findings and implications

As stated in Chapter 1, the main research question of this study is as follows:

What are the roles and responsibilities of the research library within a comprehensive workflow for data rescue?

This question's answer is preceded by a segment unpacking the question, and a brief discussion on the varying extent of library and information services involvement in data rescue. These two sections are followed by addressing the study's main research question by means of textual descriptions and clarifications and a summary diagram.

6.2.2.1 Unpacking the research question

The main research question is answered by considering, analysing and evaluating the library and information services-related findings emanating from the following activities:

- information gathered during the study's literature review,
- information gathered during the study's literature analysis,
- data collected via the online questionnaire,
- data collected via the one-on-one interviews,
- data emanating from data rescue model feedback received,
- data collected via the mini focus group session, and
- potential data management-related roles and involvement of the research library at research institutes.

In addition, the multi-faceted nature of concepts within the question itself was considered. The terms 'roles and responsibilities', the diverse nature of 'research libraries', and the varied character of 'data rescue' and its accompanying 'workflows' would therefore require unpacking. Explanatory details are supplied below.

Roles and responsibilities: In the context of this study and the workflow created, 'roles' refer to the position within the data rescue team or project, while 'responsibilities' point to the tasks and duties of the role. A hypothetical example would be a research library employee assigned the role of repository professional during a research project. Responsibilities would include studying the intended repository's requirements, assessing the metadata of the dataset, adding metadata as required, assessing the openness of the data, uploading data and its metadata to a repository, sharing the DOI or handle with relevant parties, and ensuring the completed tasks are recorded and updated in the project plan.

Research library: In the context of this study, the term 'research library' refers to a library providing services to a research institute; service provision is undertaken by the library staff employed by said institute. However, many of the roles and responsibilities discussed in the remainder of this section will be applicable to libraries at tertiary institutes of learning, and special libraries. The involvement of public libraries for certain responsibilities, especially when the rescue projects will be making use of citizen scientists, is also a scenario.

In addressing the study's main research question, this research considered library and information services positions commonly forming part of research libraries, and skills associated with those positions. The included research library positions together with their potential responsibility areas related to data rescue are listed below.

- *Library and Information Services portfolio manager/library manager/library director:*
    - approves envisaged data rescue projects managed by the research library, and involvement of Library and Information Services staff members,
    - identifies risk factors linked to the envisaged data rescue projects, and
    - approves and manages financial matters linked to data rescue projects.
- *Records manager:*
    - performs in a managerial position and oversees the data librarian, repository professional, archives technician, and intern,
    - creates awareness around records management issues, services and tools,
    - creates online training materials related to records management,
    - is responsible for institutional records management training, and
    - is co-author of procedures related to records management.

It should be noted that 'records management' is used here as an umbrella term, and that 'data management' falls under this section.

- *Data librarian/data curator/data manager:*
    - is responsible for institutional data management training,
    - creates awareness around data management issues, services and tools,
    - creates online training materials related to data management,
    - performs indexing of data and metadata onto closed and open institutional repositories,
    - is co-author of institutional data management procedure, and
    - advises on metadata standards.
- *Repository professional:*
    - indexes a range of institutional records onto closed and open institutional repositories.
- *Archive technician:*
    - is responsible for storage and preservation of paper-based records in archive,
    - creates and stores archive inventories,
    - organises cleaning of archive,
    - communicates with employees or external parties who have requested archival documents,

    - communicates with employees regarding the deposit of records in archive, and
    - indexes archival records onto institutional closed repository.
- *Digitisation clerk:*
    - is responsible for digitisation activities.
- *Information scientist/Information specialist:*
    - assists the data librarian with training of researchers, and
    - provides guidance to institutional research library employees regarding discipline-specific issues.
- *Intern (LIS graduate):*
    - undergoes on-job training in most tasks listed above.

Data rescue skills: In viewing the suggested library and information services roles and responsibilities linked to the recommended Data Rescue Workflow Model, it is crucial to keep in mind the available library and information services employees, positions at an institute, data rescue and data management skills and experience, nature of the research data in need of rescue, and institutional policies. It might prove necessary to train or upskill existing library and information services professionals prior to a data rescue project; Sections 6.4.8 and 6.4.12 discuss this issue in more detail. It is also vital that positions be needs-driven, and that the availability of current staff should not be a hindrance to data rescue.

Data rescue: As is evident from several earlier references to the state of global data rescue, the concept of 'data rescue' is one that is accompanied with diversity in complexity levels, tasks involved, scope of rescue, formats involved, and required data rescue outcomes. In addition, data rescue project features are prone to be influenced by available resources such as manpower, skills, systems, equipment and funding. These aspects contribute towards the diversity that is 'data rescue' and influence the way the research library can be involved in or contribute towards data rescue.

Comprehensive workflow: This concept refers to the workflow described in Section 6.2.1.7, and covers the summary diagram, the compact workflow, and each of the individual data rescue stages described in the section.

6.2.2.2 Recommended library and information services roles and responsibilities during data rescue

Findings included in this section refer to the data rescue activities forming part of the recommended workflow described in Section 6.2.1.7. The difference between Section 6.2.1.7 and the current section pertains to the involvement of the research library and the participation of library and information services staff. In other words, the current section reports on findings related to the combination of the following two main concepts: the LIS sector, and the recommended data rescue workflow/model. Seen in Boolean terms, this section reports on the findings emanating from the 'data rescue model' AND 'LIS' equation, while Section 6.2.1.7 did not specify which sectors of an institute, or which professional positions, would be suited to the execution of the various rescue tasks within the recommended model.

To enable an indication of findings related to specific tasks and activities, this section of the chapter will discuss library and information services roles and responsibilities in the following order:

- library and information services involvement and the data rescue summary diagram,
- library and information services involvement and Stage 1: Data rescue preparatory stage,
- library and information services involvement and Stage 2: Planning of data rescue,
- library and information services involvement and Stage 3: Storing and preserving data,
- library and information services involvement and Stage 4: Digitising or converting data,
- library and information services involvement and Stage 5: Describing the data,
- library and information services involvement and Stage 6: Sharing the data,
- library and information services involvement and Stage 7: Preserving the data (for the long term),
- library and information services involvement and Stage 8: Project closure, and
- library and information services involvement and related data rescue roles and responsibilities.

Where applicable, brief mentions will also be made of the potential involvement of an institute's non-research library sector during the stage being discussed. Examples of such sectors include an institute's SET base, the ICT division, and the Communications department.

This section – answering the study's main research question – concludes with a tabular summary of library and information services roles and responsibilities during data rescue projects and activities.

Data rescue summary diagram: Library and Information Services involvement: Figure 6.22 indicates the potential involvement of the research library in data rescue, with the blue-coloured blocks indicating rescue stages of research library involvement, and the orange-coloured blocks indicating activities of research library involvement. What is clear from Figure 6.22 is that the library and information services sector has the potential to be involved and contribute to all rescue stages.

The following points should be highlighted and explained:

- Findings regarding involvement of the research library reveal the varying levels of participation and responsibility. It would therefore be possible for the research library sector to be involved in all stages, in one stage only, or in any number of stages. The varying levels of library services involvement in data rescue (stage-wise) can also be seen to be influenced by factors such as skills and expertise of the research library employees, and available resources, including time, equipment and manpower. Data rescue involvement can therefore vary from participation in a single rescue task (e.g., repository upload) to multi-stage participation to participation in all rescue stages. However, lack of necessary data rescue skills should not be viewed as a permanent obstacle to data rescue, as library professionals can be trained, reskilled, eventually be involved in rescue ventures, and be more valuable to the organisation in areas such as the Fourth Industrial Revolution research sector.
- Just as the role of the research library may vary between involvement in all stages to involvement in one stage only, the type of involvement may also vary. A rescue project might have a research library professional managing the entire project, or be assigned as manager for a certain stage, or not be involved in a managerial capacity at all but be responsible for certain rescue tasks while reporting to a data rescue project manager emanating from the SET base. Participating as data rescue project manager would be accompanied by tasks such as the drafting of a project plan, hosting and planning project meetings, and the drafting of a final data rescue report. Appointment and suitability as project manager would depend on a range of factors, including the data discipline involved, the formats concerned and resources available. Skills and experience would also determine managerial involvement.
- The potential and feasibility of data rescue project management by the research library is bound to fluctuate between various data rescue projects. The determining factors are listed in the previous bullet.
- Minor involvement of the research library is anticipated to be more prevalent in stages where advanced technological skills are required. As stated in a previous bullet, the option of reskilling and training must be considered in order to increase the value of the library to the relevant research institute, and to the broader research environment.
- Major involvement of the research library is anticipated to be more prevalent in stages where conventional management activities form part of the stage. In addition, activities entailing supervision, training, and creation of data rescue documentation are expected to form part of the library’s responsibilities.

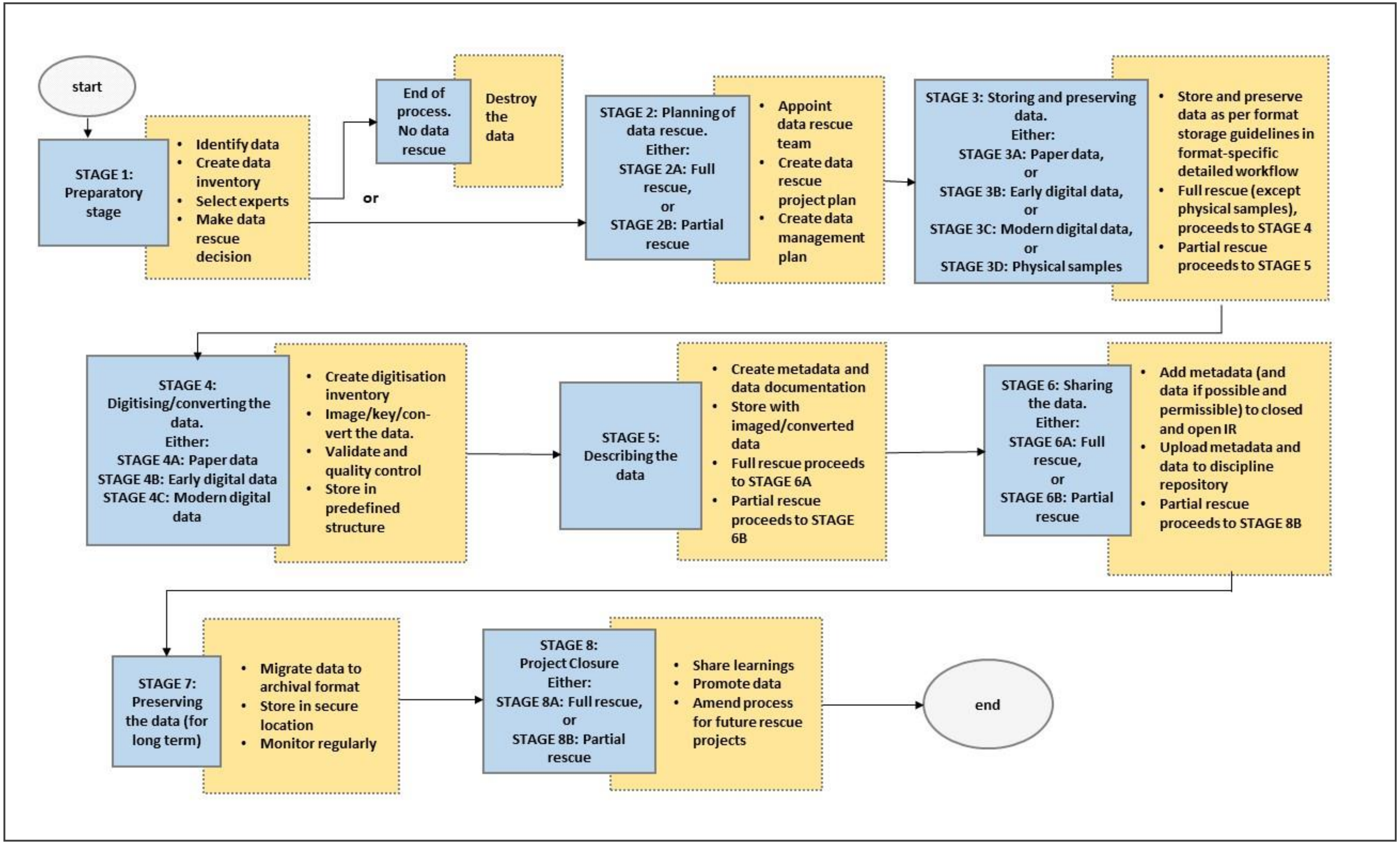

start
STAGE 1: Preparatory stage
• Identify data
• Create data inventory
• Select experts
• Make data rescue decision
or
End of process. No data rescue
Destroy the data
STAGE 2: Planning of data rescue. Either: STAGE 2A: Full rescue, or STAGE 2B: Partial rescue
• Appoint data rescue team
• Create data rescue project plan
• Create data management plan
STAGE 3: Storing and preserving data. Either: STAGE 3A: Paper data, or STAGE 3B: Early digital data, or STAGE 3C: Modern digital data, or STAGE 3D: Physical samples
• Store and preserve data as per format storage guidelines in format-specific detailed workflow
• Full rescue (except physical samples), proceeds to STAGE 4
• Partial rescue proceeds to STAGE 5
STAGE 4: Digitising/converting the data. Either: STAGE 4A: Paper data STAGE 4B: Early digital data STAGE 4C: Modern digital data
• Create digitisation inventory
• Image/key/con-vert the data.
• Validate and quality control
• Store in predefined structure
STAGE 5: Describing the data
• Create metadata and data documentation
• Store with imaged/converted data
• Full rescue proceeds to STAGE 6A
• Partial rescue proceeds to STAGE 6B
STAGE 6: Sharing the data. Either: STAGE 6A: Full rescue, or STAGE 6B: Partial rescue
• Add metadata (and data if possible and permissible) to closed and open IR
• Upload metadata and data to discipline repository
• Partial rescue proceeds to STAGE 8B
STAGE 7: Preserving the data (for long term)
• Migrate data to archival format
• Store in secure location
• Monitor regularly
STAGE 8: Project Closure Either: STAGE 8A: Full rescue, or STAGE 8B: Partial rescue
• Share learnings
• Promote data
• Amend process for future rescue projects
end

Figure 6.22: Data rescue and the involvement of the research library

- Data management currently forms part of the services provided by many research and academic libraries, and the library of the selected institute. The likelihood of a library professional, currently involved with data management who could therefore be involved with data rescue is hence a realistic possibility at many research institutes. The participation of such an incumbent to act as project manager of certain data rescue projects should be anticipated.
- In smaller data rescue projects, the simultaneous involvement as project manager and stage-specific research library expert should be expected.
- As was explained in the bullet point mentioning stage involvement, the level and type of library involvement would depend on several factors, and can include aspects such as:
    - data rescue experience of participants (SET base and library sector),
    - subject knowledge of library and information services sector,
    - data format knowledge of library and information services sector,
    - knowledge of equipment and readers involved,
    - available manpower, and
    - institutional policy and procedures.
- A potential scenario entails the research library being involved in all eight stages of the model. The participation of the research library in each of the stages is especially likely in instances where certain disciplines are no longer part of the institute, or when a subject specialist responsible for assessing data and creation of relevant data documentation cannot be traced.

In short: The involvement of the research library in data rescue can vary between participation in singular underlying tasks, to stage-specific involvement, to involvement at managerial level, and to a combination of these participation spheres. Data rescue is not a static responsibility, and cannot be allocated without examining the data, resources available and expected project outcomes. Training and upskilling should always be considered.

The next section concentrates on the potential for library and information services involvement during each stage of the recommended Data Rescue Workflow Model.

Stage specific descriptions: The next eight headings discuss and portray library and information services involvement within each of the recommended data rescue stages. Brief mentions are also made of the potential involvement of other institutional sectors.

Stage 1: Data rescue preparatory stage (refer to Figure 6.6)

The research library sector may potentially be involved in the following tasks:

- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- stage-specific training of data rescue participants:
  - the data librarian is a suitable incumbent for this task,
- identifying data that may require rescue:
  - it is anticipated that the data librarian, records manager, or a research cluster's information scientist will be approached and informed after data at risk is located,
- organising a venue/location for the preliminary storage of identified data:
  - the archives technician is a suitable incumbent for assigning a section of the library for the temporary storage of identified data (in paper-based format, or early digital format),
- creating master inventory, and adding the details of data to the master inventory:
  - it is anticipated that the data librarian will be the most suitable incumbent for this activity,
- establishing a group of experts to evaluate the data:
  - the records manager and/or the data librarian will assemble a group of experts (including themselves and selected SET-base employees familiar with the data's research discipline) who would be equipped to assess the identified data,
- evaluating resources and infrastructure:
  - the records manager and/or the data librarian, in consultation with the ICT division, evaluate available resources and infrastructure at the institute, to determine the nature of data rescue to be implemented,
- organising the destruction of data:
  - the records manager will liaise with Waste Services (for non-digital data) or ICT (electronic data) regarding the destruction of unwanted data,
- deciding whether full or partial rescue will be implemented:
  - forming part of the previous bullet point, the records manager and/or the data librarian decide whether full or partial data rescue will be implemented,
- terming the data as Stream A or Stream B:
  - all library and information services parties involved in the project are informed of this decision and will use the relevant terminology during the rescue project, and

- updating the master inventory:
  - the creator of the master inventory (as per a previous bullet) will execute this task.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with identifying data at risk, form part of the stage's group of experts, take primary responsibility for evaluating the data, and assist in assessing institutional resources and infrastructure.

During this stage, ICT services will be consulted to obtain information regarding the institute's data rescue resources and infrastructure. ICT will also be involved in the destruction of electronic data.

Waste Services and the institute's SHEQ division will be involved in the appropriate disposal of non-digital data that are no longer regarded as valuable.

A summary of the potential roles and responsibilities of the research library based on the current library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of research library involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 2: Planning for data rescue (refer to Figure 6.7 and Figure 6.8)

Planning of data rescue (full rescue)

The library and information services sector may potentially be involved in the following tasks:

- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- stage-specific training of data rescue participants:
  - the data librarian is a suitable incumbent for this task,
- creating an electronic project folder and contributing to folder contents:
  - the records manager or the data librarian are suitable incumbents for this task,
- be appointed as a data rescue project manager or data rescue project team member:
  - the records manager or the data librarian are suitable incumbents for the position of project manager,
  - team members from the library and information services sector can also include an information scientist, an archives technician, and an intern,
- appointing data rescue project team members or assisting with the selection of team members:

  - the records manager or the data librarian are suitable incumbents for the position of project manager, and for recruiting and selecting project team members,
- creating a data rescue project plan and sharing the plan with team members:
  - the records manager or the data librarian are suitable incumbents for the position of project manager and its accompanying duties and tasks, and
- creating a data management plan or assisting with the creation of the plan
  - the records manager or the data librarian are suitable incumbents for the position of project manager and its accompanying duties and tasks,
  - the data librarian to assist with completion of the data management plan if not appointed as project manager.

It is anticipated that researchers or persons with disciplinary background/knowledge (e.g., ex-researcher or students) be involved with the completion of certain sections of the data management plan.

During this stage, ICT services will be consulted to obtain information regarding the completion of certain sections of the data management plan and the project plan. The provision of secure and suitable data storage options, long-term preservation of the converted/digitised data, and conversion of proprietary modern data to a common, open, modern format are issues requiring ICT input.

A summary of the potential roles and responsibilities of the library and information services sector based on the current library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Planning of data rescue (partial rescue)

The library and information services sector may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
  - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- creating an electronic project folder and contributing to folder contents:
  - the records manager or the data librarian are suitable incumbents for this task,
- be appointed as a data rescue project manager or data rescue project team member:

  - the records manager or the data librarian are suitable incumbents for the position of project manager,
  - team members from the research library can also include an information scientist, an archives technician, and an intern,
- appointing data rescue project team members or assisting with the selection of team members:
  - the records manager or the data librarian are suitable incumbents for the position of project manager, and for recruiting and selecting project team members,
- creating a data rescue project plan and sharing the plan with team members:
  - the records manager or the data librarian are suitable incumbents for the position of project manager and its accompanying duties and tasks, and
- creating a data management plan or assisting with the creation of the plan:
  - the records manager or the data librarian are suitable incumbents for the position of project manager and its accompanying duties and tasks,
  - the data librarian to assist with completion of the data management plan if not appointed as project manager.

It is anticipated that researchers or persons with disciplinary background/knowledge (e.g., ex-researcher or students) be involved with the completion of certain sections of the data management plan.

During this stage, ICT services will be consulted to obtain information regarding the completion of certain sections of the data management plan and the project plan. The provision of secure and suitable data storage options, long-term preservation of the converted/digitised data, and conversion of proprietary modern data to a common, open, modern format are issues requiring ICT input.

A summary of the potential roles and responsibilities of the library and information services sector based on the current library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library’s involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 3: Storing and preserving data (refer to Figures 6.9, 6.10, 6.11 and 6.12)

Storing and preserving paper data (Stage 3A):

The library and information services sector may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- examining the paper-based data:
    - the data librarian, archives technician, or intern are suitable incumbents for this task,
- removing dust (from data, shelves, boxes, archive):
    - the data librarian, archives technician, or intern are suitable incumbents for this task,
- labelling boxes and shelves:
    - the data librarian, archives technician, or intern are suitable incumbents for this task,
- storing the paper-based data in boxes:
    - the data librarian, archives technician, or intern are suitable incumbents for this task, and
- monitoring progress and updating the rescue progress document:
    - to be performed by the research library expert assigned the responsibility during the planning stage and will be the data librarian or records manager.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved should the paper-based data require organising or sorting of pages.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Storing and preserving early digital data (Stage 3B)

The library and information services sector may potentially be involved in the following tasks, with the nature of involvement dependent on a research library professional being familiar with the relevant early digital format. Research library-relevant tasks are as follows:

- stage-specific training of data rescue participants:
  - the data librarian and format specialists are suitable incumbents for this task,
- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- examining the early digital data:
  - activity to be performed by any library professional familiar with the relevant early digital format,
- storing the data and its reader (if available) according to best practices:
  - activity to be performed by any library and information services professional familiar with the relevant early digital format,
- locating the metadata and recording its location:
  - the data librarian to provide guidance, and
- monitoring progress and updating the rescue progress document:
  - to be performed by the library and information services expert assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with knowledge of the relevant data format will be involved with several of the stage tasks. Furthermore, disciplinary background/knowledge will prove useful when locating metadata linked to the data.

During this stage, ICT services will be consulted to obtain information regarding conversion practices and procedures. ICT will also be approached should data storage prove to be an issue.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Storing and preserving modern digital data (Stage 3C)

This stage is concerned with the storage of data that are in a modern proprietary format; it is anticipated that the format would be understood and used by a select group of researchers and ICT staff members only. Expected involvement of the library and information services sector during this stage is minimal and is confined to the following tasks:

- stage-specific training of data rescue participants:

  - the data librarian and format specialists are suitable incumbents for this task,
- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- monitoring progress and updating the rescue progress document:
  - to be performed by the research library expert assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with knowledge of the proprietary data format be involved with several of the tasks forming part of this rescue stage. Furthermore, disciplinary background/knowledge will prove useful when locating metadata linked to the data.

During this stage, ICT services will be consulted to obtain information regarding conversion practices and procedures. ICT will also be approached should data storage prove to be an issue.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Storing and preserving physical samples (Stage 3D)

This stage is concerned with the storage of physical samples data or specimens. It is anticipated that the ideal storage conditions would be familiar to the research group in possession of the samples or specimens. Limited involvement of the research library during this rescue stage is expected. The research library may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
  - a format specialist, assisted by the data librarian, are suitable incumbents for this task,
- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- monitoring progress and updating the rescue progress document:
  - to be performed by the research library expert assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with the secure and discipline-specific storage of physical samples and specimens.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 4: Digitising/converting the data (refer to Figures 6.13, 6.14 and 6.15)

Digitising paper-based data (Stage 4A)

The library and information services sector may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- creating an imaging inventory:
    - the digitisation clerk, archives technician, intern, or the data librarian are suitable incumbents for this task,
- keying of data, creating master list of filenames and folders, creating folders, storing the spreadsheets in relevant folders, updating the inventory:
    - the digitisation clerk, archives technician, intern, or the data librarian are suitable incumbents for this task,
- validating data, quality control:
    - the data librarian or records manager are suitable incumbents for this task,
- scanning of images/data, creating a master list of filenames and folders, quality controlling of scans, validating the scans, creating folders, storing images in the created structure, updating the inventory:
    - the digitisation clerk, archives technician, intern, or the data librarian are suitable incumbents for this task,
- quality controlling of scans, validating the scans:
    - the data librarian or records manager are suitable incumbents for this task, and
- monitoring the progress and updating the rescue progress report:
    - to be performed by the research library expert assigned the responsibility during the planning stage.

It is not anticipated that the research sector will be involved with this stage of the rescue process, as digitisation and its linked activities are currently tasks performed by the institute's research library.

During this stage, ICT services will be consulted should there be concerns regarding sufficient storage space for digitised data.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Conversion of early digital data (Stage 4B)

This stage is concerned with the conversion of early digital data to a common, open and modern digital format. It is anticipated that the involvement of the research library sector during this stage will be minimal and would dependant on available expertise regarding the relevant early digital format. The research library may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
  - the data librarian, assisted by a format specialist, are suitable incumbents for this task,
- creating guidance documents and templates:
  - the data librarian is a suitable incumbent for this task,
- creating a conversion inventory or providing guidance regarding the creation of an inventory:
  - the data librarian or records manager are suitable incumbents for this task,
- there will be limited research library involvement in the actual conversion task, unless available research library employees are familiar with such conversion; the use of external conversion/digitisation entities is recommended:
  - research library professional familiar with relevant early digital format,
- creating a master list of filenames and folders:
  - the data librarian or records manager to provide guidance,
- quality controlling and validating the converted media:
  - research library professional familiar with relevant early digital format,
- creating folders as per predetermined structure:
  - research library professional familiar with relevant early digital format,
- storing the converted files in structure:
  - research library professional familiar with relevant early digital format,

- updating the inventory throughout the stage:
    - the data librarian or records manager are suitable incumbents for this task, and
- monitoring progress and updating the rescue progress report:
    - to be performed by the research library employee assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with several tasks provided there is familiarity with the relevant early digital format.

During this stage, ICT services will be consulted to obtain information regarding the provision of data storage if required, the availability of data readers if required, and other issues related to the relevant early digital format.

It is anticipated that this task might require the involvement of external parties, or the outsourcing of the entire stage to a commercial digitisation organisation.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Modern digital data (Stage 4C)

This stage is concerned with the conversion of proprietary modern digital data to a common, open and modern digital format. It is anticipated that the involvement of the research library sector during this stage will be minimal. The research library sector may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - a format specialist, assisted by the data librarian, are suitable incumbents for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- monitoring progress and updating the rescue progress report:
    - to be performed by the research library employee assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with the creation of a conversion inventory and liaising with ICT and the data librarian/records manager regarding the conversion requirements of the data.

During this stage, the input of ICT services will be vital. ICT services will be consulted to obtain information regarding the provision of data storage if required, or any issues concerning data conversion and data access.

Outsourcing of the entire conversion stage is an option that should be considered.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 5: Describing the data (refer to Figure 6.16)

The research library may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- selecting or advising on a suitable metadata standard:
    - the data librarian is the relevant incumbent for this task,
- creating a metadata template:
    - the data librarian is the relevant incumbent for this task,
    - the data librarian can also provide guidance if not personally executing the task,
- creating a metadata file:
    - the data librarian to provide guidance,
- storing metadata with the data:
    - the data librarian is the relevant incumbent for this task,
    - the data librarian can also provide guidance if not personally executing the task,
- advising on the requirements for data documentation:
    - the data librarian to provide guidance on data documentation requirements,
- storing data documentation and metadata, with the data:

    - the data librarian is the relevant incumbent for this task,
    - the data librarian can also provide guidance if not personally executing the task, and
- monitoring progress and updating the rescue progress document:
    - to be performed by person assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with most of the activities forming part of this stage.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 6: Sharing the data (refer to Figure 6.17 and 6.18)

Sharing the data: full rescue

The research library may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- determining the openness of data, and obtaining permission to share the data:
    - the data librarian to perform task or provide guidance,
- selecting suitable repositories (already stipulated in project plan and data management plan):
    - the data librarian to perform task or provide guidance,
- ensuring all repository requirements are met (from both sides):
    - the data librarian to perform task or provide guidance,
- uploading the data, metadata, and data documentation to selected repositories:
    - the data librarian or indexers to perform task,
- performing quality control of uploaded data:
    - the data librarian to review uploads of indexers,
    - records manager to review uploads of data librarian,
- obtaining DOI; sharing DOI with relevant parties:
    - the data librarian to perform task or provide guidance,

- adding all relevant repository numbers, URLs, handles or DOIs to project file:
    - The data librarian to perform task or provide guidance, and
- monitoring progress and updating rescue progress document:
    - to be performed by person assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with the upload of data to disciplinary repositories, and ensuring all requirements are met.

During this stage, ICT services will be consulted only when there are problems regarding the upload of data to the institute's repositories, or problems experienced with repository access, data display and access, and linked matters.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Sharing the data: partial rescue

The research library sector may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- determining the openness of data, and obtaining permission to share the metadata of the data:
    - the data librarian to perform task or provide guidance,
- selecting suitable repositories (if not already stipulated in project plan and data management plan):
    - the data librarian to perform task or provide guidance,
- ensuring all repository requirements are met (from both sides):
    - the data librarian to perform task or provide guidance,
- uploading the metadata to selected repositories:
    - the data librarian or indexers to perform task,

- performing quality control of uploaded metadata:
    - the data librarian to review uploads of indexers,
    - records manager to review uploads of data librarian,
- adding all relevant repository numbers, URLs, handles or DOIs to project file:
    - the data librarian to perform task or provide guidance, and
- monitoring progress and updating rescue progress document:
    - to be performed by the person assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with the upload of metadata to disciplinary repositories, and ensuring all requirements are met. This is not to say that the research library sector cannot be involved; a library and information services professional with relevant disciplinary background, or a library and information services professional who has been involved with the creation of metadata linked to the digitised data would also be a suitable candidate for this task.

During this stage, ICT services will be consulted only when there are problems regarding the upload of metadata to the institute's repositories, or problems experienced with repository access, metadata display and access, and linked matters.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 7: Preserving the data for the long term (refer to Figure 6.19)

The research library sector may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian, assisted by an ICT specialist, are suitable incumbents for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- selecting or advising on the selection of suitable and stable long-term storage option (careful consideration of archive requirements on both sides; ideally already identified during planning stage):

    - the records manager/data librarian are suitable incumbents for this task,
- advising or ensuring that data are in preservation format:
    - the records manager or data librarian to provide guidance,
- monitoring progress and updating the rescue progress document:
    - to be performed by the person assigned the responsibility during the planning stage, and
- monitoring the status and accessibility of preserved data regularly:
    - to be performed by the person assigned the responsibility during the planning stage.

During this stage, ICT services will be consulted to obtain information regarding the suitability and availability of long-term preservation options at the institute. It is likely that the ICT sector at the selected institute will be the party mostly responsible for the preservation of data.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Stage 8: Project closure (refer to Figures 6.20 and 6.21)

Full rescue (Stage 8A)

The research library may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- advising on the publishing of rescued data:
    - the data librarian can be involved in an advisory role,
- creating awareness and promoting the rescued data within the research community, the library and information services sector and the selected institute:
    - the data librarian and project manager can be involved in an advisory role,
    - the data librarian and project manager can be involved in the drafting of articles or presentations,
- thanking all involved parties:

    - to be performed by the person assigned the responsibility during the planning stage,
- creating the final project report:
    - to be performed by the person assigned the responsibility during the planning stage, and
- amending the data rescue steps for future projects if required:
    - to be performed by the person assigned the responsibility during the planning stage.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with the promotion of data via conferences presentations, data-related articles, or articles in dedicated data journals. Library and information services professionals involved with the project would also be able to author an article, or contribute towards an article mainly authored by an SET staff member.

During this stage, the institute's communications/marketing division will be involved with the publishing of intranet articles or similar promotional activities.

A summary of the potential roles and responsibilities of the library and information services sector based on current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Partial rescue (Stage 8B)

The research library may potentially be involved in the following tasks:

- stage-specific training of data rescue participants:
    - the data librarian is a suitable incumbent for this task,
- creating guidance documents and templates:
    - the data librarian is a suitable incumbent for this task,
- advising on the publishing of metadata of partially rescued data:
    - the data librarian can be involved in an advisory role,
- creating awareness and promoting the partial data rescue project within the research community, the library and information services sector and the selected institute:
    - the data librarian and project manager can be involved in an advisory role,
    - the data librarian and project manager can be involved in the drafting of articles or presentations,
- thanking all involved parties:

    - to be performed by the person assigned the responsibility during the planning stage,
- creating the final project report:
    - to be performed by the person assigned the responsibility during the planning stage,
- amending data rescue steps for future projects if required:
    - to be performed by the person assigned the responsibility during the planning stage, and
- re-assessing institutional resources/infrastructure annually to determine whether full rescue (rescue activities not yet performed) should proceed:
    - the data librarian, records manager and project manager are suitable candidates for the task.

It is anticipated that researchers or persons with disciplinary background/knowledge be involved with the promotion of the (partially rescued) data via conferences, presentations or data-related articles.

During this stage, the institute's communications division will be involved with the publishing of an intranet article regarding the data rescue project and the partially rescued data.

ICT should be consulted when performing the annual re-evaluations of the institute's data rescue resources and infrastructure.

A summary of the potential roles and responsibilities of the library and information services sector based on the current research library positions at the selected research institute is provided in Table 6.2: Recommended and potential data rescue roles of the research library. The table is placed in Section 6.2.2.3 and provides a compact overview of the research library's involvement during all stages of the recommended Data Rescue Workflow Model.

Summary: This section indicated and clarified the potential involvement of the library and information services sector during data rescue projects and made use of the recommended Data Rescue Workflow Model as a point of reference. There is potential for the library and information services sector to be involved in all stages, even though the level of involvement will differ between stages, projects and institutes. The available skills set, and disciplinary background of the research library sector is a determining factor when it comes to determining data rescue involvement and participation.

A tabular representation of library and information services roles and responsibilities based on current research library positions at the selected research institute is shown in Table 6.2: Recommended and potential data rescue roles of the research library.

The next section discusses the ways in which the library and information services sector can be involved in activities not forming part of the recommended model, but in activities indirectly contributing towards the promotion and execution of data rescue within a research institute.

6.2.2.3 Additional roles and responsibilities

Findings included in this section refer to library and information services activities related to data rescue yet not forming part of the previously discussed data rescue workflow/model. Seen in Boolean terms, this section reports on the findings emanating from the 'data rescue' NOT 'data rescue workflow model' AND 'library and information services' equation. Data rescue aspects forming part of this section include training of the library and information services sector in the concepts of data at risk and data rescue, the research library's participation in relevant surveys, and the library and information services sector's contribution towards the data rescue knowledge base via ongoing and continuous data management services. This researcher considers it crucial to mention these activities within this section, as data rescue involvement of the library and information services sector outside of the workflow model can contribute towards an institute's inclination or preparedness to conduct data rescue activities.

Data rescue manuals and guidance: Results of this study have shown that several data rescue publications were authored by LIS experts involved with data rescue. The final stage of the recommended data rescue model refers to promotion and marketing of the data rescue project via presentations, articles and a final project report. However, the library and information services sector has the potential to contribute additionally towards the topics of data at risk and data rescue by creating data rescue guidelines, manuals, or even open access online training tools and guidance material.

The implications of the potential involvement of the research library in the creation of data rescue publications are as follows:

- Research library professionals working in the areas of data management, paper-based archives, and repositories should be viewed as valuable contributors when it comes to manuals and other publications on data rescue.
- Collaboration with sectors outside of the research library should be considered; ICT and an institute's SET base are two of the sectors that would be able to provide valuable data rescue manual input.
- Roles of the research library applicable under this heading include author or editor of a data rescue manual.

- Responsibilities emanating from the mentioned role include the following:
  - as author: draft chapters or sections forming part of the incumbent's sphere of skills and experience,
  - as editor: assemble a group of authors,
  - as editor: guide the assembled authors,
  - as editor: review the written content for spelling and grammar errors, perform fact-checking where required, provide suggestions for edits when needed, and
  - as editor: approve the manual's layout, design, style and tone.

Networking and collaboration: Data rescue literature has shown the importance of the formation of interest groups related to data rescue, data rescue workshops, and collaborative data rescue projects involving the participation of more than one university library. These trends demonstrate the crucial role of collaboration during rescue projects and show the value of combined skills, effort and resources. An additional benefit of collaborative work includes the potential of mentoring and skills development.

Hosting of data rescue events: Results of the study have hinted at the prevalence of libraries and library staff being involved in the hosting of data rescue events. While many of the documented instances relate to the rescue of US federal data, the potential role of libraries being suited towards data rescue hosting cannot be ignored. Reasons for the library and information services sector being an ideal host of data rescue projects include the following:

- many research libraries already perform numerous activities forming part of data rescue, such as secure storage of older data, digitising, data upload to a repository, and creating metadata for data,
- many research and academic libraries already have a dedicated data librarian; incumbents are ideal data rescue team members, advisors and even project managers, and
- many libraries have spaces that can be used for training of participants or team member meetings.

The contribution as host of data rescue events is of relevance when the research library in question contains, or has access to one of several of the following:

- an archive for paper-based material,
- open spaces for discussions, meetings and training sessions,

- space for the temporary storage of data prior to data being inspected, assessed and inventoried,
- a range of working stations with laptops/desktop computers, and
- a scanner, or a digital camera with stand.

Additional tasks forming part of the hosting of a data rescue event may include and are not limited to:

- pre-event communications,
- arranging campus and venue access,
- arranging sufficient, secure and accessible parking,
- testing of equipment, networks and passwords,
- ensuring sufficient plug points and extension cables,
- catering, and
- performing security briefings and adhering to COVID-19 stipulations.

To summarise: the library and information services sector is often involved in data rescue, with the function of data rescue host being an ideal role considering the sector's physical work environment, spaces, and daily employee tasks.

Training, teaching and learning: The fact that the study findings have shown the potential for library and information services involvement in various data rescue workflow roles also indicates that training of LIS sector professionals is required. Training involvement of the LIS sector can include research library personnel being data rescue learners, or research library experts with data rescue experience being involved as trainers.

Implications of the library and information services sector requiring training, and being responsible for training, are stated below.

- Data rescue training should include the topics of data at risk, data rescue, and relevant aspects such as digitisation practices or the importance of metadata.
- Training by research library experts can include workshop participation, presenting at relevant conferences, creating institutional or open access online training materials and manuals, or providing one-on-one training on demand.
- It is anticipated the data rescue experts, forming part of the library and information services sector, will mostly be providing training to the library and information services sector. However, experienced research library professionals can also be involved in the training of

non-research library staff who may potentially be involved in future data rescue projects. Sector members include an institute's SET base, expert volunteers and citizen scientists.

- Library and information services sector members to receive training may involve research library professionals, non-professionals employed in the library and information services sector, and postgraduate library and information services students.
- It is vital that learners provide feedback to trainers after training events, and that sessions or training material be revised to include recommendations and suggestions.
- The designated roles of the research library applicable under this heading are:
  - trainer/teacher, and
  - learner.
- Responsibilities emanating from the mentioned roles include the following:
  - Trainers creating training tools and material, delivering training events, giving training on demand, requesting feedback, implementing feedback into training material, and updating training topics and contents as required.
  - Learners will be involved in attending of data rescue training events, working through available online training materials, and providing feedback to trainers. An eventual outcome of training will be participating in a data rescue project.

Environmental scanning: Another way in which the research library sector can contribute towards the discipline of data rescue is to become involved in research activities such as surveys, case studies and related projects aimed at gathering information about the current data rescue trends. Such activities would also establish the state of data at risk, data rescue activities performed, data rescue awareness, and the need for training and guidance. As a result of these research ventures, the scholarly depth of data rescue research emanating from the library and information services sector will be enhanced.

Within environmental scanning, the research library sector can be involved in a researcher role, or as participant.

Should a library and information services professional be involved as a researcher, the accompanying responsibilities would include:

- creating surveys and administering surveys,
- communicating with survey participants,
- obtaining ethical clearance,
- analysing data,
- reporting and publishing survey findings,

- managing survey data, and
- using findings to implement or change policy, services and tools.

Involvement as a research participant or research subject would require the research library professional to provide opinions, suggestions and recommendations to data rescue researchers, and to respond to interview questions with honesty and integrity.

An important outcome of research into data rescue performed by the library and information services sector is not only to extend scholarly research emanating from this sector, but also to use research findings to showcase trends and limitations, and in so doing potentially contributing towards improved policies, procedures, infrastructure and services related to data at risk and data rescue. Studies into data at risk and data rescue are also anticipated to provide impetus for future data rescue projects.

Summary: A perusal of the literature, combined with a consideration of the current nature of and services provided by the library and information services sector has shown that the sector is already involved in many services related to data rescue. Research libraries providing data management tools and services, even in the absence of data rescue efforts, are indirectly contributing towards aspects forming part of data rescue. Concepts included here are training and tools related to data management plans, indexing and repository services, metadata and data documentation tools and training, DOI awareness creation, data sharing awareness, best data storage practices, and long-term preservation of data.

The implications of the continuous involvement of the research library in data rescue, albeit indirectly, are as follows:

- Research library professionals, even as novice data rescuers, are often able to perform many different data rescue tasks due to their data management background.
- The undeniable and unbreakable link between data management, data at risk and data rescue implies that the concepts of data at risk and data rescue should form part of training provided, online materials available, and tools and services made available.

The next section provides tabular summaries of all previously discussed potential data rescue roles and responsibilities linked to the library and information services sector.

6.2.2.4 Summary of library and information services roles and responsibilities

The two tables below provide a summary of the research library's involvement in data rescue as discussed in the preceding sections.

- Table 6.2 indicates the potential library and information services roles and responsibilities linked to the recommended Data Rescue Workflow Model.
- Table 6.3 indicates potential additional potential research library roles and responsibilities, not directly forming part of a data rescue event or the recommended model but potentially affecting, enhancing or enabling future data events.

Table 6.2: Recommended and potential data rescue roles of the research library

(Text in *italic* indicates potential library and information services positions at a research institute)

DATA RESCUE STAGE	ROLE	RESPONSIBILITY
Stage 1: Data rescue preparatory stage	**Trainer** *Data librarian*	• Provides stage-specific training to participants • Compiles relevant training material
	Data assessor *Information scientist* *Data librarian* *Records manager*	• Assists in assessing suitability of data for rescue, and value of data • Assessment done in collaboration with assessment team members
	Venue organiser *Archives technician*	• Ensures secure and sufficient space for pre-assessed data • Communicates with current data custodian regarding delivery or collection of data
	Documentation creator *Data librarian*	• Creates relevant documentation (guidelines, templates, inventories) • Adds relevant details to inventory
	Resources assessor *Data librarian* *Records manager*	• Assists in assessing feasibility of institute's data rescue resources
	Data destruction coordinator *Records manager*	• Liaises with ICT or Waste Services regarding data destruction
	Project manager *Data librarian* *Records manager*	• Plans activities and communicates with role-players • Sets deadlines • Oversees activities of this stage • Updates project report
Stage 2: Planning for data rescue	**Trainer** *Data librarian*	• Provides stage-specific training to participants • Compiles relevant training material
	Documentation creator *Data librarian*	• Creates relevant documentation (guidelines, templates, inventories)
	Creator/co-creator of DMP and linked tasks *Data librarian*	• Assists with DMP creation • Takes responsibility for DMP editing, sharing and storage
	Team member *Data librarian* *Records manager* *Archives technician*	• If not involved as DMP creator or project manager, team members will be responsible for suitable tasks as indicated by project leader
	Project manager *Data librarian* *Records manager*	• Drafts the data rescue project plan • Manages the data rescue project team • Assigns team member roles and responsibilities • Takes responsibility for the project folder • Distributes timelines and meeting dates

DATA RESCUE STAGE	ROLE	RESPONSIBILITY
Stage 3: Storing and preserving paper data	**Project manager** *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report
	Trainer *Data librarian*	• Provides stage-specific training to participants • Compiles relevant training material
	Documentation creator *Data librarian*	• Creates relevant documentation (guidelines, templates, inventories)
	Archive cleaner/data cleaner *Data librarian* *Archive technician* *Intern*	• Removes dust from paper data, boxes and shelves. • Oversees dusting activities should a non-library cleaner be involved • Oversees dusting activities should a non-library cleaner be involved • Oversees sweeping and vacuuming of archive
	Data curator *Archive technician* *Data librarian* *Intern*	• Labels boxes • Labels shelves • Stores paper data in boxes • Stores other data formats securely
	Archivist *Records manager*	• Ensures archive meets requirements regarding humidity, temperature, pest control • Ensures archive has space and shelves • Ensures fire-suppressant system is functional and serviced regularly
Stage 3: Storing and preserving early digital data	**Project manager** *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report
	Trainer *Data librarian*	• Provides stage-specific training to participants
	Documentation creator *Data librarian*	• Creates relevant documentation (guidelines, templates, inventories)
	Format expert *Person familiar with relevant early digital format*	• Examines the data and data reader • Assists with the appropriate storage of data and reader/accompanying equipment • Locates metadata
	Data storage expert *Designated research library or SET person*	• Stores early digital data as per best practices, prior to it being digitised

DATA RESCUE STAGE	ROLE	RESPONSIBILITY
Stage 3: Storing and preserving modern data	**Trainer** *Format expert (assisted by data librarian)*	• Provides stage-specific training to participants
	Documentation creator *Data librarian (assisted by format expert)*	• Creates relevant documentation (guidelines, templates, inventories)
	Project manager *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report
	Data storage expert *Designated research library or SET person*	• Stores early digital data as per best practices, prior to it being converted to a different format
Stage 3: Storing and preserving physical samples and specimens	**Trainer** *Format expert (assisted by data librarian)*	• Provides stage-specific training to participants • Compiles relevant training material
	Documentation creator *Data librarian (assisted by format expert)*	• Creates relevant documentation (guidelines, templates, inventories)
	Project manager *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report
	Data storage expert *Designated research library or SET person*	• Stores samples and specimens as per best practices
Stage 4: Digitising/converting data	**Trainer** *Format expert (assisted by data librarian)*	• Provides stage-specific training to participants • Compiles relevant training material
	Documentation creator *Data librarian (assisted by format expert)*	• Creates relevant documentation (guidelines, templates, inventories)
	Digitiser *Digitisation clerk* *Archives technician* *Data librarian* *Research library intern*	• Creates digitisation inventory • Scans images • Keys data • Converts early digital data to common, open, modern format
	Quality controller *Data librarian* *Records manager*	• Validates that all files have been scanned/keyed, converted • Quality controls the scanned/keyed/converted data
	Project manager *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report

DATA RESCUE STAGE	ROLE	RESPONSIBILITY
Stage 5: Describing the data	**Trainer** *Format expert (assisted by data librarian)*	• Provides stage-specific training to participants • Compiles relevant training material
	Documentation creator *Data librarian (assisted by format expert)*	• Creates relevant documentation (guidelines, templates, inventories)
	Creator of metadata and related tasks *Data librarian* *Information scientist*	• Determines metadata standard required by repository • Advises on metadata creation • Advises on data documentation creation • Assists with metadata creation • Reviews metadata created by SET base • Takes responsibility for storing metadata and data documentation with data
	Project manager *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report
Stage 6: Sharing the data	**Trainer** *Data librarian*	• Provides stage-specific training to participants
	Documentation creator *Data librarian*	• Creates relevant documentation (guidelines, templates, inventories) • Compiles relevant training material
	Repository professional *Data librarian* *Indexers*	• Uploads data to selected repositories • Uploads metadata and data documentation to selected repositories • Shares DOI with relevant parties
	Project manager *Data librarian* *Records manager*	• Monitors/supervises rescue progress • Updates rescue progress report
Stage 7: Preserving the data (for the long term)	**Trainer** *Data librarian (assisted by ICT division)*	• Provides stage-specific training to participants • Compiles relevant training material
	Documentation creator *Data librarian (assisted by ICT division)*	• Creates relevant documentation (guidelines, templates, inventories)
	Archive selector *Records manager* *Data librarian*	• Provides guidance regarding the requirements for secure and stable preservation location • Assists in selecting appropriate location
	ICT collaborator *Records manager*	• Collaborates with ICT or entity selected for long-term preservation • Relays needs and requirements
	Project manager *Data librarian* *Records manager*	• Monitors rescue progress • Updates rescue progress report
	Data curator *Data librarian* *Records manager*	• Gives guidance about suitable preservation formats • Regularly monitors accessibility of preserved data

DATA RESCUE STAGE	ROLE	RESPONSIBILITY
Stage 8: Project closure	**Trainer** *Data librarian* *Records manager*	• Provides training to role-players and team members • Compiles relevant data rescue training material
	Documentation creator *Data librarian*	• Creates relevant documentation (guidelines, templates, inventories)
	Supervisor *Data librarian* *Records manager*	• Oversees relevant stage of the data rescue project
	Marketing *Data librarian* *Records manager* *Project manager to also be involved*	• Collaborates with institute's Communications Department • Promotes rescued data; creates awareness • Shares data rescue learnings
	Author/presenter *Data librarian* *Records manager*	• Write and publish article on rescued data and/or rescue project • Write and present paper on rescued data and/or rescue project • Scholarly and popular journals should be included
	Miscellaneous closure tasks *Data librarian*	• Monitor usage statistics of rescued data in repository • Provide feedback regarding workability of rescue model • Assist with implementation of feedback into updated rescue workflow if required • Annual reassessment of data not rescued, or only partially rescued
	Evaluator of rescue model *Data librarian* *Records manager* *Project manager*	• Re-evaluate rescue model • Implement feedback into updated rescue workflow if required
	Data assessor *Data librarian* *Records manager*	• Annual reassessment of data not rescued, or only partially rescued
	Project manager *Data librarian* *Records manager*	• Thank relevant parties • May be involved in marketing activities • Write and distribute final report • Organise post-rescue meeting; collate feedback

The table above shows the various roles and responsibilities of the library and information services sector during data rescue projects. While the tasks listed form part of the recommended model, and are crucial in nature, several additional related roles and responsibilities can also be identified. These are tasks not part of the recommended model, but are activities that can improve the data rescue expertise at an institution, increase the odds of a data rescue venture being successful, or increase awareness around the topic of data at risk and data rescue. These roles and responsibilities should be performed continually, and are indicated in the table below.

Table 6.3: Additional data rescue roles of the research library

(Text in *italic* indicates research library positions at a research institute)

RELATED DATA RESCUE CONCEPT	ROLE	RESPONSIBILITY
Data rescue event hosting	**Event organiser/host** *Designated research library professional*	• Pre-event communications with rescue team members • Ensures spaces for data and rescuers are available and suitable • Arranges access to premises, parking, and related matters • Establishes security measures • Ensures equipment access • Tests all equipment and tools prior to event • Organises catering • Ensures seating is comfortable, available, compliant • Arranges and tests Wi-Fi access • Ensures training area meets standard with regard to sanitation, and proximity to ablutions • Ensures there are quiet working areas • Ensures SHEQ regulations are adhered to
Training	**Trainer** *Data librarian* *Information scientist*	• Trains library and information services sector members on data rescue • Continuous rescue training of institute SET base • Implements training feedback into future trainings • Creates online manuals/guides, e.g., LibGuide
	Learner *Research library employee*	• Attends training sessions • Provides feedback to trainer • Shares learnings with colleagues
Data management services	**Data management services** *Data librarian* *Information scientist*	• Creates online guidance tools • Provides ad hoc guidance • Enhances institutional awareness • Conducts institutional training • Performs repository tasks
Data rescue publications (including manuals/guidelines)	**Author** *Data librarian*	• Contributes towards a data rescue manual/guidelines • Is involved in all documentation linked to data rescue • Creates open access online materials • May also be sole author of data rescue publications

RELATED DATA RESCUE CONCEPT	ROLE	RESPONSIBILITY
Networks and collaborations	**Committee member** *Data librarian*	• Assists in organisation of online data rescue training events, or in-person workshops and meetings
	Member *Data librarian*	• Attends meetings related to data rescue, data at risk, digitisation, and associated topics • Active participation in data rescue events
	Researcher *Data librarian*	• Conducts surveys regarding data rescue • Establishes institutional needs and challenges • Publishes and creates awareness of findings
	Participant *Research library employee*	• Participates in surveys when invited and where applicable

Section 6.2 was devoted to answering the study's main research question. Answering the question entailed indicating the roles and responsibilities of the research library within a comprehensive workflow for data rescue. An indication of potential library and information services involvement in other rescue-related spheres also formed part of this section. To provide practical examples of research library involvement, the section also referred to actual current research library positions at the selected research institute. Library and information services involvement in data rescue is bound to differ between institutes and would depend on a range of factors including data rescue skills and experience, disciplinary knowledge, size of the research library and current workload of research library employees.

The next section discusses the conclusions reached after answering the study's research questions.

6.3 Conclusions reached

This study's objective was to contribute towards the rescue of data at risk by establishing ways in which parties, other than researchers and scientists, may be involved. Furthermore, this researcher was also interested in attaining the following sub-objectives:

- to establish how data rescue is currently done, both locally and internationally,
- adding to the above: to determine who is currently involved in data rescue, and their roles,
- to establish current data rescue perceptions, needs and challenges,
- to create a data rescue workflow based on information gained via the three previous points,
- within the model: to stipulate how the library and information services professionals can be involved in data rescue, and

- to contribute towards future data rescue activities by ensuring the model is freely available to all interested parties.

By answering several research questions, this study was able to show findings which enabled the achievement of the study objectives. The main findings of this study are summarised below.

Data at risk is a global phenomenon and can be found in most research disciplines, involving a range of historic and current formats (and time periods in between). Many factors influence data to be at risk of loss, damage, or being inaccessible. Locally, data at risk at the selected institute were found to be present in most participating research groups, and were stated to include many formats and time periods of data collection. Participants, as was found in documented literature, also listed a substantial number of factors leading to data being at risk.

Similarly, projects and activities to rescue the identified data at risk portray a global presence, and extend even to the rescue of lunar, astronomical, stratospheric and oceanic data. Documented rescue projects have involved data from collection periods several centuries ago, right up to recently collected data. Data formats rescued include paper-based media, punch cards, magnetic tapes, early digital formats, microfiche and microfilm, photographic plates, and modern digital formats. Rescue projects reveal a great deal of variance, with a 'typical' rescue project, 'typical' participants, or 'typical' responsibilities and roles not readily identifiable. Results from the empirical phase of the study have shown that experienced researchers at the selected institute face a large number of obstacles and challenges when considering data rescue.

Despite the variance in the stated data rescue project attributes, broad categories of data rescue activities within the data rescue process are present in most workflows found in literature. These activities, combined with feedback obtained from study participants, have resulted in the creation of a Data Rescue Workflow Model. The recommended model indicates steps and activities forming part of the data rescue workflow and indicates roles and responsibilities. Outputs, guidance and templates also form part of the model.

Documented sources consulted during this study have shown the involvement and need for library and information services sector involvement in data rescue, as well as the valuable contribution that can be made by past researchers, expert volunteers and citizen scientists. Collaboration between different parties has proven to be beneficial to data rescue ventures.

The use of the term 'data rescue' is currently a topical issue, and it was found that current and popular terminology and issues linked to terminology are worthy of consideration.

To address the need for increased data at risk and data rescue awareness, increased understanding of data rescue activities, advocating the vital role of different parties during data rescue, promoting the newly created Data Rescue Workflow Model, and assisting the launch of a data rescue project at the selected institute, a set of recommendations has been put forward. While some of the recommendations are applicable to the selected research institute only, many of the recommendations are universal in nature and are relevant to other research institutes or tertiary institutes at the same stage of data rescue implementation.

Recommendations will be discussed fully in the next section.

6.4 Recommendations

Based on the results of this study (Chapter 5: Results and discussion), the summaries in section 6.2, and the conclusions reached in section 6.3, several recommendations can be made. These recommendations are intended to address the issue of data at risk and data rescue practices not only at the selected institute, but also for the wider library and information services community and applicable research institutes, including tertiary institutes of education.

The recommendations are categorised and address a range of topics associated with data at risk and data rescue. Recommendations are intended to speak to the following issues:

- awareness of data at risk,
- the correct handling of data at risk,
- awareness around data rescue,
- the terminology around data rescue,
- the mitigation of data rescue challenges,
- the promotion of the data rescue workflow model,
- data rescue funding,
- promotion of the role of the research library sector in data rescue,
- promotion of the role of citizen scientists in data rescue,
- promotion of the role of past researchers in data rescue,
- promotion of the role of expert volunteers in data rescue,
- the LIS curriculum and its inclusion of data rescue,
- the role of collaboration in data rescue, and
- the launch of a data rescue project.

Bearing these intentions in mind and coupled with insight gained during this study, a set of recommendations has been made and these are elaborated on below.

6.4.1 Create awareness around data at risk

The preponderance of data at risk, as established via the study's empirical findings, underlines the need for increased awareness regarding the issue. As stated during the discussion of institutional data rescue experience (see Section 5.4.9.2) and preceded by the description of Sample A (Section 4.7.1), data management at the selected institute was, at the time of writing, still in its infancy. With decades of data collection and data storage being conducted in the absence of a data management procedure, an appointed data curator, a data repository, or even rudimentary data management guidelines, it was not surprising that data were rarely managed according to best practices. As a result, data were not curated, resulting in an institutional status quo where historic paper data eventually landed up in a range of non-ideal locations, modern electronic data were often stored without metadata or in a disorganised file and folder structure, the creation of data management plans being a rarity, and data stored without any backups ever being made. There is an undeniable link between non-adherence to data management best practices, and such data being at risk.

While this study's focus is the rescue of data, an understanding of the concept of 'data at risk' should be promoted. It is vital that researchers and parties working with data, at the selected research institute and beyond, not only be aware of the data rescue model (see Section 6.2.1.7), or that a practice such as 'data rescue' exists, but also be informed about 'data at risk' and the range of factors resulting in data being at risk of loss, damage, or no longer being accessible.

It is recommended that the following 'data at risk' topics and concepts form part of institutional awareness sessions and other training events:

- definition of 'data at risk',
- inclusion of a range of formats when referring to data at risk, including paper-based data, data in an early digital format, data in modern electronic formats and physical samples data,
- discussions around the prevalence and universality of data at risk, including formats, disciplines, geographic areas, and time periods,
- training presentations to discuss examples of global and notable events featuring data at risk, as well as the effects of data at risk,
- discussions around factors leading to data being at risk: such a discussion should include factors mentioned in Section 5.4.7: Data at risk: Factors, as well as risk factors emanating from the literature review,

- demonstrations of the link between non-adherence to data management best practices, and data becoming at risk; this link should be understood by all researchers and other parties working with data,
- adding to the previous bullet: the importance of storing metadata and data documentation with data should be emphasised, with researchers to understand that linked metadata and data documentation ensure that data will be understood and interpreted by any future user,
- discussions around the benefits of applying data best practices, and the mitigating effects such practice will have on mentioned data at risk factors, and
- discussions of steps to be taken when discovering or being aware of data at risk.

While it is not possible to change the ways in which data were managed in the past, or even be able to save all data deemed to be at risk, creation of awareness around data at risk, and implementation of best data management practices can minimise the chances of current and future generated data being at risk.

6.4.2 Encourage the correct handling of data at risk

This recommendation, while forming part of the previous point, proposes the steps to be taken should institutional employees locate, receive, or become aware of data at risk. This recommendation ensures that the risk of data loss is reduced, even though a decision on the feasibility of a data rescue project would at that point not yet have been made.

The following are recommended steps to follow when handling data at risk:

- Upon discovery, location, or donation of the data at risk, the institute's library, and the data librarian and records manager in particular, should be contacted.
- Data in paper format should ideally be handled with gloves.
- Data in paper format should ideally be stored in labelled boxes in the interim.
- Data should ideally be stored data away from extreme heat or sunlight or humid areas.
- Efforts should be made to minimise the chances of pest damage, water damage and/or fire damage.
- Metadata and data documentation, if available, should be stored with the data.
- After receiving notification about the located data at risk, the designated data rescue staff members at the research library (e.g., the data librarian or the records manager) should schedule an appointment for collection/delivery and inspection/assessment of the data. Activities and processes described in Stage 1 of the recommended model (i.e., data rescue preparatory stage) will be performed.

- Plans for the assessment of data should be made; the preparation of the implementation of activities, roles and responsibilities involving the research library, as indicated, and described in Section 6.2.1.3 and summarised in Table 6.2 and Table 6.3 should commence.

Steps described above should be relayed during institution-wide awareness sessions on data at risk (see Section 6.4.1) to ensure that all researchers have a basic understanding of steps to be taken when discovering data at risk.

6.4.3 Create awareness around data rescue

A previous recommendation touched on increased awareness creation around the topic of data at risk. Similarly, data rescue as practice should be promoted, demonstrated and encouraged. Researchers and other parties included in the Data Rescue Workflow Model must be made aware that there is hope for data that have been identified as being at risk, and that data at risk do not necessarily equate with lost data.

It is recommended that the following data rescue topics and concepts form part of institutional awareness sessions and other training events:

- definition of 'data rescue',
- examples of successful data rescue projects, as documented in literature,
- discussions around the link between adherence to best data management practices and the eventual lessening of the need for data rescue to take place,
- initiation of the 'data rescue' – 'data conservation' terminology discussion (also touched on in Section 6.4.4),
- demonstration of the recommended Data Rescue Workflow Model created during this study (see Section 6.2.1.4),
- trainer to touch on the ways in which the research library sector can potentially be involved, thereby conveying the message that data rescue is a collaborative effort and not an obligation placed solely on the shoulders of an institute's SET base, and
- discussion of typical data rescue challenges or obstacles and giving advice on realistic mitigation measures.

Ideally, data rescue awareness creation and training should take place in conjunction with data management awareness sessions or training events. As already stated in Section 6.2.1.8, when giving suggestions for data rescue inclusion in the LIS curriculum, a module on data rescue should form part of data management training. In addition, data rescue should also form part of the methodology training in all disciplines, and one in which LIS departments or schools could take the lead. However, it is vital that data rescue training sessions only take place once there is an understanding of concepts such as data formats, a data management plan, metadata, data documentation, data repositories, DOIs, and data licensing. Data rescue as training topic should therefore form part of the latter section of a data management training programme.

When creating awareness regarding data at risk and data rescue at a research institute, one or several of the following information-sharing options should be considered:

- an institution-wide awareness session, with attendees ideally having already attended a prior session on data management,
- a pre-recorded video version of the awareness presentation to be accessed in viewers' own time,
- a PDF or text-based version of the session/presentation to be accessed in viewers' own time, for use when bandwidth is of concern to the viewer,
- a LibGuide covering all aspects covered in the awareness presentation to be placed on the institutional intranet, forming part of the institute's library and information services blog, and
- an article placed on the institutional intranet, featuring a call for researchers to contact the institutional data librarian or records manager should there be data at risk, and the data suitable for use in an institutional data rescue project.

The bulleted points above should also be read in conjunction with the recommendation made in Section 6.4.14 regarding the launch of a data rescue project.

6.4.4 Monitor the terminology around 'data rescue'

As discussed in the literature review chapter (see Section 2.2), use of the term 'data rescue' is one which requires consideration. It was also discovered, during the one-on-one virtual interview stage, that the term 'data rescue' at times required explanation. To some of the study participants, the term signified an activity performed by ICT services when modern digital data are not accessible or are corrupted.

In addition to researchers at the selected research institute equating 'data rescue' with the recovery of corrupt or hidden modern digital data, the global usage of the term also indicates a range of data-related settings. 'Data rescue' is used to describe the efforts by concerned scientists to save modern US federally funded climatic data from being hidden from the public, it is used to refer to the rescue of historic, vintage or legacy data in various formats, and it is also used to refer to the activities performed by ICT specialists whenever a hard drive crashes, a flash drive is damaged, or an SD card becomes unreadable.

The aspects listed in the previous paragraph show that there is a need for a term that encompasses the activities contained within this study's Data Rescue Workflow Model, with the caveat that the term should not presently be used to refer to a range of other data-related activities. As discussed in Section 2.2.2: Data conservation, the name of a renowned international interest group, formerly concerned with 'data rescue', was replaced with 'data conservation' during 2019. The name change served to indicate that the group had extended its range of interest to include the treatment of modern data, and cover aspects such as acquiring funding, and the prioritisation of rescue activities.

It is important to mention that not all global entities followed suit; research articles and conference papers (for example, Engström *et al*., 2021; Geibel *et al*., 2022) found via scholarly online search tools currently still refer to 'data rescue' when describing the treatment of data at risk. Considering the delays between publication submission and publication on a platform such as Google Scholar, it is possible that published outputs were drafted prior to the name change described above. Time will tell whether scientific outputs created after the international interest group's name change decision will incorporate the name change.

The recommendation is made to monitor the global use of the terms 'data rescue' and 'data conservation', respectively. Should there be a major shift towards the use of the latter term, it would be prudent to amend the applicable institutional guidelines, training tools and related outputs to be in line with accepted global usage. It is also possible that another scenario could arise, namely using the term 'data rescue' to refer to those activities dealing with historic data at risk, and the term 'data conservation' when dealing with current data already in a modern digital format.

6.4.5 Address and mitigate data rescue challenges

It is recommended that an effort be made to address and mitigate a research institute's data rescue challenges as much as realistically feasible. As this study collected empirical data on the data rescue challenges experienced by researchers at the selected institute, the mitigating activities listed in this section are included to minimise the challenges experienced by said researchers. However, it is anticipated that the measures listed below are applicable to a wider sphere than the selected institute only, and can be regarded as applicable to other research institutes as well.

While some of the selected institute's challenges are linked to resource constraints or non-ideal systems and infrastructure, it is estimated that most of the stated rescue challenges (see Section

5.4.10) can be reduced via researcher training and the promotion of best RDM practices. It is anticipated that many data rescue obstacles can be alleviated via the attendance of mandatory data management awareness sessions, with the implementation of the newly approved institutional data management procedure also playing a key role.

An additional alleviation measure entails the creation of an institute-wide data rescue equipment inventory, in order to facilitate the acquisition of rare and older data readers.

The listed data rescue challenges below are some of the obstacles mentioned by interviewed participants and were discussed in Section 5.4.10: Data rescue obstacles and concerns. The anticipated mitigating effect of researcher training and the promotion of best RDM practices form part of each listed obstacle.

6.4.5.1 Data rescue challenge: There is often uncertainty about whether rescuing data will be worthwhile

Mandatory creation of a DMP prior to project registration will address issues around future data use. Data assessment, forming part of the initial stages of the final Data Rescue Workflow Model, is another activity which will address uncertainty around the worth of data rescue.

6.4.5.2 There are substantial costs involved with the rescue of data

As discussed in Section 6.4.7, investigation into funders of data rescue, implementing data rescue collaboration, and making use of volunteers (citizen scientists as well as past researchers) are ways of limiting the costs associated with data rescue. Including the rescue of relevant data in an already-funded project is another way to address data rescue costs.

6.4.5.3 Specialised equipment is required for data rescue

An institution-wide data rescue equipment inventory is a method of revealing the rescue items already owned and available. In addition, collaboration with external parties, outsourcing certain data rescue activities (e.g., digitisation), and advertising for the required equipment are suggested ways of approaching the lack of data rescue equipment.

6.4.5.4 Lack of manpower, and researchers will struggle to be available for data rescue

As is mentioned in subsequent sections of this chapter (see Sections 6.4.8, 6.4.9, 6.4.10 and 6.4.11), the involvement of parties other than the SET-based component at an institute is a way of addressing the concerns about available manpower. Library and information services professionals and semi-professionals, citizen scientists, past researchers, expert volunteers, and even external collaborators are options to consider when faced with manpower and skills concerns. It should be noted that the use of mentioned parties should not only be considered when manpower is of concern, as their involvement is advantageous in its own right.

6.4.5.5 Limitation with regard to the [*organisation*] data deposit tool

With several respondents raising concerns about the often recommended [*organisation*] data deposit tool, it is important to decide on its future status as the recommended option. With the institute providing data storage upon request, and institutional researchers also able to make use of accredited data repositories, the insistence on use of the [*organisation*] data deposit tool should no longer form part of data management awareness sessions.

6.4.5.6 There is a lack of data rescue processes at the selected institute

Simply put: the creation of a final Data Rescue Workflow Model as one of the main outcomes of this study indicates that this specific data rescue obstacle should no longer be relevant. When combined with data management training, and adherence to the newly approved data management procedure, this obstacle can be considered as adequately addressed.

6.4.5.7 Researchers are lacking in data rescue skills, insight and experience

As stated in Section 6.4.3, the creation of awareness around data rescue, and the conveyance of typical data rescue steps (including a discussion of the final Data Rescue Workflow Model) are ways of mitigating this data rescue obstacle.

6.4.5.8 There is limited visibility of, and accessibility to data managers

It is anticipated that the various institutional training activities and awareness sessions around data management, data at risk, and data rescue should reduce the prevalence of this stated challenge.

6.4.5.9 Data rescuer should understand the context of the data

As is discussed further down (Section 6.4.10), promoting the role of past researchers and involving them in data rescue efforts is a study recommendation. Adding to this: creating metadata and data documentation for data and storing it with the dataset are considered vital in ensuring that future users understand the data context.

6.4.5.10 Subject knowledge is required to rescue data

As is found further down (Section 6.4.10 and Section 6.4.11), recommendations are put forward pertaining to the involvement of past researchers and expert volunteers (e.g., postgraduate students) when embarking on data rescue.

6.4.5.11 Data management is up to the individual

The newly approved institutional records management policy (mentioning that data are records), the data management procedure, and the monitoring of adherence will result in a scenario where implementation of data management will no longer be a practice left up to individual whim. Awareness training will address the importance of linked metadata and data documentation to ensure future reuse of data.

6.4.5.12 Summary

The data rescue challenges and mitigation measures listed above serve as examples of how the implementation of policies, procedures and practices can potentially mitigate or diminish data rescue obstacles experienced. These stated challenges and the potential of data management best practices should be discussed at relevant institutional awareness sessions.

6.4.6 Promote the data rescue model

The recommendation is made to license, promote and distribute the recommended Data Rescue Workflow Model. The model is a major outcome of this study and is described in detail in Section 6.2.1.4, and created to address the following research sub-question:

‘To what extent can the theory and practice be formalised in a model for a data rescue workflow?’

The aspects listed below add clarifying details regarding this recommendation.

The model is to be licensed under a Creative Commons Attribution-ShareAlike license, with users of the model able to remix, adapt and build upon this work even for commercial purposes, as long as this researcher is credited, and the new creation is licensed under identical terms. This researcher considers it vital to support open science by making it possible for different organisations, with different disciplines, budgets, resources and policies, to obtain the model freely, and adapt it as required.

The model is to be promoted as widely as possible, with avenues of publication and marketing to include:

- a research article in an LIS-related scholarly journal,
- an article in a popular journal dealing with trending research library activities and happenings,
- an information sharing session at a meeting of a relevant community of practice or interest group,
- a demonstration at a digitisation or data management workshop,

- sharing of the model on suitable social media platforms, and
- the exploration of avenues of making the international data conservation community aware of the model (a conference presentation is a possibility).

The model should also form part of lecture material, if data rescue forms part of the research methodology curriculum, for LIS students and students in other research disciplines. In addition, the model should also be referred to and demonstrated during on-the-job training provided to research library staff at research institutes, or even during SET-based data management awareness sessions at research institutes.

6.4.7 Investigate data rescue funding

It is recommended that funding for data rescue projects be investigated, and the availability of funding opportunities be fully utilised, or conveyed to relevant stakeholders.

Investigation around funding for data rescue is twofold in nature and involves the following activities:

- investigate sources of data rescue, and
- determine and implement ways in which data rescue can be performed should funding be minimal or non-existent.

As revealed in the content analysis part of the research (Section 3.3.9.2: Contribution and limitations), at least one global data rescue entity provides funding for the rescue of historical environmental data. Examples of affected rescue activities include ensuring the digitisation of environmental data funded through charitable contributions, grants and government awards, with sponsored volunteers assisting the country’s national meteorological service in readying the data rescue facility, followed by purchasing and installing computer and camera equipment, and training the data rescue personnel.

It is recommended that similar avenues for the rescue of data at risk in other disciplines be investigated. Examples of likely candidates include UNESCO[44] and the NRF[45] for the rescue of historical sociological South African data, or data related to traditional medicines or similar research fields. It is worth mentioning that the successful South African data rescue project involving apartheid era datasets benefited from funding provided by a foundation known to be the largest supporter of the humanities, arts, higher education and cultural heritage in the US.

[44] https://en.unesco.org/creativity/ifcd/apply
[45] https://www.nrf.ac.za/funding

It is also recommended that the activities mentioned below be executed to address or mitigate the limited or non-existent data rescue funding.

- Create institutional inventories of equipment that could be of use during data rescue projects, such as data readers, digital cameras, tripods, lights, stands, archive boxes, labels, archive coats and gloves, and older computers with stiffy disc/floppy disc/CD/DVD functionalities.
- Collaborate with others, both internally and externally, when considering data rescue. This not only results in shared expertise, but also lessens the premium placed on the need to purchase equipment and tools.
- Consider making use of nationwide requests when specific data rescue equipment or tools are required. An example of such a successful venture is the appeal to the global public made by the Swedish National Archives in 2019, asking whether there were institutes in possession of obsolete equipment no longer being used, and if they were willing to donate the equipment (Open Research Foundation, 2021).
- The involvement of volunteers, students, ex-researchers and interns is a tried-and-tested method of limiting the costs associated with data rescue, and addressing the vast workload associated with the rescue of huge datasets.

In addition to the aspects listed above, it is also recommended that a data rescue component be added to grant funding proposals where applicable. Funding obtained in this manner, and forming part of a normal research project, would then enhance data availability and accessibility applicable to the research project. Involved data in these instances would obviously be limited to the project targets and be relevant to the research discipline and project.

A related recommendation is that funders be made aware of the savings associated with re-using (instead of recollecting) data at risk.

6.4.8 Promote the role of LIS sector professionals and semi-professionals

As discussed in Section 6.2.2, library and information services professionals and semi-professionals may potentially be involved in all the recommended model's stages and are often able to perform various data rescue tasks. Documented sources consulted during the study's literature review on global library and information services practices stated that the sector's professionals are currently involved in various data rescue roles and capacities. Documented activities include creation of data rescue inventories, locating misplaced data documentation, managing data rescue projects, organising data rescue projects, being responsible for storage and preservation of data in different formats, contributing towards data rescue publications, participating in data rescue-related surveys, and the hosting of data rescue events. In addition, library and information services professionals are already routinely involved in daily research library activities included in the model, such as metadata creation, repository upload, data management plan guidance, and digitisation activities or management.

The US-based 'data refuge' movement (see Section 2.2.3) also showcased the vital role of libraries and library and information services professionals during these activities, with librarians and archivists either organising, managing or participating in these events. Their input and contribution have been highlighted as important aspects of the data refuge success, with librarians ensuring that data are kept under a secure chain of provenance and are free from suspicions of meddling. The addition of simple metadata to the data by the library and information services sector, such as the agency from where the data originated, when it was retrieved, and who was handling it, which was later supplemented by metadata added by scientists, also formed part of the movement's rescue activities.

This study's mini focus group session also reiterated the idea that libraries and librarians should be involved in several phases of the model demonstrated to the group (see Section 5.7). Activities included in the discussion involved higher-level tasks such as managing data rescue projects, right through to conventional research library activities such as physical archives management, indexing of data, implementing a structured file and folder structure, repository evaluation and selection, and repository uploads. Semi-professional research library activities such as scanning of paper media, manual transcription of paper data, operating of cameras, and routine duties in physical archives complete the range of data rescue activities suggested to be suited to the library and information services sector.

As a result of the findings described in preceding paragraphs, it was possible to suggest and describe the potential data rescue involvement and contributions of research library professionals and semi-professionals during all data rescue stages, and indicate the roles the library and information services sector was most likely to be involved in. These potential roles and responsibilities were discussed in detail in Section 6.2.2.

Library and information services involvement in data rescue can involve all sectors of the community, namely: pre-graduate technicians or assistants, semi-professional staff, research library professionals, and high-level managerial staff. Additionally, library and information services professionals with data management experience would be able to contribute to the areas of training, creation of guidelines, adaptation of the model, data rescue project management, and data management plan training and quality control. Collaboration with external parties regarding digitisation and disciplinary repositories are additional tasks forming part of the rescue process. Archivists and archival duties (as described in the recommended model's Stage 3: Storing and preserving the data) still form part of the service areas of many research libraries and can be considered a research library responsibility.

The role of library and information services professionals should further be promoted via the inclusion of a data rescue module during postgraduate LIS studies, as elaborated on in Section 6.4.13. In addition, ensuring the details of data rescue projects involving library and information services sectors are published in relevant LIS journals, and data rescue activities are included in suitable workshops and symposia are also bound to promote library and information services awareness and involvement. On-the-job training of members forming part of the current library and information services sector would be a precursor to the sector's inclusion in future data rescue projects.

The practice of automatically assigning certain data rescue responsibilities or activities to researchers should be re-evaluated, with the expansion of the research library's technological skills a recommended step. This not only guarantees that the research library sector is more valuable to the institute, but also ensures that certain data rescue stages are not viewed as the sole domain of an institute's SET environment or ICT sector.

6.4.9 Promote the role of citizen scientists

The study's literature review revealed the vital role played by volunteers and citizen scientists in a range of data rescue projects. While prominent and acclaimed ventures such as Old Weather or the DRAW project make use of volunteers to transcribe historic weather data, recent data rescue campaigns in the USA, rescuing federally funded modern digital data, have also involved volunteers. One of the benefits of having volunteers on the data rescue team is the enthusiasm and eagerness brought to the projects by stated participants.

It is important to make a distinction between 'citizen scientists' and 'expert volunteers'. In the context of the study recommendations, the former group comprises volunteering members of the public without any relevant disciplinary or research knowledge, while the latter are volunteers who have certain skills, knowledge or experience required in certain rescue projects.

Citizen scientists used in data rescue pursuits can contribute in many ways, including the following tasks:

- key data onto electronic spreadsheets,
- electronic transcription of handwritten data,

- clean and vacuum clean physical data archives,
- clean and remove dust from paper records, boxes and shelves,
- store paper data in boxes,
- label archive boxes and shelves,
- operate paper scanners, and
- operate digital cameras.

While this study's Data Rescue Workflow Model is not aligned to the US 'data refuge' activities described elsewhere (see Section 2.2.3), it is important to take note of the activities performed by volunteers during documented ventures. As mentioned, 'citizen scientists' as well as 'expert volunteers' took part in data refuge projects; the recommendation regarding the participation of the latter group is found in Section 6.4.11.

Examples of non-expert volunteer tasks performed during data refuge events include:

- provision of physical space during data rescue projects,
- provision of computer space for data storage,
- working as a 'seeder'; collecting more than 1 000 URLs from the US Fish and Wildlife Service pages and nominating them to the Internet Archive[46], and
- working as a 'storyteller'; this activity entailed creating signs for the March for Science, making a visualisation of relevant tweets over time, and working on redesigning relevant government websites.

The recommendation discussed in this section underlines the fact that there is great potential for citizen scientists to be involved in a range of data management tasks, with involvement ranging from immediate participation to tasks requiring a certain amount of demonstration and training provided by fellow data rescuers.

6.4.10 Promote the role of past researchers

It is recommended that retired or previously involved researchers be considered and included in data rescue projects. The valuable role to be played by employees who had previously been involved with the data concerned cannot be ignored. This is especially pertinent when the data at risk is older, and when members of the original research team are no longer employed by the institute where the data at risk reside.

[46] Internet Archive: a non-profit digital library that has saved more than 286 billion webpages

As established via the findings emanating from both the study's interview stage and model feedback stage, input from researchers familiar with the data is considered important, as it provides context to the data. The significant role of retired researchers is also mentioned in literature, with the WMO including retired climatologists as parties often involved in climate data rescue (WMO, 2016: 11). It is anticipated that ex-researchers, previously involved with the project, would be able to assist with the following tasks included in the final Data Rescue Workflow Model:

- assess the data, and determine its value,
- indicate the readers and equipment required to read the data,
- advise on how to operate older data readers and equipment,
- indicate the software required to read the data,
- advise on how to handle the data, should it be in proprietary or uncommon format,
- supply details of projects that had generated the data (e.g., project title, project objectives, relevant period, related scientific outputs, main findings),
- assist with the creation of metadata required for repository upload,
- assist with the creation of data documentation required for repository upload and reuse of data, and
- indicate parties who might be interested in the data.

Past researchers are not the only parties who would be able to assist with the reading of the data, and the correct use of such older readers. Current researchers who had previously dealt with such a data format or data reader should also be regarded as suitable stakeholders. Similarly, library and information services professionals or archival experts, experienced in dealing with the readers and data concerned, should also be considered and approached. The input of the research library, including archival staff, is discussed in Section 6.4.8 and Section 6.4.11.

6.4.11 Promote the role of expert volunteers

Adding on to the recommendation that the role of citizen scientists should be promoted (see Section 6.4.9) is the recommendation that the crucial input of expert volunteers should be actively encouraged and pursued. In the context of this study, the concept of 'expert volunteers' refers to members of the public with skills, experience or knowledge relevant to the data rescue project, the data, and its required activities.

The contributions of expert volunteers during US data refuge events are documented widely, with participants comprising a range of skilled professionals. Software engineers, programmers, environmental studies professors, scientists and historians, the coding community, web developers and archivists, and digital librarians were the most documented skilled participants. Their data rescue contributions entailed:

- backing up datasets and documents from applicable websites dealing with environment and climate data,
- experts from the coding community downloading and scraping datasets,
- 'harvesters' writing code in three different languages to scrape data on a range of environmental topics,
- 'storytellers' redesigning relevant government websites, and
- 'storytellers' creating a visualisation of relevant tweets over time.

Within the context of this study, it is anticipated that the expert volunteers would include the following sectors or professions:

- academia,
- archive experts,
- the technology sector,
- non-profit organisations,
- retired researchers, and
- retired library and information services professionals.

The anticipated contributions of expert volunteers would comprise the following:

- assess the value of the data,
- assist an institute's data librarian in the decision-making process regarding the feasibility of data rescue,
- suggest potential users of the rescued data,
- advise on the correct handling and use of the data format and its linked readers or equipment,
- assist with deciphering the filenames, folders and structure of metadata linked to the data at risk,
- assist with the creation of new metadata if required,
- assist with the creation of updated filenames and folder structure of rescued data,
- advise on the selection of a suitable data repository, and
- volunteers with exceptional marketing skills, writing skills or social media publishing experience will be involved in the promotion of the project and the rescued data.

The bulleted lists above demonstrate the potential for expert volunteers to be involved in a range of data rescue tasks. The participation and contribution of these individuals should certainly be considered in an environment where institutional researchers are under tremendous workload, with data rescue projects potentially being viewed as an intrusion on research resources.

6.4.12 Amend and adapt the LIS curriculum to include data rescue

It is recommended that the topic of 'data rescue' (including 'data at risk') forms part of a relevant early postgraduate LIS module. In the South African context, such a module would ideally form part of the honours course and be included in the latter segment of the Research Data Management module.

With the crucial role of library and information services experts already indicated in the final Data Rescue Workflow Model, a next step requires that staff be trained in the required data rescue tasks. While some of the tasks will show similarities with tasks already being performed by library and information services professionals (e.g., uploading data to a repository, ensuring metadata are complete), many of the roles indicated in the workflow entail the research library professional understanding the entire data rescue process. In addition, certain activities will require either on-the-job training (such as discussed in Section 6.4.8), or more formal training during pre-employment years at a tertiary institute.

It is therefore recommended that the topics of data at risk and data rescue form part of the LIS curriculum. It is further recommended that the data rescue training be done at a level where students would have already dealt with research data or been exposed to a module related to research data management. It is not advisable to implement data rescue as an LIS topic earlier in the curriculum; experience with and an understanding of research data is crucial. Students should already have come across and experienced concepts such as file formats, secure data storage, data licensing, metadata, data documentation and data repositories before being presented with a lecture on data at risk and data rescue.

A final recommendation pertains to the administering of a test to evaluate learners' grasp of the topic. Recurring and identical gaps in data rescue insights among most learners will also indicate that the topic presentation requires modification and adaptation.

6.4.13 Emphasise the role of collaboration in data rescue

Collaboration and sharing of resources, manpower and skills are often mentioned during documented data rescue outputs. It is therefore recommended that future data rescue projects at research institutes consider the option of collaborative efforts, especially since this study's interview findings revealed that participants were concerned about data rescue costs, manpower, time and expertise. Benefits emanating from collaborative data rescue efforts include an increase in data rescue expertise and a reduction in the need to procure equipment and infrastructure. The easing of demands on available data rescue resources, such as manpower and time, is another benefit.

Collaboration during a rescue project can involve many shapes and forms; several options are listed as follows:

- consider the outsourcing of certain data rescue activities to a vendor (e.g., making use of professional digitisation services),
- consider involving different sectors within the same institute in a rescue project (e.g., SET base, library and information services sector, ICT services),
- consider implementing research projects involving different institutes yet having certain research disciplines in common,
- consider sharing rescue equipment within the same institute (e.g., scanners, digital cameras, tripods),
- consider making use of citizen scientists when the rescue project includes tasks that do not require rescue-specific expertise, or when understanding and mastery of tasks can be conveyed via training, and
- consider making use of past researchers or expert volunteers (e.g., postgraduate students) when the rescue project requires subject understanding or expertise.

The collaborative nature during US data refuge events underlines the beneficial effects of involvement of a range of experts and skills, with such events often including library spaces, digital librarians, scientists, programmers, software engineers, archivists, social media experts, journalists, data rescue advocates, students and citizen scientists. It can be argued that the community-driven and collaborative nature of the projects was a factor contributing to the success and popularity of the movement.

In SA, the successful and well-publicised rescue of apartheid era datasets involved the combined participation of three different university libraries and a university labour studies unit, while the project was managed by a prominent South African data services entity.

At the selected institute, the only instance of data rescue collaboration ascertained involved historic audio data digitisation performed by the digitisation unit of a nearby university, and the digitised data uploaded to a discipline repository hosted by the national discipline entity.

6.4.14 Launch a data rescue project

Launching a data rescue project is an important recommendation and is a natural progression of activities after demonstrating the Data Rescue Workflow Model to relevant parties at awareness sessions. The prevalence of data at risk at not only the selected institute, but also globally and in a range of research disciplines has led to the recommendation that institutes should launch a data rescue project and make use of this study's recommended Data Rescue Workflow Model.

The implementation of such a project is anticipated to consist of two main steps, namely the active location of data at risk at an institute, and the launch of a data rescue pilot project. More details about each of these main steps are provided below.

- Locate data at risk:
    - these activities should be preceded by the data at risk awareness session (See Section 6.4.1),
    - discovery or knowledge of data at risk should be conveyed to the data librarian, who would also be the person responsible for training during data at risk awareness sessions,
    - correct handling of the data at risk should be paramount at this stage of data discovery, to minimise loss of or damage to data,
    - the data librarian should assemble a data assessment team, who would be responsible for assessing the data, and
    - following data assessment, the feasibility of a data rescue project, considering the available resources (skills, equipment and infrastructure), should be determined.
- Launch the data rescue pilot project:
    - should data be worthy of rescue and data rescue resources available, the rescue should commence,
    - library and information services sector involvement is expected, recommended, and crucial for most of the data rescue stages; details about involvement (positions, roles and responsibilities) can be viewed in Table 6.2,
    - the first data rescue project should be viewed as a pilot run and ideally include the full rescue of one dataset, and the partial rescue of a second dataset, and
    - the data rescue project should be included in the strategic plan of an institute's research library.

Future additions, amendments and adaptations to the model are anticipated to be a natural outcome of the implementation of the pilot project. These potential future changes should be viewed in the same light as the need for further research – an issue discussed in the next section (Section 6.5).

6.5 The potential and need for further research

This research paves the way for several follow-up studies or projects and the following are recommendations for future research topics:

- The model should be tested at different research institutes, and the way in which disciplinary differences influence the relevance or practicality of the model should be established.
- Emanating from the previous point: data rescue research that has made use of the model (or even evaluated the model) should indicate the stages and steps that would need to be adapted, based on their institute size and workforce, research library size and workforce, budget and available resources.
- The model will need to be adapted to accommodate the unique data-related requirements of certain research disciplines.
- Additional further research is anticipated to include collaborative data rescue projects, thereby involving researchers and library and information services staff outside of the selected research institute with rescue ventures.
- Research addressing the limitations of this study is another potential follow-up opportunity.
- A data rescue workflow model integrated with the institute's electronic workflow system and other bespoke online platforms is a long-term project envisaged by this researcher.

Two of these potential future outcomes, namely discipline-specific adaptations and the automation of the model, are addressed in the remainder of this section. These outcomes were selected as they are regarded as the options most likely to be addressed and implemented by the selected institute in the near future.

Discipline-specific adaptions and versions of the model will most likely contain guidelines, recommendations and stipulations varying between helpful information and mandatory practice. Examples of beneficial discipline-specific details will include details about funding and assistance, details about digitisation best practices, and details about metadata standards and data repositories. For climate data rescue, details about funding and assistance provided by IEDRO (2014), ACRE (2017) or the European Union (Brandsma, 2006) will form part of a discipline-specific rescue model. The transcription efforts of satellite data rescuers will be significantly boosted when making use of a software program that was developed using a game-like interface to keep participants engaged while automating as much of the work as possible (Gallaher, 2015). For soil data rescue, details about the metadata standard described by Arrouays (2017) will prove helpful. Biodiversity data rescue projects will benefit from knowledge regarding the reBIND data rescue workflow (Güntsch *et al*., 2012). Seismic data rescue projects need to be informed about the number of commercial (e.g., ImageToSEG-Y) and open-source (e.g., IMAGE2SEGY or Seismic Unix and NetPBM) software that exist to convert seismic images to digital SEG-Y data (Diviacco *et al*., 2015). For hydrometric data, the development and availability of the NUNIEAU software since 2006, being a virtual digitiser with no paper size limit, is an international reference in tide gauge, river and rainfall records (Pons *et al*., 2016).

While the subject-specific guidance and recommendations linked to funding, data formats, data rescue software and data repositories are helpful, the implications of non-adherence are anticipated to be less severe than non-adherence to another data rescue activity, namely, sharing of data. The importance of a discipline-specific data rescue model containing specifications around the sharing of data cannot be underestimated, as transgressions linked to non-permissible data sharing can have dire contractual implications, legal ramifications and loss of research credibility. As such, the question, 'How can a data rescue model accommodate sensitive data?' is one that needs to be considered carefully.

Limitations and restrictions on data sharing requirements often feature in disciplines such as the health sciences, ethnographic studies or sub-sections of ecology. With the recommended data rescue model featuring the sharing of data as one of its main rescue stages, it is anticipated that future adaptations of the model will make provision for the treatment of data that may not be uploaded to an open access repository and may not be shared with the general public.

According to Author Services, the following data should not be shared (2023):

- when sharing data conflicts with a need to protect personal identities,
- when the researcher does not have ownership of the data,
- where data are commercially sensitive or protected by competition laws or market regulation,
- where release of the data poses a security risk,
- when data are *sub judice*, and
- when data can compromise the safety of threatened species.

Should any of the data listed above form part of a data rescue project, such data should not be shared to an open access platform. Efforts should be made to obtain data documentation linked to the sharing limitations, contractual agreements, consent forms, or metadata providing details around data access and data sharing. Obtaining details about the rescued data's original funder may also enable the data rescue team to make an informed data sharing decision.

With the NIH one of the first entities to mandate the creation of DMPs it is worthwhile examining their policy regarding data sharing (National Institutes of Health, 2022). For research involving human participants, the NIH has specific requirements for research staff, and policies regarding research conduct, safety monitoring, and reporting of information about research progress. The NIH also states that applicants need to follow all applicable federal, tribal, state, and local laws, regulations, statutes, guidance, and institutional policies that govern research involving human participants and the sharing and use of scientific data derived from human participants. The NIH also respects tribal sovereignty, even in the absence of written tribal laws or policies. Researchers are expected to take steps to protect the privacy, rights and confidentiality of participants via de-identification, certificates of confidentiality, and other protective measures. With regard to data sharing, the NIH stipulates that researchers should assess limitations on subsequent use of data and communicate these limitations to the individuals or repositories preserving and sharing the data. Furthermore, it is stated that researchers should consider whether access to shared scientific data derived from humans should be controlled, even when the data are de-identified and lacking explicit limitations on subsequent use.

When rescuing data at risk that are linked to indigenous parties or minorities it is anticipated that such a project's data sharing requirements will be unique and project-specific. Kowal, Llamas and Tishkoff state that when it comes to data sharing, the generic broad-consent models used for human studies, which leave decisions on data sharing to the researchers, may not be appropriate for work with indigenous peoples (2017). Carroll *et al*., in a COVID-19 data study, add to this idea by mentioning that governments, non-profits, researchers and other institutions must collaborate with indigenous peoples on their own terms to improve access to and use of data (2021).

An article by Welz explains how the sharing of threatened species data can put species at further risk (2017). The author explains that even giving species data to a trusted person means that the data are at risk, as there have been many instances of fake biologists seeking data on rare species. Additional recommendations linked to sharing of such data include the urgent unlearning of the 'centuries-old publishing culture' and to rethink the benefits of publishing location data and habitat descriptions for rare and endangered species (2017). The implications of this article for data rescue are the following: while valuable data about threatened species can be digitised and converted, the value of the data to poachers and collectors should also be considered. In addition, the removal of sensitive location and habitat data is a way of preparing the data for a wider audience. Lastly, and also applicable to sensitive human data, the rescue activities of converting, describing and preserving data, albeit in an archive not accessible to the public, ensure that the data can be consulted by the researcher should verification be required, or if the data are deemed to be shareable in the future.

Meyer has provided several tips regarding the ethical sharing of data when the consent form either is silent about data sharing or explicitly promised participants that the data would not be shared (2018). It is likely that this issue will be encountered during data rescue, as the data documentation linked to older data at risk is not always located or clear. Meyer proposes that data sharing in these circumstances be determined on a case-by-case basis, and that in general, the argument for sharing will be stronger the more of the following conditions are present:

- the original consent form was merely silent about data sharing, and did not include a promise not to share data,
- the data are not especially sensitive (i.e., re-identification would be unlikely to cause significant harm),
- the data are not individually identified and are not especially likely to be re-identified,
- the shared data will be accessible only under restricted conditions, protected by agreements prohibiting re-identification,
- sharing will be limited to secondary research purposes that fall within the scope of the research described in the original consent form, and
- sharing will be limited to secondary research purposes participants are not known to object to.

The study of Dong, Ilieva and Medeiros reports on an interesting development that involves unlocking the potential of historical patient data (2018). Beginning in 2016, historians and archivists from several institutions (Johns Hopkins University, Harvard University, Western Michigan University, the Wellcome Trust, Yale University and the New York Academy of Medicine Library) came together to brainstorm about the possibility of developing a scalable database and discovery tools for older patient data. The project aims to make digitised historical records more readily discoverable and accessible to a wide variety of users, including staff, patients' family members, researchers and the general public. Differing levels of access to the records form part of the project and will depend on the type of information requested and the credentials of the information seeker. The implications of this study for data rescue is the realisation that although the data are potentially sensitive in nature, processes and steps can be put in place, thereby enabling reuse of the valuable information contained within the data. Clearly, decisions around the sharing of the data will be made during the earlier stages of any data rescue project, during the assessment of data and more formally when the DMP for the data involved in the project is drafted.

The examples above give an idea of how different disciplines, be it ethnic studies, health studies, ecology, environmental studies or psychology, view the sharing of sensitive data. The examples have been included here to illustrate how this important data rescue activity – sharing of rescued data – is not identical in all disciplines and that future adaptations of the model will need to include disciplinary guidance around the ethical sharing of sensitive information. With the majority of published data rescue studies involving the use of open access archives and the sharing of data with the public, the sharing of sensitive data is not a topic yet addressed in data rescue literature. This researcher anticipates that while future adaptations of the model may include helpful guidance regarding discipline-specific data rescue aspects such as funding, assistance and locating data, or recommended guidance around discipline-specific metadata standards and repositories, the most crucial changes to the model will involve (i) the data sharing decisions made during the assessing and planning phases of data rescue, and (ii) implementation of the sharing limitations and restrictions during the sharing stage of the rescue project.

The recommended data rescue workflow model displays potential to be automated and is a course of action that will be investigated at the selected institute once the transition period between current and future document management systems, and current and future closed institutional repository platforms has ended. Systems and processes that could potentially be included in the automated version of the model are listed in the table below.

Table 6.4: Potential institutional systems to be accommodated during model automation

SYSTEM, TOOL OR PLATFORM	DATA RESCUE ACTIVITY
Institutional 'Workflow' system (not to be confused with the workflow demonstrated in the data rescue model)	• Used by researchers to initiate a new data rescue request • Used by researchers to add metadata linked to the data at risk • Used by the researcher to provide links to uploaded SharePoint images/video/audio • Used by research library to request additional metadata • Used by research library to indicate progress of rescue project • Used by research library to reassign or rework tasks
SharePoint	• Contains rescue project folder (with project plan, DMP, details about team members) • Contains data inventories • Contains digitised/converted data • Contains any documentation linked to the data that cannot be accommodated via Workflow
MS Teams	• Default platform for online discussions and meetings regarding the project • Platform is also used to schedule in-person visits
Inmagic: TOdB Old	• Closed institutional repository for the metadata of scientific outputs created before 2004
Inmagic: TOdB New	• Closed institutional repository for the metadata of scientific outputs created during and after 2004
DSpace	• Open access repository for institutional scientific outputs • Records contain the uploaded record with its metadata, or the metadata only • A unique handle is created by the system
DMP tool	• Default platform used when creating a DMP for the data • DMP will be downloaded and stored in SharePoint
Institutional intraweb news page	• To be used for internal promotion of rescued data
Institutional social media platforms	• To be used for external promotion

It is anticipated that an automated data rescue workflow will include all of the systems, tools and platforms listed in the table, and that a typical data rescue stage will involve moving back and forth between the various systems. A hypothetical first stage of data rescue, comprising preparing for data rescue, is shown in the figure below and indicates the potential inclusion and interplay of various institutional systems in an automated process.

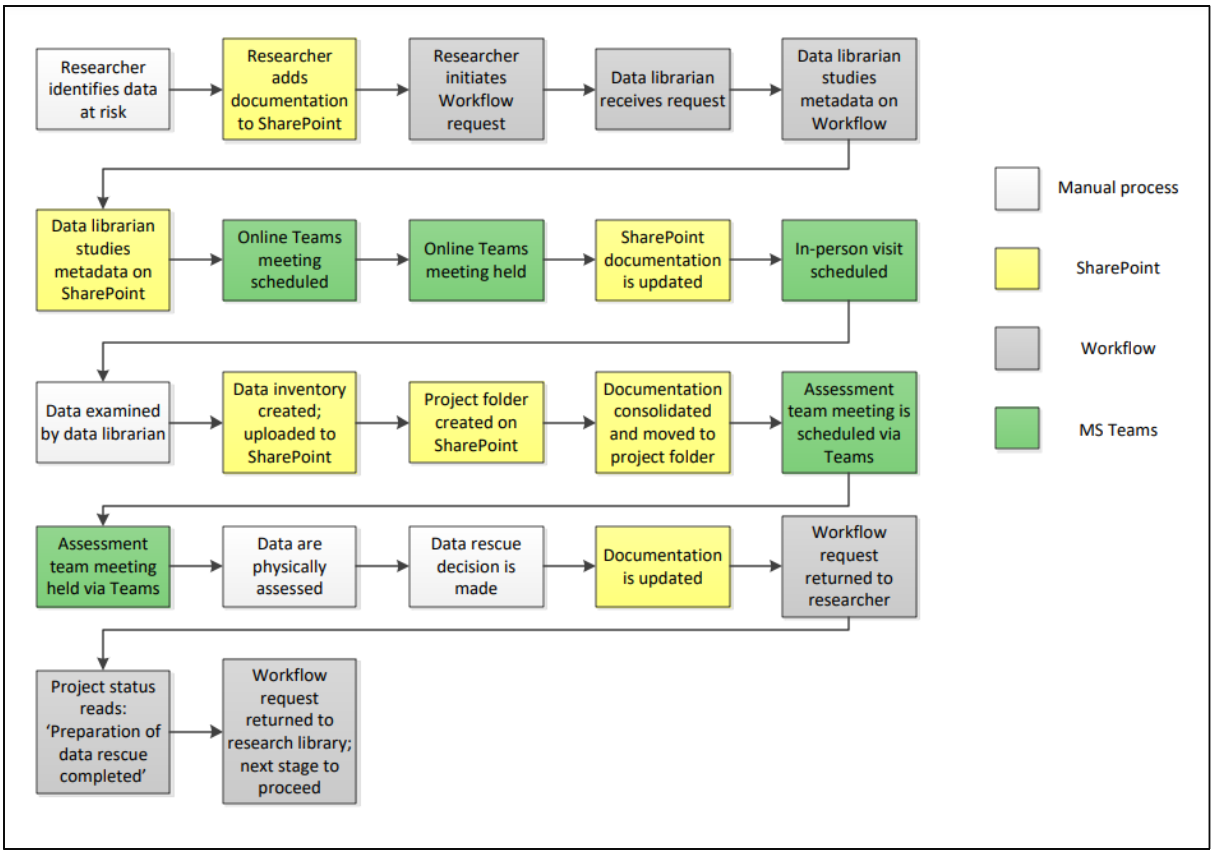

Researcher identifies data at risk
Researcher adds documentation to SharePoint
Researcher initiates Workflow request
Data librarian receives request
Data librarian studies metadata on Workflow
Data librarian studies metadata on SharePoint
Online Teams meeting scheduled
Online Teams meeting held
SharePoint documentation is updated
In-person visit scheduled
Data examined by data librarian
Data inventory created; uploaded to SharePoint
Project folder created on SharePoint
Documentation consolidated and moved to project folder
Assessment team meeting is scheduled via Teams
Assessment team meeting held via Teams
Data are physically assessed
Data rescue decision is made
Documentation is updated
Workflow request returned to researcher
Project status reads: 'Preparation of data rescue completed'
Workflow request returned to research library; next stage to proceed
Manual process
SharePoint
Workflow
MS Teams

Figure 6.23: Example of interplay of various institutional systems

The figure above is a hypothetical portrayal of Stage 1 of eight stages of a generic data rescue project, and shows the potential inclusion of three of the systems listed in Table 6.4. As a simple figure it does not yet show the inclusion of a digital data assessment checklist, or any links to external systems that might be required. Automation of the process will require that the platform integrates a range of different institutional systems, should at least direct participants to external platforms (e.g., a disciplinary repository or long-term data archive), should enable the various data rescue participants to move seamlessly between systems, will prompt participants about next steps, allow the use of proxies, should contain 'rework' or 'reassign' features and contain clickable links to guidance and mandatory requirements. The automation of the data rescue workflow will be a welcome upgrade to the recommended model, as it is bound to simplify the many pages of diagrams and instructions currently used to convey the various data rescue stages, activities and outputs.

The next section comprises a conclusion of the study, presenting a summary of the research and explaining the study's relevance to the LIS field.

6.6 Study conclusion

This research endeavoured to establish the current data rescue activities and experiences of researchers and library and information services staff, both globally and at a selected research institute in South Africa. This investigation was done to put forward suggestions and recommendations on how to address and improve the data rescue (and data at risk) status quo, and devise a Data Rescue Workflow Model, where the roles and responsibilities of various stakeholders, including research library personnel, were indicated. To ascertain the data rescue experiences and data workflows used, it was necessary to study (a) data at risk at the selected institute and globally, (b) data rescue practices at the selected institute, elsewhere in SA, and globally, and (c) data rescue workflows or frameworks used at the selected institute, elsewhere in SA, and globally. By doing so, it was possible to create a Data Rescue Workflow Model indicating the roles and responsibilities of various stakeholders, including library and information services staff.

The main objective of this study was to contribute towards the rescue of data at risk by establishing ways in which parties, other than researchers and scientists, may be involved. Before this objective, and the various sub-objectives could be addressed, several research questions had to be answered. The research findings enabled the answering of the study's research questions, with this study reporting on the following:

- the prevalence and nature of data at risk globally as well as at the selected research institute,
- global data rescue experiences and data rescue experiences at the selected institute,
- state of data rescue as established globally, in SA, and at the selected institute,
- data rescue frameworks and models used globally and in SA, and
- library and information services involvement in data rescue and in workflows.

The study's literature review also enabled the drafting of an initial Data Rescue Workflow Model and reported on the feedback received from the selected institute's SET-based component after scrutinising the model. After revising the initial model, the amended version was reviewed by research library experts at the selected research institute. Based on all feedback received, a recommended Data Rescue Workflow Model was created. This data rescue model addresses the study's main research question, namely:

'What are the roles and responsibilities of the research library within a comprehensive workflow for data rescue?'

In general, it was found that the prevalence and nature of data at risk at the selected institute showed much similarity with global trends. While the literature review showed that data rescue projects had been implemented all over the globe, the situation at the selected research institute was different, with only one example of such a project detected through various empirical data collection methods. Documented data rescue projects of other South African entities were found, with university researchers and research departments, a well-known South African research data service, and a national discipline-specific research centre being the involved stakeholders. Regarding library and information services involvement globally, it was found that the research library community, and libraries, were involved in many data rescue activities and studies. In addition, the recommended involvement of library and information services professionals in data rescue activities, and on several levels were documented. Research library involvement in data rescue was not present at the selected research institute, however, documented proof of academic libraries' participation was mentioned in a data rescue project involving sociological datasets, with university libraries being gatekeepers of the data at risk, and involved in the actual data rescue efforts.

While 'complete' data rescue was not common at the selected research institute, participants were able to provide rudimentary details regarding loose-standing data rescue activities performed, data rescue challenges and obstacles experienced, and the aspects they require from a usable data rescue workflow model. Library and information services experts at the selected research institute were also involved in the study and provided recommendations and suggestions regarding anticipated and realistic research library involvement when reviewing the revised model.

When considering the findings of the research, the research questions asked and answered, and the related topics and concerns emanating from these questions, it was necessary to put forwards several suggestions and recommendations. These recommendations are related to data at risk, data rescue practices, and the use of the final Data Rescue Workflow Model. Most of the recommendations are applicable to a range of research or library environments and are put forward to ensure that research institutes apply best practices when locating data at risk and performing data rescue activities. Suggested recommendations are anticipated to minimise data rescue challenges, reduce the prevalence of data at risk, and lessen the need for data rescue. While the recommendations emanate from empirical data collected from researchers linked to the selected research institute, their applicability is anticipated to be of value to the wider research environment, the wider library and information services community, and to relevant research institutes (including tertiary institutes of education).

Recommendations entail an increased awareness around the topic of data at risk, understanding of factors leading to data being at risk, and exposure to the correct handling of such data prior to data rescue. Recommendations also touch on heightened awareness and understanding of the process of data rescue, the terminology used, and in conveying the message that adherence to good data management practices will minimise the need for data rescue. Awareness of these concepts should be seen as a crucial skill in the research environment, even when no rescue ventures are imminent.

With the creation of the final Data Rescue Workflow Model being a major outcome of this study, it was crucial that a recommendation be made to implement a data rescue project making use of the model. The inclusion of library and information services professionals and semi-professionals in various stages of the data rescue model is one of the model's major features, with this sector anticipated to be involved with activities such as training, creating documentation, conventional data management activities, and supervising the relevant data rescue stages. As a result of this involvement of the library and information services sector, a recommendation regarding the importance of data rescue in the LIS curriculum was also made. Additional recommendations regarding the importance of expert volunteers, past researchers, citizen scientists and collaborative parties were presented.

Suggestions for further research emanating from the current study include addressing the limitations of the study and involving researchers with data rescue experience, even outside of the selected research institute, in drafting discipline-specific versions of the data rescue model. Aspects such as the recommended metadata standard, discipline repositories to be used, and intra-disciplinary promotion of the rescued data will feature in discipline-specific models. Additional further research is anticipated to include collaborative data rescue projects, thereby involving researchers and library and information services staff outside of the selected research institute with the venture. Reassessment and evaluation of the data rescue model created during this study will be a mandatory activity following any future data rescue projects.

Several study limitations could be found, the most pressing of which relates to the lack of data rescue experience of most researchers and research library experts involved in the empirical phase of the study. Many of the participants acknowledged the need for data rescue, even though only a single instance of a fully-fledged data rescue project at the selected institute could be ascertained via the one-on-one interviews with RGLs. Furthermore, although the findings reveal a lack of frequent or numerous data rescue projects at the selected institute, evidence was found of singular activities forming part of a data rescue project. The lack of data rescue experience meant that much of the feedback and recommendations regarding a usable data rescue workflow model were hypothetical or theoretical in nature, and not based on having participated in data rescue ventures. Apart from much of the feedback being theoretical in nature, the less-than-anticipated feedback volume after reviewing the initial data rescue model can also be ascribed to lack of data rescue undertakings.

This researcher also foresees follow-up research emanating from the recommendations put forward. A machine-actionable implementation of the Data Rescue Workflow Model, and adapted versions of the model based on discipline-specific requirements are examples of future work on the topic at hand.

The significance of this study lies in the filling of many current gaps in knowledge, as it reports on the factors leading to data at risk at a selected multidisciplinary research institute, and the prevalence of data at risk at said institute. The study also delves into the data rescue obstacles and challenges experienced by experienced researchers. Furthermore, after reporting on a range of data rescue workflows and frameworks documented globally and incorporating the requirements and suggestions of different data rescue parties, the study produced a Data Rescue Workflow Model indicating the roles and responsibilities of various participatory parties, including the library and information services sector.

Recommendations put forward involve the promotion of awareness around data at risk and data rescue, leading to a greater institutional understanding of the role and value of collaboration during data rescue ventures. These steps should be followed by the launch of a data rescue project. Linked to this recommendation is the inclusion of the library and information services sector in data rescue projects, with the valuable input of past researchers, citizen scientists and expert volunteers also proposed.

This study was successful in that it managed to address all the research questions, attained the research objectives put forward, and in doing so contributed towards the theoretical knowledge base of the LIS discipline. The study managed to show how data rescue is currently carried out, both locally and internationally. It was also possible to determine who is currently involved in data rescue, and their roles. Current data rescue perceptions, needs and challenges were addressed via the findings, and a Data Rescue Workflow Model was created based on the information gained via the previous points. Additional significance of this study is ascribed to the model indicating options for either full or partial data rescue. Furthermore, within the model, it was stipulated how library and information services professionals can be involved in data rescue. Launching a pilot project is one of the recommendations of the study, with an adaptation of the model based on project outcomes envisaged. Lastly, a contribution towards future data rescue will be made, as the model will be assigned a Creative Commons Attribution-ShareAlike license, thereby making it free for use and adaptation by data rescue stakeholders.

It is anticipated that the recommendations of this study, coupled with the availability of a freely available Data Rescue Workflow Model, will encourage the undertaking of data rescue ventures. Adding to this is the anticipation of similar research at comparable institutes, with adaptations, amendments and improvements to the model suggested and implemented. It is also foreseen that the study findings will not only serve as guidance during data rescue projects, but will also encourage the active participation of the research library in future data rescue ventures.

REFERENCES

ADAMS, K.A. & LAWRENCE, E.K. 2015. *Research Methods: Statistics, and Applications*. Thousand Oaks, California: Sage.

ADMIN ADMIN. 2021. Emerging Trends & Technologies in Library & Information Services. *Algorhythms Consultants,* 10 August. Available from: https://slimkm.com/blog/emerging-trends-technologies-in-library-information-services/. Accessed on 8 January 2023.

ADOLPHUS, M. 2021. *How to…Undertake case study research*. Bingley, UK: Emerald Publishing. Available from: https://www.emeraldgrouppublishing.com/how-to/research-methods/undertake-case-study-research. Accessed on 4 June 2021.

ADU, P. 2017. *Difference between Delimitations, Limitations, and Assumptions*. SlideShare. Available from: https://www.slideshare.net/kontorphilip/difference-between-delimitations-limitations-and-assumptions-73285262. Accessed on 8 December 2019.

AFRIBARY. 2020. *Conceptual Framework - Meaning, Importance and How to Write it*. Available from: https://afribary.com/knowledge/conceptual-framework/. Accessed on 28 December 2022.

AGUINIS, H. & SOLARINO, A.M. 2019. Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, vol. 40(8): 1291–1315. Available from: https://doi.org/10.1002/smj.3015.

ALLAN, R., BROHAN, P., COMPO, G.P., STONE, R., LUTERBACHER, J. & BRÖNNIMANN, S. 2011. The International Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative. *Bulletin of the American Meteorological Society,* vol. 92(11): 1421–1425. Available from: https://doi.org/10.1175/2011BAMS3218.1.

ALLAN, R. & CROUTHAMEL, R. 2013. Data to the rescue. *International Innovations*, vol. 124: 6–11. Available from: http://iedro.org/wordpress/wp-content/uploads/2014/01/ACRE-IEDRO-InternationalInnovation.pdf. Accessed on 26 January 2022.

ALLAN, R. & WILLETT, K. 2017. Overview of the C3S Data Rescue Service. *C3S DRS Capacity Building Workshop, Auckland, New Zealand, 4–8 December 2017.* Available from: http://about-c3s-dr.eu/pages/workshop_material/AllanRWillettK_C3SDRS_CB1Dec2017.pdf. Accessed on 26 January 2022.

ALLEN, C., ALLTON, J., LOFGREN, G., RIGHTER, K. & ZOLENSKY, M. 2011. *Curating NASA's extra-terrestrial samples – past, present, and future*. Available from: https://www.lpi.usra.edu/meetings/sssr2011/pdf/5006.pdf. Accessed on 18 January 2022.

ALLEN, L., STEWART, C. & WRIGHT, S. 2017. Strategic open data preservation: Roles and opportunities for broader engagement by librarians and the public. *College and Research Libraries News*, vol. 78(9): 482–485. Available from: https://doi.org/10.5860/crln.78.9.482.

AMERICAN LIBRARY ASSOCIATION, ALA STORE. 2014. *Data Management for Libraries: A LITA Guide*. Available from: https://www.alastore.ala.org/content/data-management-libraries-lita-guide. Accessed on 15 January 2023.

AN, V.A., OVTCHINNIKOV, V.M., KAAZIK, P.B., ADUSHKIN, V.A., SOKOLOVA, I.N., ALESCHENKO, I.B., MIKHAILOVA, N.N., KIM, W-Y., RICHARDS, P.G., PATTON, H.J. *et al*. 2015. A digital seismogram archive of nuclear explosion signals, recorded at the Borovoye Geophysical Observatory, Kazakhstan, from 1966 to 1996. *GeoResJ*, vol. 6: 141–163. Available from: https://doi.org/10.1016/j.grj.2015.02.014.

ANDRIKOPOULOU, A., ROWLEY, J. & WALTON, G. 2022. Research Data Management (RDM) and the Evolving Identity of Academic Libraries and Librarians: A Literature Review. *New Review of Academic Librarianship*, vol. 28(4): 349–365. Available from: https://doi.org/10.1080/13614533.2021.1964549.

ANTELL, K., FOOTE, J.B., TURNER, J. & SHULTS, B. 2014. Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions. *College & Research Libraries*, vol. 75(4): 55–574. Available from: https://crl.acrl.org/index.php/crl/article/view/16374. Accessed on 7 January 2023.

ANTICO, A., AGUIAR, R.O. & AMSLER, M.L. 2018. Hydrometric Data Rescue in the Paraná River Basin. *Water Resources Research*, vol. 54(2): 1368–1381. Available from: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017WR020897.

ANTUÑA, J.C., FONTE, A., ESTEVAN, R., BARJA, B., ACEA, R. & ANTUÑA, J.C (Jnr). 2008. Solar Radiation Data Rescue at Camagüey, Cuba. *Bulletin of the American Meteorological Society*, vol. 89(10): 1507–1512. Available from: https://www.jstor.org/stable/26220899. Accessed on 26 January 2022.

APPEL, R. 2017. DataRescue Philly: Environmental Data Archiving, Workflows, and Description. *ProjectARCC*, 26 January. Available from: https://projectarcc.org/2017/01/26/datarescue-philly-environmental-data-archiving-workflows-and-description/. Accessed on 26 January 2022.

APPLIED DOCTORAL CENTER. 2022. *Trustworthiness of the Data*. Available from: https://resources.nu.edu/c.php?g=1013606&p=8394398. Accessed on 28 December 2022.

ARROUAYS, D., LEENAARS, J.G.B., RICHER-DE-FORGES, A.C., ADHIKARI, K., BALLABIO, C., GREVE, M., GRUNDY, M., GUERRERO, E., HEMPEL, J., HENGL, T. *et al*. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ,* vol. 14: 1–19. Available from: https://doi.org/10.1016/j.grj.2017.06.001.

ASHCROFT, L. 2016. Historical climate data rescue activities in Australia. *Historical Climatology,* 10 July. Available from: https://www.historicalclimatology.com/projects/linden-ashcroft-historical-climate-data-rescue-activities-in-australia. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2010. *Chilean Chapter*. Available from: http://www.met-acre.org/data-projects-and-regional-chapters/chilean-chapter. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2016a. *ACRE China*. Available from: http://www.met-acre.org/data-projects-and-regional-chapters/acre-china. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2016b. *Data Rescue: Citizen Science*. Available from: http://www.met-acre.net/citsci.htm. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2017. *ACRE Worldwide*. Available from: http://www.met-acre.net/chapters.htm. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2019. *Data Rescue: ACRE's Commitment*. Available from: http://www.met-acre.net/data%20rescue.htm. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2020a. *2019–2020 ACRE Regional Data Rescue activities*. Available from: https://sites.google.com/a/met-acre.org/acre/Home/2019-2020%20ACRE2%20%20.png?attredirects=0&d=1. Accessed on 26 January 2022.

ATMOSPHERIC CIRCULATION RECONSTRUCTIONS OVER THE EARTH (ACRE). 2020b. *About ACRE*. Available from: http://www.met-acre.org/. Accessed on 26 January 2022.

AUER, I., CHIMANI, B. & the EUMETNET Data Rescue Expert Team. 2014. The EuMetNet Data Rescue portal. *14th EMS Annual Meeting & 10th European Conference on Applied Climatology (ECAC), Prague, Czech Republic, 6–10 October 2014.* Available from: https://www.ems2014.eu/ems2014_programme_book.pdf. Accessed on 26 January 2022.

AUTHOR SERVICES. TAYLOR & FRANCIS GROUP. 2023. *A short introduction to data sharing ethics*. Available from: https://authorservices.taylorandfrancis.com/data-sharing/data-sharing-ethics/. Accessed on 8 January 2023.

AXELSSON, A-L., ÖSTLUND, L. & HELLBERG, E. 2002. Changes in mixed deciduous forests of boreal Sweden 18661999 based on interpretation of historical records. *Landscape Ecology*, vol. 17(5): 403–418. Available from: https://doi.org/10.1023/A:1021226600159.

BARBROW, S., BRUSH, D. & GOLDMAN, J. 2017. Research data management and services: Resources for novice data librarians. *College & Research Libraries News*, vol. 78(5): 274. Available from: https://crln.acrl.org/index.php/crlnews/article/view/16660.

BARNARD LIBRARY & ACADEMIC INFORMATION SERVICES. 2019. *A different kind of data rescue: Digitizing & Preserving Obsolete Analog Video*. Available from: https://library.barnard.edu/events/Different-Kind-Data-Rescue-Digitizing-Preserving-Obsolete-Analog-Video. Accessed on 26 January 2022.

BBC NEWS. 2016. Toronto 'guerrilla' archivists to help preserve US climate data. *BBC News,* 15 December. Available from: https://www.bbc.com/news/world-us-canada-38324045. Accessed on 26 January 2022.

BEELER, C. 2017. People around the world are helping the US save its climate data. *The World from PRX,* 26 January. Available from: https://www.pri.org/stories/2017-01-26/people-around-world-are-helping-us-save-its-climate-data. Accessed on 26 January 2022.

BEZUIDENHOUT, L. & CHAKAUYA, E. 2018. Hidden concerns of sharing research data by low/middle-income country scientists. *Global Bioethics*, vol. 29(1): 39-54. Available from: https://doi.org/10.1080/11287462.2018.1441780.

BHANDARI, P. 2022. Triangulation in Research: Guide, Types, Examples. *Scribbr*, 3 January. Available from: https://www.scribbr.com/methodology/triangulation/. Accessed on 31 December 2022.

BIODIVERSITY INFORMATICS @ BGBM. 2011. *reBiND Movie*. Available from: https://www.youtube.com/watch?v=rh7bM1kfRsE. Accessed on 6 February 2022.

BOCHNICEK, O., FAŠKO, P., MARKOVIČ, L., PAL'UŠOVÁ, Z. & KAJABA, P. 2017. Data Rescue Approach (Visualisation, Methodology, Examples and Homogenization Scheme) in Slovakia. *11th EUMETNET Data Management Workshop, Zagreb, Croatia, 18–20 October 2017*. Available from: http://meteo.hr/DMW_2017/presentations/posters.pdf. Accessed on 26 January 2022.

BOEHM, R., ABRAHAMS, J., IGNATOWSKI, D., PATTON, M. & CHIU, C. 2017. *DMD Data Rescue Documentation Space/Process Documentation*. Center for Open Science. Available from: https://osf.io/95svk/. Accessed on 1 November 2021.

BRADSHAW, E., RICKARDS, L. & AARUP, T. 2015. Sea level data archaeology and the Global Sea Level Observing System (GLOSS). *GeoResJ*, vol. 6: 9–16. Available from: https://doi.org/10.1016/j.grj.2015.02.005.

BRANDALL, B. 2021. What is a Workflow? A Simple Guide to Getting Started. *Process Street,* 20 October. Available from: https://www.process.st/what-is-a-workflow/. Accessed on 28 October 2021.

BRANDSMA, T. 2007. Data rescue and digitization: tips and tricks resulting from the Dutch experience. *International Workshop on Rescue and Digitization of Climate Records in the Mediterranean Basin, Tarragona, Spain, 28–30 November 2007*. Available from: http://edepot.wur.nl/311376. Accessed on 14 March 2022.

BRAUN, V., CLARKE, V. & RANCE, N. 2015. How to use thematic analysis with interview data. In: Vossler, A. & Moller, N. (*eds*). *The counselling and psychotherapy research book.* London: Sage, pp. 183–197.

BRAUN, V., CLARKE, V. & WEATE, P. 2016. Using thematic analysis in sport and exercise research. In Smith, B. & Sparkes, A.C. (*eds*). *Routledge handbook of qualitative research in sport and exercise.* London: Routledge, pp. 191–205.

BROHAN, P. 2009. Why historical climate and weather observations matter. *ACRE Data and Visualisation meeting, Met Office, Exeter, UK, 15 September 2009*. Available from: https://slideplayer.com/slide/17460565/. Accessed on 26 January 2022.

BROHAN, P., COMPO, G.P., BRÖNNIMANN, S., ALLAN, R.J., AUCHMANN, R., BRUGNARA, Y., SARDESHMUKH, P.D. & WHITAKER, J.S. 2016. The 1816 'year without a summer' in an atmospheric reanalysis. *Climate of the Past Discussions*, vol. 0: 1–11. Available from: https://www.mendeley.com/catalogue/ae13fbd3-e573-35de-949e-9cab66fc51cf/. Accessed on 19 February 2022.

BRÖNNIMANN, S., ANNIS, J., DANN, W., EWEN, T., GRANT, A.N., GRIESSER, T., KRÄHENMANN, S., MOHR, C., SCHERER, M. & VOGLER, C. 2006. A guide for digitizing manuscript climate data. *Climate of the Past*, vol. 2: 137–144. Available from: https://doi.org/10.5194/cp-2-137-2006.

BRÖNNIMANN, S., BRUGNARA, Y., ALLAN, R.J., BRUNET, M., COMPO, G.P., CROUTHAMEL, R.I., JONES, P.D., JOURDAIN, S., LUTERBACHER, J., SIEGMUND, P. et *al*. 2018. A roadmap to climate data rescue services. *Geoscience Data Journal*, vol. 5(1): 28–39. Available from: https://doi.org/10.1002/gdj3.56.

BRUNET, M., BRUGNARA, Y., NOONE, S., STEPHENS, A., VALENTE, M.A., VENTURA, C., JONES, P., GILABERT, A., BRÖNNIMANN, S., LUTERBACHER, J. et *al*. 2020a. *Best Practice Guidelines for Climate Data and Metadata Formatting, Quality Control and Submission of the Copernicus Climate Change Service Data Rescue Service*. Copernicus Climate Change Service. Available from: https://doi.org/10.24381/kctk-8j22.

BRUNET, M., GILABERT, A., JONES, P. & SIEGMUND, P. 2020b. *Identifying Data Rescue gaps and issues. C3S Data Rescue Service Deliverable Number: DC3S311a_Lot 1.1.2.1.* Copernicus Climate Change Service. Available from: http://www.c3.urv.cat/docs/publicacions/2020/C3S_DC3S311a_Lot1.1.2.1_2020_Identification_of_gaps_report.pdf. Accessed on 26 January 2022.

BRUNET, M. & JONES, P. 2011. Data rescue initiatives: bringing historical climate data into the 21st century. *Climate Research*, vol. 47(1–2): 29–40. Available from: https://doi.org/10.3354/cr00960.

BURNETT, P. 2013. *What is the role of a librarian in Research Data Management?* Available from: https://blog.inasp.info/research-data-management-role-librarians/. Accessed on 7 January 2023.

BWESTRA@UOREGON.EDU. 2017. Rescuing unloved data. *UO Research Data Management Blog,* 17 February. Available from: https://blogs.uoregon.edu/datamanagement/2017/02/17/lydweek-2017-5/. Accessed on 19 February 2017.

CAIRNS, I. 2020. Using virtual interviews in your project. *Victoria University of Wellington,* 17 April. Available from: https://www.wgtn.ac.nz/__data/assets/pdf_file/0009/1845963/using-virtual-interviews-in-your-project-HEC-guidance.pdf. Accessed on 29 January 2022.

CALDWELL, P. 2003. NOAA support for global sea level data rescue. *Earth System Monitor*, vol. 14(1): 1–6. Available from: https://www.nodc.noaa.gov/media/pdf/esm/ESM_SEP2003vol14no1.pdf. Accessed on 26 January 2022.

CALDWELL, P.C. 2012. Tide gauge data rescue. In: *Proceedings of The Memory of the World in the Digital age: Digitization and Preservation, UNESCO Conference, Vancouver, Canada, 26–28 September 2012*. Available from: https://www.sonel.org/IMG/pdf/caldwell_2012unesco.pdf. Accessed on 26 January 2022.

CALMA, J. 2021. How scientists scrambled to stop Donald Trump's EPA from wiping out climate data. *The Verge,* 8 March. Available from: https://www.theverge.com/22313763/scientists-climate-change-data-rescue-donald-trump. Accessed on 26 January 2022.

CAPOZZI, V., COTRONEO, Y., CASTAGNO, P., DE VIVO, C. & BUDILLON, G. 2020. Rescue and quality control of sub-daily meteorological data collected at Montevergine Observatory (Southern Apennines), 1884–1963. *Earth System Science Data*, vol. 12: 1467–1487. Available from: https://doi.org/10.5194/essd-12-1467-2020.

CAREY, M. 2017. University Hosts Federal Climate Data Rescue Event. *The Hoya,* 23 February. Available from: https://www.thehoya.com/university-hosts-federal-climate-data-rescue-event/. Accessed on 26 January 2022.

CARROLL, S.R., AKEE, R., CHUNG, P., CORMACK, D., KUKUTAI, T., LOVETT, R., SUINA, M. & ROWE, R.K. 2021. Indigenous Peoples' Data During COVID-19: From External to Internal. *Frontiers in Sociology*, vol. 6: 617895. Available from: https://doi.org/10.3389/fsoc.2021.617895.

CASCONE, S. 2022. How Tech Experts in the West Are Rushing to Save the Digital Archives of Ukraine's Museums. *Artnet Worldwide Corporation,* 14 March. Available from: https://news.artnet.com/art-world/saving-ukrainian-cultural-heritage-online-2084036. Accessed on 21 March 2022.

CHARLESWORTH AUTHOR SERVICES. 2022. *Conceptual framework vs. Theoretical framework – and constructing each*. Available from: https://www.cwauthors.com/article/conceptual-framework-versus-theoretical-framework-in-research. Accessed on 28 December 2022.

CHASSANOFF, A. 2017. Guest Post: Data Rescue Boston@MIT Wrap up. *Micah Altman's Blog,* 2 March. Available from: https://drmaltman.wordpress.com/2017/03/02/guest-post-alex-chassanoff-on-datarescue-bostonmit-wrap-up/. Accessed on 26 January 2022.

CHRONOPOULOU, A., CROSS, K.J., KING, D.M. & SALIMI, E. 2016. Using case studies to enhance the critical thinking skills of IE students. *2016 ASEE Annual Conference & Exposition, New Orleans, Louisiana, 26–29 June 2016.* Available from: https://peer.asee.org/using-case-studies-to-enhance-the-critical-thinking-skills-of-ie-students. Accessed on 8 November 2021.

CLARKE. C.T. & SHIUE, H.S.Y. 2020. *Final Report and Recommendations of the Data Rescue Project at the National Agricultural Library*. University of Maryland, University Libraries, Digital Repository at the University of Maryland (DRUM). Available from: https://drum.lib.umd.edu/bitstream/handle/1903/26363/Final%20Report%20and%20Recommendations%20of%20the%20Data%20Rescue%20Project%20at%20the%20National%20Agricultural%20Library.pdf?sequence=1&isAllowed=y. Accessed on 29 March 2022.

COLUMBIA UNIVERSITY, MAILMAN SCHOOL OF PUBLIC HEALTH. 2019. *Content Analysis*. Available from: https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis. Accessed on 16 November 2021.

COMMITTEE ON DATA OF THE INTERNATIONAL SCIENCE COUNCIL (CODATA). 2013. *Data at Risk*. Available from: https://codata.org/initiatives/task-groups/previous-tgs/data-at-risk/. Accessed on 26 January 2022.

COMMITTEE ON DATA OF THE INTERNATIONAL SCIENCE COUNCIL (CODATA). 2015. *DAR – Past Achievements*. Available from: https://codata.org/initiatives/task-groups/previous-tgs/data-at-risk/dar-past-achievements/. Accessed on 26 January 2022.

COOPER, J. 2015. Rescue, Archive and Stewardship of Weather Records and Data. *Bulletin*, vol. 64(1): 28–30. Available from: https://public.wmo.int/en/resources/bulletin/rescue-archival-and-stewardship-of-weather-records-and-data-0. Accessed on 26 January 2022.

COPERNICUS CLIMATE CHANGE SERVICE (C3S). 2017. *C3S Data Rescue Service Capacity Building Workshop and 10th ACRE Workshop – Dec 2017*. Available from: https://datarescue.climate.copernicus.eu/node/26. Accessed on 26 January 2022.

COPERNICUS CLIMATE CHANGE SERVICE (C3S). 2018. *The Copernicus Climate Change Service Data Rescue Service*. Available from: https://insitu.copernicus.eu/news/the-c3s-data-rescue-service. Accessed on 7 September 2022.

CORTI, L. 2018. *Show Me the Data: research reproducibility in qualitative research*. Available from: https://blog.ukdataservice.ac.uk/show-me-the-data/. Accessed on 7 September 2022.

COSTA, K. 2020. *Systematic Guide to Qualitative Data Analysis: within the COSTA Postgraduate Research Model*. OSF Preprints. Available from: https://scholar.google.co.za/citations?view_op=view_citation&hl=en&user=7tHFG2sAAAAJ&citation_for_view=7tHFG2sAAAAJ:UeHWp8X0CEIC. Accessed on 18 November 2021.

COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR). 2022. *The CSIR in Brief*. Available from: https://www.csir.co.za/csir-brief#:~:text=The%20CSIR%20was%20established%20through,Higher%20Education%2C%20Science%20and%20Technology. Accessed on 20 February 2020.

COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR). 2023a. *Clusters*. Available from: https://www.csir.co.za/. Accessed on 7 January 2023.

COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR). 2023b. *CSIR through the Years*. A Selection of Highlights from our Research and Technological Innovation Journey. Available from: https://www.csir.co.za/our-history. Accessed on 7 January 2023.

COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR). 2023c. *Explore the CSIR*. Available from: https://www.csir.co.za/about-us. Accessed on 7 January 2023.

COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR). 2023d. *Facts and Figures*. Available from: https://www.csir.co.za/facts-and-figures. Accessed on 7 January 2023.

COWING, K. 2014. Earliest Satellite Images of Antarctica. *Moonviews,* 6 November. Available from: https://moonviews.com/?p=754#more-754. Accessed on 29 December 2022.

COX, A. M. & PINFIELD, S. 2014. Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*, vol. 46(4): 299–316. Available from: https://doi.org/10.1177/0961000613492542.

CREAMER, A., MORALES, M.E., CRESPO, J., KAFEL, D. & MARTIN, E.R. 2012. An Assessment of Needed Competencies to Promote the Data Curation and Management Librarianship of Health Sciences and Science and Technology Librarians in New England. *Journal of eScience Librarianship* 1(1): 4. Available from: https://doi.org/10.7191/jeslib.2012.1006.

CRESWELL, J.W. 1999. Mixed-Method Research: Introduction and Application. In: Cizek, G. (*ed*). *Handbook of educational policy***.** Cambridge, USA: Academic, pp. 455–472.

CRESWELL, J.W. 2012. *Educational Research: Planning, Conducting, and Evaluation Quantitative and Qualitative Research*. Boston, Massachusetts: Pearson.

CRESWELL, J.W. 2013. *Qualitative inquiry and research design: choosing among five approaches*. Thousand Oaks, California: Sage. Available from: http://www.ceil-conicet.gov.ar/wp-content/uploads/2018/04/CRESWELLQualitative-Inquary-and-Research-Design-Creswell.pdf. Accessed on 7 November 2021.

CRESWELL, J.W. 2014. *Research design: Qualitative, quantitative, and mixed methods approaches*. London: Sage. Available from: http://englishlangkan.com/produk/E%20Book%20Research%20Design%20Cressweell%202014.pdf. Accessed on 8 December 2019.

CROSSLEY, J. & JANSEN, D. 2021. Saunders' Research Onion: Explained Simply. *Grad Coach,* 26 January. Available from: https://gradcoach.com/saunders-research-onion/. Accessed on 24 January 2022.

CROWE, S., CRESWELL, K., ROBERTSON, A., HUBY, G., AVERY, A. & SHEIKH, A. 2011. The case study approach. *BMC Medical Research Methodology*, vol. 11: 1–10. Available from: https://doi.org/10.1186/1471-2288-11-100.

CUEVA, D. 2022. *How to Make a Conceptual Framework (with Samples).* Available from: https://topnotcher.ph/how-to-make-a-conceptual-framework/. Accessed on 30 December 2022.

CURRIN, G. 2021. Citizen Scientists Digitized Centuries of Handwritten Rain Data. *Wired Science,* 24 June. Available from: https://www.wired.com/story/citizen-scientists-digitized-centuries-of-handwritten-rain-data/. Accessed on 26 January 2022.

CURRY, A. 2011. Rescue of Old Data Offers Lesson for Particle Physicists. *Science*, vol. 331(6018): 694–695. Available from: DOI: 10.1126/science.331.6018.694.

DATAFIRST. 2018. *DataFirst Launches Rescued Historical South African Data*. Available from: https://www.datafirst.uct.ac.za/about-us/latest-news?highlight=WyJyZXNjdWUiXQ==. Accessed on 26 January 2022.

DATAFIRST. 2022. *DataFirst and the Rescue of Data at Risk: Partnering to enable Research Access to Historical African Data*. Available from: https://www.datafirst.uct.ac.za/services/data-rescue?highlight=WyJyZXNjdWVkIl0=. Accessed on 12 October 2022.

DATA RESCUE: ARCHIVAL AND WEATHER (DRAW). 2019. *What is DRAW?* Available from: https://citsci.geog.mcgill.ca/en/about/draw. Accessed on 26 January 2022.

DATARESCUEDENTON. 2017. *Data Rescue Denton: Help preserve climate change data before it's too late*. Available from: http://datarescuedenton.com/. Accessed on 26 January 2022.

DATARESCUE WUSTL. 2017. *DataRescue WUSTL: April 14, 2017 at Washington University in St. Louis*. Available from: https://datarescuewu.github.io/. Accessed on 26 January 2022.

DATT, S. & CHETTY, P. 2016. *8-Step procedure to conduct qualitative content analysis in a research.* Available from: https://www.projectguru.in/qualitative-content-analysis-research/. Accessed on 5 January 2023.

DAVIES, L., LECLAIR, K.L., BAGLEY, P., BLUNT, H., HINTON, L., RYAN, S. & ZIEBLAND, S. 2020. Face-to-Face Compared with Online Collected Accounts of Health and Illness Experiences: A Scoping Review. *Qualitative Health Research*, vol. 30(13): 2092–2102. Available from: https://doi.org/10.1177/1049732320935835.

DELL´OLIO, L., IBEAS, A., DE ONA, J. & DE ONA, R. 2018. Public Participation Techniques and Choice of Variables. In: DELL´OLIO, L., IBEAS, A., DE ONA, J. & DE ONA, R. *Public Transportation Quality of Service: Factors, Models, and Applications*. Amsterdam, Elsevier. Available from: https://doi.org/10.1016/B978-0-08-102080-7.00003-3.

DEVRIENDT, T., BORRY, P. & SHABANI, M. 2021. Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers. *PLoS One*, vol. 16(7): e0254202. Available from: https://doi.org/10.1371/journal.pone.0254202.

DEWITT WALLACE LIBRARY. 2021. *Data Module #1: What is Research Data?* Macalester College, Dewitt Wallace Library. Available from: https://libguides.macalester.edu/c.php?g=527786&p=3608639. Accessed on 3 July 2021.

DJS RESEARCH. 2022. *Mini Groups*. Available from: https://www.djsresearch.co.uk/glossary/item/Mini-Groups. Accessed on 29 December 2022.

DIGITAL PRESERVATION COALITION. 2018. *Bit List of Digitally Endangered Species Revision 2.* Available from*:* https://www.dpconline.org/docs/miscellaneous/advocacy/%0b1932-bitlist2018-final/file. Accessed on 26 January 2022.

DIVIACCO, P., WARDELL, N., FORLIN, E., SAULI, C., BURCA, M., BUSATO, A., CENTONZE, J. & PELOS, C. 2015. Data rescue to extend the value of vintage seismic data: The OGS-SNAP experience. *GeoResJ*, vol. 6: 44–52. Available from: https://doi.org/10.1016/j.grj.2015.01.006.

DIWAKAR, A.S., KULKARNI, A.A. & TALWAI, P. 2008. Rescue and preservation of climate data by extraction and digitization from autographic weather charts using image processing tools. *SSIP'08: Proceedings of the 8th conference on Signal, Speech and image processing, Santander, Cantabria, Spain, 23–25 September 2008.* Available from: https://dl.acm.org/citation.cfm?id=1503945. Accessed on 26 January 2022.

DONG, L., ILIEVA, P., & MEDEIROS, A. 2018. Data dreams: planning for the future of historical medical documents. *Journal of the Medical Library Association*, vol. 106(4): 547-551. Available from: https://doi.org/10.5195/jmla.2018.444.

DOWNS, R.R. 2015. Data Rescue at a Scientific Data Center. *Best Practices Exchange, Pennsylvania State Archives, Harrisburg, Pennsylvania, USA, 19–21 October 2015.* Available from: https://bpexchange.files.wordpress.com/2017/01/bpe2015_20151020_downs_datarescuescidatactr.pdf. Accessed on 26 January 2022.

DOWNS, R.R. 2018. Understanding Risks to the Use of Research Data. *ESIP 2018 Winter Meeting, Bethesda, Maryland, USA, 9–11 January 2018.* Available from: http://wiki.esipfed.org/index.php/File:DownsUnderstandingRiskstotheUseofResearchData20180110.pptx. Accessed on 26 January 2022.

DOWNS, R.R. & CHEN, R.S. 2017. Curation of Scientific Data at Risk of Loss. Data Rescue and Dissemination. In: Johnston, L.R. *Curating research data: practical strategies for your digital repository, pp.* 263–277. Available from: https://academiccommons.columbia.edu/doi/10.7916/D8W09BMQ. Accessed on 26 January 2022.

DRACHEN,T.M., ELLEGAARD, O., LARSEN, A.V. & DORCH, S.B.F. 2016. Sharing Data Increases Citations. *Liber Quarterly*, vol. 26(2): 67–82. Available from: https://doi.org/10.18352/lq.10149.

DRUMMOND, C. 2016. Embracing diversity: when is a librarian not a librarian? *The Australian Library Journal*, vol. 65(4): 274–279. Available from: https://doi.org/10.1080/00049670.2016.1233600.

DRYAD. 2020. *Dryad for your research data*. Available from: https://datadryad.org/stash/. Accessed on 26 January 2022.

DUBLIN CORE METADATA INITIATIVE. 2020. *Dublin Core Metadata Initiative*. Available from: https://dublincore.org/. Accessed on 26 January 2022.

DUDOVSKIY, J. 2016. *Pragmatism Research Philosophy*. Business Research Methodology. Available from: https://research-methodology.net/research-philosophy/pragmatism-research-philosophy/. Accessed on 21 March 2022.

DUDOVSKIY, J. 2019. *Purposive sampling*. Business Research Methodology. Available from: https://research-methodology.net/sampling-in-primary-data-collection/purposive-sampling/. Accessed on 8 December 2019.

DU PLESSIS, S.H. 2020. Email to L.H. Patterton, 18 September 2020.

EARLS, A.C., CLARY, E., GREENBERG, J., KIRSCHENFELD, A., MURILLO, A.P., DAVENPORT ROBERTSON, W., SWAUGER, S. & ANDERSON, W.L. 2013. The Data-at-Risk Initiative: A Metadata Scheme for Documenting Data Rescue Activities. *10th International Conference on Preservation of Digital Objects (iPRES 2013), Lisbon, Portugal, 2–6 September 2013.* Available from: https://cci.drexel.edu/mrc/wp-content/uploads/2017/10/8-iPRES_2013_DARI_FINAL_6-26.pdf. Accessed on 26 January 2022.

EDINA AND DATA LIBRARY, UNIVERSITY OF EDINBURGH. 2013. *Do-It-Yourself Research Data Management Training Kit for Librarians.* Available from: https://mantra.ed.ac.uk/libtraining.html. Accessed on 12 January 2023.

EDMONDS, W.A. & KENNEDY, T.D. 2017. *An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods*. Los Angeles: Sage.

EDUCATION STUDIES, UNIVERSITY OF WARWICK. 2017. *What is pragmatism*? Available from: https://warwick.ac.uk/fac/soc/ces/research/current/socialtheory/maps/pragmatism/. Accessed on 21 March 2022.

EKE, K. 2017. *How libraries can help rescue data*. Available from: https://www.slideshare.net/kimeke/how-libraries-can-help-levels-of-data-rescue?qid=460c012e-76e2-4e1f-b142-7292775eb3b4&v=&b=&from_search=26/16/engaging-i. SlideShare. Accessed on 26 January 2022.

ELSEVIER. 2016. *International Data Rescue Award in the Geosciences*. Available from: https://www.elsevier.com/awards/international-data-rescue-award-in-the-geosciences. Accessed on 17 July 2022.

ENDANGERED DATA WEEK. 2019. *About Endangered Data Week*. Available from: https://endangereddataweek.org/about/. Accessed on 21 January 2020.

ENDANGERED DATA WEEK. 2020. *About Endangered Data Week*. Available from: https://endangereddataweek.org/about/. Accessed on 28 August 2022.

ENGSTRÖM, E., AZORÍN-MOLINA, C., WERN, L., HELLSTRÖM, S., STURM, C., JOELSSON, M., ZHANG, G., MINOLA, L., DENG, K. & CHEN, D. 2021. Advances in the data rescue and digitization of historical wind speed observations in Sweden; the WINDGUST project. *EGU General Assembly 2021, 19–30 April 2021*. Available from: https://doi.org/10.5194/egusphere-egu21-5848.

ESURV. 2019. *Make online surveys for free*! Available from: https://esurv.org/. Accessed on 8 December 2019.

ETIKAN, I., MUSA, S.A. & ALKASSIM, R.S. 2016. Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, vol. 5(1): 1–4. Available from: https://www.sciencepublishinggroup.com/journal/paperinfo?journalid=146&doi=10.11648/j.ajtas.20160501.11. Accessed on 8 December 2019.

EUROPEAN SPACE AGENCY (ESA). 2016. ESA and the Vatican join forces to save data in the digital age. *European Space Agency,* 4 November. Available from: https://www.esa.int/Applications/Observing_the_Earth/ESA_and_the_Vatican_join_forces_to_save_data_in_the_digital_age. Accessed on 8 January 2023.

EUROPEAN SPACE AGENCY (ESA). 2018. ESA and Vatican work to preserve heritage data. *European Space Agency,* 4 May. Available from: https://www.esa.int/Our_Activities/Observing_the_Earth/ESA_and_Vatican_work_to_preserve_heritage_data. Accessed on 26 January 2022.

EVELETH, R. 2014. The quest to scan millions of weather records. *The Atlantic,* 25 August. Available from: https://www.theatlantic.com/technology/archive/2014/08/the-quest-to-scan-millions-of-weather-records/378962/. Accessed on 26 January 2022.

THE FARNSWORTH GROUP. 2022. *How to Achieve Trustworthiness in Qualitative Research*. Available from: https://www.thefarnsworthgroup.com/blog/trustworthiness-qualitative-research. Accessed on 28 December 2022.

FASKO, P., BOCHNÍČEK, O., ŠVEC, M., PALUŠOVÁ, Z. & MARKOVIČ, L. 2016. Data rescue for precipitation network in Slovak Republic. *EGU General Assembly 2016, Vienna, Austria, 17–22*

April 2016. Available from: https://adsabs.harvard.edu/abs/2016EGUGA..18.2099F. Accessed on 26 January 2022.

FAY, B. 2017. DataRescue at MIT. *MIT Libraries,* 1 February. Available from: https://libraries.mit.edu/news/datarescue-mit/24116/. Accessed on 26 January 2022.

FEDERER, L. 2016. Research Data Management in the Age of Big Data: Roles and Opportunities for Librarians'. *Information Services & Use*, vol. 36(1–2): 35–43. Available from: https://content.iospress.com/articles/information-services-and-use/isu797. Accessed on 7 January 2023.

FEDERER, L. 2018. Defining Data Librarianship: A Survey Of Competencies, Skills, And Training. *Journal of the Medical Library Association,* vol. 106(3): 294–303. Available from: https://doi.org/10.5195/jmla.2018.306.

FRASER-ARNOTT, M. 2021. Embracing the Non-Traditional: Incorporating Non-Traditional Elements into Library Identity. *Urban Library Journal*, vol. 27(1): Article 3. Available from: https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1221&context=ulj. Accessed on 7 January 2023.

FREDERICK, A. & RUN, Y. 2019. The Role of Academic Libraries in Research Data Management: A Case in Ghanaian University Libraries. *Open Access Library Journal*, vol. 6: 1–16. Available from: doi: 10.4236/oalib.1105286.

FRY, M. 2010a. Hydrological Data Rescue – the current state of affairs. In: Servat, E., Demuth, S., Dezetter, A. & Daniell, T. (*eds). Global Change: Facing Risks and Threats to Water Resources. IAHS,* vol. 340: 459–464. Available from: https://www.rd-alliance.org/sites/default/files/Fry.%202010.%20Hydrological%20Data%20Rescue%20%E2%80%93%20the%20current%20state%20of%20affairs.pdf. Accessed on 26 January 2022.

FRY, J. 2010b. Data Rescue in Canada, A case study. *IASSIST Conference, Ithaca, New York, USA, 2 June 2010*. Available from: https://www.slideserve.com/kerryn/data-rescue-in-canada-a-case-study. Accessed on 26 January 2022.

GALLAHER, D., CAMPBELL, G.G., MEIER, W., MOSES, J. & WINGO, D. 2015. The process of bringing dark data to light: The rescue of the early Nimbus satellite data. *GeoResJ*, vol. 6: 124–134. Available from: https://doi.org/10.1016/j.grj.2015.02.013.

GARCIA, K. 2017. Meet the Data Rescuers. *Penn Program in Environmental Humanities,* 28 February. Available from: https://ppehlab.squarespace.com/blogposts/2017/2/28/data-rescue-dc. Accessed on 26 January 2022.

GARDNER, K. 2021. What are Population and Sample in Statistics? *Study.com*, 25 September. Available from: https://study.com/learn/lesson/what-is-difference-between-population-vs-sample-in-statistics-example-of-sample-population-in-statistics.html. Accessed on 20 March 2022.

GAUDIN, S. 2017. Coders and librarians team up to save scientific data. *ComputerWorld*, 20 March. Available from: https://www.computerworld.com/article/3182384/data-storage/coders-and-librarians-team-up-to-save-scientific-data.html. Accessed on 26 January 2022.

GEIBEL, L., HUSS, M., KURZBÖCK, C., HODEL, E, BAUDER, A. & FARINOTTI, D. 2022. Rescue and homogenisation of 140 years of glacier mass balance data in Switzerland. *Earth System Science Data*. [preprint]. Available from: https://doi.org/10.5194/essd-2022-56.

GIBNEY, E. 2017. Citizen scientists to rescue 150 years of cosmic images. *Nature:* https://doi.org/10.1038/nature.2017.21702.

GIUSTI, M. 2022. *C3S Data Rescue Service User Forum*. Available from: https://confluence.ecmwf.int/pages/viewpage.action?pageId=171412166. Accessed on 7 September 2022.

GLOBAL BIODIVERSITY INFORMATION FACILITY. 2022. *Free and open access to biodiversity data*. Available from: https://www.gbif.org/. Accessed on 29 December 2022.

GLOBAL SURFACE AIR TEMPERATURE (GLoSAT). 2021. *Data Rescue*. The National Oceanography Centre (NOC), Global Surface Air Temperature. Available from: https://www.glosat.org/data-rescue. Accessed on 26 January 2022.

GOBEN, A. & RASZEWSKI, R. 2015. Research Data Management Self-Education for Librarians: A Webliography. *Issues in Science and Technology Librarianship*, no. 82: https://doi.org/10.29173/istl1666.

GRAB, S. 2018a. *ACRE South Africa.* Atmospheric Circulation Reconstructions over the Earth. Available from: http://www.met-acre.org/Home. Accessed on 26 January 2022.

GRAB, S. 2018b. *C3S DRS under ACRE South Africa – progress report – 28 August 2018*. The International Data Rescue (I-DARE) Portal. Available from: https://idare-portal.org/sites/default/files/C3S%20DRS%20under%20ACRE%20South%20Africa.pdf. Accessed on 26 January 2022.

GRAB, S. 2021a. Email to Louise Patterton, 23 August 2021.

GRAB, S. 2021b. *Weather Climate Science for Service Partnership South Africa. End of Contract Activity Report.* Atmospheric Circulation Reconstructions over the Earth. Available from: http://www.met-acre.org/data-projects-and-regional-chapters. Accessed on 26 January 2022.

GRIFFIN, E. 2006. Rescuing and recovering lost or endangered data. *Data Science Journal*, vol. 4: 21–26. Available from: https://datascience.codata.org/articles/abstract/354/. Accessed on 26 January 2022.

GRIFFIN, E. 2017. Rescue old data before it's too late. *Nature*, vol. 545: 267. Available from: https://doi.org/10.1038/545267a.

GRIFFIN, R.E. 2005. The Detection and Measurement of Telluric Ozone from Stellar Spectra. *Publications of the Astronomical Society of the Pacific*, vol. 117: 885–894. Available from: https://iopscience.iop.org/article/10.1086/431935. Accessed on 17 July 2022.

GRIFFIN, R.E. 2015. When are old data new data? *GeoResJ*, vol. 6: 92–97. Available from: https://doi.org/10.1016/j.grj.2015.02.004.

GROSSMAN, D. 2017. Long-lost Congo notebooks may shed light on how trees react to climate change. *The Guardian,* 25 September. [updated]. Available from: https://www.theguardian.com/environment/2017/sep/22/long-lost-congo-notebooks-shed-light-how-trees-react-to-climate-change. Accessed on 17 February 2022.

GUEST BLOGGER. 2017. How data refuge works, and how YOU can help save federal open data. *Sunlight Foundation,* 6 February. Available from: https://sunlightfoundation.com/%0b2017/02/06/how-data-refuge-works-and-how-you-can-help-save-federal-open-data/. Accessed on 26 January 2022.

GÜNTSCH, A, D. FICHTMÜLLER, D, KIRCHHOFF, A. & BERENDSOHN, G. 2012. Efficient rescue of threatened biodiversity data using reBiND workflows. *Plant Biosystems – An International Journal Dealing with all Aspects of Plant Biology*, vol. 146(4): 752–755. Available from: https://doi.org/10.1080/11263504.2012.740086.

GUTMANN, M., SCHÜRER, K., DONAKOWSKI, D. & BEEDHAM, H. 2004. The Selection, Appraisal, and Retention of Digital Social Science Data. *Data Science Journal*, vol. 3: 209–221. Available from: https://www.digitalpreservation.gov/partners/documents/data-pass_selection_data.pdf. Accessed on 13 January 2023.

HACHILEKA, E. 2015. Why we need to save Africa's historical climate data. *United Nations Development Programme,* 14 October. Available from: https://reliefweb.int/report/world/why-we-need-save-africa-s-historical-climate-data. Accessed on 29 December 2022.

HAMAMURAD, Q.H., MAT JUSOH, N. & UJANG, U. 2022. Factors That Affect Spatial Data Sharing in Malaysia. *ISPRS International Journal of Geo-Information*, vol. 11(8): 446. Available from: https://doi.org/10.3390/ijgi11080446.

HAMMARBERG, K., KIRKMAN, M. & DE LACEY, S. 2016. Qualitative Research Methods: When to Use them and how to judge them. *Human Reproduction*, vol. 31(3): 498–501. Available from: https://doi.org/10.1093/humrep/dev334.

HANCOCK, D.R. & ALGOZZINE, B. 2006. *Doing Case Study Research. A Practical Guide for Beginning Researchers.* New York: Teachers College Press. Available from: https://student.cc.uoc.gr/uploadFiles/192-%CE%A3%CE%A0%CE%91%CE%9D104/HANCOCK%20and%20ALGOZZINE%20Case%20Study%20Research%202.pdf. Accessed on 28 January 2022.

HANNA, P. 2012. Using internet technologies (such as Skype) as a research medium: a research note. *Qualitative Research*, vol. 12(2): 239–242. Available from: https://doi.org/10.1177/1468794111426607.

HARMON, A. 2017. Activists Rush to Save Government Science Data – If They Can Find It. *The New York Times,* 6 March. Available from: https://www.nytimes.com/2017/03/06/science/donald-trump-data-rescue-science.html. Accessed on 28 January 2022.

HARRISON, H., BIRKS, M., FRANKLIN, R. & MILLS, J. 2017. Case Study Research: Foundations and Methodological Orientations. *Forum: Qualitative Social Research*, vol. 18(1): Article 19. Available from: https://doi.org/10.17169/fqs-18.1.2655.

HARVARD MEDICAL SCHOOL. RESEARCH DATA MANAGEMENT. 2018. *NSF Data Management Plan*. Available from: https://datamanagement.hms.harvard.edu/plan/data-management-plans/nsf-data-management-plan. Accessed on 14 January 2023.

HAWKINS, S.J., FIRTH, L.B., McHUGH, M., PLOCZANSKA, E.S., HERBERT, R.J.H., BURROWS, M.T., KENDALL, M.A., MOORE, P.J., THOMPSON, R.C., JENKINS, S.R. *et al.* 2013. Data rescue and re-use: Recycling old information to address new policy concerns. *Marine Policy,* vol. 42: 91–98. Available from: https://doi.org/10.1016/j.marpol.2013.02.001.

HAYS, L. & STUDEBAKER, B**.** 2019. Academic Instruction Librarians' Teacher Identity Development Through Participation in the Scholarship of Teaching and Learning. *International Journal for the Scholarship of Teaching and Learning*, vol. 13(2): Article 4. Available from: https://doi.org/10.20429/ijsotl.2019.130204.

HEALE, R. & TWYCROSS, A. 2018. What is a case study? *Evidence-Based Nursing*, vol. 21(1): 7–8. Available from: https://ebn.bmj.com/content/21/1/7. Accessed on 31 May 2022.

HEDGES, J. & FELLOUS-SIGRIST, M. 2017. Training subject librarians in Research Data Management. In: *LEARN Toolkit of Best Practice for Research Data Management*. Available from: https://discovery.ucl.ac.uk/id/eprint/1546591/. Accessed on 7 January 2023.

HESSE-BIBER, S.N. 2010. *Mixed Methods Research: Merging Theory with Practice*. New York: Guilford Press.

HESSE-BIBER, S.N. & LEAVY, P. (*eds*). 2004. *Approaches to Qualitative Research: A reader on theory and practice.* New York: Oxford University Press.

HESSE-BIBER, S.N. & LEAVY, P. 2006. *The Practice of Qualitative Research*. London: Sage.

HILLS, D. 2016. A Story of Hope: Efforts to Bring Geological Samples into the 21st Century. *CODATA/RDA workshop on The Rescue of Data at Risk***,** *Boulder, Colorado, USA, 8–9 September 2016*.

HILLS, D.J. 2015. Let's make it easy: A workflow for physical sample metadata rescue. *GeoResJ*, vol. 6: 1–8. Available from: https://doi.org/10.1016/j.grj.2015.02.007.

HILLS, D.J. 2019. Data Rescue IG. *Research Data Alliance,* 14 August. Available from: https://www.rd-alliance.org/groups/data-rescue.html. Accessed on 28 January 2022.

HOFFMAN, K.M., CLARKE, C.T., SZU YIN SHIUE, H., NICHOLAS, P., SHAW, M. & FENLON, K. 2020. *Data Rescue: An assessment framework for legacy research collections*. University of Maryland, University Libraries, Digital Repository at the University of Maryland (DRUM). Available from: https://drum.lib.umd.edu/bitstream/handle/1903/26475/DataRescue_whitepaper_final_NALDC.pdf?sequence=1&isAllowed=y. Accessed on 28 January 2022.

HSU, L., LEHNERT, K., CARBOTTE, S., FERRINI, V., DELANO, J., GILL, J.B. & TIVEY, M. 2013. Rescue of long-tail data from the ocean bottom to the Moon. *AGU Fall Meeting, San Francisco, USA, 9–13 December 2013.* SlideShare. Available from: https://www.slideshare.net/hsuleslie/hsu-agu2013?qid=e164d577-cec1-45ae-9dc4-ef38c3c47652&v=&b=&from_search=22. Accessed on 28 January 2022.

HSU, L., LEHNERT, K.A., GOODWILLIE, A., DELANO, J.W., GILL, J.B., TIVEY, M.A., FERRINI, V.L., CARBOTTE, S.M. & ARKO, R.A. 2015. Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards. *GeoResJ*, vol. 6: 108–114. Available from: https://doi.org/10.1016/j.grj.2015.02.012.

HUETTICH, J. 2020. What is a workflow? Types, benefits, and examples. *MindManager Blog,* 2 June. Available from: https://blog.mindmanager.com/blog/2020/06/02/202006202005your-guide-to-the-different-types-of-workflows/. Accessed on 30 October 2021.

HYGEN, H.O., ELO, C.A. & GJELTEN, H.M. 2021. Data rescue and digitization through image recognition. *EMS Annual Meeting 2021, [online], 6–10 September 2021*. Available from: https://doi.org/10.5194/ems2021-43.

IGAD CLIMATE PREDICTION & APPLICATIONS CENTRE (ICPAC). 2016. *Data rescue activities at ICPAC (2014–2015)*. Available from: https://www.icpac.net/publications/data-rescue-activities-icpac/. Accessed on 31 March 2022.

ILLINOIS TECH LIBRARY GUIDES. 2017. *Endangered Data Week: April 17th–April 21st, 2017*. Illinois Tech Library. Available from: https://guides.library.iit.edu/govdocs/events/edw2017. Accessed on 28 January 2022.

INDEED. 2021. 6 Ways To Streamline Business Processes and Workflows. *Indeed,* 17 March. Available from: https://www.indeed.com/career-advice/career-development/streamline-processes-and-workflows. Accessed on 31 March 2022.

INTERGOVERNMENTAL OCEANOGRAPHIC COMMISSION (IOC). 2003. *Proceedings of the International Global Oceanographic Data Archaeology and Rescue (GODAR) Review Conference. IOC Workshop Report, Silver Spring, Maryland, USA, 12–15 July 1999*. UNESCO/IOC Project Office for IODE. Available from: https://oceanexpert.org/document/954. Accessed on 27 March 2022.

INTERGOVERNMENTAL OCEANOGRAPHIC COMMISSION (IOC). 2020. *Workshop on Sea Level Data Archaeology, Paris, France, 10–12 March 2020.* Available from: https://nora.nerc.ac.uk/id/eprint/527878/1/373327eng.pdf. Accessed on 28 January 2022.

INTERNATIONAL DATA RESCUE (I-DARE) PORTAL. 2019. *France semaphore and lighthouse meteorological data rescue 1879–1930*. Available from: https://www.idare-portal.org/data/france-semaphore-and-lighthouse-meteorological-data-rescue-1868-1940. Accessed on 28 January 2022.

INTERNATIONAL DATA RESCUE (I-DARE) PORTAL. 2021. *The International Data Rescue (I-DARE) Portal*. Available from: https://www.idare-portal.org/. Accessed on 28 January 2022.

INTERNATIONAL ENVIRONMENTAL DATA RESCUE ORGANIZATION (IEDRO). 2010. *Be Part of the Solution*. SlideShare. Available from: https://www.slideshare.net/dottuta/iedro-presentation-with-music?qid=460c012e-76e2-4e1f-b142-7292775eb3b4&v=&b=&from_search=10. Accessed on 28 January 2022.

INTERNATIONAL ENVIRONMENTAL DATA RESCUE ORGANIZATION (IEDRO). 2014. *The Data Rescue Process.* Available from: https://iedro.org/data-rescue-process/. Accessed on 31 May 2022.

INTERNATIONAL ENVIRONMENTAL DATA RESCUE ORGANIZATION (IEDRO). 2015. *Weather Wizards*. Available from: https://iedro.org/activities/weather-wizards/. Accessed on 7 September 2022.

INTERNATIONAL ENVIRONMENTAL DATA RESCUE ORGANIZATION (IEDRO). 2016. *Where We've Been*. http://iedro.org/activities/drd-sites/. Accessed on 28 January 2022.

INTERNATIONAL ENVIRONMENTAL DATA RESCUE ORGANIZATION (IEDRO). 2018. *Dare Activities Report.* Available from: https://www.idare-portal.org/sites/default/files/IEDRO_projects.pdf. Accessed on 28 January 2022.

INTERNATIONAL ENVIRONMENTAL DATA RESCUE ORGANIZATION (IEDRO). 2020. *Uzbekistan Data Rescue Project*. Available from: https://iedro.org/activities/uzbekistan/. Accessed on 28 January 2022.

INTERNATIONAL OCEANOGRAPHIC DATA AND INFORMATION EXCHANGE. 2013. *Global Oceanographic Data Archaeology and Rescue (GODAR)*. Available from: https://www.iode.org/index.php?option=com_content&view=article&id=18&Itemid=100087. Accessed on 28 January 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. 2015. *ISO 16245:2009 Information and documentation – Boxes, file covers and other enclosures, made from cellulosic materials, for storage of paper and parchment documents.* Available from: https://www.iso.org/standard/45988.html. Accessed on 28 January 2022.

INTRAC. 2017. *Focus Group Discussions*. Available from: https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Focus-group-discussions.pdf. Accessed on 3 January 2022.

JANGHORBAN, R., ROUDSARI, R.L. & TAGHIPOUR, A. 2014. Skype interviewing: The new generation of online synchronous interview in qualitative research. *International Journal of Qualitative Studies on Health and Well-being,* vol. 9(1): 1–3. Available from: https://doi.org/10.3402/qhw.v9.24152.

JANZ, M. 2018. Data Rescue. *University of Massachusetts and New England Area Librarian e-Science Symposium, University of Massachusetts, Worcester, MA, USA, 5 April 2018.* Available from: https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1182&context=escience_symposium. Accessed on 28 January 2022.

JISC DIGITAL MEDIA. 2014. *Introduction to using a copy stand*. Available from: https://www.youtube.com/watch?v=udiRErQ1CnA. Accessed on 28 January 2022.

JOHAR, A. 2017. Internet security 101: Six ways hackers can attack you and how to stay safe. *The Economic Times,* 30 October. Available from: https://economictimes.indiatimes.com/tech/internet/internet-security-101-six-ways-hackers-can-attack-you-and-how-to-stay-safe/articleshow/61342742.cms. Accessed on 8 January 2023.

JOHNSTON, L. R. (*ed*). 2017. *Curating research data: Volume one: Practical strategies for your digital repository [e-book].* American Library Association, Association of College and Research Libraries. Available from: https://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf. Accessed on 28 January 2022.

JONES, B. 2015. Historical Canadian Climate Data: Volunteer Data Rescue Project. *International Data Rescue (I-DARE) Portal,* 15 May. Available from: https://www.idare-portal.org/sites/default/files/Canada_historical_climate_data_project.pdf. Accessed on 31 March 2022.

JONES, P.D. 2008. Benefits from undertaking data rescue activities. In: *Proceedings of the International Workshop on Rescue and Digitization of Climate Records in the Mediterranean*

Basin, University Rovira I Virgili, Tarragona, Spain, May 2007, pp. 9–13. Available from: https://library.wmo.int/doc_num.php?explnum_id=9382. Accessed on 27 March 2022.

JONES, S., GUY, M. & PICKTON, M. 2013. *Research Data Management for librarians*. Available from: https://www.dcc.ac.uk/sites/default/files/documents/events/RDM-for-librarians/RDM-for-librarians-booklet.pdf. Accessed on 7 January 2023.

JOWETT, A. 2020. Carrying out qualitative research under lockdown – Practical and ethical considerations. *London School of Economics and Political Science,* 20 April. Available from: https://blogs.lse.ac.uk/impactofsocialsciences/2020/04/20/carrying-out-qualitative-research-under-lockdown-practical-and-ethical-considerations/. Accessed on 28 January 2022.

KAPISZEWSKI, D. & KARCHER, S. 2021. Transparency in Practice in Qualitative Research. *PS: Political Science & Politics*, vol. 54(2): 285–291. Available from: doi:10.1017/S1049096520000955.

KARCHER, S. 2019. *The Limits of Reproducibility: Strategies for Transparent Qualitative Research*. Available from: https://figshare.com/articles/presentation/The_Limits_of_Reproducibility_Strategies_for_Transparent_Qualitative_Research/7637129/1. Accessed on 7 September 2022.

KASPAR, F., TINZ, B., MÄCHEL, H. & GATES, L. 2015. Data rescue of national and international meteorological observations at Deutscher Wetterdienst. *Advances in Science and Research*, vol. 12: 57–61. Available from: https://doi.org/10.5194/asr-12-57-2015.

KENNEDY, M. 2017. *Guidelines for marine species occurrence data rescue - The OBIS Canada Cookbook*. COINAtlantic. Available from: https://www.coinatlantic.ca/_files/ugd/cf2ff9_82e7008749294b83b31c4a8a9ecd99cf.pdf. Accessed on 28 March 2022.

KHWELA, Z. 2018. Data rescue for future researchers. *Rhodes, Latest News,* 27 November. Available from: https://www.ru.ac.za/latestnews/datarescueforfutureresearchers.html. Accessed on 28 January 2022.

KIJAS, A.E. 2018. Engaging in Small Data Rescue. *Humanities Commons,* 9 February. Available from: https://hcommons.org/?s=small+data+rescue. Accessed on 1 April 2022.

KIMANI, J. 2008. WMO REGION 1: AFRICA. *WMO Report of the meeting of the CCl Expert Team on the Rescue, Preservation and Digitization of Climate Records, Bamako, Mali, 13–15 May 2008*. Available from: https://library.wmo.int/doc_num.php?explnum_id=9411. Accessed on 31 May 2022.

KITZINGER, J. 1995. Qualitative Research: Introducing focus groups. *British Medical Journal*, vol. 311: 299–302. Available from: https://doi.org/10.1136/bmj.311.7000.299.

KLEIN, S. & LENART, B. 2020. In Search of Shifting and Emergent Librarian Identities: A Philosophical Approach to the Librarian Identity Problem. *Partnership: The Canadian Journal of Library and Information Practice and Research*, vol. 15(1): 1–27. Available from: https://doi.org/10.21083/partnership.v15i1.5113.

KNAPP, K.R., BATES, J.J. & BARKSTROM, B. 2007. Scientific Data Stewardship: Lessons learned from a Satellite-Data Rescue Effort. *Bulletin of the American Meteorological Society,* vol. 88(9): 1359–1361. Available from: https://doi.org/10.1175/BAMS-88-9-1359.

KNOCKAERT, C., TYBERGHEIN, L., GOFFIN, A., VANHAECKE, D., ONG'ANDA, H., WAKWABI, E.O. & MEES, J. 2019. Biodiversity data rescue in the framework of a long-term Kenya-Belgium cooperation in marine sciences. *Scientific Data*, vol. 6(85): 1–6. Available from: https://doi.org/10.1038/s41597-019-0092-8.

KOCHTUBAJDA, B., HUMPHREY, C. & JOHNSON, M. 1995. Data Rescue: experiences from the Alberta Hail Project. *IASSIST Quarterly*, vol. 19(1): https://doi.org/10.29173/iq558.

KOSMALA, M., WIGGINS, A., SWANSON, A. & SIMMONS, B. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, vol. 14(10): 551–560. Available from: https://doi.org/10.1002/fee.1436.

KOTHARI, C.R. 2004. *Research Methodology Methods and Techniques*. New Delhi: New Age International Publishers.

KOWAL, E., LLAMAS, B. & TISHKOFF, S. 2017. Data-sharing for indigenous peoples. *Nature, vol.* 546: 474. Available from: https://doi.org/10.1038/546474a.

KRIUKOW, J. 2018. *Validity and Reliability in Qualitative research*. Available from: https://drkriukow.com/validity-and-reliability-in-qualitative-research/. Accessed on 7 November 2021.

KROTZ, D. 2011. From Dusty Punch Cards, New Insights into Link Between Cholesterol and Heart Disease. *Berkeley Lab, News Center*, 4 January 2011. Available from: https://newscenter.lbl.gov/featurestories/2011/01/04/cholesterol-heart-disease/. Accessed on 31 May 2022.

KROUWEL, M.J., JOLLY, K. & GREENFIELD, S. 2019. Comparing Skype (video calling) and in-person qualitative interview modes in a study of people with irritable bowel syndrome – an exploratory comparative analysis. *BMC Medical Research Methodology*, vol. 19(1): 1–9. Available from: https://doi.org/10.1186/s12874-019-0867-9.

KUADA, J. 2015. *Thesis without Tears: A Guide for African University Students.* London: Adonis & Abbey.

KUMAR, R. 2011. *Research Methodology: A Step-by-step guide for beginners*. Thousand Oaks, California: Sage.

LAERD DISSERTATION. 2012. *Purposive Sampling*. Lund Research Ltd. Available from: https://dissertation.laerd.com/purposive-sampling.php. Accessed on 31 May 2022.

LAVRAKAS, P.J. (*ed*). 2008. *Encyclopedia of Survey Research Methods*. Thousand Oaks, California: Sage.

LEEDY, P.D. & ORMROD, J.E. 2005. *Practical Research: Planning and Design*. Upper Saddle River, N.J.: Pearson.

LEVITUS, S. 2012. The UNESCO-IOC-IODE "Global Oceanographic Data Archaeology and Rescue" (GODAR) Project and "World Ocean Database" Project. *Data Science Journal*, vol. 11: 46–71. Available from: https://datascience.codata.org/articles/abstract/48/. Accessed on 28 January 2022.

LIBER WORKING GROUP ON E-SCIENCE/RESEARCH DATA MANAGEMENT. 2012. *Ten recommendations for libraries to get started with research data management: Final report of the LIBER working group on E-Science / Research Data Management*. Available from: https://www.fosteropenscience.eu/sites/default/files/original/681.pdf. Accessed on 8 January 2023.

LIBRARIES PLUS NETWORK. 2017. *Libraries+ Network Meeting, Washington, D.C., USA, 8–9 May 2017*. Available from: https://libraries.network/may-meeting/. Accessed on 28 January 2022.

LIBRARY JUICE ACADEMY. 2023. *New course on Data Management*. Available from: https://libraryjuiceacademy.com/new-course-on-data-management/. Accessed on 7 January 2023.

LINCOLN, Y.S. & GUBA, E.G. 1985. *Naturalistic Inquiry*. Newbury Park, California: Sage.

LINDSAY, R. 2017. Data Rescue Movement Seeks to Fight Trump Administration's Anti-Science Agenda. *GBH News, Science and Technology,* 9 May. Available from: https://www.wgbh.org/news/2017/05/09/science-and-technology/data-rescue-movement-seeks-fight-trump-administrations-anti. Accessed on 28 January 2022.

LUCID CONTENT TEAM. 2021. *What Is a Workflow? Benefits and Examples of Repeatable Processes*. Lucid. Available from: https://www.lucidchart.com/blog/what-is-workflow. Accessed on 28 October 2021.

LUMEN LEARNING. 2016. *Flowchart Symbols (8.3.13).* Lumen Learning. Available from: https://courses.lumenlearning.com/ivytech-sdev-dev-1/chapter/flowchart-symbols-8-3-13/. Accessed on 28 October 2021.

LYNCH, A. 2019. Flowchart benefits. *Edrawsoft,* 14 April. Available from: https://www.edrawsoft.com/flowchart-benefits.php. Accessed on 8 December 2019.

MAGUIRE, M. & DELAHUNT, B. 2017. Doing a Thematic Analysis: A Practical, Step-by-Step Guide for Learning and Teaching Scholars. *All Ireland Journal of Higher Education*, vol. 9(3): 3351–33514. Available from: https://ojs.aishe.org/index.php/aishe-j/article/view/335/553. Accessed on 25 March 2022.

MAHDI, O., NASSAR, I. & ALMUSLAMANI, H. 2020. The Role of Using Case Studies Method in Improving Students' Critical Thinking Skills in Higher Education. *International Journal of Higher Education*, vol. 9(2): 297–308. Available from: https://files.eric.ed.gov/fulltext/EJ1248562.pdf. Accessed on 8 November 2021.

MANDELBAUM, R.F. 2017. Rescuing Government Data from Trump Has Become a National Movement. *Gizmodo,* 21 February. Available from: https://gizmodo.com/rescuing-government-data-from-trump-has-become-a-nation-1792582499. Accessed on 28 January 2022.

MANESS, J.M., DUER, R., DULOCK, M., FETTERER, F. & HICKS, G. 2017. Revealing Our Melting Past: Rescuing Historical Snow and Ice Data. *GeoResJ*, vol. 14: 92–97. Available from: https://doi.org/10.1016/j.grj.2017.10.002.

MARINE ENVIRONMENTAL DATA AND INFORMATION NETWORK (MEDIN). 2018. [Twitter]. 30 April. Available from: https://twitter.com/medin_marine/status/990918635903504384. Accessed on 6 February 2022.

MATEUS, C., POTITO, A. & CURLEY, M. 2021. Engaging secondary school students in climate data rescue through service-learning partnerships. *Weather*, vol. 76(4): 113–118. Available from: https://doi.org/10.1002/wea.3841.

MATTSON, R. & SCHELL, J. 2018. The Many Forms of Endangered Data: Notes from Michigan. *Digital Library Federation,* 4 April. Available from: https://www.diglib.org/the-many-forms-of-endangered-data-notes-from-michigan/. Accessed on 28 January 2022.

MAVRAKI, D., FANINI, L., TSOMPANOU, M., GEROVASILEIOU, V., NIKOLOPOULOU, S., CHATZINIKOLAOU, E., PLAITIS, W. & FAULWETTER, S. 2016. Rescuing biogeographic legacy data: The "Thor" Expedition, a historical oceanographic expedition to the Mediterranean Sea. *Biodiversity Data Journal*, vol. 4: e11054. Available from: https://doi.org/10.3897/BDJ.4.e11054.

MAXWELL, J. A. 2005. *Qualitative research design: An interactive approach.* Thousand Oaks, California: Sage.

MAYERNIK, M.S., DOWNS, R.R., DUERR, R., HOU, C-Y, MEYERS, N., RITCHEY, N., THOMER, A. & YARMEY, L. 2017. *Stronger together: the case for cross-sector collaboration in identifying and preserving at-risk data*. NCAR/UCAR Research: User Guides and Manuscripts. Available from: https://opensky.ucar.edu/islandora/object/manuscripts%3A948. Accessed on 28 January 2022.

MAYERNIK, M.S., BRESEMAN, K., DOWNS, R.R., DUERR, R., GARRETSON, A., HOU, C.-Y., ENVIRONMENTAL DATA GOVERNANCE INITIATIVE (EDGI) & EARTH SCIENCE INFORMATION PARTNERS (ESIP) DATA STEWARDSHIP COMMITTEE. 2020. Risk Assessment for Scientific Data. *Data Science Journal*, vol. 19(1): 15pp. Available from: https://datascience.codata.org /articles/10.5334/dsj-2020-010/. Accessed on 19 February 2022.

MAYNOOTH UNIVERSITY. ICARUS CLIMATE RESEARCH CENTRE. 2017. *Data Rescue in the classroom*. Maynooth University. Available from: https://www.maynoothuniversity.ie/icarus/data-rescue-classroom. Accessed on 6 February 2022.

McCOMBES, S. 2022. Case Study: Definition, Examples and Methods. *Scribbr,* 7 February. [revised]. Available from: https://www.scribbr.com/methodology/case-study/. Accessed on 6 February 2022.

McCRINDLE, C.M.E. 2017. *Choosing and Using Quantitative Research Methods and Tools*. Past Presentation: Research Guide: Research Methods. Available from: https://library.up.ac.za/ld.php?content_id=34337840. Accessed on 8 December 2019.

MC GILL LIBRARY. 2019. Digitization services for research needs. Available from: https://www.mcgill.ca/library/services/research/digitization. Accessed on 8 January 2023.

McGOVERN, N. 2017. Data rescue: observations from an archivist. *ACM SIGCAS Computers and Society,* vol. 47(2): 19–26. Available from: https://doi.org/10.1145/3112644.3112648.

METHODSPACE. 2021. *Are We Too Limited on Group Size? What About 2 or 3 Person "Mini-Groups"?* Available from: https://www.methodspace.com/blog/are-we-too-limited-on-group-size-what-about-2-or-3-person-mini-groups. Accessed on 29 December 2022.

MEYER, M.N. 2018. Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, vol. 1(1): 131–144. Available from: https://doi.org/10.1177/2515245917747656.

MIT LIBRARIES. 2019. *Data Management*. Available from: https://libraries.mit.edu/data-management/plan/write/. Accessed on 8 January 2023.

MODINE, A. 2008. Vintage IBM tape drive in Apollo moon dust rescue. *The Register, Science,* 11 November. Available from: https://www.theregister.com/2008/11/11/vintage_ibm_tape_drive_moon_dust_data/. Accessed on 6 February 2022.

MOHAJAN, H.K. 2017. Two Criteria for Good Measurements in Research: Validity and Reliability. *Annals of Spiru Haret University*, vol. 17(3): 58–82. Available from: https://mpra.ub.uni-muenchen.de/83458/1/MPRA_paper_83458.pdf. Accessed on 7 November 2021.

MOHAPI, J. 2019. *Procedures and Processes of the CSIR Research Ethics Committee.* Unpublished presentation delivered on 2 October 2019.

MONAHAN, K. 2017. Tufts Spatial Data Rescue: Crawling at-risk Government Data. *Free and Open Source Software for Geospatial* (*FOSS4G*)*, Boston, USA, 14–19 August 2017*. SlideShare. Available from: https://www.slideshare.net/KyleMonahan1/tufts-spatial-data-rescue-crawling-atrisk-government-data?qid=e164d577-cec1-45ae-9dc4-ef38c3c47652&v=&b=&from_search=3. Accessed on 10 February 2022.

MONDAY.COM. 2021. What is workflow? A step-by-step guide for beginners. *Monday.com*, 5 June. Available from: https://monday.com/blog/project-management/what-is-workflow/. Accessed on 28 October 2021.

MORRIS, D.Z. 2017. Hackers Scrambling to Save Climate Data from Trump Administration. *Fortune,* 22 January. Available from: https://fortune.com/2017/01/22/climate-data-trump-admin-hackers/. Accessed on 10 February 2022.

MUKHONDIA, M. 2017. Historical Climate Data Rescue and Digitization Efforts in Africa**.** *IEDRO UNDP CIRDA Workshop "Towards Sustainability for Climate Services", Lusaka, Zambia, 29 November 2017–1 December 2017.* Available from: https://www.adaptation-undp.org/sites/default/files/resources/6._iedro_data_rescue.pdf. Accessed on 6 February 2022.

MUKHONDIA, M., SWASWA, M. & BOJANG, T. 2017. Session 4: Using historical data: data rescue efforts to enhance climate information services. *IEDRO UNDP CIRDA Workshop "Towards Sustainability for Climate Services", Lusaka, Zambia, 29 November 2017–1 December 2017.* Available from: https://www.adaptation-undp.org/sites/default/files/resources/5._data_digitization_intro_mtadross.pdf. Accessed on 6 February 2022.

MULLER, C. 2012. Data at Risk: The Duty to Find, Rescue, Preserve. *UNESCO Conference, "Memory of the World" conference, Vancouver, Canada, 26–28 September 2012.* SlideShare. Available from: https://www.slideshare.net/ctm0608/data-at-risk-poster-for-unesco-conference. Accessed on 6 February 2022.

MULLER, C. 2015a. Rescuing early digital assets and preserving data rescue capabilities. *Best Practices Exchange, Pennsylvania State Archives, Harrisburg, USA, 19–21 October 2015*. SlideShare. Available from: https://www.slideshare.net/ctm0608/data-rescue-and-preserving-dr-capabilities?qid=460c012e-76e2-4e1f-b142-7292775eb3b4&v=&b=&from_search=1. Accessed on 6 February 2022.

MULLER, C. 2015b. The Digital Vapor Trail: Why early digital assets merit special attention. *Against the Grain,* vol. 27(4): 18–22. Available from: https://doi.org/10.7771/2380-176X.7128.

MURILLO, A.P. 2014. Data at Risk Initiative: Examining and Facilitating the Scientific Process in Relation to Endangered Data. *Data Science Journal*, vol. 12: 207–219. Available from: https://datascience.codata.org/articles/abstract/51/. Accessed on 6 February 2022.

NATIONAL ARCHIVES. 2016. *Archive Principles and Practice: an introduction to archives for non-archivists.* Available from: https://www.nationalarchives.gov.uk/documents/archives/archive-principles-and-practice-an-introduction-to-archives-for-non-archivists.pdf. Accessed on 23 March 2022.

NATIONAL ARCHIVES. 2019. *Environmental Management*. Available from: https://www.nationalarchives.gov.uk/documents/information-management/environmental-management.pdf. Accessed on 23 March 2022.

NATIONAL INSTITUTES OF HEALTH. 2022. *Data Sharing Approaches*. Available from: https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/data-sharing-approaches. Accessed on 13 January 2023.

NEUENDORF, K. A. 2019. Content analysis and thematic analysis. In Bough, P. (*ed*). *Research methods for applied psychologists: Design, analysis and reporting*. New York: Routledge, pp. 211–223. Available from: https://academic.csuohio.edu/kneuendorf/vitae/NeuendorfCA_TA_19.pdf. Accessed on 6 February 2022.

NEUMAN, W.L. 2014. *Social Research Methods: Qualitative and Quantitative approaches*. Essex: Pearson Education Limited.

NEWBERRY GEOTHERMAL ENERGY. 2016. *Sample and Core Curation Plan*. Available from: https://www.energy.gov/sites/prod/files/2016/09/f33/Sample%20and%20Core%20 Curation%20Plan_%28Newberry%20Volcano%2C%20OR%29.pdf. Accessed on 18 January 2022.

NEWMAN, E. 2018. Advantages and limitations of flowcharts. *Yonyx*, 20 April. Available from: https://corp.yonyx.com/customer-service/advantages-and-limitations-of-flowcharts/. Accessed on 8 December 2019.

NORDLING, L. 2010. Researchers launch hunt for endangered data. *Nature*, vol. 468: 17. Available from: https://doi.org/10.1038/468017a.

NORTHEAST DOCUMENT CONSERVATION CENTER (NEDCC). 1999. *The Environment: Temperature, Relative Humidity, Light, and Air Quality: Basic Guidelines for Preservation*. Available from: https://www.nedcc.org/free-resources/preservation-leaflets/2.-the-environment/2.1-temperature,-relative-humidity,-light,-and-air-quality-basic-guidelines-for-preservation. Accessed on 23 March 2022.

NYUMBA, T.O., WILSON, K., DERRICK, C.J. & MUKHERJEE, N. 2018. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution*, vol. 9(1): 20–32. Available from: https://doi.org/10.1111/2041-210X.12860. Accessed on 26 March 2022.

OCUL DATA COMMUNITY DATA RESCUE GROUP. 2020. *Data Rescue & Curation Best Practices Guide*. OCUL Data Community Dataverse. Available from: https://doi.org/10.5683/SP2/Y8MQXV.

OPEN PRESERVATION FOUNDATION. 2021. *"A Call for Help" – Collecting obsolete equipment and playback devices*. Available from: https://openpreservation.org/events/a-call-for-help-collecting-obsolete-equipment-and-playback-devices/. Accessed on 6 February 2022.

THE OPEN UNIVERSITY. 2023. *Data Management Plans*. Available from: https://www.open.ac.uk/library-research-support/research-data-management/data-management-plans. Accessed on 8 January 2023.

PAGE, C.M., NICHOLLS, N., PLUMMER, N., TREWIN. B., MANTON, M., ALEXANDER, L., CHAMBERS, L.E., CHOI, Y., COLLINS, D.A., GOSAI, A. *et al*. 2004. Data Rescue in the Southeast Asia and South Pacific Region Challenges and Opportunities. *Bulletin of the American Meteorological Society,* vol. 85(10): 1483–1489. Available from: https://doi.org/10.1175/BAMS-85-10-1483.

PALMER, C.L., WEBER, N.M. & CRAGIN, M.H. 2011. The Analytic Potential of Scientific Data: Understanding Re-use Value. *Proceedings of the American Society for Information Science and Technology*, vol. 48(1): 1–10. Available from: https://doi.org/10.1002/meet.2011.14504801174.

PALYS, T. 2008. Purposive sampling. In: Given, L.M. (*ed*). *The Sage Encyclopedia of Qualitative Research Methods, vol. 2.* Thousand Oaks, California: Sage, pp. 687–698.

PARK, E.G., BURR, G., SLONOSKY, V., SIEBER, R. & PODOLSKY, L. 2018. Data rescue archive weather (DRAW): Preserving the complexity of historical climate data. *Journal of Documentation*, Vol. 74(4): 763–780. Available from: https://doi.org/10.1108/JD-10-2017-0150.

PARKER, A. & TRITTER, J. 2006. Focus group method and methodology: Current practice and recent debate*. International Journal of Research & Method in Education*, vol. 29: 23–37. Available from: https://doi.org/10.1080/01406720500537304.

PATIL, S. & ADITYA. 2020. *Research Methodology in Social Sciences*. New Delhi, New India Publishing Agency.

PATTERTON, L. 2013. *The Rookie Research Data Manager: 3 personal accounts*. Available from: https://nedicc.files.wordpress.com/2013/07/the-rookie-research-data-manager-copy1.pptx. Accessed on 11 January 2023.

PATTERTON, L. 2014. *An introduction to the basics of Research Data*. Available from: https://www.youtube.com/watch?v=q2aiDJzJPuw. Accessed on 11 January 2023.

PATTERTON, L.H. 2014. *Research data management at the CSIR: an exploratory survey*. (Unpublished).

PATTERTON, L.H. 2016. *Research data management practices of emerging researchers at a South African research council.* MIS Dissertation, University of Pretoria, Pretoria. Available from: https://repository.up.ac.za/handle/2263/59502?show=full. Accessed on 6 February 2022.

PATTERTON, L., BOTHMA, T.J.D. & VAN DEVENTER, M.J. 2018. From planning to practice : an action plan for the implementation of research data management services in resource-constrained institutions. *South African Journal of Libraries and Information Science*, vol. 84(2): 14–26. Available from: https://hdl.handle.net/10520/EJC-141a9acd34. Accessed on 11 October 2022.

PhDSTUDENT. 2019a. *Diving Deeper into Limitations and Delimitations*. Available from: https://www.phdstudent.com/thesis-and-dissertation-survival/research-design/diving-deeper-into-limitations-and-delimitations/. Accessed on 8 December 2019.

PERRIER, L., BLONDAL, E. & MACDONALD, H. 2018**.** Exploring the experiences of academic libraries with research data management: A meta-ethnographic analysis of qualitative studies. *Library & Information Science Research*, vol. 40(3–4): 173–183. Available from: https://doi.org/10.5281/zenodo.1324412.

PhDSTUDENT. 2019b. *Stating the Obvious: Writing Assumptions, Limitations, and Delimitations*. Available from: https://www.phdstudent.com/Choosing-a-Research-Design/stating-the-obvious-writing-assumptions-limitations-and-delimitations. Accessed on 8 December 2019.

PHIFFER, D. 2017. Grabbing government data before it's destroyed. *Source,* 14 February. Available from: https://source.opennews.org/articles/data-rescue/. Accessed on 6 February 2022.

PICKARD, A.J. 2013. *Research Methods in Information*. London: Facet.

PICAS, J. & GRAB, S. 2017. ACRE South Africa. *C3S Data Rescue Capacity Building Workshop, NIWA, Auckland, New Zealand, 4–6 December 2017*. Available from: https://www.dropbox.com/s/texp3ahukbj5et4/ACRE%20SA%20Presentation%202017%20%28updated%29.pptx?dl=0. Accessed on 6 February 2022.

PICKELL, D. 2021. Qualitative vs Quantitative Data – What's the Difference? *G2,* 14 May. Available from: https://www.g2.com/articles/qualitative-vs-quantitative-data#what-is-quantitative-data. Accessed on 3 July 2021.

PIWOWAR, H.A., DAY, R.S. & FRIDSMA, D.B. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, vol. 2(3): e308. Available from: https://doi.org/10.1371/journal.pone.0000308.

PIZARRO-TAPIA, R., GONZÁLEZ-LEIVA, F., VALDÉS-PINEDA, R., INGRAM, B., SANGÜESA, C. & VALLEJOS, C. 2020. A Rainfall Intensity Data Rescue Initiative for Central Chile Utilizing a Pluviograph Strip Charts Reader (PSCR). *Water*, vol. 12(7): https://doi.org/10.3390/w12071887.

PLANETTOGETHER. 2020. Advantages and Disadvantages of Flowcharts. *PlanetTogether,* 24 August. Available from: https://www.planettogether.com/blog/advantages-and-disadvantages-of-flowcharts. Accessed on 6 February 2022.

PONS, F., LAROCHE, C., TRMAL, C., PUECHBERTY, R. & BAILLON, M. 2016. Hydrometry data rescue, a stake for the future. *E3S Web of Conferences*, vol. 7: 04–021. Available from: https://www.e3s-conferences.org/articles/e3sconf/pdf/2016/02/e3sconf_flood2016_04021.pdf. Accessed on 6 February 2022.

POPKIN, G. 2019. Data sharing and how it can benefit your scientific career. *Nature*, vol. 569: 445–447. Available from: https://www.nature.com/articles/d41586-019-01506-x. Accessed on 6 February 2022.

PRICE, G. 2018. South Africa: DataFirst Rescues Historical Apartheid-Era Datasets. *INFOdocket*, 12 December. Available from: https://www.infodocket.com/2018/12/12/south-africa-datafirst-rescues-historical-apartheid-era-datasets/. Accessed on 6 February 2022.

PRICE, P.C., JHANGIANI, R. & CHIANG, I. A. 2015. *Research Methods in Psychology – 2nd Canadian Edition.* Victoria, B.C.: BCcampus. Available from: https://opentextbc.ca/researchmethods/. Accessed on 7 November 2021.

QUESTIONPRO. 2021. *Focus Group research: Steps to conduct a focus group*. Available from: https://www.questionpro.com/blog/focus-group/. Accessed on 4 June 2021.

RAMBO, N. 2015. *Research Data Management: Roles for Libraries*. Available from: https://doi.org/10.18665/sr.274643.

RAPPERT, B. & BEZUIDENHOUT, L. 2016. Data sharing in low-resourced research environments. *Prometheus*, vol. 34(3–4): 207–224. Available from: https://doi.org/10.1080/08109028.2017.1325142.

RAUGHLEY, L. 2017. Library Advances Data Rescue Effort. *University of Michigan Library, News*, 30 January. Available from: https://www.lib.umich.edu/news/library-advances-data-rescue-effort. Accessed on 8 December 2019.

RAVITCH, S.M. & RIGGAN, M. 2017. *Reason & Rigor: How Conceptual Frameworks Guide Research*. Thousand Oaks, California: Sage.

REBIND. BIODIVERSITY NEEDS DATA. 2015. *reBiND Movie*. Available from: https://rebind.bgbm.org/rebind_movie. Accessed on 6 February 2022.

REGIONAL CENTRE FOR MAPPING OF RESOURCES FOR DEVELOPMENT (RCMRD). 2016. *Tanzania Meteorological Agency Climate Data Rescue Assessment Workshop held in Tanzania*. SERVIR, Regional Centre for Mapping of Resources for Development. Available from: https://rcmrd.org/about-servir/2-uncategorised/313-tanzania-meteorological-agency-climate-data-rescue-assessment-workshop-held-in-tanzania. Accessed on 6 February 2022.

REGMI, P.R., WAITHAKA, E., PAUDYAL, A., SIMKHADA, P. & VAN TEIJLINGEN, E. 2016. Guide to the design and application of online questionnaire surveys. *Nepal Journal of Epidemiology*, vol. 6(4): 640–644. Available from: https://doi.org/10.3126/nje.v6i4.17258.

RESEARCH DATA ALLIANCE (RDA), DATA CONSERVATION IG. 2019. *Data Conservation IG*. Available from: https://www.rd-alliance.org/groups/data-conservation-ig. Accessed on 6 February 2022.

RESEARCH DATA ALLIANCE (RDA), DATA RESCUE IG. 2015. *Charter: Data Rescue Interest Group (DR-IG)*. Available from: https://rd-alliance.org/sites/default/files/rda-dr.pdf. Accessed on 10 February 2022.

RESEARCH DATA ALLIANCE (RDA). DATA RESCUE IG. 2017a. *Guidelines for Data Rescue*. Available from: https://www.rd-alliance.org/guidelines-data-rescue-0. Accessed on 10 February 2022.

RESEARCH DATA ALLIANCE (RDA), DATA RESCUE IG. 2017b. *Two of us at NAGARA 2017 in Boise.* Available from: https://www.rd-alliance.org/group/data-rescue-ig/post/two-us-nagara-2017-boise. Accessed on 10 February 2022.

RESEARCH DATA ALLIANCE (RDA), DATA RESCUE IG. 2019. *Data Rescue IG*. Available from: https://www.rd-alliance.org/groups/data-rescue.html. Accessed on 10 February 2022.

RESEARCH DATA MANAGEMENT LIBRARIAN ACADEMY. 2019. *Research Data Management Librarian Academy (RDMLA)*. Available from: https://rdmla.github.io/. Accessed on 7 January 2023.

RICE, R. & MACDONALD, S. 2013. *Research Data Management Training for Librarians - An Edinburgh Approach*. Available from: https://www.slideshare.net/edinadocumentationofficer/research-data-management-training-for-librarians. Accessed on 7 January 2023.

RITCHEY, N. 2017. Risk evaluation for Data Rescue*. 2017 AGU Fall Meeting, New Orleans, USA, 1115 December 2017.* Available from: https://docs.google.com/presentation/d/1WbCu4aZShRz-vOCBLRX8psodp0xxPELUuotHNCJQxzc/edit#slide=id.p. Accessed on 6 February 2022.

ROBSON, C. 2002. *Real World Research: A Resource for Social Scientists and Practitioner-Researchers.* Oxford, UK: Blackwell.

ROSS, S. & GOW, A. 1999. *Digital archaeology: rescuing neglected and damaged data resources*. A JISC/NPO study within the Electronic Libraries (eLib) Programme on the preservation of electronic materials. Project Report. Library Information Technology Centre, South Bank University, London. Available from: https://eprints.gla.ac.uk/100304/1/100304.pdf. Accessed on 6 February 2022.

ROUNTREE, R.A., PERKINS, P.J., KENNEY, R.D. & HINGA, K.R. 2002. Sounds of Western North Atlantic Fishes – Data Rescue. *Bioacoustics*, vol. 12(2–3): 242–244. Available from: https://doi.org/10.1080/09524622.2002.9753710.

ROYAL CONTENT RESEARCH SERVICES. 2021. *Types of Research Philosophy*. Available from: https://www.rcrservices.in/latest-update/selecting-an-appropr/42. Accessed on 10 February 2022.

THE ROYAL SOCIETY. 2012. *Science as an open enterprise*. Available from: https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf. Accessed on 31 December 2022.

RUDIN, C., ERTEKIN, Ş, PASSONNEAY, R., RADEVA, A., TOMAR, A., XIE, B., LEWIS, S., RIDDLE, M., PANGSRIVINIJ, D. & McCORMICK, T. 2014. Analytics for Power Grid Distribution Reliability in New York City. *INFORMS Journal on Applied Analytics*, vol. 44(4): 364–383. Available from: https://doi.org/10.1287/inte.2014.0748.

RYAN, C., BRODERICK, C., CURLEY, M., DALY, C., DUFFY, C., THORNE, P., TREANOR, M., WALSH, S. & MURPHY, C. 2017. Integrating Data Rescue into the Classroom. *European Geoscience Union (EGU) General Assembly 2017, Vienna, Austria, 23–28 April 2017*. Available from: https://library.wmo.int/doc_num.php?explnum_id=3583. Accessed on 23 March 2022.

RYAN, C., DUFFY, C., BRODERICK, C., THORNE, P.W., CURLEY, M., WALSH, S., DALY, C., TREANOR, M. & MURPHY, C. 2018. Integrating Data Rescue into the Classroom. *Bulletin of the American Meteorological Society*, vol. 99(9): 1757–1764. Available from: https://doi.org/10.1175/BAMS-D-17-0147.1. Accessed on 10 February 2022.

SALO, D. 2022. *Teaching and training*. Available from: https://dsalo.info/teaching/. Accessed on 7 January 2023.

SARWONO, J. 2022. *Quantitative, Qualitative and Mixed Method Research Methodology*. Washington: Amazon.

SAUNDERS, M.N.K., LEWIS, P. & THORNHILL, A. 2019. *Research Methods for Business Students*. Harlow: Pearson Education Limited.

SCHELL, J. 2018. Data rescue one year later: Sustaining a movement. *Endangered Data Panel Discussion, ASIS&T WSU*, 3 March. Available from: https://www.youtube.com/watch?v=5iyl6U2hVA4&t=9s. Accessed on 30 March 2022.

SCHLANGER, Z. 2017a. Rogue Scientists Race to Save Climate Data from Trump. *WIRED,* 19 January. Available from: https://www.wired.com/2017/01/rogue-scientists-race-save-climate-data-trump/. Accessed on 6 February 2022.

SCHLANGER, Z. 2017b. Hackers downloaded US government climate data and stored it on European servers as Trump was being inaugurated. *Quartz Media,* 21 January. Available from: https://qz.com/891201/hackers-were-downloading-government-climate-data-and-storing-it-on-european-servers-as-trump-was-being-inaugurated/. Accessed on 10 February 2022.

SCHMIDT, G. 2017. Data rescue projects. *RealClimate: Climate science from climate scientists,* 17 August. Available from: https://www.realclimate.org/index.php/archives/2017/08/data-rescue-projects/. Accessed on 10 February 2022.

SCHOCH, K. 2019. Case Study Research. In: Burkholder, G.J., Cox, K.A., Crawford, L.M. and Hitchcock, J.H. (*eds*). *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*. Thousand Oaks, California: Sage.

SCHROEDER, K. 2018. Fast Forward 10 Years…Where are your Assets? *Endangered Data Panel Discussion, ASIS&T WSU*, 3 March. Available from: https://www.youtube.com/watch?v=5iyl6U2hVA4&t=9s. Accessed on 30 March 2022.

SCHUMACHER, J. & VANDECREEK, D. 2015. Intellectual Capital at Risk: Data Management Practices and Data Loss by Faculty Members at Five American Universities. *International Journal of Data Curation*, vol. 10(2): 96–109. Available from: https://doi.org/10.2218/ijdc.v10i2.321.

SCOTT, C. 2013. Getting started with user focus groups. *American Library Association,* 5 September. Available from: https://www.ala.org/llama/sites/ala.org.llama/files/content/9-4-13%20Slides.pdf. Accessed on 4 June 2021.

SEADLE, M. 2018. *Replication testing*. Humboldt-Elsevier Advanced Data and Text Centre (HEADT Centre). Available from: https://headt.eu/column-on-information-integrity. Accessed on 7 November 2021.

SEADLE, M. & RÜGENHAGEN, M. 2018. *Replication in Qualitative Research*. Humboldt-Elsevier Advanced Data and Text Centre (HEADT Centre). Available from: https://headt.eu/Replication-in-Qualitative-Research/. Accessed on 7 November 2021.

SHEBLE, L. 2018. Data: Endangered & endangering. *Endangered Data Panel Discussion, ASIS&T WSU*, 3 March. Available from: https://www.youtube.com/watch?v=5iyl6U2hVA4&t=9s. Accessed on 30 March 2022.

SHIUE, H., CLARKE, C. T. & FENLON, K. 2020. Data Rescue at the U.S. National Agricultural Library: Case Studies of 3 Hybrid Collections. *Research Data Alliance 2020, Costa Rica, 9–12 November 2020 [virtual]*. Available from: https://drum.lib.umd.edu/handle/1903/26691. Accessed on 6 February 2022.

SHIUE H.S.Y., CLARKE C.T., SHAW M., HOFFMAN K.M. & FENLON K. 2021. Assessing Legacy Collections for Scientific Data Rescue. In: Toeppe, K., Yan, H., and Chu, S.K.W. (*eds*). *Diversity, Divergence, Dialogue. iConference 2021. Lecture Notes in Computer Science*, vol. 12646: https://doi.org/10.1007/978-3-030-71305-8_25.

SIMON, M.K. & GOES, J. 2013. *Assumptions, Limitations, Delimitations, and Scope of the Study.* Scribd. Available from: https://www.scribd.com/document/360403048/Assumptions-Limitations-Delimitations-and-Scope-of-the-Study. Accessed on 8 December 2019.

SIPES, J.B.A., ROBERTS, L.D. & MULAN, B. 2019. Voice-only Skype for use in researching sensitive topics: a research note. *Qualitative Research in Psychology*, vol. 19(1): 204–220. Available from*:* https://doi.org/10.1080/14780887.2019.1577518.

SLONOSKY, V. 2011. Rescuing data… *RealClimate: Climate science from climate scientists,* 7 April. Available from: https://www.realclimate.org/index.php/archives/2011/04/rescuing-data/. Accessed on 6 February 2022.

SLONOSKY, V. 2021. Bringing Past Weather Observations into Current Climate Change Research: the McGill DRAW (Data Rescue: Archives and Weather) Project. *Historical Climatology,* 18 March. Available from: https://www.historicalclimatology.com/projects/bringing-past-weather-observations-into-current-climate-change-research-the-mcgill-draw-data-rescue-archives-and-weather-project. Accessed on 6 February 2022.

SLONOSKY, V., SIEBER, R., BURR, G., PODOLSKY, L., SMITH, R., BARTLETT, M., PARK, E., CULLEN, J. & FABRY, F. 2019. From books to bytes: A new data rescue tool. *Geoscience Data Journal*, 6(1): 58–73. Available from: https://doi.org/10.1002/gdj3.62.

SMARTDRAW. 2021. *Flowchart Symbols*. Available from: https://www.smartdraw.com/flowchart /flowchart-symbols.htm. Accessed on 28 October 2021.

SMITH, M.J. 2015. Use of legacy data in geomorphological research. *GeoResJ*, vol. 6: 7480. Available from: https://doi.org/10.1016/j.grj.2015.02.008.

SMITHSONIAN INSTITUTION ARCHIVES. 2011. *Setting up paper files*. Available from: https://siarchives.si.edu/what-we-do/setting-paper-files. Accessed on 6 February 2022.

SOCIAL SCIENCE ENVIRONMENTAL HEALTH RESEARCH INSTITUTE. 2017. *Northeastern hosts 'archive-a-thon' to preserve federal science data.* Northeastern University. Available from: https://www.northeastern.edu/environmentalhealth/northeastern-hosts-archive-a-thon-to-preserve-federal-science-data/. Accessed on 6 February 2022.

STAHL, N.A. & KING, J.R. 2020. Expanding Approaches for Research: Understanding and Using Trustworthiness in Qualitative Research. *Journal of Developmental Education*, vol. 44(1): 26–28. Available from: https://files.eric.ed.gov/fulltext/EJ1320570.pdf. Accessed on 28 December 2022.

STANLEY, V., KOPPERS, A., STONER, J., CHESEBY, M., FRITZ, C. & MINNETT, R. 2020. Relocating the Antarctic Core Collection: The Story of a Large-Scale Data Rescue Initiative for an Historic

Collection of Geologic Samples. *Earth and Space Science Open Archive*: https://doi.org/10.1002/essoar.10501988.1.

STRAUSS, A. & CORBIN, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (2nd ed.). Thousand Oaks, California: Sage.

STUDYLIB. 2019. *Limitations and Delimitations of Research*. Available from: https://studylib.net/doc/6652765/limitations-and-delimitations-of-research. Accessed on 8 December 2019.

TANG, R. & HU, Z. 2019. Providing Research Data Management (RDM) Services in Libraries: Preparedness, Roles, Challenges, and Training for RDM Practice. *Data and Information Management*, vol. 3(2): 84–101. Available from: https://doi.org/10.2478/dim-2019-0009.

TECH DIFFERENCES. 2019. *Differences between algorithm and flowchart*. Available from: https://techdifferences.com/difference-between-algorithm-and-flowchart.html. Accessed on 8 December 2019.

TEHERANI, A., MARTIMIANAKIS, T., STENFORS-HAYES, T., WADHWA, A. & VARPIO, L. 2015. Choosing a Qualitative Research Approach. *Journal of Graduate Medical Education*, vol. 7(4): 669–670. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4675428/pdf/i1949-8357-7-4-669.pdf. Accessed on 26 February 2022.

TESOL INTERNATIONAL ASSOCIATION. 2021. *Qualitative Research: Case Study Guidelines*. Available from: https://www.tesol.org/read-and-publish/journals/tesol-quarterly /tesol-quarterly-research-guidelines/qualitative-research-case-study-guidelines. Accessed on 6 February 2022.

THOMPSON, C.A., DAVENPORT ROBERTSON, W. & GREENBERG, J. 2014. Where Have All the Scientific Data Gone? LIS Perspective on the Data-At-Risk Predicament. *College & Research Libraries*, vol. 75(6): 842–861. Available from: https://doi.org/10.5860/crl.75.6.842.

THOMPSON, K. 2017. Documentation as Data Rescue: Restoring a Collection of Canadian Health Survey Files. *Against the Grain*, vol. 29(6): 33–35. Available from: https://doi.org/10.7771/2380-176X.7876.

THORNE, P.W., HUTCHINSON, R., ALLAN, R., CROUTHAMEL, R, ANGEL, W., BRUNET, M., BRONNIMANN, S., LAWRIMORE, J., MACHEL, H., LUTERBACHER, J. *et al*. 2011. *The International Surface Temperature Initiative: Data Rescue.* International Surface Temperature Initiative.

Available from: https://4310b1a9-a-85778f51-s-sites.googlegroups.com/a/surfacetemperatures.org/home/databank/data-rescue-task-team/ISTI_data_rescue_poster.pdf?attachauth=ANoY7crRnoWFVxzCeVhEtKEB47dRE6OZkAkjFCzRdBPiQFK7yaHvWiwo1QuhQIZandJVXcJdVtUsVmovrpweO4j6ccCDEHVXFIozwgpbE8Dg8tRlKuX6BZVMhot0VZSl07338YFxzzqBpK8UZMJI0Q9fZf2lqXDUwDP6oGD1X4DbMj6cOrB9aTDtWLpEus1Po280YUkirNDTeEsBvSEzpiE9DpWFOIpEtwKxnMex-7UfxwDxMWBvDP0pGY5X00I366XljFGDNK6jM42salqaAoJtbHuNTiaotw%3D%3D&attredirects=0. Accessed on 6 February 2022.

THORNTON, S. 2017. What are the benefits of a flowchart? *Bizfluent,* 26 September. Available from: https://bizfluent.com/about-5455875-workflow-diagram.html. Accessed on 8 December 2019.

TROTMAN, A.R. 2012. *Why data? Climate Monitoring, Sectoral Applications and More….* Available from: https://www.mona.uwi.edu/physics/sites/default/files/physics/uploads/Trotman-%20Examination%20of%20Regional%20Rainfall%20and%20Drought%20Projections%20Part%202.pdf. Accessed on 6 February 2022.

TUCKER, V.M. 2021. Becoming an information architect: The evolving librarian's skillset, mindset, and professional identity. *Education for Information,* vol. 37(2): 1–16. Available from: https://doi.org/10.3233/EFI-211558.

TUTOR2U. 2021. Content and Thematic Analysis. *Tutor2u,* 22 March. Available from: https://www.tutor2u.net/psychology/reference/content-and-thematic-analysis. Accessed on 18 November 2021.

UCD LIBRARY. 2019. *Research Data Management: Why Share Research Data*. Available from: https://libguides.ucd.ie/data/share. Accessed on 6 February 2022.

UKESSAYS. 2021. The Advantages and Disadvantages of Case Study Research. *UKEssays,* 19 July. Available from: https://www.ukessays.com/essays/psychology/the-advantages-and-disadvantages-of-case-study-research-psychology-essay.php?vref=1. Accessed 8 November 2021.

UNAIDS. 2010. *An introduction to triangulation*. Available from: https://www.unaids.org/sites/default/files/sub_landing/files/10_4-Intro-to-triangulation-MEF.pdf. Accessed on 31 December 2022.

UNIDATA. 2016. Upcoming Workshop: The Rescue of Data at Risk. *Unidata,* 29 June. Available from: https://www.unidata.ucar.edu/blogs/news/entry/upcoming-workshop-the-rescue-of. Accessed on 6 February 2022.

UNISA, LIBRARY. 2022. *Research Data Management*. Available from: https://www.unisa.ac.za/sites/corporate/default/Library/Library-services/Research-support/Research-Data-Management. Accessed on 14 January 2023.

UNITED NATIONS ECONOMIC AND SOCIAL COMMISSION FOR WESTERN ASIA. 2013. *Sub-regional Training Workshop on Climate Data Rescue and Digitization*. United Nations Economic and Social Commission For Western Asia. Available from: https://www.unescwa.org/events/sub-regional-training-workshop-climate-data-rescue-and-digitization. Accessed on 10 February 2022.

UNITED STATES GEOLOGICAL SURVEY. 2004. *Physical Sample Preservation and Curation*. Available from: https://www.usgs.gov/programs/national-geological-and-geophysical-data-preservation-program/physical-sample-preservation. Accessed on 18 January 2022.

UNITED STATES GEOLOGICAL SURVEY (USGS). 2019a. *Data at Risk Project*. Available from: https://apps.usgs.gov/ldi/data-at-risk-project. Accessed on 8 December 2019.

UNITED STATES GEOLOGICAL SURVEY (USGS). 2019b. *LDIRS Evaluations and Reporting*. Available from: https://apps.usgs.gov/ldi/evaluations-reports. Accessed on 8 December 2019.

UNIVERSITY COLLEGE DUBLIN, UNIVERSITY LIBRARY. 2022. *Research Data Management: Why Share Research Data.* Available from: https://libguides.ucd.ie/data/share. Accessed on 27 September 2022.

UNIVERSITY OF CAPE TOWN, UCT LIBRARIES. 2023. *Research Data Management*. Available from: https://lib.uct.ac.za/digitalservices/services/research-data-management. Accessed on 14 January 2023.

UNIVERSITY OF EDINBURGH. REFLECTION TOOLKIT. 2020. *What? So what? Now what?* Available from: https://www.ed.ac.uk/reflection/reflectors-toolkit/reflecting-on-experience/what-so-what-now-what. Accessed on 7 January 2023.

UNIVERSITY OF EDINBURGH, RESEARCH DATA SERVICE. 2019. *Benefits of writing a DMP*. Available from: https://www.ed.ac.uk/information-services/research-support/research-data-service/before/benefits-of-writing-a-dmp. Accessed on 8 December 2019.

UNIVERSITY OF HAWAII, SEA LEVEL CENTER. 2018. *Data Rescue and Digitization*. Available from: https://uhslc.soest.hawaii.edu/data-rescue/. Accessed on 6 February 2022.

UNIVERSITY OF MARYLAND, MARYLAND INSTITUTE FOR TECHNOLOGY IN THE HUMANITIES. 2018. *Endangered Data Week*. Available from: https://mith.umd.edu/research/endangered-data-week/. Accessed on 10 February 2021.

UNIVERSITY OF MINNESOTA LIBRARIES. 2017. *DataRescue-Twin Cities*. Available from: https://www.facebook.com/events/223317941409060/. Accessed on 6 February 2022.

UNIVERSITY OF NEVADA, UNIVERSITY LIBRARIES. 2017. *DATA Rescue Event*. Available from: https://unr.libcal.com/event/3232847. Accessed on 10 February 2022.

UNIVERSITY OF NOTRE DAME, HESBURG LIBRARIES. 2019. *Community-Based Data Rescue Nomination Tool.* Available from: https://library.nd.edu/event/love-data-week-2019-lightning-tutorials. Accessed on 10 February 2022.

UNIVERSITY OF PRETORIA, DEPARTMENT OF LIBRARY SERVICES. 2021. *Digitisation.* Available from: https://library.up.ac.za/digi. Accessed on 8 January 2023.

UNIVERSITY OF PRETORIA, DEPARTMENT OF LIBRARY SERVICES. 2022. *Research Data Management (RDM): Home*. Available from: https://library.up.ac.za/c.php?g=356288. Accessed on 14 January 2023.

UNIVERSITY OF PRETORIA, INFORMATION SCIENCE. 2022. *List of Postgraduate modules presented by the Department of Information Science*. Available from: https://www.up.ac.za/media/shared/117/Study%20Postgraduate%20Programmes/postgraduate-modules-presented-by-dept-information-science-01.zp198825.pdf. Accessed on 11 January 2023.

UNIVERSITY OF VIRGINIA, LIBRARY SERVICES. 2018. *Endangered Data Week (Feb 26–Mar 2, 2018).* Available from: https://data.library.virginia.edu/endangered-data-week-2018/. Accessed on 10 February 2022.

UTRECHT UNIVERSITY, DIGITAL HUMANITIES LAB. 2019. *Endangered Data Week 2019*. Available from: https://dig.hum.uu.nl/endangered-data-week-2019/. Accessed on 29 January 2022.

VAISMORADI, M., TURUNEN, H. & BONDAS, T. 2013. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & Health Sciences*, vol. 15(3): 398–405. Available from: https://doi.org/10.1111/nhs.12048.

VAN DER VELDE, T., MILTON, D.A., LAWSON, T.J., WILCOX, C., LANSDELL, M., DAVIS, G., PERKINS, G. & HARDESTY, B.D. 2017. Comparison of marine debris data collected by researchers and citizen scientists: Is citizen science data worth the effort? *Biological conservation*, vol. 208: 127–138. Available from: https://doi.org/10.1016/j.biocon.2016.05.025.

VARINSKY, D. 2017. Scientists across the US are scrambling to save government research in 'Data Rescue' events. *Business Insider*, 11 February. Available from: https://www.businessinsider.com/data-rescue-government-data-preservation-efforts-2017-2?IR=T. Accessed on 29 January 2022.

VARPIO, L., PARADIS, E., UIJTDEHAAGE, S. & YOUNG, M. 2020. The Distinctions Between Theory, Theoretical Framework, and Conceptual Framework. *Academic Medicine,* vol. 95(7): 989–994. Available from: https://journals.lww.com/academicmedicine/fulltext/2020/07000/the_distinctions_between_theory,_theoretical.21.aspx. Accessed on 19 January 2023.

WALLIMAN, N. 2011. *Research Methods: The Basics*. New York: Routledge.

WALWYN, D. 2017. *Session Four: Techniques for Data Gathering and Analysis*. University of Pretoria, Department of Library Services. Available from: http://up-za.libguides.com/ld.php?content_id=34337838. Accessed on 8 December 2019.

WARGO, W.G. 2015. *Identifying Assumptions and Limitations for your Dissertation*. Academic Info Center. Available from: https://www.academia.edu/33174930/How_to_Write_Assumptions_for_a_Thesis. Accessed on 31 March 2022.

WEARE, W.H. 2013. Focus Group Research in the Academic Library: An Overview of the Methodology. *Qualitative and Quantitative Methods in Libraries*, vol. 1: 47–58. Available from: http://qqml.net/papers/March_2013_Issue/216QQML_Journal_2013_Weare_1.47-58.pdf. Accessed 4 June 2021.

WEISBERG, H.F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press. Available from: https://www.academia.edu/31859666/Weisberg_2005_TheTotal_Survey_Error_Approach?auto=download. Accessed on 8 December 2019.

WELLINGTON, J. & SZCZERBINSKI, M. 2007. *Research Methods for the Social Sciences*. London: Bloomsbury.

WELMAN, C., KRUGER, F. & MITCHELL, B. 2012. *Research Methodology.* Cape Town: Oxford University Press.

WELZ, A. 2017. Unnatural Surveillance: How Online Data Is Putting Species at Risk. *Yale Environment 360*, 6 September. Available from: https://e360.yale.edu/features/unnatural-surveillance-how-online-data-is-putting-species-at-risk. Accessed on 13 January 2023.

WERN, L. 2017. Data rescue of Swedish data. *11th EUMETNET Data Management Workshop, 18–20 October 2017, Zagreb, Croatia*. Available from: http://meteo.hr/DMW_2017/presentations/Wednesday03.pdf. Accessed on 29 January 2022.

WIEBE, P.H. & ALLISON, M.D. 2015. Bringing dark data into the light: a case study of the recovery of Northwestern Atlantic zooplankton data collected in the 1970s and 1980s. *GeoResJ*, vol. 6: 195–201. Available from: https://doi.org/10.1016/j.grj.2015.03.001.

WILKINSON, C., BRÖNNIMANN, S., JOURDAIN, S., ROUCATE, E., CROUTHAMEL, R., IEDRO TEAM, BROHAN, P., VALENTE, A., BRUGNARA, Y., BRUNET, M. *et al*. 2019. *Best Practice Guidelines for Climate Data Rescue.* International Data Rescue (I-DARE) Portal. Available from: https://idare-portal.org/sites/default/files/BestPracticeGuidelines_Part1_v1_NoBranding_0.pdf. Accessed on 29 January 2022.

WILLIAM & MARY SCHOOL OF EDUCATION, DOCTORAL DOCUMENTS. 2019. *Sample Chapter 1 and 3 Outlines*. Available from: https://education.wm.edu/documents/doctoral/Sample%20Chapter%201%20and%203%20Outlines.pdf. Accessed on 8 December 2019.

WILLIAMS, R. 2017. Michigan web developers and archivists join race to back up federal agency data. *Michigan Radio,* 29 January. Available from: https://www.michiganradio.org/environment-science/2017-01-29/michigan-web-developers-and-archivists-join-race-to-back-up-federal-agency-data. Accessed on 29 January 2022.

WILLIAMSON, F., ALLAN, R., SWITZER, A., CHAN, J.C. L., WASSON, R.J., D'ARRIGO, R. & GARTNER, R. 2015. New directions in hydro-climatic histories: observational data recovery, proxy records and the atmospheric circulation reconstructions over the earth (ACRE) initiative in Southeast Asia. *Geoscience Letters,* vol. 2(2): 1–12. Available from: https://doi.org/10.1186/s40562-015-0018-z.

WIPPICH, C. 2012. *Preserving science for the ages--USGS data rescue*. Fact Sheet 2012-3078. United States Geological Survey. Available from: https://pubs.er.usgs.gov/publication/fs20123078. Accessed on 29 January 2022.

WOOD, C.M., HOWARD, D.C., HENRYS, P.A., BUNCE, R.G.H. & SMART, S.M. 2012. *Countryside Survey: measuring habitat change over 30 years: 1978 data rescue – final report.* Lancaster, NERC/Centre for Ecology & Hydrology, 18pp. (CEH Project Number: C03689). Available from: http://nora.nerc.ac.uk/id/eprint/16880/. Accessed on 29 January 2022.

WOOLFREY, L. 2016. The Rescue of “at Risk” data on forced resettlement in South Africa. *RDA/CODATA Workshop on the Rescue of Data at Risk, National Center for Atmospheric Research, Boulder, Colorado, USA, 8–9 September 2016*.

WOOLFREY, L. 2021. Email to Louise Patterton, 17 August 2021.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2012*. International Workshop on Data Rescue and Digitization of Climate Records for countries in West Africa, 19–23 November 2012, Accra, Ghana*. Available from: https://library.wmo.int/index.php?lvl=notice_display&id=14002#.YUBnMp0zbxp. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2014. *Guidelines for Hydrological Data Rescue.* Available from: https://library.wmo.int/index.php?lvl=notice_display&id=16792#.YXJuBxpBxAU. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2016. *Guidelines on Best Practices for Climate Data Rescue.* Available from: https://library.wmo.int/doc_num.php?explnum_id=3318. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2016. *Summary Report: Indian Ocean Data Rescue (INDARE) Capacity building workshop on quality control and data homogenisation to support climate assessments and services in the Indian Ocean rim countries and islands.* Available from: https://gfcs.wmo.int/sites/default/files/Report_Arusha_final.pdf. Accessed on 12 October 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2017. Uzbekistan rescues historical meteorological and hydrological data*. World Meteorological Organization,* 1 August. Available from: https://public.wmo.int/en/media/news/uzbekistan-rescues-historical-meteorological-and-hydrological-data. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2018. *Guide to Climatological Practices.* Available from: https://library.wmo.int/doc_num.php?explnum_id=5541. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2019a. *Joint C3S Data Rescue Service Capacity Building Workshop and 12th ACRE Meeting, 8–12 April 2019, Buenos Aires, Argentina*. Available from: https://public.wmo.int/en/events/meetings/joint-c3s-data-rescue-service-capacity-building-workshop-and-12th-acre-meeting. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2019b. *Making a Difference on the Ground: Data Rescue in Burkina Faso, Mali and Niger*. Available from: https://library.wmo.int/doc_num.php?explnum_id=3582. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION (WMO). 2020. *Mauritius Data Rescue Workshop (POSTPONED).* Available from: https://public.wmo.int/en/events/workshops/mauritius-data-rescue-workshop-postponed. Accessed on 29 January 2022.

WYBORN, L., HSU, L., LEHNERT, K. & PARSONS, M.A. 2015. Guest Editorial: Special issue Rescuing Legacy data for Future Science. *GeoResJ*, vol. 6: 106–107. Available from: https://doi.org/10.1016/j.grj.2015.02.017.

WYLIE, S. 2021. How scientists scrambled to stop Donald Trump's EPA from wiping out climate data. *Northeastern University, College of Social Sciences and Humanities,* 3 September. Available from: https://cssh.northeastern.edu/how-scientists-scrambled-to-stop-donald-trumps-epa-from-wiping-out-climate-data/. Accessed on 29 January 2022.

XU, Z. 2022. Research Data Management Practice in Academic Libraries. *Journal of Librarianship and Scholarly Communication*, vol. 10(1): eP13700. Available from: https://doi.org/10.31274/jlsc.13700.

YALE UNIVERSITY. HARVEY CUSHING/JOHN HAY WHITNEY MEDICAL LIBRARY. 2019. *Endangered Data Week 2019 at CWML.* Available from: https://library.medicine.yale.edu/blog/medical-library/endangered-data-week-2019-cwml. Accessed on 29 January 2022.

YIN, R.K. 2014. *Case Study Research: Design and Methods*. Thousand Oaks, California: Sage.

ZAINAL, Z. 2007. Case study as a research method. *Jurnal Kemanusiaan*, vol. 5(1): 6pp. Available from: https://core.ac.uk/download/pdf/11784113.pdf. Accessed on 4 June 2021.

ZAVERI, M., GATES, G. & ZRAICK, K. 2019. The Government Shutdown Was the Longest Ever. Here's the History. *The New York Times*, 25 January. Available from: https://www.nytimes.com/interactive/2019/01/09/us/politics/longest-government-shutdown.html. Accessed on 29 January 2022.

ZENK-MÖLTGEN, W., AKDENIZ, E., KATSANIDOU, A., NAẞHOVEN, V. & BALABAN, E. 2018. Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal* *of Documentation*, vol. 74(5): 1053–1073. Available from: https://doi.org/10.1108/JD-09-2017-0126.

ZHANG, Y. & WILDEMUTH, B.M. 2009. *Qualitative Analysis of Content*. Available from: https://www.ischool.utexas.edu/~yanz/Content_analysis.pdf. Accessed on 5 January 2023.

ŽUKAUSKAS, P., VVEINHARDT, J. & ANDRIUKAITIENĖ, R. 2018. Philosophy and Paradigm of Scientific Research. In: Žukauskas, P., Vveinhardt, J. & Andriukaitienė, R. *Management Culture and Corporate Social Responsibility.* Available from: https://www.intechopen.com/chapters/58890. Accessed on 29 January 2022.

APPENDICES

Important notice regarding appendices:

- All appendices, except the letters containing details of institutional research approval and ethical clearance, were drafted by this researcher.
- Appendices either contain details regarding the study methodology, or provide guidelines regarding activities forming part of the data rescue workflow model.
- Where necessary, documents are redacted.
- Redaction is indicated via the following shape: [redacted]

APPENDIX 1 E-mail informing Sample A about upcoming web-based questionnaire

Dear Research Group Leader

My name is Louise Patterton, I am the CSIR Data Librarian, and I am writing to ask your help in a project investigating the prevalence of data at risk at the CSIR, and data rescue activities performed. This project forms part of my PhD studies in Information Science at the University of Pretoria. The main research objective of this study is to assuage and contribute towards the rescue of data at risk.

Sub-objectives of the study are the following:

- establishing how data rescue is currently done, both locally and internationally,
- adding to the above: determining who is currently involved in data rescue, and their roles,
- establishing current data rescue perceptions, needs, and challenges,
- creating a data rescue workflow based on information gained via the three previous points,
- within the model: stipulating the how library and information science professionals can be involved in data rescue,
- adapting a data rescue model, created by this researcher model, based on data collection outcomes, and
- contributing towards future data rescue activities by ensuring the model is freely available to all interested parties.

A first part of this study entails the completion of a short web-based questionnaire by CSIR Research Group Leaders (RGLs). It is thought that RGLs are most likely to have knowledge or information concerning:

- data at risk in their research group, and/or
- data rescue activities performed by their research group.

Information supplied by RGLs via the short questionnaire will provide a good overall glimpse of the current state of data at risk, and data rescue at the CSIR. Managerial approval from each cluster director to approach RGLs has been obtained. Ethical clearance has been obtained from the University of Pretoria, and the Research Ethics Committee of the CSIR.

Kindly take note of the following:

- This online questionnaire is not anonymous, as this researcher will be involving a next sample of RGLs during the next phase of this study, based on their questionnaire responses. As such, your name and name of research group will be required during completion of this questionnaire.
- Confidentiality is guaranteed. The online survey software, eSurv, guarantees that online responses are visible to the survey administrator/author of this study only. No one other than this researcher will know your individual answers to this questionnaire. Exported data will be stored in a password-protected folder.
- Results will be anonymised and de-identified before being published or shared.

Your participation is this research project is completely voluntary. There are no known risks to participation beyond those encountered in every life. However, you can help me very much by taking a few minutes to share your data management habits and requirements. If possible, please complete the questionnaire by ***date***.

The questionnaire contains eight questions and will take a maximum of 15 minutes to complete. Kindly answer as truthfully as possible; there are no correct or incorrect answers. Respondents are free to refuse to answer any question, or to withdraw at any time. No information supplied will be held against you, now or at any time in the future.

Upon completion of the study, I will be providing you with a handle of the published thesis, the DOIs of articles, and the DOI of the published data.

Kindly view the consent form following this letter, and click YES if in agreement. A YES response will lead you to the questionnaire part of the tool, while a NO response will result in termination of the session (i.e. survey tool will not progress to question section).

If you have any additional questions about this questionnaire, or about the study, feel free to contact me using the details below. Should you wish to discuss or query details of the study with someone other than this researcher, kindly contact:

- [redacted]

or

-

Thank you for your assistance in this important endeavour.

Sincerely,

Louise Patterton

APPENDIX 2 Web-based questionnaire

This appendix contains the text used in the web-based questionnaire. Kindly consult Section 4.5.1: Web-based questionnaire, for more information about this data collection instrument.

A. INTRODUCTION:

Dear Research Group Leader

Thank you for agreeing to take part in this survey measuring the prevalence of data at risk, and data rescue activities in the CSIR.

The survey will take approximately 15 minutes to complete.

The questionnaire is not anonymous, but be assured that all answers you provide will be kept in the strictest confidentiality.

Please click 'next' to begin. Clicking 'next' will take you to the consent statement. The questionnaire will be accessed after completing the consent statement.

B. CONSENT STATEMENT

INFORMED CONSENT FORM

- I hereby voluntarily grant my permission for participation in the project as explained to me by Louise Patterton.
    - The nature, the objective, and potential safety and health implications have been explained to me and I understand them.
    - I understand my right to choose whether to participate in the project and that the information offered will be managed confidentially. I am aware that the results of the investigation may be used for the purposes of publication.

Multiple choice, one answer only, 'yes' is required to proceed to the questions of the questionnaire

C. QUESTIONNAIRE

RESPONDENT DETAILS:

Name: *text box supplied*

Research Group: *text box supplied*

QUESTIONS

1. DOES YOUR RESEARCH GROUP HAVE DATA AT RISK?

(For the purposes of this study, data at risk is defined as research data in any format at risk due to:

- deterioration of the media (e.g. paper/microfilm damage, pest damage, war/unrest, loss of inventories)
- catastrophic loss (e.g. only one set of records, vulnerability to fire/floods/disasters, obsolete storage media, obsolete file format, archival destruction, loss of human knowledge related to use of the dataset)

Answer:

- *Yes*
- *No*
- *Unsure*

2. IF YOUR RESEARCH GROUP HAS DATA AT RISK, KINDLY INDICATE THE FORMAT(S) OF THE DATA. CLICK ALL THAT APPLY.

Answer:

- *Paper*
- *Magnetic tape*
- *Microfiche*
- *Early digital format*
- *Modern electronic format*
- *Physical samples*
- *Photographic plates*
- *Other formats*

If 'other formats', kindly supply details below.

(Text box for answer appears below)

3. IF YOUR RESEARCH GROUP HAS DATA AT RISK, KINDLY SUPPLY A BRIEF DESCRIPTION OF THE DATA.

Example:

- **Historical data, paper format, collected in the 1970s. Contains measurements of dam levels in South Africa**
- **Data collected in the early 50s, on microfilm. Unsure about the exact subject area, but probably mining-related. Data are on several tape reels, stored in five boxes.**

(Text box for answer appears below)

4. WHERE ARE THE DATA AT RISK CURRENTLY LOCATED?

(Text box for answer appears below)

5. HAS YOUR RESEARCH GROUP EVER PERFORMED DATA RESCUE ACTIVITIES?

Such activities include, but are not limited to:

- Creating an inventory of the data
- Imaging of paper media
- Keying paper data values onto an electronic spreadsheet
- Adding metadata to historical data
- Sharing previously private data on a data repository

Answer:

- *Yes*
- *No*
- *Unsure*

6. DOES YOUR RESEARCH GROUP HAVE ANY DOCUMENTATION (SUCH AS PROCEDURES, GUIDELINES, INSTRUCTIONS, AND PRINCIPLES) RELATED TO THE RESCUE OF DATA AT RISK?

Answer:

- *Yes*
- *No*
- *Unsure*

7. IF 'YES' TO QUESTION 6, IS THE DOCUMENTATION SHAREABLE WITH THE DATA LIBRARIAN AT A LATER STAGE?

Answer:

- *Yes*
- *No*
- *Unsure*

8. **IF YOU ARE AWARE OF ANY CSIR EMPLOYEE, RESEARCH GROUP, IMPACT AREA, OR CLUSTER WHO MIGHT HAVE DATA AT RISK, OR HAVE PARTICIPATED IN DATA RESCUE ACTIVITIES, KINDLY PROVIDE CONTACT DETAILS BELOW.**

   (Text box for answer appears below)

Thank you for completing this questionnaire. Your participation provides a valuable contribution towards this project and will enable this researcher to have a clearer understanding of data at risk in the CSIR, and data rescue activities performed.

For any problems experienced during the completion of the questionnaire, kindly contact me at lpatterton@csir.co.za

Once again, thank you for your time, input, and contribution.

Louise Patterton
(CSIR Data Librarian)

APPENDIX 3 Cover letter

Dear Research Group Leader

My name is Louise Patterton, I am the CSIR Data Librarian, and I am writing to ask your help in a project investigating the prevalence of data at risk at the CSIR, and data rescue activities performed. This project forms part of my PhD studies in Information Science at the University of Pretoria; the title being 'Data Rescue: Defining a Comprehensive Workflow that includes the Roles and Responsibilities of the Research Library'.

The main research objective of this study main is to contribute towards the rescue of historical data-at-risk by establishing objective ways in which parties, other than researchers and scientists, may be involved.

Sub-objectives of the study are the following:

- establishing how data rescue is currently performed, both locally and internationally,
- adding to the above: determining who is currently involved in data rescue, and their roles,
- establishing current data rescue perceptions, needs, and challenges,
- creating a data rescue workflow based on information gained via the three previous points,
- within the model: stipulating the how library and information science professionals can be involved in data rescue,
- adapting a data rescue model, created by this researcher model, based on data collection outcomes, and
- contribute towards future data rescue activities by ensuring the model is freely available to all interested parties.

A first part of this study entails the completion of a short web-based questionnaire by CSIR Research Group Leaders (RGLs). It is thought that RGLs are most likely to have knowledge or information concerning:

- data at risk in their research group, and/or
- data rescue activities performed by their research group.

Information supplied by RGLs via the short questionnaire will provide a good overall glimpse of the current state of data at risk, and data rescue at the CSIR.

Managerial approval from each cluster director to approach RGLs has been obtained. Ethical clearance has been obtained from the University of Pretoria, and the Research Ethics Committee of the CSIR.

Kindly take note of the following:

- This online questionnaire is not anonymous, as this researcher will be involving a next sample of RGLs during the next phase of this study, based on their questionnaire responses. As such, your name and name of research group will be required during completion of this questionnaire.
- After sampling is completed, this data will be anonymised, and the names of participants/groups not given back to the CSIR to avoid participation pressure.
- Confidentiality is guaranteed. The online survey software, eSurv, guarantees that online responses are visible to the survey administrator/author of this study only. No one other than this researcher will know your individual answers to this questionnaire. Exported data will be stored in a password-protected folder.
- Results will be anonymised and de-identified before being published or shared.

Your participation is this research project is completely voluntary. There are no known risks to participation beyond those encountered in every life. However, you can help me very much by taking a few minutes to share your data management habits and requirements. If possible, please complete the questionnaire by ***date***.

The questionnaire contains eight questions and will take a maximum of 15 minutes to complete. Kindly answer as truthfully as possible; there are no correct or incorrect answers. Respondents are free to refuse to answer any question, or to withdraw at any time. No information supplied will be held against you, now or at any time in the future.

Upon completion of the study, I will be providing you with a handle of the published thesis, the DOIs of articles, and the DOI of the published data.

Kindly view the consent form following this letter, and click YES if in agreement. A YES response will lead you to the questionnaire part of the tool, while a NO response will result in termination of the session (i.e. survey tool will not progress to question section).

APPENDIX 4 Consent form

I, ……………………..*name of participant*…………………… hereby voluntarily grant my permission for participation in the project as explained to me by Louise Patterton.

The nature, objective, possible safety and health implications have been explained to me and I understand them.

I understand my right to choose whether to participate in the project and that the information furnished will be handled confidentially. I am aware that the results of the investigation may be used for the purposes of publication.

Click YES if in agreement with the consent statement. Clicking YES will grant entry to the questionnaire.

Click NO if not agreeing to consent. Clicking NO will terminate this questionnaire session.

APPENDIX 5 Email to Sample A respondents regarding interview

Dear *****

Thank you for taking the time to complete the web-based questionnaire on data at risk, and data rescue activities. A preliminary analysis of the data has revealed that your research group either has data at risk, and/or has participated in data rescue activities. Based on the responses supplied, I am inviting you to form part of the next phase of this study.

The next phase entails a virtual interview with each RGL (or their proxy), who either has data at risk, or has performed data rescue activities. I am hereby requesting a Skype audio interview with you to discuss the following topics:

- data at risk in your research group,
- data rescue activities performed by your group (if applicable),
- your data rescue requirements, expectations, challenges, and suggestions, and
- a demonstration/brief discussion of the first draft of a Data Rescue Workflow Model created by me.

Your participation will be a valuable addition to this research, and findings could lead to greater understanding of data rescue within a research institute.

Interview details are the following:

- Due to COVID regulations, interviews will have to be virtual in nature, using Skype audio. A web camera is not required.
- Interviews will be 20–40 minutes in length.
- There will be no right or wrong answers; the objective of the interview is to gain more information on data at risk in the group, data rescue performed (if applicable), and data rescue requirements, challenges, expectations, and suggestions.

If you are willing to participate in a Skype audio interview, kindly suggest a date and time that suits you.

The time frame is 7 September 2020–29 September 2020. If such an interview is logistically impossible or inconvenient, would you be willing to consider providing text-based answers after being sent the questions forming part of the interview schedule?

If you have any questions, please do not hesitate to ask.

Regards

Louise Patterton

(CSIR Data Librarian/ PhD student)

APPENDIX 6 Email to non-Sample A respondents regarding interview

Good day

I suspect [redacted] might have informed you about an interview I wish to conduct with RGLs or their proxy, regarding data at risk in their group. CSIR RGLs completed a short web questionnaire some weeks ago, and those who had indicated to have data at risk, are invited for interviews.

Kindly find the interview invitation letter that was distributed, below. Let me know which date and time would suit you for a Skype audio interview. If that is inconvenient or not possible, supplying responses via text after being supplied with the interview schedule is also an option.

Regards

Louise Patterton

****

Dear *****

Thank you for taking the time to complete the web-based questionnaire on data at risk, and data rescue activities. A preliminary analysis of the data has revealed that your research group either has data at risk, and/or has participated in data rescue activities. Based on the responses supplied, I am inviting you to form part of the next phase of this study.

The next phase entails a virtual interview with each RGL (or their proxy), who either has data at risk, or has performed data rescue activities. I am hereby requesting a Skype audio interview with you to discuss the following topics:

- data at risk in your research group,
- data rescue activities performed by your group (if applicable),
- your data rescue requirements, expectations, challenges, and suggestions, and
- a demonstration/brief discussion of the first draft of a Data Rescue Workflow Model created by me.

Your participation will be a valuable addition to this research, and findings could lead to greater understanding of data rescue within a research institute.

Interview details are the following:

- Due to COVID regulations, interviews will have to be virtual in nature, using Skype audio. A web camera is not required.
- Interviews will be 20–40 minutes in length.
- There will be no right or wrong answers; the objective of the interview is to gain more information on data at risk in the group, data rescue performed (if applicable), and data rescue requirements, challenges, expectations, and suggestions.

If you are willing to participate in a Skype audio interview, kindly suggest a date and time that suits you.

The time frame is 7 September 2020–29 September 2020. If such an interview is logistically impossible or inconvenient, would you be willing to consider providing text-based answers after being sent the questions forming part of the interview schedule?

If you have any questions, please do not hesitate to ask.

Regards

Louise Patterton

(CSIR Data Librarian)

APPENDIX 7 Interview schedule

This appendix contains the interview guide used during virtual one-on-one interview with Sample B members. More information on the data collection tool and the activity can be obtained in Section 4.5.2: Virtual one-on-one interview, and Section 4.9.21: Conduct virtual one-on-one interviews.

The interview was semi-structured, which means that the following information is not included in the basic interview schedule:

- prompts used
- follow-up questions in the case of vague or ambiguous replies
- any other data rescue-related topics discussed

CONTENTS:

1. Welcome and thanks (1 minute)

2. Background: (3 minutes)

- Provide study outline
- Convey study objectives
- Explain why respondent was selected
- Explain purpose of this interview
- List topics to be covered, briefly
- Explain that interview is semi-structured
- Explain that interview will be audio-recorded; interviewer will also be making notes
- Interview length is scheduled for 60 minutes
- Confidentiality is guaranteed
- Ask for permission to record interview
- Explain that audio-recording will be downloaded to password-protected folder on Vibe
- Explain that audio-recording to be transcribed by this researcher, and that the transcription will be saved in password-protected folder on Vibe
- Explain that paper-based interview notes will be locked in personal closet in researcher's office
- Explain that study results will be made available if required/interested
- Ask if there are any questions? (Address these questions, if any)

- (*Switch on recording device*)

3. Data at risk section of interview (10–15 minutes)

You have indicated on your submitted web-based questionnaire that you have data at risk. I would like more information about this data:

Tell me more about:

- Discipline:
- Specific subject:
- Project details:
- Study objective:
- Format:
- Software/reader required:
- Metadata present:
- Data documentation:
- Scope:
- Location:
- Condition:
- Who knows about the data:
- Its usefulness:
- Possible users:

4. Data rescue activities (10–15 minutes)

- Has this ever been done within this research group?
- Have you ever done data rescue when in another research group?
- Do you have data rescue documentation? Guidance? Procedures?
- If yes, then:
    - Into which formats
    - Repository if applicable
    - Team members, tasks, experience, number
    - Length of rescue, also timeline
    - Any documentation about the rescue? Flowchart?
- If no, then:
    - Reasons why no rescue done

  - Enablers/resources/services required, or suggestions

5. **Data rescue workflow model (10 minutes)**

- Demonstrate and describe
- Participant to study and provide feedback
- Which format would work for you, to receive model? Paper? Electronic? Both
- Feedback required, one month from now
- Which format for feedback to this researcher? Email? Virtual interview? Word-doc? Telephone?
- Feedback guide to be emailed to participant following this interview

6. **Snowballing (2 minutes)**

   - Are you aware of data at risk elsewhere in institute?
   - Are you aware of data rescue activities elsewhere in institute?
   - Are you aware of any researchers I should contact regarding data rescue or having data at risk?

7. **Any other data at risk/data rescue questions or concerns? (1–5 minutes)**

8. **Interview conclusion (1 minute)**

APPENDIX 8 Email requesting Data Rescue Workflow Model feedback

Dear ***********

You will recall that we discussed a newly created Data Rescue Workflow Model during our recent Skype discussion. As part of this study, the feedback of RGLs (or their proxy) regarding this model is vital, and forms part of phase 3 of the study. The first phase entailed the web survey distributed during July; the second phase was the Skype discussion during September.

To provide feedback on the Data Rescue Workflow Model, kindly follow the following steps:

- View the feedback schedule that will serve as guidance towards providing feedback on the Data Rescue Workflow Model. This document is attached below.

- Kindly work through the feedback schedule, and provide feedback on the topics listed in the document by viewing the respective documents referred to in the feedback schedule document.

- At a minimum, kindly view and critique the Data Rescue Workflow Model summary (attached below) and its guidance (attached below).

I am cognizant of the fact that CSIR employees are under heavy workload and time constraints. Respondents are therefore under no obligation to provide feedback on all sections listed in the Feedback schedule.

Feedback is due by *********.

INFORMED CONSENT:

Kindly view the following paragraph, and take note that informed consent is implied when supplying feedback:

Confidentiality is guaranteed. Results will be anonymised and de-identified before being published or shared. The feedback guide as well as the informed consent form has been viewed and approved by the Ethics Research Committee of the University of Pretoria and the Research Ethics Committee of the CSIR.

There are no correct or incorrect answers. There are no risks involved. Information supplied and/or not supplied will not be held against you.

I, _________________, hereby voluntarily grant my permission for participation in the project as explained to me by Louise Patterton. The nature, objective, potential safety and health implications have been explained to me and I understand them. I understand my right to choose whether to participate in the project and that the information furnished will be handled confidentially. I am aware that the results of the investigation may be used for the purposes of publication.

Thank you for your participation in this project.

Kind regards

Louise Patterton

(CSIR Data Librarian)

APPENDIX 9 Data Rescue Model feedback guide

Dear Research Group Leader/RGL proxy

Thank you for agreeing to view and critique the data rescue flowcharts. The feedback task will take between 45 and 120 minutes to complete. Please follow the instructions below:

- Kindly comment on flowchart sections and categories listed below
- Links are provided to the documents to be viewed and critiqued
- All drawings and documents are on Vibe; informants have been given access to all relevant documents
- **Fields may be left empty. Respondents are free to skip sections. At a MINIMUM, kindly view and critique the Data Rescue summary and its single guidelines document (in other words, point 1 below). Kindly critique the model in terms of its suitability towards the RESCUE OF PAPER-BASED DATA AT RISK.**
- Feedback may can also include aspects not listed below
- Feedback can be submitted in any format e.g., electronic text, Skype discussion, feedback in the Vibe comments block, to mention a few options. You are welcome to provide feedback in any format preferred by you.
- Kindly provide feedback by 30 October 2020

DATA RESCUE MODEL FEEDBACK TOPICS TO BE ADDRESSED:

1. Feedback specific to flowchart titled 'Data Rescue Workflow Summary'

- **Kindly consult flowchart:**[redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Guidelines to the drawing:
        - Guidelines_DataRescueWorkflow.pdf : [redacted]
    - Expertise required (research discipline expertise vs library/information science expertise vs other expertise)

- Any other comments

2. Feedback specific to flowchart titled 'Stage 1_Project Initiation'

- **Kindly consult flowchart :** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 1 outputs/deliverables
    - Guidelines to the drawing:
        - Guidelines_DataAtRisk.pdf [redacted]
        - Guidelines_DataManagementPlan.pdf : [redacted]
        - Guidelines_DataRescueProjectPlan [redacted]
    - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
    - Any other comments

3. Feedback specific to flowchart titled 'Stage 2_Storage and Preservation'

- **Kindly consult flowchart:** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 2 outputs/deliverables
    - Guidelines to the drawing:
        - Guidelines_PaperStorageArchive.pdf : [redacted]
        - Guidelines_SHEQandArchives.pdf : [redacted]
        - Guidelines_ArchiveLabelling.pdf : [redacted]
    - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
    - Any other comments

4. **Feedback specific to workflow titled 'Stage3_Creating an Electronic Inventory'**

- **Kindly consult flowchart :** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 3 outputs/deliverables
    - Guidelines to the drawing:
        - Guidelines_TemplatePaper.pdf : [redacted]
        - Guidelines_TemplateMaster.pdf : [redacted]
    - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
    - Any other comments

5. **Feedback specific to flowchart titled 'Stage 4_Imaging the Paper Media'**

- **Kindly consult flowchart :** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 4 outputs/deliverables
    - Guidelines to the drawing
        - Guidelines_TemplateMasterImagesInventory.pdf : [redacted]
        - Guidelines_Imaging.pdf : [redacted]
    - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
    - Any other comments

6. **Feedback specific to flowchart titled 'Stage 5_Digitisation of Data Values'**

- **Kindly consult flowchart:** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps

    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 5 outputs/deliverables
    - Guidelines to the drawing
        - Guidelines_Digitisation.pdf : [redacted]
    - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
    - Any other comments

7. Feedback specific to flowchart titled 'Stage 6_Describe the Data'

- **Kindly consult flowchart :** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 6 outputs/deliverables
    - Guidelines to the drawing
        - Guidelines_Metadata.pdf : [redacted]
    - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
    - Any other comments

8. Feedback specific to flowchart titled 'Stage 7_Making the Data Discoverable'

- **Kindly consult flowchart :** [redacted]
- Any feedback can be provided, including (but not limited to):
    - Complexity/simplicity/logic of steps
    - Complexity/simplicity/logic of text
    - Complexity/simplicity/logic of shapes and arrows
    - Logic of activities
    - Stage 7 outputs/deliverables
    - Guidelines to the drawing
        - Guidelines_DataRepositories.pdf : [redacted]

  - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
  - Any other comments

9. Feedback specific to flowchart titled 'Stage 8_Archiving the Data'

- **Kindly consult flowchart:** [redacted]
- Any feedback can be provided, including (but not limited to):
  - Complexity/simplicity/logic of steps
  - Complexity/simplicity/logic of text
  - Complexity/simplicity/logic of shapes and arrows
  - Logic of activities
  - Stage 8 outputs/deliverables
  - Guidelines to the drawing
    - No guidelines; kindly comment on the lack of guidelines
  - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
  - Any other comments

10. Feedback specific to flowchart titled 'Stage 9_Project Closure'

- **Kindly consult flowchart:** [redacted]
- Any feedback can be provided, including (but not limited to):
  - Complexity/simplicity/logic of steps
  - Complexity/simplicity/logic of text
  - Complexity/simplicity/logic of shapes and arrows
  - Logic of activities
  - Stage 9 outputs/deliverables
  - Guidelines to the drawing
    - No guidelines; kindly comment on the lack of guidelines
  - Expertise required for this step (research discipline expertise vs library/information science expertise vs other expertise)
  - Any other comments

11. General feedback:

- General flowchart look/appearance, e.g. colours, shapes, clarity

- Guideline documents: complexity/simplicity/comprehensibility
- Flowcharts: complexity/simplicity/comprehensibility
- Discipline-specific concerns and suggestions
- Any other comments not covered under any previous heading

Thank you for your contribution.

APPENDIX 10 Focus group semi-structured schedule

FOCUS GROUP SESSION: OUTLINE OF DISCUSSION TOPICS

- Welcome
- Explanation of **ground rules** for participation
- Short discussion regarding **informed consent**: participants are free to refuse to answer questions and prompts, and may refuse to participate in any discussion taking place during the session
- Short declaration regarding **ethical treatment of data**: confidentiality, anonymisation of data in future publications and on all open platforms
- Short explanation of **data at risk**
- Short explanation of **data rescue**
- **Initial data rescue workflow model**
    - Explain how model was created
    - Discuss **Project Initiation**
    - Discuss **Storage and Preservation**
    - Discuss **Create Inventories**
    - Discuss **Imaging of Media**
    - Discuss **Digitisation of Media**
    - Discuss **Describing the Data**
    - Discuss **Making Data Discoverable**
    - Discuss **Archive the Data**
    - Discuss **Project Closure**
- **Amended data rescue workflow model**
    - Explain how model was created
    - Discuss **Data Rescue Preparatory Stage**
    - Discuss **Data Rescue Planning**
    - Discuss **Data Storage and Preservation**
    - Discuss **Digitisation**
    - Discuss **Documenting the Data**
    - Discuss **Data Sharing**
    - Discuss **Long-Term Preservation**
    - Discuss **Project Closure Stage**
- Additional topics

- Appreciation
- Reminder to send additional feedback to facilitator

APPENDIX 11 Guidance on data assessment

This document serves as guideline when determining whether paper-based media/data are at risk.

RISK DETERMINATION FACTORS

Paper-based data are at risk, and in need of rescue, when the following conditions exist:

1. The scientific data are valuable

- Consider its potential use
  - Institutional, national, and international importance of data
  - Operational importance
  - Scientific importance
  - Likelihood of data to be used in future
  - Risk of physical loss of records
  - Risk of loss of knowledge about the data or their accuracy
  - Impact of loss of records
- Consider its potential users, as well as researchers currently interested in the data
- Document the previous use (publications, policies, etc.)

2. The research data cannot be recollected, or recollection will be costly

- Consider potential costs of recollection (if recollection is possible)
- Consider the efficiency of recollection (if recollection is possible)
- Identify and considers cost of data rescue
- Consider the trade-offs with each route (rescue vs recollect)

3. No modern electronic version of the data are available.

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

APPENDIX 12 Guidance on data rescue project planning

PURPOSE OF DOCUMENT

This document serves as guideline for the drafting of a Data Rescue Project Plan (DRPP), and accompanying document/s. It is for use as a guide only, as the complexity of DRPPs can vary, depending on factors such as the research subject discipline, size of data rescued, team members available, budget, and so forth.

The target audience for this document are researchers embarking on a data rescue project, as well as other parties at a research institute/tertiary institute, involved with the rescue of data at risk.

DATA RESCUE PROJECT PLAN (DRPP) SUMMARY

This guideline suggests the drafting of the following **two pre-data rescue documents**:

1. A document containing the Data Rescue Project Plan (DRPP)
2. A document containing details about the envisaged data rescue activities and responsibilities

Alternatively, the two documents can be amended to consist of a single document.

Each of the two documents is described below. An example of each document, based on a fictitious data rescue project, is supplied.

DATA RESCUE PROJECT PLAN (DRPP)

This document is to be drafted and completed prior to the rescue endeavour. It describes the data rescue project and serves as a guide throughout the process. The template used below, as well as the table format, serves as an example only. A non-tabular format, or a project plan combining both tables displayed here, are viable alternatives.

The table is an example of a typical project plan for the rescue of paper data. It should be viewed as a generic plan; actual data rescue could include more or fewer headings/categories, or even headings being subdivided.

A well-drafted and completed project plan is also a helpful tool when drafting the project's final report.

TABLE 1: DATA RESCUE PROJECT PLAN (FICTITIOUS EXAMPLE):

Title of Data Rescue Project	Rescue of paper-based historical climatology data, measured at Hartbeespoort dam, North-West Province, South Africa, December 1975.
ID	DR NRE 0001-2020
Data of DRPP creation	22 March 2020
Research unit	Unit formerly known as CSIR: Natural Resources and Environment Unit
Current location of paper data	Building 33, Room 105, CSIR, Pretoria
Data ownership	CSIR
Abstract **to include:** • **Brief description of project** • **Reason for rescue** • **Description of formats** • **What needs to be done to secure the data** • **Interested parties**	This project entails the rescue of paper-based historical climatology data collected during December 1975 at Hartbeespoort dam. The data were found to be at risk, mainly because: • no digital versions of the same measurements are available, and • the data are currently stored in an environment not ideal for storing paper data. The data will be transported to an archive, an inventory created, will be imaged, saved in electronic folders, and backed up to a DVD. Metadata will be added. The dataset will be uploaded to Tambora.org, an open access data repository for historical climatology data. Rescued data will be of value to researchers involved with longitudinal South African water quality studies
Rescue period	June 2020–December 2020
Project dates	Locate, preamble, DRPP: June 2020 Storing: July 2020 Inventory: July 2020 Imaging: August 2020 Digitising: August 2020 Describe: September 2020 Make data findable (repositories): October 2020 Archive: November 2020 Promote: December 2020
Team members	[redacted]
Activities and responsibilities	See table with activities checklist below
Current state of data	Paper data are in a good condition Some of the data are dusty

Current format of data	Paper
Estimated current data scope/volume	8 archive boxes (approximately 100 spreadsheets, 250 photos, 8 paper maps)
Estimated electronic data size	Less than 50MB
Expected outcome	• Paper data safely stored under ideal archival conditions • Imaged media will be uploaded to a discipline-specific repository • Metadata to be added to CSIR closed institutional repository • Metadata to be added to open access institutional repository
Intended repository	• Metadata: CSIR's TOdB, also ResearchSpace • Dataset upload to Tambora.org (Historical Climatology database)
Metadata standard	Dublin Core plus a fixity field
Main keywords	Water temperature, Water level, Sediment, Hartbeespoort dam, Water quality
Value of data	Rescued data will be of value to researchers involved with longitudinal South African water quality studies
Expected use of data	Rescued data will be of value to researchers involved with longitudinal South African water quality studies
Cost (including time)	No additional equipment will be purchased; all equipment and supplies are currently in stock. An estimate of 200 working hours required (2 CSIRIS-subsidised interns X 100 hours)
Equipment/supplies required (in priority order)	• PC/laptop with spreadsheet and word processing applications/software • Scanner • Two (2) DVDs • Four (4) acid-free archival boxes • One (1) paper duster • Two (2) pairs of archive gloves • Two (2) DVD labels • Four (4) box labels • One (1) paper duster • Two (2) archive coats • Paper, pens
Expertise and skills required	• Archival knowledge • Working knowledge of Excel, Word (or similar) • Data indexing expertise • Knowledge of climatology terms and subjects
Permission required	• Permission from unit manager to transport data to CSIR Archive in Building 22, and initially to CSIR Library • Permission from unit manager to upload dataset to Tambora.org • Permission from CSIRIS manager to store paper data in Archive • Permission from CSIRIS manager to make use of two current CSIRIS interns, 4 hours per week, for 5 months

Links to any supporting information	Several papers have been presented, and several articles published. Details will be added to metadata on all repositories
Project Status	Locate, preamble, DRPP completed: Storing completed: Inventory completed: Imaging completed: Digitising completed: Describe completed: Make data findable (repositories) completed: Archiving completed: Project closure completed:

CHECKLIST OF PLANNED DATA RESCUE ACTIVITIES (FICTITIOUS EXAMPLE)

The following table is an example of typical paper data rescue activities that should be executed. Provision is also made for stipulating the responsible party for each activity. A comments block, which can be used before, during, or after the activity is performed, also forms part of the table. The rightmost column is filled in upon completion of the activity. It should be viewed as a generic table; actual data rescue could include more or fewer activities, or even activities being subdivided.

The document can also be used as a checklist to monitor progress of the data rescue project.

DATA RESCUE ACTIVITY	RESP	COMMENTS/PROGRESS REMARKS	COMPLETION DATE
Drafting of DRPP			
Creation of Data Management Plan (DMP)			
DRPP storing, backups, updates, changes			
DMP updates, editing, communication			
DR team communications, meeting convening, etc.			
Updating DR status			
Data examining/assessment			
Data risk assessment			
SHEQ communications			
Arrange for archive boxes and supplies during storing phase			
Dusting/vacuuming of documents/artefacts			
Dusting/vacuuming of shelves			
Dusting/vacuuming of boxes			
Dusting/vacuuming of archive			
Box labelling			

Shelf labelling			
Paper data storing			
Creating electronic inventory of paper records			
Creating electronic master inventory			
Updating electronic master inventory			
Create master imaging directory			
Arrange scanner or camera			
Arrange transport of data			
Imaging of paper records			
Check imaged media for readability			
Check imaged media for number of images			
Renaming imaged media			
Data validation			
Creating file structure, folders			
Write to CDs or DVDs*			
Label CDs or DVDs*			
*OR: Choose safe storage medium in consultation with Archivist/Records Manager. Transfer data to safe storage medium			
Create master list of filenames			
Update master list of filenames			
Digitise spreadsheets			
Metadata creation			
Metadata upload			
Data indexed to institutional closed repository			
Data indexed to institutional open repository			
Data indexed to discipline/generalist repository			
Data archiving (including technology migration plan)			
DMP created			
Data article/publish external			
Data article/publish internal			

Letter of thanks			
Create recommendations document. Draft final report			
Updated DRPP template			
Sharing of knowledge			
Backups of all digital media			
Data migration			Perpetual

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

APPENDIX 13 Guidance on data management plans

This document contains guidelines for creating a Data Management Plan (DMP) by way of a list of frequently asked questions.

1. What is a data management plan?

A data management plan (or a **DMP**) is a formal document that outlines how you will handle your data both during research, and after the project is completed.

2. Why is it important to have a DMP?

A DMP, being a systematic approach in planning the lifecycle of your data, will assist researchers with the following:

- a DMP will help alert researchers on possible data collection and management issues, and enabling researchers to prepare for issues before the project begins, and
- a DMP ensures that there is continuity of the research project when staff leave or new staff start work.

Often, a DMP is also a funder requirement, and/or part of institutional policy and procedures.

3. What should be included in a DMP?

A DMP usually addresses the following main categories:

- Background to the study/project
- Data collection
- Data documentation and metadata
- Ethics and legal compliance
- Data storage and backup
- Selection and preservation
- Responsibilities and resources

4. What needs to be included in each DMP section?

The details to be provided within each of the DMP categories are listed below.

Background/administrative information

This section is a brief description of the research project.

- Name of project
- Name of contact person for project
- Name of contact person for data
- Description of the research
- Funders
- Project duration
- The version of the DMP

Data Collection

This section is a brief description of the data to be collected or generated.

- Description of type of data that will be produced, e.g., experimental measures, whether the data are quantitative or qualitative in nature
- Description of data formats
- Details regarding the data volume and data growth rate
- Details regarding the use of pre-existing data (if applicable) and its source
- Details regarding data collection methods, i.e., instrumentation, hardware or software used

Data documentation and metadata

This section describes how data will be documented to help future users understand and reuse the data.

- Description of the metadata that will be created
- Description of the metadata standard that will be used
- Description of data documentation that will be added, and the contextual details needed to make the data meaningful to future users
- Information regarding the issue of persistent citation, i.e., DOIs for datasets

Ethics and legal compliance

This section describes how ethical issues, as well as copyright and intellectual property rights issues, will be managed.

- Description of ethical and privacy issues linked to the data
- Description of measures to be used to resolve ethical and privacy issues
- Description of owner of data copyright, and owner of data intellectual property rights, if

applicable

- Details regarding the embargo period before the data are opened up to wider use (if applicable)

Data storage and backup

This section describes the ways in which the researcher ensures that data are stored securely, and regularly backed up.

- Description of data storage location, and data storage media
- Details regarding data back-ups
- Details regarding back-up regularity
- Details regarding the enforcement of permissions, restrictions and embargoes
- Details regarding the management of data access arrangements and data security
- Description of any other data security issues

Selection and preservation

This section describes the strategy for preservation of the data.

- Description of the criteria used for data preservation selection
- Description of length of data preservation beyond the life of the project
- Description of disposition and transfer of sensitive data
- Details regarding the archive/repository/central database/data centre identified for data preservation
- Details regarding the transformations necessary to prepare data for preservation
- Stating whether data will be shared
- Description of reasons not to share or reuse data?
- Details regarding the transformations necessary to prepare data for sharing
- Description of bodies/groups likely to be interested in the data
- Description regarding the foreseeable contemporary or future uses for the data
- Details of any embargo periods for political/commercial/patent reasons

Responsibilities and resources

This section contains details regarding the responsibilities surrounding the data, and resources required to manage the data and implement the DMP.

- Name the person responsible for implementation of this plan/DMP compliance
- Name the person responsible for each of the data management activities listed
- Description of resources required to deliver the DMP

5. I have seen different DMP formats. Why are they not all identical?

There are indeed many different DMP templates available, resulting in DMPs created in many different formats. Basically these plans are the same: a supplementary document, usually no longer than two pages, that describes the kind of data to be produced, how data will be managed, and how research results will be made accessible to other researchers.

6. I have heard of online DMP tools. What are these?

Examples of free online DMP tools include the following:

- DMPTool is a service of the California Digital Library, a division of the University of California: https://dmptool.org/
- DMPonline helps you to create, review, and share data management plans, and is provided by the Digital Curation Centre (DCC): https://dmponline.dcc.ac.uk/
- The SA-DMP Tool is developed and provided by the Data Intensive Research Initiative of South Africa (DIRISA): https://secure.dirisa.ac.za/SADMPTool/

7. Is creating and submitting a DMP a mandatory task?

This depends on the institute’s data management policy. However, creating and adhering to a data management plan is considered part of good research practice.

8. When should a DMP be created?

Researchers should ideally create a DMP at the start of a research project. It can be created when the funding application is made, or when the project plan is drafted. In some cases though, a DMP can also be drafted during the data rescue process (e.g., when historical paper-based data at risk is being digitised, described, and archived).

9. How long should a DMP be?

A DMP has no fixed length, but two pages is the average length. In certain instances a DMP could be longer than two pages should the data management practices of the project require extensive or intricate description.

10. What should I do with my DMP once it has been drafted?

Your completed DMP should be sent to the following entities:

- the research funder,
- the relevant Research Ethics Committee(s), and
- all research project members.

If data storage requirements are out of the ordinary, ICT should be informed of the storage requirements and a copy forwarded to them as well. The completed DMP should also be stored with other project documents.

11. I have created a DMP but suspect that some of the information might change over time. Is this a problem?

This is not a problem at all. In fact, DMPs are dynamic in nature, and it is expected that some information might change over time as the project progresses. Should these changes occur, the DMP should be updated (or a newer version created) and the updated plan distributed to the relevant parties.

12. What about DMPs and confidentiality? May I share my DMP with other parties?

If your research is secret, your DMP should only be shown to people froming part of your project, and to parties in need of viewing the DMP. DMPs of non-confidential projects can be shared freely.

13. How should I manage my DMP?

Treat your DMP as an important part of research,as it is a document forming part of your data documentation. Make sure it is sent to required parties, stored securely, backed up, and updated as needed.

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

APPENDIX 14 Guidance on storage of paper data

This document serves as a guideline for the storing of paper records in archive boxes.

REQUIREMENTS FOR ARCHIVE BOXES

Paper records should be stored in archive boxes that are:

- Acid free
- Pest free
- Dust free
- Free from moisture
- Not physically damaged

In addition to the requirements above, the use of low-lignin boxes is recommended for storing:

- Photographic records
- Paper records with high intrinsic value

FILLING ARCHIVE BOXES

- Boxes should not be under-filled (as this will result in paper records bending or slumping).
- Boxes should not be overfilled (as this will result in paper records forced in and out).
- Corrugated acid-free spacer boards should be used in partially filled boxes.
- Boxes should be large enough to accommodate the paper records.

STORING AND LABELLING OF ARCHIVE BOXES, SHELVES, AND CABINETS

Fixed labelling rules do not exist, as the activity can be influenced by aspects such as the specific discipline, or years and volume of paper records. It is important that a logical project-related labelling system be used.

The following are suggested practices for climate data on paper, as per the World Meteorological Organization (2016):

- Store paper data in acid free boxes or filing cabinets by:
  - Type (form or chart type)

  - Station, year, month
- The number of years of data may determine the most logical manner of labelling
- Do not add labels to the paper data itself, as this enhances deterioration of the media
- After labelling the boxes:
  - Add storage location (box number) to inventory
  - Add Station/year/month to inventory

The table below is a fictitious example of an inventory for labelled boxes:

Box number	**Title**	**Number of boxes**	**Contents**	**Coverage**	**Condition**
000001-000007	Apples and Pears	7	Business records	Not known	Good
000008-0000012	Another example	5	Photographs	1935–1935	Several photographs are bent/folded
0000013-0000020	Another example	8	Papers	C1890–1950	Slight water damage

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

ADDITIONAL READING

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. 2015. ISO 16245:2009 *Information and documentation – Boxes, file covers and other enclosures, made from cellulosic materials, for storage of paper and parchment documents.* Available from: https://www.iso.org/standard/45988.html. Accessed on 23 March 2022.

SMITHSONIAN INSTITUTION ARCHIVES. 2011. *Setting up Paper Files*. Available from: https://siarchives.si.edu/what-we-do/setting-paper-files. Accessed on 23 March 2022.

WORLD METEOROLOGICAL ORGANIZATION. 2016. *Guidelines on Best Practices for Climate Data Rescue*. Available from: https://library.wmo.int/doc_num.php?explnum_id=3318. Accessed on 23 March 2022.

APPENDIX 15 Guidance on archives and SHEQ

This document serves as guideline when using an archive for storing paper records. It intends to describe the archival conditions putting paper data at risk, and conditions which are beneficial to the storage of paper media.

RELATIVE HUMIDITY AND TEMPERATURE

Archives containing paper records should be free of the following:

- **Dampness**
  - Dampness encourages pest activity.
  - Dampness encourages mould growth.
- **Excessive dryness**
  - Dryness can cause archival material to become brittle.
- **Fluctuations in humidity**
  - Rapid loss and absorption of moisture can cause microscopic structural damage, contributing to deterioration of paper.
- **Too warm conditions**
  - Chemical reactions are faster in warmer conditions.

An ideal temperature is between 13–20 degrees Celsius.

An ideal relative humidity level is between 35% and 60%.

AIR QUALITY

- Lack of air movement encourages the growth of mould.
- Shelving should be ventilated; open shelving is preferable.
- Archive ventilation should not compromise ideal temperature and humidity.
- Paper records should be stored away from the floor and ceiling.
- Pollutants should not be introduced where outside air is drawn into the archive.
- Air should be filtered.

LIGHT

- Ideally, and archive should have as little constant light as possible.
- Light causes fading of paper records.

- Light energy creates heat.
- Sunlight is particularly damaging to archival items.
- Windows should be as small as possible.
- Newly designed archives should ideally be without windows.
- Windows should be positioned to avoid direct sunlight.
- Install blinds where required.
- Electric lights to be switched off when not in use.

ADDITIONAL CONTROL MEASURES

- Close archive windows and doors.
- Use conservation-quality archive boxes.
- Make use of heating, ventilating and air conditioning systems.
- A dehumidifier can be used inside an archive.
- Fans may be used inside an archive.

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

ADDITIONAL READING

NATIONAL ARCHIVES. 2016. *Archive Principles and Practice: an introduction to archives for non-archivists.* Available from: https://www.nationalarchives.gov.uk/documents/archives/archive-principles-and-practice-an-introduction-to-archives-for-non-archivists.pdf. Accessed on 23 March 2022.

NATIONAL ARCHIVES. 2019. *Environmental Management*. Available from: https://www.nationalarchives.gov.uk/documents/information-management/environmental-management.pdf. Accessed on 23 March 2022.

NORTHEAST DOCUMENT CONSERVATION CENTER (NEDCC). 1999. *The Environment: Temperature, Relative Humidity, Light, and Air Quality: Basic Guidelines for Preservation*. Available from: https://www.nedcc.org/free-resources/preservation-leaflets/2.-the-environment/2.1-temperature,-relative-humidity,-light,-and-air-quality-basic-guidelines-for-preservation. Accessed on 23 March 2022.

APPENDIX 16 Guidance on storage of physical samples and specimens

This document serves as a guideline for the storing of physical samples and specimens

Specimen storage

Physical samples are a basic element for reference, study, and experimentation in research.

Tests and analysis are conducted directly on samples, including:

- biological specimens,
- rock or mineral specimens,
- soil or sediment cores,
- plants and seeds,
- water quality samples,
- archaeological artefacts, and
- DNA and human tissue samples.

Other physical objects, such as maps or analogue images are also direct objects of study, and, if digitised, may become a source of digital data.

There is an urgent need for better integrating these physical objects into the digital research data ecosystem, both in a global and in an interdisciplinary context to support search, retrieval, analysis, reuse, preservation, and scientific reproducibility.

Like analogue records, physical sample archives require environmental controls that exclude the ambient environment and maintain optimal conditions to preserve the longevity of their holdings. Issues such as projected use of the samples, access frequency, and nature of research can dictate the requirements needed to appropriately store these collections.

The sections below describe the ideal storage conditions and environmental control for diverse types of specimens.

Ice, Marine and Well Water Samples and Cores

- Ice cores should be stored in facilities that maintain temperatures far below freezing (archival temperature is maintained at -35º C) with little fluctuation to avoid melting and loss of potential chemical and paleo atmospheric information.

- Marine and lacustrine soft sediment cores have refrigerated or freezing storage requirements, as fluctuations in air temperature and humidity can degrade the viability of the samples for geochemical and geophysical properties research, as well as promote organic growth.
- Preserved water well cores require storage in sealed, air-evacuated tubes containing native groundwater with chemically enhanced reducing conditions or are desiccated under reducing conditions.

Rock Cores and Samples

- Rock cores with little to no moisture content must be protected from natural hazards (i.e., extreme weather).
- The assurance of protected indoor storage is widely recognised as crucial for any sample archive.
- Large rock core collections should be stored in low-humidity, climate-controlled warehouses.
- Should facilities be lacking, rock core collections can be stored in non-controlled shipping containers. While this is not ideal, the materials are not directly exposed to weather, which preserves their integrity for research.

Geothermal energy samples

- Cuttings should ideally be preserved by storage in a sample repository in indexed and barcoded trays on labelled storage racks.
- The trays should be labelled to represent discrete downhole depth sections.
- Split core samples should be stored in the repository in split inner barrels, in indexed trays holding multiple core sections, on indexed shelves.
- Fluid samples should be individually labelled, barcoded, and preserved by storage in a secure repository in original sample bottles where feasible, on indexed and barcoded trays and indexed shelves, or where required, placed in refrigerated storage.
- In all cases, half of all curated cutting and fluid samples are to be preserved for archival storage, and half made available for distribution.

Unique and Hazardous Samples

- Samples that require special and unique methods of preservation are abundant.

- Radioactive samples require isolation, which may include systematically controlled ventilation.
- Samples of gas hydrate require extreme cold (stored in liquid nitrogen dewars) to preserve their integrity or must be kept under high pressure in Parr vessels.

ADDITIONAL READING:

ALLEN, C., ALLTON, J., LOFGREN, G., RIGHTER, K. & ZOLENSKY, M. 2011. *Curating NASA's extra-terrestrial samples – past, present, and future*. Available from: https://www.lpi.usra.edu/meetings/sssr2011/pdf/5006.pdf. Accessed on 18 January 2022.

NEWBERRY GEOTHERMAL ENERGY. 2016. *Sample and Core Curation Plan*. Available from: https://www.energy.gov/sites/prod/files/2016/09/f33/Sample%20and%20Core%20Curation%20Plan_%28Newberry%20Volcano%2C%20OR%29.pdf. Accessed on 18 January 2022.

UNITED STATES GEOLOGICAL SURVEY. 2004. *Physical Sample Preservation and Curation*. Available from: https://www.usgs.gov/programs/national-geological-and-geophysical-data-preservation-program/physical-sample-preservation. Accessed on 18 January 2022.

APPENDIX 17 Sample templates for the creation of data inventories

DATA INVENTORY WORKSHEET					**Page**	**of**
CLUSTER			**IMPACT AREA**			
PERSON RESPONSIBLE FOR DATA			**CONTACT DETAILS**			
ROOM						
Series	**Description**	**File location**	**Media type**	**Years covered**	**Scope/Volume**	**Remarks/Notes**

MASTER INVENTORY WORKSHEET					**Page**	**of**
CLUSTER			**IMPACT AREA**			
Station	**Date**	**Media type**	**Number**	**Periods of record**	**Estimate page numbers**	**Remarks/Notes**

APPENDIX 18 Guidance on digitisation of paper data

This document serves as guideline when digitising paper records.

DIGITISING DEFINED

Digitising involves the scanning or photographing of paper records to create a digital image. Digitising of paper records is performed after the records have been collected, and an electronic inventory of the paper records created. A digital camera or optical scanner is used to digitise the data.

BENEFITS OF DIGITISING PAPER DATA

- Digitising creates a digital version of the record, which can be preserved.
- Digitising creates a digital version of the record, thereby removing the need to handle the (fragile) paper record.
- Digitising allows the option of resizing a document.
- A digitised record's legibility can be enhanced.

THE DIGITISING PROCESS

The main digitising steps for paper records are the following (as described by the World Meteorological Organization, 2016):

- **Create a master digitising inventory** from the inventory created during the storing process
  - Master digitising inventory to stipulate number of pages to be digitised
  - Master digitising inventory to be updated as digitising progresses (also add computer directory/cd location
- **Validate the digitised data**
- **Rename filenames to relate to content**
  - Good filenames include those with indications of location, form type, date, and page number
  - Filename extension will be .png, .jpg, or .tif
- **Store digitised content in a logical predefined structure**
- **Maintain a master list of filenames**
- **Update the master digitising inventory**
  - Compare pages counted with number of digitised pages
  - Enter directory name or applicable data location
- **Document the entire process**

PRIORITISING THE DIGITISING OF PAPER RECORDS

Due to the declining of paper and ink quality in recent years, it is often best to start with newest paper media first.

If paper and ink quality is not a factor, start with paper media carrying the most importance.

All pages should be digitised, to capture all metadata, descriptions, calibrations, practices, and procedures.

SCANNERS

- Two types of scanners exist: flatbed scanners (for scanning books) and feeding scanners (for scanning loose pages)
- An advantage of scanners: scanning requires less equipment than imaging via camera
- Disadvantages of scanners:
    - may damage book bindings
    - scanned data closest to bindings may be out of focus
    - requires maintenance (often a contract also)
    - specialised training often required for scanners
    - may be slow and cumbersome

CAMERAS USED FOR DIGITISING

Another way of digitising paper-based data is by using a **copy stand with a digital camera**. Cameras can be used for both loose and bound paper data, as well as for oversized documents.

The camera is held steady at a fixed distance above the document. Some copy stands come with fixed lighting equipment, thereby enabling adjustable and maintainable illumination of the paper-based data.

The following camera functions and photographic tools are useful during the digitising process:

- an AC adapter allowing unlimited pictures to be taken, to prevent draining of batteries,
- spare batteries and charger if an adapter is not available,

- a camera allowing batteries to be removed without moving camera from the stand,
- a thread holding the camera on the camera stand,
- software and a cable to transmit images to a computer,
- large memory card if the above software is not available,
- a remote control,
- software enabling the automatic renaming of digitised files,
- a zoom lens, or lens with macro mode,
- option to change camera settings (i.e., cancel flash, change speed sensitivity, or white balance), and
- option to select size of image.

VALIDATION OF DIGITISED MEDIA

Validation entails quality controlling the digitised records. After digitising the paper media, the images are checked to ascertain whether:

- images are in focus,
- images are not too light, and that penned/pencilled text is visible,
- images are not too dark,
- parts of images have not been obliterated by a flash, or too bright a light, and
- all pages have been digitised; the number of pages should correspond.

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

REFERENCES:

JISC DIGITAL MEDIA. 2014. *Introduction to using a copy stand*. Available from: https://www.youtube.com/watch?v=udiRErQ1CnA. Accessed on 28 January 2022.

WORLD METEOROLOGICAL ORGANIZATION. 2014. *Guidelines for Hydrological Data Rescue*. Available from: https://library.wmo.int/pmb_ged/wmo_1146_en.pdf. Accessed on 29 January 2022.

WORLD METEOROLOGICAL ORGANIZATION. 2016. *Guidelines on Best Practices for Climate Data Rescue.* World Meteorological Organization. Available from: https://library.wmo.int/doc_num.php?explnum_id=3318. Accessed on 29 January 2022.

APPENDIX 19 Guidance on the use of metadata

GUIDELINE: METADATA ELEMENT SET FOR SCIENTIFIC DATA

This document serves as a guideline for the creation of metadata elements for scientific data. It includes the listing of two popular metadata standards, namely the Dublin Core Metadata Element Set, and the Dryad Data Object Module for Scientific Data. Kindly consult the webpages of the listed metadata standards for more practical information on implementing the specific standard (see REFERENCES).

DUBLIN CORE

The Dublin Core Metadata Element Set is one of the simplest and most widely used metadata schema. Originally developed to describe web resources, Dublin Core has been used to describe a variety of physical and digital resources.

It is comprised of the following 15 core metadata elements:

Element	**Use**
Title	A name given to the resource
Subject	The topic of the resource
Description	An account of the resource
Creator	An entity primarily responsible for making the resource
Publisher	An entity responsible for making the resource available
Contributor	An entity responsible for making contributions to the resource
Date	A point or period of time associated with an event in the lifecycle of the resource
Type	The nature or genre of the resource
Format	The file format, physical medium, or dimensions of the resource
Identifier	An unambiguous reference to the resource within a given context
Source	A related resource from which the described resource is derived
Language	A language of the resource
Relation	A related resource
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant
Rights	Information about rights held in and over the resource

DRYAD DATA OBJECT MODULE FOR SCIENTIFIC DATA

This module includes 21 properties, with the main difference with Dublin Core being the inclusion of a fixity field. Fixity is the assurance that a digital file has remained unchanged, i.e., fixed.

Name	Obligation
Header	Required
Type	Required
Author	Required
Contributing author	Optional
Dataset title	Optional
Dataset handle	Required
DOI of published article	Required
Depositor	Required
Primary contact	Required
Rights statement	Required
Description	Optional
Keywords	Required
Taxonomic name	Optional
Locality	Optional
Date range	Optional
Date of issue	Required
Embargo date	Optional
Date modified	Required
File format	Required
File size	Required
Fixity	Required

CONTACT DETAILS (for this document)

Louise Patterton
lpatterton@csir.co.za

REFERENCES:

DUBLIN CORE METADATA INITIATIVE. 2020. Available from: https://dublincore.org/. Accessed on 26 January 2022.

DRYAD. 2020. *Dryad for your research data*. Dryad. Available from: https://datadryad.org/stash/. Accessed on 26 January 2022.

APPENDIX 20 Guidance on use of data repositories

This document serves as guideline when deciding on a suitable data repository for depositing research data.

It intends to answer the following commonly asked questions:

‘Where can I deposit my research data?’

‘How do I make my data visible to the research community?’

‘I still have valuable datasets from previous projects and would like to share it with interested parties. How do I go about this?’

BACKGROUND

Researchers often voice concerns about the storage of datasets and tend to be unaware of options available to them. It is recommended that researchers consider using an established data repository for long-term curation and sharing of data.

Data repositories can be general or discipline specific. The advantage of using a disciplinarydata repository is that researchers will be making their research data accessible to their community of interest.

BENEFITS OF DATA SHARING

- Data sharing can fulfil funder requirements (add reference)
- Sharing data can raise interest in one’s your research; sharing detailed research data is associated with an increased citation rate (add reference
- Data sharing can lead to new collaborations (add reference)
- Data sharing generates goodwill between researchers (add reference)
- Sharing data allows for follow-on research (add reference)

FINDING A SUITABLE DATA REPOSITORY

Preserving research data in data centres or repositories which are managed by trusted entities for long-term access is the most common way to share data.

Re3Data (https://re3data.org/) is a tool that can help researchers or data depositors find disciplinary data repositories across hundreds of research areas. This vetted source helps answers questions such as who can deposit data and if there are any specific requirements for depositing data.

If there is not an appropriate disciplinary repository for your data, consider using a general data repository. Below are examples of data repositories that accept data from all disciplines.

- ○ **DataONE dash** (https://www.dataone.org/software-tools/dash)
- ○ **Dataverse** (https://dataverse.org/)
- ○ **Figshare** (https://figshare.com/)
- ○ **Open Science Framework** (https://osf.io/)
- ○ **Zenodo** (https://zenodo.org/)

Kindly contact your information specialist or the data librarian (lpatterton@csir.co.za) for more information on data repositories.

CONTACT DETAILS (for this document)

Louise Patterton

lpatterton@csir.co.za

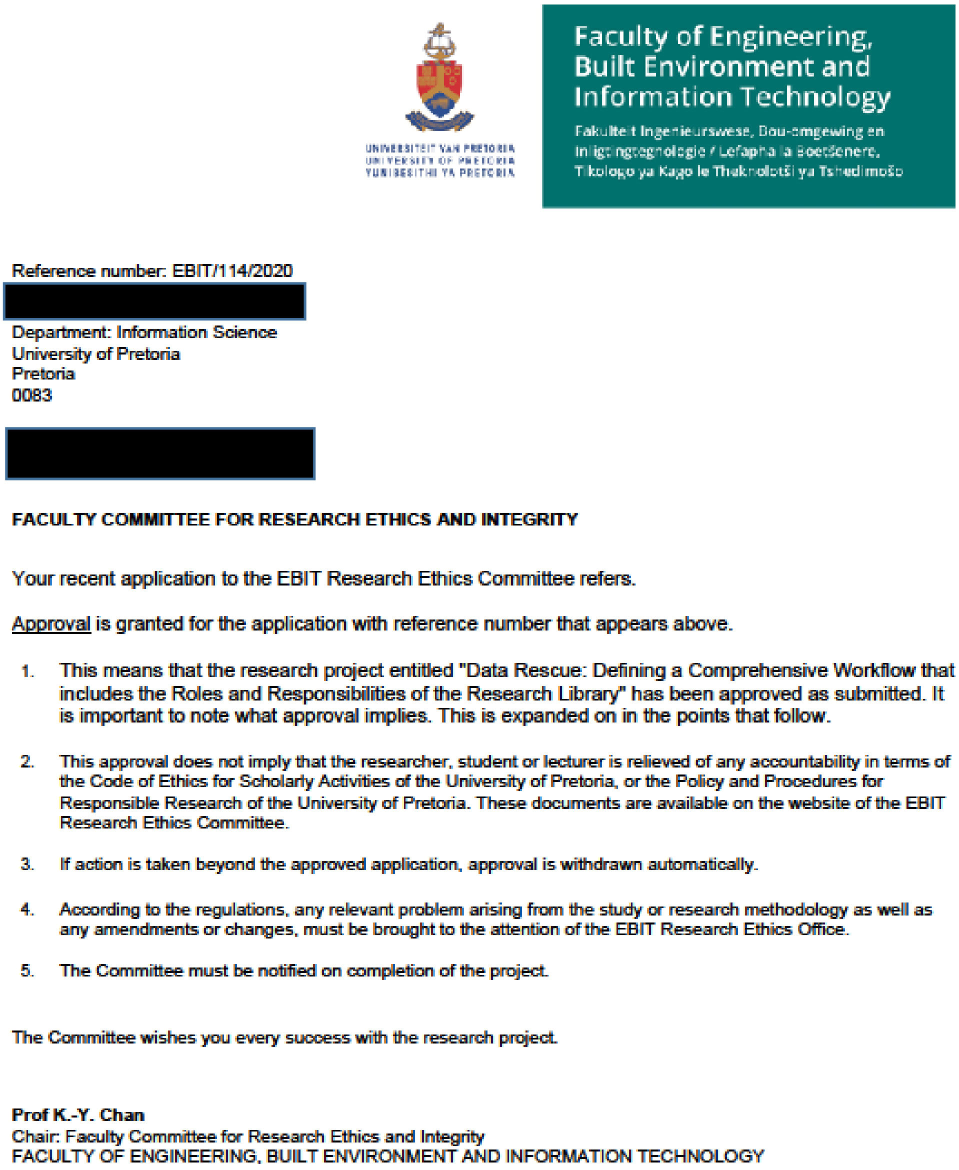
ADDITIONAL READING

PIWOWAR, H.A., DAY, R.S. & FRIDSMA, D.B. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE,* vol. 2(3): e308. Available from: https://doi.org/10.1371/journal.pone.0000308.

POPKIN, G. 2019. Data sharing and how it can benefit your scientific career. *Nature*, vol. 569: 445–447. Available from: https://doi.org/10.1038/d41586-019-01506-x.

UNIVERSITY COLLEGE DUBLIN, UNIVERSITY LIBRARY. 2022. *Research Data Management: Why Share Research Data.* Available from: https://libguides.ucd.ie/data/share. Accessed on 27 September 2022.

APPENDIX 21 Proof of Ethics Approval



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Reference number: EBIT/114/2020
[redacted]
Department: Information Science
University of Pretoria
Pretoria
0083

[redacted]

FACULTY COMMITTEE FOR RESEARCH ETHICS AND INTEGRITY

Your recent application to the EBIT Research Ethics Committee refers.

Approval is granted for the application with reference number that appears above.

1. This means that the research project entitled "Data Rescue: Defining a Comprehensive Workflow that includes the Roles and Responsibilities of the Research Library" has been approved as submitted. It is important to note what approval implies. This is expanded on in the points that follow.

2. This approval does not imply that the researcher, student or lecturer is relieved of any accountability in terms of the Code of Ethics for Scholarly Activities of the University of Pretoria, or the Policy and Procedures for Responsible Research of the University of Pretoria. These documents are available on the website of the EBIT Research Ethics Committee.

3. If action is taken beyond the approved application, approval is withdrawn automatically.

4. According to the regulations, any relevant problem arising from the study or research methodology as well as any amendments or changes, must be brought to the attention of the EBIT Research Ethics Office.

5. The Committee must be notified on completion of the project.

The Committee wishes you every success with the research project.

Prof K.-Y. Chan
Chair: Faculty Committee for Research Ethics and Integrity
FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY

APPENDIX 22 Data Management Plan

PLAN DESCRIPTION: Data management plan for data collected during PhD studies in Information Science.

DATE OF PLAN: January 2022

STUDY TITLE: Data Rescue: Defining a Comprehensive Workflow that includes the Roles and Responsibilities of the Research Library

RESEARCHER: Louise Patterton

ORCID:

FUNDER: CSIR

DEGREE and INSTITUTE: PhD in Information Science, University of Pretoria

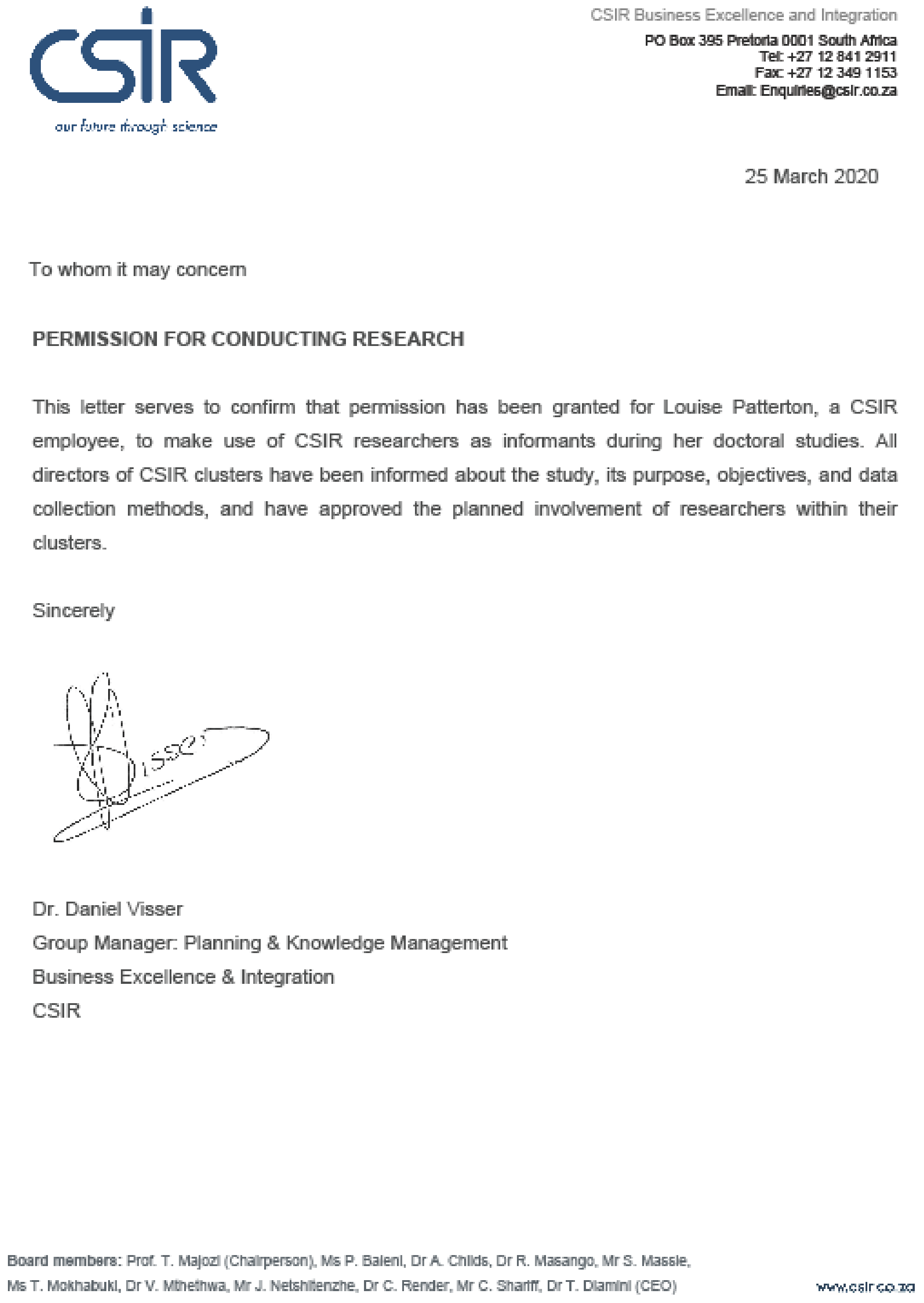
CONTACT DETAILS: 073-XXXXXXX/ lpatterton@csir.co.za

DATA COLLECTION	
What data will be collected or created? **How will data be collected or created?**	Diverse types of data will be collected: • text data from web-based questionnaires • audio data from personal interviews • text data from electronic and web-based communications (email and Skype) • text data from notes made during focus group session Text data will be less than 250MB. Audio data volume will depend on relevance of question to respondent, and length of response given. A total of 1GB of audio data is anticipated. Text data will be collected via a web-based questionnaire created using the eSurv tool. Audio data will be collected via an audio recorder or smartphone appliance, during personal interviews. A second round of text data will be collected, and this will be in the form of textual feedback sent to this researcher by means of emails or verbal feedback using Skype. A third round of text data will be collected, and this will be in the form of text notes made by this researcher during a focus group session. Data folders will be named as follows: Questionnaire data: Questionnaire question number, and date e.g., Q_q1_2020 Interview data: Interview number and data e.g., Int9_2020 Summarised interview data: Question/topic and date e.g., Data rescue challenges_2020 Feedback data: Connection will be made to the interview number e.g., F_Int9_2020 Summarised feedback data:

	Focus group data: Topic and date is crucial, e.g., FG_data_storage_2021 Separate folders will be created for each stage of data collection (web-based questionnaire, interviews, feedback, focus group)
1. DOCUMENTATION AND METADATA	
What documentation and metadata will accompany the data?	**Data documentation** will be the following: • a list of web-based questionnaire questions asked • objective of web-based questionnaire • characteristics and number of researchers involved with web-based questionnaire • an interview schedule • characteristics and number of researchers interviewed • objective of interview • objective of feedback requested from researchers • number of researchers supplying feedback • number of focus group participants • expertise of focus group participants • topics discussed during focus group sessions **Metadata** will adhere to the Dublin Core standard. All documentation will take the form of a MS Word document
2. ETHICS AND LEGAL COMPLIANCE	
How will ethical issues be managed? How will copyright and Intellectual Property Rights (IPR) issues be managed?	Ethical approval for data collection and data sharing has been obtained. All data will be anonymised. It is not anticipated that the study will generate sensitive data. If sensitive data happens to be generated (e.g., incriminatory interview responses) the data will be anonymised and de-identified. Sensitive data will be stored in password-protected folders on Groupwise Vibe prior to being made ready for sharing, and while being anonymised. The data are owned by the University of Pretoria. Shareable data will be uploaded to an open access repository as required by the funder. There will be no restrictions on re-use; citing of the data is required. Data will be published as soon as the thesis has been accepted for degree purposes.
3. STORAGE AND BACKUP	
How will the data be stored and backed up during the research? How will access and security be managed?	As a registered user on the institutional Vibe, sufficient storage is available to this researcher to store the anticipated data. Data stored in Vibe folders will be automatically backed up. ICT bears the responsibility for daily backups.

How will the data be stored and backed up during the research? How will access and security be managed?	Data will be stored in a password-protected folder on Vibe. Study supervisors will act as collaborators. The window between interview data collection and data upload is brief (10 minutes) as both locations are on work premises, meaning that no data loss is anticipated.
4. SELECTION AND PRESERVATION	
Which data are of long-term value and should be retained, shared, and/or preserved? What is the long-term preservation plan for the dataset?	All data will be retained, shared, and preserved. It is assumed that the data repository will be preserving the data in perpetuity. Format migration will be performed by the repository.
5. DATA SHARING	
How will the data be shared? Are any restrictions on data sharing required?	Data will be shared via the data repository of the University of Pretoria. A DOI will be obtained and supplied to interested parties. Data will be made available (i.e., uploaded to the repository) upon submission of the thesis for examination. No restrictions or challenges regarding data sharing is anticipated. Data citation will be required.
6. RESPONSIBILITIES AND RESOURCES	
Who will be responsible for data management? What resources are required to deliver this plan?	The main researcher bears full responsibility for implementing the DMP. All activities are managed by the researcher, except where mentioned elsewhere in the plan. The following resources are required: • Space on an archival server • Archival assistance from the institutional records manager or similar experienced employee • All other resources required (e.g., access to Vibe) are already available to the employee and used daily.

APPENDIX 23 CSIR Permission Letter



our future through science

CSIR Business Excellence and Integration

PO Box 395 Pretoria 0001 South Africa
Tel: +27 12 841 2911
Fax: +27 12 349 1153
Email: Enquiries@csir.co.za

25 March 2020

To whom it may concern

PERMISSION FOR CONDUCTING RESEARCH

This letter serves to confirm that permission has been granted for Louise Patterton, a CSIR employee, to make use of CSIR researchers as informants during her doctoral studies. All directors of CSIR clusters have been informed about the study, its purpose, objectives, and data collection methods, and have approved the planned involvement of researchers within their clusters.

Sincerely

Visser

Dr. Daniel Visser
Group Manager: Planning & Knowledge Management
Business Excellence & Integration
CSIR

www.csir.co.za