

Developing a framework for data quality-focused RDM to guide the data management planning process in a Social Sciences Research Organisation

by

Glenn Tiisetso Tshweu
29346836

Submitted in fulfilment of the requirements for the degree

Doctor of Philosophy in Information Science

in the

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

at the

UNIVERSITY OF PRETORIA

Supervisors:

Dr MJ van Deventer

Prof TJD Bothma

March 2024

Declaration regarding plagiarism

I, Glenn Tiisetso Tshweu, student number 29346836, in respect of the thesis: Developing a framework for data quality-focused RDM to guide the data management planning process in a Social Sciences Research Organisation, declare the following:

- I understand what plagiarism entails and am aware of the University's policy in this regard.
- I declare that this thesis is my own, original work. Where other people's work has been used (whether from a printed source, the Internet or any other source) due
- acknowledgement has been given and the source referenced according to departmental requirements.
- I have not used work previously produced by another student or any other person to hand in as my own.
- I have not allowed, and will not allow, anyone to copy my work with the intention of presenting it as his/her own work.

Signature

Date

Editing declaration

PO Box 1439

Wingate Park

0153

22 February 2024

Editing of Thesis: Mr Glenn Tiisetso Tshweu

Student Number: 29346836

This letter serves to confirm that I have performed the English language editing on the thesis to be submitted by Mr Glenn Tshweu in accordance with the requirements for the degree of DPhil (Information Science) in the Department of Information Science, Faculty of Engineering, Built Environment and Information Technology at the University of Pretoria and entitled:

Developing a framework for data quality-focused RDM to guide the data management planning process in a Social Sciences Research Organisation

While I am a permanent employee of the Directorate: Language Services at the University of South Africa in Pretoria – where I serve as an editor and translator – the editing of Mr Tshweu’s thesis was undertaken in my personal capacity.

Yours faithfully



CE Baudin

082 803 3355

Acknowledgements

I would like to express my sincere appreciation to:

- Professor Theo Bothma and Dr Martie van Deventer for their valuable advice, direction, leadership and input as study leaders
- Dr Lucia Lotter for her support and assistance during this study
- My mother and father for their encouragement and support and also for contributing financially towards my studies
- My beautiful girlfriend Kelebogile Motlopye for her continuous support and advice.

Developing a framework for data quality-focused RDM to guide the data management planning process in a Social Sciences Research Organisation

Abstract

The aim of the study was to conceptualise how RDM (research data management) can facilitate best practice and the production of quality data in a social sciences research organisation, with this to be achieved by focussing on the full research data lifecycle (i.e. all the different phases) and the activities in each stage of the data lifecycle. The goal was to create a practical, quality-focused RDM framework that can be utilised to direct the data quality management planning process in a Social Sciences Research Organisation.

In the first part of the study, a literature review was conducted, focussing on ten topics: research data management as a theoretical concept; data quality as a theoretical concept; data quality as documented by international RDM role-players; data quality as recognised by significant RDM initiatives; data quality as an interest for large research funders; data quality as expressed in evaluation systems; quality recommendations as expressed in books on social science methodology; quality as an element of data management planning; data quality evaluation in South African RDM initiatives; and RDM refocused to incorporate data quality. The reason for reviewing these topics was to discuss and understand how the concepts of RDM, data quality and quality assurance are perceived and addressed in the literature. The literature review chapter was a necessary point of departure for this study, making it possible to identify gaps in the knowledge in the field and bringing to light opportunities to contribute to the development of a quality-focused RDM framework. The reviewed literature facilitated the identification of measuring indicators that are used to assess how quality assurance is maintained as part of the different phases in the research data lifecycle. The criteria thus identified are social sciences data measuring criteria based on selected data quality dimensions.

In the second part of the study, the framework development was outlined. The chosen strategy used to develop the quality-focused RDM framework employed a theoretical approach. The theory reviewed in the literature review chapter was used firstly as the foundation of the quality-focused RDM framework and, secondly, as the measuring criteria to assess how data is managed in the different stages of the data lifecycle.

In the third part of the study, the research methodology was defined. The selected methodology involved a hybrid approach consisting of a non-empirical section comprising a literature review, and an empirical section consisting of a case study from a social sciences organisation. The case study was selected because the social sciences institution concerned already had a mature data curation system in place – one that prioritised quality assurance. Investigating how the concept of ‘data quality’ is maintained – using RDM services – was another reason why this social sciences organisation was chosen. This study

is regarded as a qualitative rather than a quantitative study, as the primary aim was to establish and explore the utility for a social sciences organisation of a developed framework based on data quality.

The contribution of this study lies primarily in the presentation of a revised quality process model RDM framework that includes anecdotal suggestions (from research participants) on how to better manage research data. These suggestions were used to produce a concentrated quality-focused RDM framework that is based on theoretical best practices and the refined practical measures that researchers use and trust.

Keywords

Research data management (RDM), quality assurance, research data lifecycle, data quality, quality dimensions, theoretical framework, measuring criteria, social sciences data.

TABLE OF CONTENTS

Declaration regarding plagiarism.....	i
Editing declaration.....	ii
Acknowledgements.....	iii
Abstract.....	iv
Keywords.....	v
List of figures.....	xvi
List of tables.....	xvii
List of abbreviations	xviii
CHAPTER 1: DATA QUALITY MANAGEMENT IN CONTEXT	20
1.1 Introduction.....	20
1.2 Research problem.....	23
1.3 Research objectives.....	24
1.4 Research questions.....	25
1.5 Relevance of study for the subject field.....	25
1.6 Existing knowledge - literature overview	26
1.7 Research methodology	29
1.7.1 Research paradigm/philosophy	29
1.7.2 Research approach	29
1.7.3 Research design.....	30
1.7.4 Method of data collection	30
1.7.5 Methods of data analysis.....	31
1.7.6 Clarification of key terms	31
1.8 Target population and sampling.....	32
1.9 Research scope and limitation.....	32

1.10	Description of chapters	32
1.11	Conclusion	34
CHAPTER 2: LITERATURE REVIEW		35
2.1	Introduction.....	35
2.2	Research data management as a theoretical concept.....	36
2.2.1	Defining research data management	36
2.2.2	Benefits of RDM.....	40
2.2.3	Generic challenges to managing research data	42
2.3	Data quality as a theoretical concept.....	48
2.3.1	Defining data quality.....	51
2.3.2	Scientific quality vs technical quality of data	53
2.3.3	Data quality dimensions.....	54
2.3.4	Data ‘integrity’, ‘authenticity’ and ‘validity’	56
2.3.5	Data quality guidelines.....	57
2.3.6	Impact of data quality	64
2.4	Data quality as documented by international RDM role-players.....	65
2.5	Data quality as recognised by significant RDM initiatives.....	66
2.6	Data quality as an interest for large research funders	67
2.7	Data quality as expressed in evaluation systems.....	67
2.7.1	International Standards Organization (ISO).....	68
2.7.2	Core Trustworthy Data Repositories Requirements (Core Trust Seal).....	69
2.8	Quality recommendations as expressed in books on social science methodology.....	71
2.8.1	First book on social sciences methodological methods considered	72
2.8.2	Second book on social sciences methodological methods considered.....	74
2.8.3	Third book on social sciences methodological methods considered.....	75
2.8.4	Fourth book on social sciences methodological methods considered.....	77
2.8.5	Fifth book on social sciences methodological methods considered.....	79

2.8.6	Sixth book on social sciences methodological methods considered	80
2.8.7	Seventh book on social sciences methodological methods considered.....	80
2.8.8	Eighth book on social sciences methodological methods considered.....	80
2.8.9	Ninth book on social sciences methodological methods considered	81
2.8.10	Tenth book on social sciences methodological methods considered	81
2.9	Quality as an element of data management planning.....	81
2.9.1	Quality assurance (QA) and quality control (QC)	82
2.9.2	File Formats	83
2.9.3	Documentation.....	83
2.9.4	Metadata.....	84
2.9.5	Short-term storage and data management infrastructure.....	85
2.10	Data quality evaluation in South African RDM initiatives	86
2.11	RDM refocused to incorporate data quality	86
2.11.1	Address the human side of change systematically	86
2.11.2	Change starts at the top and begins on day one	87
2.11.3	Real change happens at the bottom.....	88
2.11.4	Confront reality, demonstrate faith and craft a vision.....	88
2.11.5	Create ownership, not just buy-in	88
2.11.6	Practice targeted over-communication.....	89
2.11.7	Explicitly address culture and attack the cultural centre.....	90
2.11.8	Assess the cultural landscape early	90
2.11.9	Prepare for the unexpected.....	91
2.11.10	Speak to the individual as well as to the institution	91
2.12	A consolidated view of the data quality indicators	92
2.13	Summary	92
CHAPTER 3: A RELEVANT FRAMEWORK TO GUIDE DATA QUALITY		94
3.1	Introduction.....	94

3.2	Defining frameworks and framework development	94
3.2.1	Frameworks.....	94
3.2.2	Identifying assumptions	96
3.3	Frameworks typology	96
3.3.1	Theoretical framework.....	97
3.3.2	Conceptual framework.....	97
3.3.3	Process model as a framework.....	99
3.4	Principles for the development of a framework.....	100
3.5	Framework development strategy deployed: University of Kansas process framework	104
3.6	Framework documentation	108
3.7	Relevant frameworks	111
3.7.1	UK Data Archive model as a framework.....	111
3.7.2	Inter-university Consortium for Political and Social Research (ICPSR) model.....	118
3.7.3	The South African Statistical Quality Assessment Framework (SASQAF)	119
3.8	A process model and quality assessment framework for Social Sciences data.....	120
3.8.1	A quality-focused RDM for Social Sciences data.....	121
3.8.2	The intended uses of the model and the framework.....	121
3.8.3	The objectives of the model and the framework.....	121
3.8.4	Developing the process model	122
3.8.5	Developing the quality measuring framework.....	123
3.8.6	Components included in the quality review framework	124
3.8.7	The components included in the quality process model	125
3.8.8	Depicting the quality process model.....	127
3.8.9	Checking the completeness of the quality measuring framework.....	131
3.8.10	Putting the framework to use	132
3.8.11	Revising the model.....	133
3.9	Summary	133

CHAPTER 4: RESEARCH METHODOLOGY	135
4.1 Introduction.....	135
4.1.1 Short overview and goal of the study.....	135
4.2 Research paradigm.....	137
4.2.1 Interpretivist paradigm.....	137
4.2.2 Positivist paradigm.....	138
4.3 Research approach	138
4.3.1 Qualitative research.....	138
4.3.2 Quantitative research.....	142
4.3.3 Mixed-method research.....	144
4.4 Research design	145
4.4.1 Case study	145
4.4.2 Justification for using a case study design	149
4.5 Target population	150
4.6 Sampling	150
4.6.1 Probability sampling	151
4.6.2 Non-probability sampling	151
4.7 Data collection methods.....	153
4.7.1 Qualitative data collection methods.....	154
4.8 Pilot study	156
4.9 Data collection tools	157
4.9.1 Questionnaires.....	157
4.9.2 Interview schedule	158
4.10 Selection of research location	159
4.11 Data analysis and interpretation.....	159
4.12 Data management plan (DMP)	161
4.13 Validity and reliability	162

4.14	Ethics approval by the university	164
4.15	In summary	165
CHAPTER 5: RESEARCH FINDINGS AND ANALYSIS		166
5.1	Introduction.....	166
5.2	Background of the case study and data collection	166
5.3	Structure of the data analysis	166
5.4	Findings: Section A - Demographics of the research participants	167
5.4.1	Educational background of the research participants.....	167
5.4.2	Expertise as a researcher and data curator across time	168
5.5	Findings: Section B - Measuring data quality.....	169
5.5.1	Findings: Quality dimensions to add (Question 3) – Researchers	169
5.5.2	Findings: Quality dimensions to add (Question 3) – Curators.....	171
5.5.3	Analysis and discussion of the findings	171
5.5.4	Interview schedule Question 4: Scientific quality – Research participants	173
5.5.5	Interview schedule Question 5: Technical quality – Research participants.....	175
5.6	Findings: Section C - Measuring the Data Quality Dimension compliance - Research participants.....	178
5.6.1	Accessibility.....	179
5.6.2	Accuracy	180
5.6.3	Comparability	181
5.6.4	Completeness	181
5.6.5	Consistency	182
5.6.6	Flexibility of the data	183
5.6.7	Integrity.....	184
5.6.8	Interpretability.....	185
5.6.9	Methodological Soundness	186
5.6.10	Relevance	188
5.6.11	Standards and Policies	188

5.6.12	Timeliness	189
5.6.13	Uniqueness	189
5.6.14	Usability of the data	190
5.6.15	Validity	191
5.6.16	In conclusion and after considering all the measures collected	192
5.7	Personal experience and preferences in data quality management planning	192
5.7.1	Quality dimensions and measuring criteria that researchers address	193
5.7.2	Quality dimensions, measuring criteria and evaluation standards curators use when curating research data	195
5.7.3	Quality procedures that researchers use/are aware of	195
5.7.4	Quality control processes that curators use	197
5.7.5	Planning data quality at the same time as data management planning	198
5.7.6	Procedure in place to confirm data quality preparation completed - Curators	198
5.7.7	When to measure the scientific quality of data - Researchers	199
5.7.8	When to measure the scientific quality of the data - Curators	200
5.7.9	When to measure the technical quality of data – Researchers	201
5.7.10	When to measure the technical quality of data - Curators	201
5.8	Findings: Section C - Opinions about the proposed quality process model	202
5.8.1	Opinions expressed by Researchers	203
5.8.2	Opinions about the proposed model - Curators	204
5.8.3	Required changes to the model	204
5.8.4	In which areas will the model assist? - Researchers	206
5.8.5	In which areas will the model assist? - Curators	208
5.8.6	The data management areas covered well by the model - Researchers	208
5.8.7	The data management areas covered well by the model - Curators	209
5.8.8	Areas of concern that the model will NOT address – Researchers	210
5.8.9	Areas of concern that the model will NOT address - Curators	211
5.8.10	Anything else regarding the quality of research data - Researchers	212
5.8.11	Anything else, regarding the quality of research data - Curators	212

5.9	Chapter summary	213
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS		214
6.1	Introduction.....	214
6.2	Overview of the study	214
6.3	Defining the term ‘data quality’	215
6.4	Stakeholders addressing data quality issues.....	217
6.4.1	Large international RDM role-players	217
6.4.2	Large research funders	218
6.4.3	Evaluation systems.....	219
6.4.4	South African RDM initiatives	219
6.4.5	South African academic and government institutions.....	220
6.5	The characteristics of a framework for data quality-focussed RDM	221
6.6	Measuring data quality using the data quality dimensions	224
6.6.1	Data quality assurance	224
6.6.2	Quality control	224
6.6.3	Data ‘integrity’ and ‘authenticity’	225
6.6.4	Data validation methods.....	225
6.6.5	How each data quality dimension assists with data quality	226
6.7	Remedies for the proposed quality-focused process model	231
6.7.1	The characteristics of a model for data quality	231
6.7.2	Features of an effective, quality-focused framework.....	232
6.7.3	Remedies for the revised quality-focused process model - Planning.....	234
6.7.4	Remedies for the revised quality-focused process model – data collection.....	242
6.7.5	Processing data.....	247
6.7.6	Analysing data.....	250
6.7.7	Preserving data.....	253
6.7.8	Giving access to data.....	256

6.7.9	Reusing data.....	259
6.7.10	The final quality process framework.....	263
6.8	An overview of the functioning of the quality process framework.....	265
6.9	Features of an effective, quality-focused framework.....	265
6.10	Change management recommendation for the social sciences institution.....	267
6.11	Suggestions for further research	271
6.12	Contribution of this study to the subject field.....	272
6.13	Critical evaluation of the research methodology	273
6.14	Chapter summary	274
	References.....	276
	Appendix 1: Extended background information	312
	Appendix 2: Quality Process Framework	329
	Appendix 3: Data lifecycles with an emphasis on quality	335
	Appendix 4: Measurements for each quality dimension.....	346
	Appendix 5: Criteria used to identify relevant models	363
	Appendix 6: Flow diagram symbol description – initial Quality Process Framework	364
	Appendix 7: GlennTshweu – DMPDCCDMPOnline	365
	Appendix 8: Interview schedule - Researchers.....	367
	Appendix 9: Interview schedule – Data curators	377
	Appendix 10: Research participant handout	387
	Appendix 11: The initial quality process model	393
	Appendix 12: Results of the criteria used to identify relevant role players/models	394
	Appendix 13: Grouping of quality dimensions with stages of the research data lifecycle that form the quality process framework.....	407

Appendix 14: UK Data Quality Dimensions	409
Appendix 15: Data quality evaluations by different stakeholders	412
Appendix 16: Augmented UK DAMA data quality dimensions	435

List of figures

Figure 1. 1 - UK Data Archive data lifecycle	24
Figure 3. 1 - Conceptual framework for data-driven decision-making skills	101
Figure 3. 2 - Center for Community Health and Development framework (Author's depiction)	104
Figure 3. 3 - UK Data Archive research data lifecycle (UK Data Archive, 2017)	112
Figure 3. 4 - Activities influencing data quality in the first two stages of the research data lifecycle	115
Figure 3. 5 - ICPSR Data Lifecycle	118
Figure 3. 6 - Overview of the quality review process	123
Figure 3. 7 - Establishing the quality review framework.....	124
Figure 3. 8 - Proposed quality process model	128
Figure 3. 9 - Reviewing data quality (Technical measures).....	130
Figure 3. 10 - Reviewing data quality (Scientific measures)	130
Figure 3. 11 - Creating Data phase quality dimensions as part of the quality review framework	129
Figure 4. 1 - Basic differences between quantitative and qualitative research	143
Figure 6. 1 – The initial; quality process framework	222
Figure 6. 2 - Revised final version of Data Management Planning phase of quality process model..	241
Figure 6. 3 - Revised final version - Creating Data phase of the quality process model	246
Figure 6. 4 - Revised final version - Processing Data phase of the quality process model.....	249
Figure 6. 5 - Revised final version - Analysing Data phase of the quality process model	252
Figure 6. 6 - Revised final version - Preserving Data phase of the quality process model	255
Figure 6. 7 - Revised final version - Giving Access to Data phase of the quality process model	258
Figure 6. 8 - Revised final version of the Reusing Data phase of the quality process model	261
Figure 6. 9 - Final quality process framework	264

List of tables

Table 2. 1 - Key Elements of Behaviour Change (Witte, 2015)	43
Table 2. 2 - Data errors that may occur during a research project (Chapman, 2005)	53
Table 2. 3 - Recommended code and coding guidelines for researchers	61
Table 3. 1 - The differences between a theoretical framework and a conceptual framework.....	98
Table 3. 2 - Typical process models	99
Table 3. 3 - Example of the SASQAF dimension (Statistics South Africa, 2010)	120
Table 4. 1 - Steps of a phenomenological research study	140
Table 4. 2 - Case study typology.....	147
Table 4. 3 - Types of research questions that a research strategy should answer	149
Table 4. 4 - Validity and Reliability comparison (Surbhi, 2017).....	162
Table 5. 1 - Academic qualifications of participants	168
Table 5. 2 - Years of experience	168
Table 5. 3 - Alignment between suggested and available quality dimensions.....	172
Table 5. 4 - Quality dimensions that can determine scientific quality	173
Table 5. 5 - Quality dimensions that can determine technical quality	176
Table 5. 6 - Dimensions excluded from the discussion	193

List of abbreviations

Acronym	Phrase in full
ADA	Australian Data Archive
ARC	Australian Research Council
ARDC	Australian Research Data Commons
CARDIO	Collaborative Assessment of Research Data Infrastructure and Objective
CESSDA	Consortium of European Social Science Data Archives
CIESIN	Center for International Earth Science Information Network
CODATA	Committee on Data for Science and Technology
CPUT	Cape Peninsula University of Technology
CSIR	The Council for Scientific and Industrial Research
DAMA	Data Management Association
DAMA UK	Data Management Association UK
DataONE	Data Observation Network for Earth
DCC	Digital Curation Centre
DDF	Data Deposit Form
DEA	Department of Environmental Affairs – RSA
DIRISA	Data Intensive Research Infrastructure for South Africa
DMP	Data Management Plan
DOH	Department of Health
DOI	Digital Object Identifier
DPE	Digital Preservation Europe
DQR	Data Quality Rules
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
DSA	Data Seal of Approval
DST	Department of Science and Technology
ERC	European Research Council
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
IR	Institutional Repositories
ISC	International Science Council
ISO	International Standards Organization
JISC	Joint Information Systems Committee
LIS	Library and Information Science

Acronym	Phrase in full
NISS	U.S. National Institute of Statistical Sciences
NRF	National Research Foundation
OECD	Organization for Economic Cooperation and Development
RDM	Research data management
RI	Research Infrastructures
LIASA	The Library and Information Association of South Africa
MRC	The South African Medical Research Council
NeDICC	The Network of Data and Information Curation Communities
NICIS	The National Integrated Cyber Infrastructure System
NSF	The National Science Foundation
OAIS	The Open Archival Information System
RTF	Rich Text Format
SADA	The South African Data Archive
SAEON	The South African Environmental Observation Network
SASQAF	South African Statistical Quality Assessment Framework
SEDAC	Socioeconomic Data and Applications Center
Stats SA	Statistics South Africa
TLA	The Language Archive
UCT	The University of Cape Town
UKDA	UK Data Archive
UP	University of Pretoria
USAID	United States Agency for International Development
WDS	World Data System

CHAPTER 1: DATA QUALITY MANAGEMENT IN CONTEXT

1.1 Introduction

Exceptional Research Data Management (RDM) is the cornerstone of secondary data use but RDM on its own does not guarantee good data quality. According to Nguyen (2023:52), “the interrelation between data quality and information use can be seen as [both] a vicious and virtuous data cycle.” The quality of data depends on the processes and procedures used in data quality management and are meant to guarantee and enhance data quality during the course of the data lifecycle. The goal of this research study is to provide a quality process framework that will connect data quality, data planning, data collection, data processing, and all other management tasks involved in the research data lifecycle.

Data quality management procedures include data cleaning, validation, profiling, and continuous monitoring. To preserve and improve the quality of their data assets, organisations should create data governance policies and use data quality solutions. Using this approach “data must be recognized as a valued and strategic enterprise asset that must be managed effectively” (Wichita State University, 2023:2).

Data quality is a crucial factor in the social sciences since the validity, reliability, and precision of the data collected determine research findings, and conclusions upon which resolutions and problem-solving are based. To improve the general quality of social science data, researchers are required to frequently use reliability tests, validation checks, and other quality control and quality assurance procedures. Information on the issues pertaining to the challenges of managing research data and the impact of data quality is provided in sections, 2.2.3 and 2.3.5.

According to Zhou et al. (2023:2) “over the past ten years, researchers have encountered many challenges related to managing data. Academic librarians have noticed these issues, and in response, many academic libraries have established RDM services and conducted research on RDM needs and best practices”. The following are the activities that represent what RDM services entail:

- Ensures reliability: a measurement is considered reliable if it consistently yields results with the same values. It assesses a study's reliability, accuracy, reproducibility, and consistency. It shows how bias-free (error-free) the measurement is, which guarantees consistent measurement over time and among the different instrument items (the observed scores) (Haradhan, 2017).
- Data Management Planning: RDM requires a researcher to complete a DMP. According to van der Graaf (2023:6), "guidelines and formats for data management plans (DMPs) of research funders and/or the institution" must be followed by researchers in order to complete their DMPs.

- Establishing successful collaboration: Working with non-library stakeholders in fields like information technology and research administration is frequently a requirement of research data management for librarians. However, if these stakeholders are unaware of the library's value proposition in RDM, this might breed scepticism or disinterest. Consequently, it is crucial to establish confidence with cooperative stakeholders outside of other libraries (Bryant et al., 2023:8).
- Infrastructures and services for data sharing and data storage during the research project: Without specific RDM services, universities, faculties, and research departments have infrastructures in place for data storage and data sharing for researchers (van der Graaf, 2023:8).
- Manage research data throughout the research data lifecycle: this means that for each research project, researchers should consider and make decisions about how they manage data during data collection, data analysis, and results preparation as well as what happens with research data after the end of the research project (Wilkinson et al., 2016).

Research data quality, accessibility, and long-term usability are all dependent on effective RDM. Effective research data management improves data quality and reproducibility and makes it easier for scientists to collaborate and share data, which is why the aforementioned RDM services are helpful.

A snapshot of what has been occurring globally in the RDM landscape is given by Rod et al. (2023:2) “over the past 10 years, research data management (RDM) services have been increasingly defined and formalized in tandem with the growth of new policies, norms, and principles regarding the role of research data in scientific reproducibility and transparency”. The RDM landscape in a developing country is different from that of countries in first world countries. Van Deventer and Pienaar (2015:2) reported that “... in South Africa, researchers (even at our own institutions) were initiating many disconnected, small projects to enable participation in the global research network”. Contrary to what is happening in a developing country, Subaveerapandiyan (2023:13) points out that academic institutions (such as the University of Glasgow) have “developed their institutional research management systems to meet the evolving needs of funders, policymakers, managers, and researchers. They emphasise the importance of open science and the challenges of supporting researchers in implementing research data management processes that align with their workflows. The universities have employed integrated models, utilising standards and fostering an ongoing stakeholder dialogue”. Researchers and institutions must embrace recognised RDM best practice guidelines, standards, and frameworks and extensively practice them in order to attain the level at which international role-players (such as those included in Section 2.4) operate; otherwise, such small projects will not provide any fruitful outcomes.

As researchers collect more and more data they are held increasingly more accountable for stewarding their research data for the long-term. There are many factors to consider when attempting to change how researchers manage their data and how they can better produce quality data. Chapter 2 discusses several topics that need to be carefully considered in order to try to enhance the way researchers manage their data and produce higher-quality data, and these issues will be covered in this research study. Thorough preparation, rigorous procedures, and close attention to detail are needed to address these issues at every stage of the research process, from study design to data collecting, analysis, and interpretation.

Data management planning and the application of a Data Management Plan (DMP) are crucial elements of RDM. A research project's data management plan (DMP) "details how data will be collected, processed, analyzed, described, preserved, and shared" (National Library of Medicine, 2023:1). Comprehensive information, such as the kinds of data generated, the metadata standards applied, the access and sharing regulations, and the strategies for data archiving and preservation to ensure its accessibility over time, must all be included in a DMP. A data management plan is necessary to support the long-term use of research data, ensure regulatory compliance, encourage collaboration, and promote excellent research practices. The explicit planning that the DMP brings with it enhances the general calibre and significance of scientific research.

Just as a DMP is an essential part of RDM the effective use of secondary data relies heavily on proper RDM. RDM is an important activity that researchers acknowledge when carrying out research work. Research institutions and associated researchers are bound by RDM policies and the mandates of funders (Patel, 2016). The quality of the data must be ensured or secondary use will not be desirable; maintaining the quality of research data is essential for ensuring data reuse and shareability, as well as for consolidating social data with other types of data for other purposes. The primary emphasis is that the organisation faces significant risks, making data quality non-negotiable. RDM is an essential task that researchers must consider when doing their research. When conducting scientific research, researchers need to follow a set of methodical principles to ensure that study findings are correct and verifiable. According to Awre et al. (2015), "managing research data has long been a concern for researchers - however, the quantity and complexity of digital data make managing it effectively increasingly challenging". The management of research data is a basic concept that all researchers should be aware of in their work. Curdt and Hoffmeister (2015:495) claim that "evidence-based research is created and sustained by effective research data management" because evidence is highly appreciated in the scientific community. When RDM is used, research data can be disseminated and made accessible for further use.

Sharing data has a number of advantages. When data is shared, Patel (2016:227) states that “data can be reused by other researchers and applied in various other contexts; when data is shared the authenticity and objectivity of the data will be reaffirmed by usage.” The various components that go into managing research data combine to form a comprehensive process. According to Curdt and Hoffmeister (2015:495), “the technical framework to collect, describe, and provide research data is an RDM system”. Data collection is a crucial component of maintaining data quality. According to Taherdoost (2021:11), “data collection as a main stage in research can overshadow the quality of achieving results by decreasing the possible errors which may occur during a research project. Therefore, alongside a good design for the study, plenty of quality time should be spent in the collection of data to gain appropriate results since insufficient and inaccurate data prevents assuring the accuracy of findings”. Researchers can handle their research data properly with the help of an RDM system. According to University of Pittsburgh (2024:1), the solution “describe[s] the organization, storage, preservation, and sharing of data collected and used in a research project”. The research problem of this study is examined in the following section.

1.2 Research problem

The Social Sciences Research Organisation that will be examined conducts both quantitative and qualitative research investigations. That being said, the suggested RDM framework would have to accommodate research programs that are both qualitative and quantitative. The research data life cycle offers a high-level overview of the steps necessary for the effective management and preservation of data for use and reuse when applied in connection to the data lifecycle (DataOne, 2017:1). The data management concerns that this study focuses on include research behaviour, copyright issues (linked especially to the rights of disseminating of data), data licensing, data security, and the re-use of research data. These concerns are discussed in Section 2.2.3. The phases that research data passes through throughout a typical research project are depicted in Figure 1.1, below, which depicts the data lifecycle:

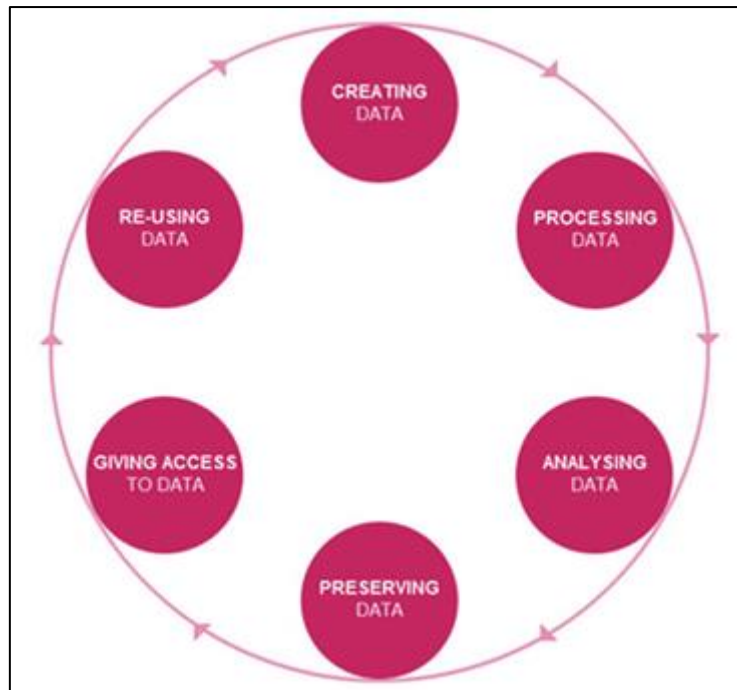


Figure 1.1 - UK Data Archive data lifecycle

A thorough assessment of researchers' current data management procedures is necessary to ensure the practicality of creating a viable RDM guiding framework. These data management techniques will be contrasted with the data management standards, RDM best practices, and some of the challenges faced by researchers (2.9.1, 2.9.2, 2.9.3, 2.9.5, 2.2.3). Recommendations for raising quality at each stage of the research data lifecycle are provided in the conclusion.

1.3 Research objectives

The idea behind this research is to consider how RDM may support best practices and improve the calibre of data output that a Social Sciences Research Organisation produces. Improving the calibre of data generated by a Social Sciences Research Organisation is a crucial component of this study.

Creating an RDM framework with an emphasis on quality is the primary goal of this study. As part of the data lifecycle, this framework will be employed as a mechanism for improving the quality of the data in order to solve the issues that come up during research data management planning procedures at a Social Sciences Research Organisation. By combining optimal RDM practices, guidelines, and benchmarking against other RDM frameworks, such a framework will be developed. Because this method will be used to reflect on issues in terms of best practice data management standards, it is envisioned that it has a positive effect on the quality of the data. The goal of the proposed study is to improve the quality of the data produced while utilising an existing data lifecycle model, which includes all of the lifecycle's phases.

The purpose is to use the defined framework to provide recommendations for strategy and process improvements to the organisation. Enhancing quality across the research data lifecycle requires starting with improving the efficiency of social science research data collection and management while following RDM best practices. In Chapter two of the dissertation, several RDM frameworks, guidelines, and best practices will be thoroughly reviewed.

1.4 Research questions

To be able to develop a quality-focused RDM guiding framework, the following main research question was formulated: “What would an effective, quality-focused framework, to guide the data management activities of a Social Sciences Research Organisation, look like?”

In order to address the main research question, the following sub-questions were identified:

1. What is meant by the term ‘data quality’?
2. How does the network of stakeholders, involved in research data management, address data quality issues?
3. What will be the characteristics of a framework for quality-focused data management?
4. Considering the data quality dimensions, how is data quality measured?
5. What are the unique remedies that the proposed quality-focused framework could incorporate?
6. What recommendations regarding change management would be applicable to the changing of RDM practices at South African Social Sciences Research Organisations so as to include data quality?

The main research question and sub-questions will be used to develop a framework for data quality-focused RDM to guide the data management planning process in a Social Sciences Research Organisation. The next section discusses the Relevance of study for the subject field.

1.5 Relevance of study for the subject field

Since RDM is so important in research, researchers should seek ways to consistently improve their RDM skills and competence. Policies and procedures must be in place to enlighten both researchers and data users from an organisational standpoint. For RDM techniques to be adopted, organisational policies must be matched with acknowledged data quality standards. This would increase the quality of RDM and ensure that research activities are carried out to a higher standard.

RDM is a global phenomenon that South African research data producers must constantly enhance in order to achieve research data quality criteria. RDM is still a (relatively) new idea in South Africa (De

Jager, 2016). Consequently, it is critical to promote RDM best practices at the organisational level in order for such practices to become developed in research-generating organisations.

The importance of this study lies in the development of a quality-focused framework that will introduce new approaches to RDM and allow data curators to play a more active role in assisting researchers to be more cognisant and responsive when conducting research, all in the pursuit of producing high-quality data.

1.6 Existing knowledge - literature overview

RDM has grown in importance as a means of guaranteeing effective data management. Wilms et al. (2018:4411) emphasise that “research data emerge over the entire research lifecycle, including the collection, analysis and evaluation of data”. As a result, producing high-quality data and accurately managing data throughout the various phases of the data research lifecycle becomes crucial. Such data have great importance because they aid in the discovery of novel ideas and breakthroughs. Mining the research data that is readily available is of great importance to researchers, educational institutions, and governments alike since it expedites the advancement of research in various disciplines of study (Wilms et al., 2018).

Curdt and Hoffmeister (2015:495) previously reported that “many national and international institutions” (e.g. The National Science Foundation (NSF), The Organisation for Economic Co-operation and Development (OECD), The UK Research Council) have emphasised the importance of research data management (RDM). Data that have been generated in the context of public funding should be archived and reusable for future research questions”. In order for publicly funded data to be reusable it has to meet certain quality standards that will ensure the integrity of the data. Data quality dimensions contribute to the goal of making research data accurate, consistent and reliable. Sengupta et al. (2023:6) reported that “literature on data quality reveals that researchers follow a different methodology to evaluate data quality”, but Wang et al. (2023) reported that data quality refers to “... the degree of data usability”. Arriving at a working definition of the term ‘data quality’ required an extensive review and understanding of what ‘quality’ entails in today’s data and information society. According to McCord et al. (2022:17), “data quality is more than data management. Ensuring data quality requires 1) clear communication among team members; 2) appropriate sample design; 3) training of data collectors, data managers, and data users; 4) observer and sensor calibration; and 5) active data management”. Further research in this area is needed to uncover what factors contribute to establishing data quality in social sciences research. It is the viewpoint of Shanmugam et al., (2023) that “numerous researchers have attempted to define data quality and identify its measures, typically focusing on accuracy. However, research has shown that data quality encompasses various other dimensions as well.”. This study intends to highlight different quality dimensions and assign measuring criteria to each quality dimension to arrive at useful assessment guidelines for managing social science

research data. It is doubtful that a single set of data quality metrics can be applied to all data, despite the efforts of some academics to do so. This is because different data contexts have different characteristics. Accuracy, completeness, timeliness, and consistency are the four generally used data quality metrics; however, enhancing one may compromise another (Shanmugam et al., 2023).

In one study Masinde et al. (2021:413) reported that “when research data is appropriately managed, it can ensure more productive and efficient science as new knowledge is created by building on discoveries, innovations, and open scientific inquiry”. Reusing high-quality research data can save costs by lowering redundancy, improving decision-making, enabling compliance with funding requirements, and making the data publicly available. A key goal of this study was to design a framework that would address the planning issues related to data management and provide guidance on data quality management practices. Partelow (2023) is of the view that “Frameworks are important research tools across nearly all fields of science. They are critically important for structuring empirical inquiry and theoretical development”. Frameworks are not generic; rather, they are discipline-specific and should serve a specific purpose because of their unique traits. Frameworks are not generic; rather, they are discipline-specific and should serve a specific purpose because of their unique traits. As a result, Ban and Cox (2017) believe that it is unclear what makes an effective scientific framework and what its objective is. Shanmugan et al.(2023) proposed a framework for data quality, which includes “...top-level direction and guidance for data quality, a management function to determine how to implement the approach, and a system for executing and verifying data quality control”. In the context of this research study, an organisation might use a systematic approach, set of guidelines, and protocols known as a framework for data quality to ensure that the data used for its operations is accurate, dependable, consistent, and pertinent.

The LinkedIn Corporation (2023) explains that “the exponential growth in data volume and variety has presented a significant challenge to organizations seeking to maintain data quality. Traditional methods of manually inspecting and cleansing data are no longer feasible given the sheer scale of data being generated. To address this challenge, enterprises are adopting a scalable approach to data quality”. One such approach is the development of a framework for data quality-focused RDM to guide the data management planning process in a Social Sciences Research Organisation, which is the ultimate objective of this research study.

The objectives of the literature review (Chapter 2) were to establish a theoretical foundation for the empirical investigation and assist in defining the key concepts.

The following sections make up the literature review:

- Research data management as a theoretical concept (Van den Eynden, 2011; Asher et al., 2013; Kahn et al., 2014; Australian National Data Service, 2015; Domingus, 2015; Tripathi et al., 2017; European Research Council (ERC), 2018; Borghi et al., 2018; University of Johannesburg, 2023; McMaster University, 2023; Bryant et al., 2023).
- Data quality as a theoretical concept (Cooper and Kahn, 2012; Sidi et al., 2012; Doorn and Dillo, 2017; Fischer et al., 2023; Gutema, 2023; Rudolph et al., 2023; Sengupta et al., 2023; Woods et al., 2023).
- Data quality is documented by international RDM role-players (World Data Systems, 2015; The Research Data Alliance, 2016; DAMA UK, 2017; Organisation for Economic Co-operation and Development (OECD), 2017; UK Data Archive, 2017; Inter-university Consortium for Political and Social Research, 2018); Victoria State Government, 2018; UK Data Service, 2019; UK Data Service, 2023).
- Data quality as recognised by significant RDM initiatives (InterPARES research projects, 2008; The Australian Research Council (ARC) now known as Australian Research Data Commons ARDC, 2016; California Digital Library (CDL), 2020; Socioeconomic Data and Applications Center (SEDAC), 2020; Digital Repository of Ireland (DRI), 2023).
- Data quality as an interest for large research funders (World Bank Group, 2015; European Research Council (ERC), 2018; United States Agency for International Development (USAID), 2019; The National Science Foundation (NSF), 2024).
- Data quality as expressed in evaluation systems (The World Data System (WDS) – Data Seal of Approval (DSA), 2016; International Standards Organization (ISO), 2018).
- Quality recommendations as expressed in books on social science methodology (Singleton et al., 1993; Frankfort-Nachmias and Nachmias, 1996; Scarbrough and Tanenbaum, 1998; Monette et al., 2002; Maydanchick, 2007; Goertz and Mahoney, 2012; Maggetti et al., 2013; Creswell, 2014; Silverman, 2017).
- Quality as an element of data management planning; important authors consulted included Van den Eynden (2011); Cooper and Kahn (2012); Pienaar (2014); Wissik and Durco, (2015); Bishop et al. (2023); Subaveerapandiyani (2023).
- Data quality evaluation in South African RDM initiatives (The Human Sciences Research Council (HSRC), 2014; The Network of Data and Information Curation Communities (NeDICC), 2014; The Association for Library Professionals of South Africa (LIASA), 2017; The National Research Foundation (NRF), 2017; The University of Cape Town (UCT), 2018; Data-Intensive Research Infrastructure for South Africa (DIRISA), 2019; The Council for Scientific and Industrial Research (CSIR), 2019; The South African Data Archive (SADA)

2020; DataFirst, 2020; The South African Medical Research Council (MRC) 2022; Department of Science and Innovation (DSI), 2020; University of Pretoria (UP), 2023).

- Data Quality Guidelines; notable writers consulted were Wiggins et al., 2013; Abedjan et al., 2018; Galli, 2021; Rohman et al., 2023.
- A consolidated view of the data quality indicators (Laliberté et al., 2004; Oracle, 2009; Statistics South Africa, 2010; Vale, 2010; Russell, 2011; Sidi et al., 2012; Asher et al., 2013; UK-DAMA Working Group, 2013; Buckey, 2016; Jayawardene, et al., 2016; Open Risk Academy, 2016; Charles et al., 2017; OECD, 2017; ICPSR, 2018; Bauman, 2019; Black and van Nederpelt, 2020; Nasr et al., 2020; UNICEF, 2021; Hassenstein and Vanella, 2022; Torana Inc, 2023; Sengupta et al., 2023).

The identification of the gaps in the body of knowledge on data quality and RDM were made possible by the literature review, as were the data quality recommendations. This study required the use of a research methodology framework in order to address its primary research issue.

1.7 Research methodology

A research methodology offers a guide for carrying out a study so that it may be understood and repeated by others. It also aids in assessing the calibre and reliability of the research's results. A research methodology describes the steps, methods, and resources used by researchers to gather, examine, and evaluate data in an organised and rigorous way. Ensuring the validity and reliability of study findings requires a clearly defined research methodology (Taherdoost, 2021). A brief overview of the methodology used for this research is provided here. Refer to Chapter 4 for a detailed discussion of the methodology used.

1.7.1 Research paradigm/philosophy

A research paradigm is a framework or viewpoint that guides how researchers approach the process of inquiry, conceive their work, and comprehend the nature of reality. A research paradigm offers a framework for the way in which researchers view the world, the techniques they employ, and the standards by which they assess information (Kivunja & Kuyini, 2017). The entire research process is impacted by the choice of research paradigm, including the creation of research questions, the choice of techniques, and the interpretation of results. The Interpretivist paradigm is the research paradigm that was applied in this study; further information on this paradigm is provided in Chapter 4 (Section 4.2).

1.7.2 Research approach

In order to investigate RDM, quality assurance, and data quality, the author used a qualitative approach. According to Austin and Sutton (2014:436), “qualitative research involves asking participants about their experiences of things that happen in their lives. It enables researchers to obtain insights into what

it feels like to be another person and to understand the world as another experiences it”. Using this approach, a researcher can investigate and comprehend human behaviour, experiences, and events using a qualitative approach –as opposed to quantitative research, which emphasises statistical analysis and numerical data. (Refer to Section 4.3 for more details).

1.7.3 Research design

An outline of the research design is given in this section (see Chapter 4, Section 4.4 for a more thorough description).

An essential part of the research process, the research design acts as a guide that specifies the structure and methodology of a study. The study design serves as a guide for researchers, describing the procedures they will follow to address their research inquiries or evaluate hypotheses. It guarantees that the investigation is carried out in an orderly and methodical manner (Kaplan, 2004). Because it enabled researchers to thoroughly examine the phenomenon of RDM practices in a social scientific organisation and its current RDM, quality assurance, and data quality procedures, the explanatory case study research approach was chosen for this study.

Chapter 5 covers the empirical portion of the study, which focuses on a single case study (the social sciences organisation). In a case study, the researcher investigates a particular entity or phenomenon and explains the interactions and/or actions of a person maintained by an outside observer (Creswell, 2009). A case study enables a researcher to narrow down to a specific issue and go into great detail about it, which is precisely what this study set out to achieve.

In Chapter 5, results from a formative evaluation method are presented, and then results from a final assessment procedure using interview questions are reviewed. The data used in the case study were gathered using semi-structured interviews as a data collection tool and through face-to-face interviews as a strategy.

1.7.4 Method of data collection

There are several ways to collect data for research, and the approach that is chosen often relies on the study's objectives, research topic, and available resources. The method of collecting data should be chosen with the goals of the study and the characteristics of the phenomenon under investigation in mind (Austin & Sutton, 2014). The different types of data collection methods - including the one employed in this research study (i.e. Semi-structured interviews) - are discussed in Chapter 4 (Section 4.7.1).

1.7.5 Methods of data analysis

Researchers can analyse data more precisely and accurately if their data analysis techniques are clearly specified. This is necessary in order to derive legitimate findings and trustworthy inferences from the gathered data (University of Kansas, 2023).

Content analysis was one of the techniques used to examine the findings from the semi-structured interviews. The collected data was grouped into "Measures", "Themes", "Criteria", "Approaches", "Processes", "Options", "Opinions" and "Concern areas" which indicate categorised thoughts about particular subjects and are then analysed to determine their significance.

1.7.6 Clarification of key terms

This study relies on several fundamental terms, whose definitions enable the study to be contextualised with respect to the primary research question. The main concepts used in this study are defined in this section.

Framework

Any kind of technique creation must acknowledge and refer to an outside body of knowledge, referred to as a framework. Jabareen (2009:50) defines a framework as "characteristics identified from previous inquiry that provide an internal structure that provides a starting point for observations which may result into processes of theorization". A framework is a collection of particular, accepted concepts, guidelines, and procedures that will guide the creation of a research data management framework with an emphasis on quality.

Research data

Research data, according to the University of Leicester, is "recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; all research data is included irrespective of the format in which it is created" (University of Leicester, 2015). The majority of such data is created in digital format. Since research data is intentionally gathered to solve particular research issues and is useful to the scientific community, it differs from other types of data (Cambridge University, 2017).

Research Data Management (RDM)

RDM encompasses all the procedures carried out on research data to guarantee that the data are appropriately arranged from the time the data enters the research cycle until the important findings are shared and archived (Whyte & Tedds, 2011). While there is a lot of overlap in the word "research data management," in the end it refers to a process that "aims to make the research process as efficient as possible, and meet expectations and requirements of the university, research funders, and legislation" (Whyte & Tedds, 2011).

Researcher

The researcher, who bears ultimate responsibility for the results of their research, is at the core of RDM. A researcher is "someone whose job it is to study a subject carefully, especially in order to discover new information or understand the subject better" according to Cambridge University Press & Assessment (2024)). There are many other topic areas under the umbrella of scientific research; for the purposes of this study, a researcher is someone who carries out academic or scientific research in the social sciences.

1.8 Target population and sampling

The target population for this study were researchers who conduct both qualitative and quantitative research projects and data curators in a social science institution. The researcher's familiarity with the participating researchers, their responsibilities and attributes within the social sciences organisation, and the nature of their research projects led to the decision to utilise purposive sampling to find participants.

Purposive sampling was used in this study as a type of non-probability sampling. Purposive sampling "is a sampling technique in which researcher relies on his or her own judgment when choosing members of a population to participate in the study" (Quinn & Cochran, 2002:10). The population of this study was selected and identified based on their actual RDM experience working for a social scientific organisation.

1.9 Research scope and limitation

The research study was carried out within a social sciences context, with social sciences researchers and excluded researchers from different research fields. The research study investigated only one South African Social Sciences Research Organisation, which prevents the proposed quality-focused RDM framework from being generalised. This is a limitation because it means that the investigation will have to be repeated at several similar organisations before generalised recommendations would be valid.

1.10 Description of chapters

The contents of each of the six chapters that make up this research study are described in the section that follows.

Chapter 1: General introduction

This chapter provides an overview of the context of the research problem, followed by the research problem itself and its sub-questions. The scope and limitations of the study are then described, followed by the relevance of the study to the subject field.

Chapter 2: Literature review

This chapter focuses on the matter of quality throughout the practical implementation of RDM. Data quality is discussed from the perspective of international RDM role-players and stakeholders. The chapter also looks at measures aimed at improving the quality of data in South Africa. The advantages of refocusing RDM to include data quality are described, as well as the need for, and efforts to implement, the change management activities essential to supporting the refocused RDM initiative. Various measuring criteria will be identified in this chapter and will then be used to assess the data quality dimensions established.

Chapter 3: A relevant framework to guide data quality

The aim of this chapter is to conceptualise the knowledge gained in Chapter 2. The data from Chapter 2 was used to develop a theoretical framework for determining data quality in a social science organisation. The ideas acquired from the researched literature were blended to match the environment in which the concepts were to be employed to develop an effective measuring framework. Following that, a process framework was created to include quality evaluation.

Chapter 4: Research methodology

The objective of this chapter is to discuss the research methodologies and the design used to investigate how researchers and data curators think about data quality assurance and data quality. The chapter opens with an overview of the aims and goals of the study, followed by a discussion of the research paradigm and methodology. There is also a discussion of the research design employed for the study, as well as the target population, data collection methods, and the pilot study. The chapter closes with an explanation of data collection instruments, study data sampling, research location selection, data analysis and interpretation, the data management plan, and validity and reliability.

Chapter 5: Research findings and analysis

The purpose of this chapter is to describe the key findings of the semi-structured interviews in order to provide interpretation and meaning. This chapter also intends to explain the acquired data in relation to the study topics indicated in Chapter 1. The data analysis technique for this study is based on categorising data into themes.

Chapter 6: Conclusions and recommendations

The goal of this chapter is to reflect on the aims and research questions of the study (as stated in chapter one) and how they relate to the literature review (chapter two) and the research findings (as explained in chapter five). This chapter also discusses the conclusions and recommendations made based on the obtained and analysed data and on the literature review. The chapter concludes with a conclusion to the research investigation.

1.11 Conclusion

Through contextualising the research problem and questions, this chapter has provided an overview of the research study. The chapter has also outlined the scope and limitations of the research, both of which aid in developing a solid explanatory context for the study. The relevance of this study to the topic field has been highlighted in order to demonstrate the importance that the study may have for the social sciences organisation at which it was conducted. The chapter also includes summary explanations of each of the other chapters. The reviewed literature is discussed in the next chapter.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Researchers and research organisations are faced with the reality of data management during their operations (Bloom et al., 2014; Patel, 2016). The research process can be made more efficient and error production can be reduced by proper research data management (RDM). According to the University of Sheffield, good data management:

- Increases the visibility of the study material and the number of citations.
- Provides opportunities for interaction inside the field with other researchers or with other disciplines.
- Facilitates academic data exchange and reuse for future research studies.
- Reduces the possibility of data loss by safeguarding and securing the analysis of data.
- Demonstrates the integrity of science and the validation of research findings.
- Ensures conformity with the standards and policies of both funders and institutional research data (University of Sheffield, 2019).

The benefits of data management demonstrate how maintaining good research practices and satisfying reuse requirements become easier when data is managed efficiently. However, strong RDM alone does not guarantee better data quality. Poor research data is worthless for curating, making available, and archiving since, whether it is stored on the cloud or in an archive, its poor quality will never go away. The matter of data quality is covered in more detail in Section 2.3. The trustworthiness of research findings and the quality of their related data are critical in the context of this investigation. The significance of quality assurance (across the whole data lifecycle) and the requirement that researchers continually produce high-quality data cannot be stressed enough if RDM is to achieve the benefits listed above. The quality of the data handled for this study is evaluated using a modified version of an existing and well-known data lifecycle model – the UK Data Archive – and a framework of data quality dimensions (see Chapter 3, Section 3.7.1).

The emphasis in this chapter is on quality throughout the actual application of RDM. Section 2.4 points to the sixteen appendices to this document, which describe how international RDM stakeholders and role-players are concerned about data quality. Thereafter, initiatives to raise the quality of South African data come into focus. Finally, the necessity for the required change management activities to support the refocused RDM drive, as well as the steps to implement them (see Section 2.11), are examined in depth. The benefits of refocusing RDM to include data quality are then discussed. The following two sections go over RDM and data quality as theoretical ideas so as to provide the necessary background for the study. The quality process model is based on measuring criteria that were collected from the

literature. Where appropriate, a measuring criterion will be identified along with whichever quality dimension it was used to measure.

2.2 Research data management as a theoretical concept

Research is an important aspect of higher education (Sanjeeva, 2018). RDM is a management activity that is tied to the governance of the data created throughout the research process, and research data is the foundation for new knowledge development. RDM can, therefore, be seen as a broad term that includes all of the tasks necessary to guarantee that data collection is not only transparent, easy to repeat, and linear, but also reusable and of integrity. Borghi et al. (2018:2) support this point of view, stating that “related guidance and standards highlight that proper data management is a key factor in enabling effective data sharing which is itself a key factor in establishing research transparency and reproducibility”. This viewpoint is supported by the University of Johannesburg (2023) “RDM facilitates the sharing of research data and, where appropriate, the re-use of data, in order to foster collaborative practices amongst scholars that maximises the financial and intellectual benefit of research data collected and makes the most effective use of public funds”. For the purposes of this study, it was vital to define RDM as a concept in order to contextualise its importance in relation to data quality.

2.2.1 Defining research data management

According to Tripathi et al. (2017:417), “research data is the data which is generated when the researchers undertake or execute any research activity or project”. When handled properly, research data produced during research operations can be preserved, accessed, and used again for later scientific purposes. Data may be in any one or more of the following formats: text, images, sound recordings, music compositions, spoken word, simulations, visual art, computer code, and experimental readings (Carlson et al., 2015).

To properly comprehend research data management (RDM), it is important to understand that RDM is an acronym that encompasses a variety of data management tasks. Processing data, disseminating data, and long-term preservation of research data are all examples of these tasks. Knight (2015:431) elaborates that “research data management is increasingly perceived as a key component of good research practice that should be considered early to ensure opportunities are identified, risks are mitigated and appropriate resources are allocated before work commences”. Donner (2022) confirmed this viewpoint by pointing out that “an RDM strategy lays the foundation for policies, guidelines as well as the allocation of resources”. This viewpoint is used as a measuring criterion for the Relevance quality dimension.

The literature contains various definitions of RDM and a summary of these is provided here as foundational to the study. This summary highlights similarities and differences between these definitions, as the concept of RDM can be confusing for both researchers and academics. To provide a

general overview, an initial description and classification of RDM-related topics is provided. According to Sanjeeva (2018:3), research data management “put simply, refers to the effective handling of data that is created in the course of research. Managing research data is usually an integral part of the research process and extends over the entire lifecycle of the data”. This explanation demonstrates the primary objectives of RDM, which are direct and concise. RDM encompasses all duties and procedures that enable proper research data recording, coordination, storage, archiving, and curation (Domingus, 2015). To maximise the quality of the final data, RDM processes are used throughout the data lifecycle. To ensure that data for secondary use is understandable and practical, data processing needs the application of realistic and reliable methodologies (Domingus, 2015). For the Comparability quality dimension, the viability and user understanding data are employed as measuring criteria. Van den Eynden (2011:1) views RDM as “a practice that allows reliable verification of results and permits new and innovative research built on existing information”; while the University of Wollongong Australia (2017:7) explains that “research data must be carefully managed throughout the data lifecycle to enable it to be preserved and accessible by a controlled audience for current and future research”. RDM enhances the research process by facilitating the documentation and collection of meaningful data. It contributes to knowledge growth by providing a solid foundation for research. RDM includes data classification, setup, documentation, replication and storage, data preparation for assessment, sharing, and long-term data storage. (Sanjeeva, 2018). For the comparability quality dimension, the development of data documentation is employed as a criterion.

According to Whyte and Tedds (2011:2), a thorough description of RDM may be found in their earlier description (2011:2) stating that research data management “concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information”. The description, in its complete form, explains how data is managed from generation to archiving, focusing on various activities Research Data Management (RDM) ensures consistent checking of results, a crucial aspect of this description. Verification of research findings goes beyond reproducing the same outcomes; it also involves assessing the reliability and authenticity of research findings. Measures taken to validate study findings are also important since they must follow best practices (Whyte & Tedds, 2011); besides validating study findings McMaster University (2023) provided more aspects of using RDM, “applying good RDM practices improves the efficiency and impact of research, increases research visibility, facilitates collaboration, protects intellectual property, enables reuse and verification of research results, and supports a culture of reproducibility”. RDM is a feasible method based on established standards and guidelines, and it involves making decisions about data transfer and maintenance post-project, such as storing data in a repository for long-term preservation and access (University of Leicester, 2017). For the Accessibility quality dimension, the type of repository where data will be stored is employed as a criterion.

In the context of this study, the goals of RDM include the following:

- *“Enhances the visibility of research data and increases the number of citations for the research”* (University of Sheffield, 2019). Data that is visible can be retrieved, used, and cited; visibility makes data more accessible. Therefore, this RDM goal serves as the foundation for the assessment criteria for the Accessibility quality dimension.
- *“Provides opportunities for collaboration with other researchers within a discipline, or even across disciplines”* (University of Sheffield, 2019). It is necessary to assess data from different disciplines to determine its applicability in a collaborative endeavour. This can be achieved by assessing the precision and calibre of the data. Hence, this RDM objective serves as a measure for the Relevance quality dimension.
- *“Supports open data, which is a movement that advocates for researchers to share their data in order to boost the development of knowledge”* (York University, 2018). For the Timeliness quality dimension, the manner in which research collaborations are facilitated is employed as a measuring criterion.
- *“Helps to make research collaboration between institutions, groups and individuals more efficient by reducing duplication of effort and avoiding data loss”* (Kahn et al. 2014:297); The University of California (2023) subsequently clarified that “Data management planning improves efficiency in research, ensures information is protected, allows results to be checked by others, improves reproducibility, improves exposure via sharing, and allows compliance with funders' policies”. RDM is a starting point for improving collaboration across institutions, groups, and individuals, as well as preventing research duplication and data loss. As a result, this RDM objective is employed as a measuring criterion for the Methodological soundness quality dimension.
- *“Safeguards research data by establishing appropriate storage, back-up and management procedures”* (York University, 2018). The Interpretability quality dimension is measured by establishing research methodologies for data management procedures.

RDM simplifies the distribution and preservation of research data, ensuring proper recording, coordination, and processing before and during a research project. It also ensures that data can be recovered and used in the future. RDM is concerned with data preparation before the start of a research endeavour, ensuring proper recording, formatting, and processing. However, it requires a results review before self-describing data sets can be formed. These functions overlap due to the validation aspect provided by RDM. Data preservation is another key activity when managing research data. The RDM definitions of data preservation, which include archiving data and guaranteeing long-term data preservation, differ dramatically from the idea of data preservation. Although they are frequently used

synonymously, these words don't always have the same meaning. It is not acceptable to use the terms "archiving", "storage" and "preservation" interchangeably.

Archiving refers to the very basics of placing records in storage, generally to preserve them indefinitely or for a finite or indefinite period of time as proof of a historical or legal record (Reitz, 2002). The definition of archiving is also explained by The European Commission (2023:2) as “the appropriate measures taken to ensure that the information becomes, and stays, sustainably accessible”. Storage means placing the data in or leaving the data at a location (such as a warehouse or library, or in computer memory) for use but with no active management activities (Merriam-Webster, 2020); preservation refers to the act of maintaining anything or preventing it from being destroyed for an extended period of time in its original or current state (Cambridge University Press, 2020).

RDM (by its very nature) represents the common goal of delivering adequate research data management. RDM's main objective is the creation of high-quality data (Domingus, 2015), which is the focus of this study. The scope of methodological dependability, whether the data is correct and whether it is truthful, determines data quality (Clarke, 2014); this viewpoint is utilised as a measurement criterion for the Relevance quality dimension. RDM is crucial for maintaining appropriate management practices and ensuring high-quality data. When RDM and data quality are combined, a "united front" perspective is revealed, and a practical data management plan is created, based on best-practice standards and guidelines. This study focuses on the use of Research Data Management (RDM) to support data generation and management. RDM aims to improve research efficiency, produce error-free data, and meet user demands by utilising best practices and ensuring high-quality research outputs.

Policies are implemented only if they are declared and adopted inside an organisation. A critical examination of the RDM policy process is necessary to situate RDM within the appropriate institutional framework. The next section discusses the RDM policy framework.

To begin the process of changing RDM in an institution, it is crucial to “strengthen the institution’s policy framework to ensure that research data management requirements are addressed; enhance infrastructure to support data management activities; and improve data management practice among researchers within the institution” (Knight, 2015:426). Prior to creating quality assurance procedures, it is imperative that researchers comprehend the importance of RDM and how to use it. This is due to the fact that RDM influences how researchers see data quality and assurance, so any modifications to RDM should take this into consideration. An RDM policy should be created and then coordinated with the content of other organisational strategy documents (such as an IT policy) to prevent disputes and needless issues from arising. (Bloom et al., 2014).

According to Hodson and Molloy (2014), an RDM policy regulates the management of research data at an institution and ensures that all the research data produced are effectively and efficiently controlled and curated to support the institution's research goals. Youngson-Baines and Lee (2023:2) confirm that

an RDM Policy “help promote good practice around research data, with note to ensure that it is stored securely and preserved in order to ensure long term accessibility and utilisation, this includes ensuring that it remains identifiable and retains contextual information alongside the raw data”. The RDM policy framework is crucial for managing data management operations, outlining procedures and ensuring a unified and understandable context. It should include concepts related to quality control and data management solutions:

- Openness: Publicly funded research data is a public asset that is produced in the public interest and should be freely available in a timely and open manner, with as few restrictions as possible (Hodson & Molloy, 2014). Important data has to be kept up to date, preserved, and remain easily available for upcoming studies.
- To ensure that others may readily collect and retrieve data, adequate metadata should be registered and explicitly made available so that other researchers can enjoy the capabilities of the data with respect to analysis and reuse (Hodson & Molloy, 2014). For the Interpretability quality dimension, the notion of retrievability is applied as a measuring criterion.
- Data quality: An RDM Policy Structure is a high-level document that specifies how all RDM-related actions should be carried out. To tackle quality assurance and data quality challenges, it is vital that the writing process of the RDM policy structure be thorough and that it be written according to best RDM techniques and recommendations.

Before staff members gladly adhere to the policy, they must be persuaded of the inherent value of RDM operations. The advantages of RDM are discussed in the next section.

2.2.2 Benefits of RDM

It is important to remember that RDM offers advantages that support data analysis in addition to streamlining procedures across the research data lifecycle. The ways in which RDM initiatives benefit and promote knowledge growth, transparent scientific research, and long-term data preservation – as identified in this study – are detailed below.

Advances knowledge

Researchers' RDM efforts should be guided by recognised data management concepts and guidelines, with understanding the data and its context being the key rationale for RDM. According to Domingus (2015), Peters (2016), Chiware and Mathe (2015), it is possible to uncover fresh interpretations and discoveries in this way, allowing for the generation of new knowledge. This understanding was later backed-up by Algonquin College (2023:4) in that “RDM, founded on equity and diversity, is an essential part of research excellence that advances knowledge mobilization and strengthens economic and social impact” However, the data must be of good quality for secondary usage.

Transparency of scientific research

One of the main reasons for RDM is that it ensures accountability during the verification process. Domingus (2015:1) clarifies further by saying “documenting key aspects during the whole research process provides transparency, for instance for reproduction and validation purposes and even in cases where questions have arisen with regards to scientific integrity”. The Interpretability quality dimension uses ‘documentation of crucial components of the research process’ as a measuring criterion. Transparency is vital in scientific research as it enables secondary users to examine the criteria and techniques used by the researcher. Edge Hill University (2013) stated that “transparent and thorough data management are crucial components of responsible research where open access data activities are encouraged. This is verified in addition by McMaster University (2023) “Good management and stewardship of research data supports research excellence by improving efficiency and integrity, enabling new types of exploration, and supporting research transparency and reproducibility”. If data is not open access, there should be clarity and responsibility about the conditions controlling data access and reuse (Australian National Data Service, 2015). A researcher is expected to maintain high ethical standards in order to ensure that research is transparent. Professional groups, host universities, and funding organisations typically set ethical criteria for research (Norwegian National Research Ethics Committees, 2019). Researchers should obtain ethical guidelines from their institutions to avoid breaches of ethical research conduct. This component is used as a criterion for the Quality Check Processes quality dimension. Scientific research openness simplifies data availability to the public for verification and validation (Coulehan & Wells, 2012); according to Kabanda et al. (2023:1) “the open science movement has the potential to revolutionise scientific research and improve its transparency and potential for collaboration” which is in line with this point of view of scientific research.

Ensures the long-term preservation of data

The Preservation phase is a crucial aspect of the research data lifecycle, ensuring long-term benefits and ensuring the longevity of the data (Flores et al., 2015). Preserving data entails protecting it from destruction, tampering, or obsolete technology. According to the Texas State Library and Archives Commission (2023) “preserving electronic records in a distributed manner helps reduce potential technical threats faced by digital materials and ensures longevity”.

The importance of adequate documentation in promoting data retention is continually emphasised by researchers (UK Data Archive, 2011), this is also confirmed by the MRC Regulatory Support Centre (2022:3) “it should be easy for those authorised to access data and/or samples, to be able find and access what they need (i.e. via meta-data or data documentation for electronic format or an inventory or database for samples or other physical records)”. For the Timeliness quality dimension, the preparation of preservation documentation is employed as a criterion. A vital aspect of performing research is being able to obtain stored data promptly.

'Preservation' can also refer to a process used to address hardware and software problems. RDM preservation sometimes requires a combination of multiple controlled and ongoing operations to guarantee long-term access to digital content (Decman & Vintar, 2013). Organisations should incorporate activities to prevent file obsolescence and degradation, handle format issues, and enable future access to authentic information when preserving research data (RDM), as this effectively maintains data quality (Decman & Vintar, 2013). To be able to create “a digital preservation programme, a solid understanding of corporate drivers is required. This may include collection development, use access, information reuse, legal and regulatory compliance, and efficiencies and savings” (Pinnick, 2017:177); Karimova (2023:10) also mentioned that “the best practice accepted by the scientific community is to use recommended, open, standard, and common file formats for preservation”.

Researchers should be aware of the benefits of excellent RDM, but they should also be aware of the general issues involved with research data management. The difficulties are explored in the next section.

2.2.3 Generic challenges to managing research data

Managing an organisation well has a number of unique challenges. The organisation's human, technological, and infrastructure resources, as well as the management of the research data itself, all have an impact on the use of best practices. According to Abiteboul et al. (2016:7) “...the landscape of available data has dramatically changed - while the huge amount of available data is perceived as a clear asset, exploiting this data meets many challenges”. Understanding and addressing the challenges in managing research data is crucial to developing effective solutions. Common issues include researcher behaviour, copyright, data licensing, data security, and ensuring quality for secondary use. These areas are discussed below:

2.2.3.1 Researcher behaviour

Researcher behaviour is influenced by human motivations, activities, psychological processes and interactions and all of these have an effect on how a researcher carries out research tasks (University of Southern California, 2016). Unal et al. (2019:2) provided a different perspective regarding data sharing and research behaviour, according to the authors “data sharing is regarded as a good behaviour, it appears that researchers invest little time in metadata and documentation rendering a great deal of data undiscoverable”. Overcoming old habits and adopting new methods can be challenging, especially with respect to research data management techniques. Possession and inability to adapt to research management procedures are significant challenges. Researchers invest significant effort, and time to analyse data in conducting research, making it difficult for them to agree that their data can be used for secondary purposes. Accepting this fact, there needs to be “a major shift in the perception of carrying out research and sharing of scholarly information” (Patel, 2016:228).

One of the primary objectives of this research (the overall goal being to improve quality assurance and the quality of data) is to determine ways to influence the behaviour of researchers. Anderson (2014) argues that, in order for people to develop and maintain progressive work habits, they must be motivated to do so. This study proposes the implementation of innovative, efficient, and quality-focused RDM solutions as a framework, highlighting the challenging task of influencing researchers to behave differently and examining various areas to achieve this goal. Understanding the numerous factors that influence and maintain behavioural improvement is a difficult but necessary task, and regardless of what people state they want to change, procrastination and acceptance of the status quo are widespread (Anderson, 2014). When it comes to incorporating new practical approaches into people's daily routines, it can be incredibly difficult to get them to 'walk the walk'.

Witte (2015) developed the *Key Elements of Behaviour Change* (see Table 2.1) to understand the full complexity of behaviour change. These factors are beneficial since they clarify the variables that are required to comprehend behaviour change.

Table 2.1 - Key Elements of Behaviour Change (Witte, 2015)

Key element	Definition	Strategies for behaviour change
Threat	A hazardous event or anything that may or may not be disclosed to individuals	Focus on the essence of the threat and how people are vulnerable to it and cultivate awareness of the threat.
Fear	Emotional excitement induced by a person's understanding of a serious and personal danger	Fear can significantly influence behaviour, and if it is effectively guided, it can motivate people to seek knowledge. It can also, however, lead individuals to dispute the possibility that they are at risk.
Effectiveness of response	The principle that a proposed reaction works against the risk	Providing examples to show that the suggested reaction would escape the danger
Self-confidence	An individual's feeling that he/she should behave according to a suggested response	Growing people's confidence so that they can respond to the danger and help prevent it.

The shift in the organisational approach to change management and how it affects researchers' actions is discussed in greater detail in Section 2.11. The issue of copyright is dealt with in 2.2.3.2, below.

2.2.3.2 Copyright

Researchers often consider the information (findings) they produce to be their own intellectual property. However, this is not always the case. According to Coulehan and Wells (2012:3) “understanding data ownership, who can possess data, and who can publish books or articles about it are often complicated issues, related to questions of project funding, affiliations, and the sources and forms of the research itself”. The way data ownership and copyright are viewed ensures that data ownership and distribution are both possible. A research contract between the researcher and the funder should include this criterion. This aspect of copyright is used as a measuring criterion for the Methodological Soundness quality dimension. Moreover, Patel (2016:227) points out that “the question of copyright in the data collected during a research project is complex because it involves several stakeholders, such as the researchers, data collectors, data analysts, institution or university, and the funding agencies”. This aspect of copyright is used as a measuring criterion for the Standards and Policies quality dimension. Essentially, determining who owns the copyright to the data can be a complex and time-consuming task if not addressed at the start of a research project.

According to the ERC Scientific Council (2018), research data should adhere to the FAIR principles. The ERC Scientific Council (2018) stated that to be FAIR, research data should be findable, accessible, interoperable and reusable (collectively known as the FAIR principles). This viewpoint is confirmed by Sananikone et al. (2022) “Data quality and documentation are at the core of the FAIR (Findable, Accessible, Interoperable, Reusable) principles”. Essentially, data should be:

- “*Identified in a persistent manner using community conventions, and described using sufficiently rich meta-data*”. This principle is used as a measuring criterion for the Completeness quality dimension.
- “*Stored in such a way that they can be accessed by humans and machines*”.
- “*Structured in such a way that they can be combined with other data sets*”.
- “*Licensed or have terms-of-use that spell out how they can be used by others*” (ERC Scientific Council, 2018); this principle is used as a measuring criterion for the Accessibility quality dimension.

Many stakeholders (including academics, data analysts, and data collectors) add to the dataset in different ways before a project is completed. All phases of the data lifecycle are covered by these activities. Because researchers who adhere to the FAIR principles consider re-use from the beginning, the principles address copyright. This means that data citation – rather than ownership – is the problem. The larger the research team, the more sense it makes to use FAIR as the copyright default. Since not all data are created equal, it may be preferable in some circumstances to make data FAIR but not fully reusable. The next matter to resolve would be the issue of data licensing.

2.2.3.3 *Data licensing*

"Open data" is another area that has an impact on research data accessibility. According to Suber (2009), open data are those that are digital, downloadable, cost-free, and unrestricted by copyright or licensing. There are no restrictions on access to this kind of data and it is, therefore, readily available. The majority of open data uses Creative Commons licences since it is difficult to apply restrictive licences to it due to its transparency and public availability. A group of copyright-based licences, known as the Creative Commons (CC) licenses, control how creative works are distributed and reused (Friesike, 2017; Hagedorn et al., 2011; Vetro et al., 2016). Patel (2016:227) provides an overview of data licensing, "the issues that need to be addressed are the terms of use – restrictions on use, redistribution of data, permissions, terms and conditions for derivative works"; this definition serves as a criterion for the Accessibility quality dimension. The responses to these factors will inform data policy and practice when dealing with data licensing concerns (Borgman, 2015). Patel (2016) provided an overview of data licensing, in a more recent publication Margoni and Schirru (2023:15) explained that "when "data" is protected by copyright and/or SGDR (Sui Generis Database Rights), there are two main avenues to reusability: (i) reuse within the scope of statutory permissions, e.g., exceptions and limitations to copyright provided by law; or (ii) reuse within the scope of a contractual permission, e.g., copyright licenses" The method according to which data licensing issues are handled is utilised as a measuring criterion for the Methodological Soundness quality dimension. Scientific data from a variety of sources might be merged and utilised for unintended reasons if the correct data license is not in place.

RDM process helps custodians "to identify the various forms of data licensing before the research project starts, also keeping in mind the anticipated quality of the data" (Patel, 2016:227). Resolving data licensing issues provides clear knowledge for organisations and researchers about data sharing types, sharing permissions, and sharing conditions. Without data security, even the use of all licences combined may not guarantee data access and usage, highlighting the importance of understanding data sharing. Leading on from this, the next section briefly discusses the data security issue.

2.2.3.4 *Data security*

Researchers must comprehend the importance of data security. In order to accomplish this, "an organisation should have a well-defined and well-documented policy for research data security" (Patel, 2016:227). The success of RDM guarantees that policies address the issue of data security. A security policy usually lays out the procedures that must be taken to protect research data. Unethical access to and use of research data is possible if the data is not properly secured. Patel (2016:228) states that "serious or casual hackers may try to breach the security of servers – in such a situation, the security of data is a crucial issue which needs to be addressed". Farid et al. (2023) later reported that "some academic libraries have developed DISM policies on data protection, data backup, information security (IS) systems, the development of hardware and software, the training of library staff, data protection from malware and social engineering, and data security and privacy". An entity must have unique data

protection policies, based on organisational resources and the technological infrastructure, to protect research data at all times (JISC, 2017). When it comes to data protection, masking, data encoding, tokenisation, and access controls have to be considered. When looking at data security techniques from a technological and realistic standpoint, the most commonly utilised strategies include access controls, data encryption, passwords, and masking (CESSDA/ERIC, 2017). The accessibility and licensing of a data source are used as a criterion for the Accessibility quality dimension. According to JISC (2017), in order to bolster resistance to security breaches, companies must evaluate their sensitive data risk, which includes where the data is held, how it is used, when it is used, its relevance, and how it is safeguarded. RDM provides researchers with a comprehensive understanding of data security threats due to its strict control over the analysis process and the guided planning function, enabling them to prepare for potential security concerns.

2.2.3.5 Re-use of research data

Sarpawari (2014) emphasises the importance of data reuse in science, with such reuse requiring a thorough understanding of the data as well as the use of relevant metadata and supplementary materials to make research data understandable. Schneider et al. (2023:73) explored the concept of data reuse from the perspective of data mining; “a data scientist should structure computation on data so that results of identical operations on the same data during preparation and modelling are reused and are thus only executed once”. According to Stiles and Boothroyd (2015:6) “ethical issues associated with the growth in electronic data are many”; hence, ethical secondary data use requires specialised knowledge and skills. These skills ought to be encouraged at the institutional level (via training initiatives). Some of the ethical competencies and abilities needed when dealing with data are problem-solving, analysis and synthesis (data ethics), stakeholder relationship management, applied understanding of social sciences, data communication skills, and facilitating decisions and risks (Rajhans, 2018).

Research data can be shared and made available for secondary uses when RDM is applied effectively. However, data management does not ensure high-quality data. Primary data is information obtained for a specific purpose by the researcher (UK Essays, 2017). Primary data is distinct in that it is information that has been gathered for the first time and has not been made available to others. Secondary data refers to information that has been released for re-use: “secondary data is data that was collected by someone else for another primary purpose. The utilisation of this existing data provides a viable option for researchers who may have limited time and resources” (Johnston, 2014:619). Secondary data is easier to come by and less expensive to obtain than primary data. The requirement for secondary data is now more important than ever. Andrews et al. (2012:619) make the point that “vast amounts of data are being collected and archived by researchers all over the world, and the practicality of utilising existing data for research is becoming more prevalent”; while Ewim et al. (2023:6) reported that “the quality and relevance of the secondary data analysis are contingent on the availability of suitable data sources”. The ability to manage primary data efficiently, according to Johnston (2014:623), is the most important

factor because “the secondary analyst researcher to obtain all documentation of the processes and protocols followed by the primary researchers, including the questionnaire, all coding materials, and any publications that are related to the data”. A secondary user may have access to a raw, accurate data set in order to do new analyses and investigate new options, but without documentation that explains the context and purpose of the original research and/or how a dataset was acquired, the data may be of limited use (Johnston, 2014). RDM has been adopted by institutions, NGOs, research organisations, and big international organisations as part of their business processes. Several programmes have been established to assist data custodians in developing their abilities. The next section considers two of these (MANTRA and RDMRose).

2.2.3.6 Skills development

The manner in which RDM is implemented affects its assessment and quality assurance. This study explores the RDM practices of the researchers and data curators taking part so as to enhance their RDM proficiency and practical training needs. It considers the quality of training materials as a criterion for evaluating data product quality. Academic institutions have recognised the need for RDM education, leading to the creation of training programmes. Two of these training courses and how they approach quality control are considered below:

2.2.3.7 MANTRA

MANTRA is a free, unassessed, online course with guidance to help users understand and learn how to manage the digital data they collect during their research (MANTRA, 2017). This online course aims to assist future researchers and early career researchers in matters of data management (Rice et al., 2011) and consists of eight units, as detailed below:

1. “Understand the nature of research data in a variety of disciplinary settings
2. Create a data management plan and apply it from the start to the finish of your research project – this unit was used as a measuring criterion for the Consistency quality dimension.
3. Name, organise, and version your data files effectively – this unit was acknowledged as a measuring criterion for the Consistency quality dimension.
4. Gain familiarity with different kinds of data formats and know how and when to transform your data – this unit was acknowledged as a measuring criterion for the Integrity quality dimension.
5. Document your data well for yourself and others, learn about metadata standards and cite data properly – this unit was acknowledged as a measuring criterion for the Quality Check Processes quality dimension.
6. Know how to store and transport your data safely and securely (backup and encryption)
7. Understand legal and ethical requirements for managing data about human subjects; manage intellectual property rights – this unit was acknowledged as a measuring criterion for the Methodological Soundness quality dimension.

8. Understand the benefits of sharing, preserving and licensing data for re-use; this unit was acknowledged as a measuring criterion for the Standards and Policies quality dimension.
9. Improve your data handling skills in one of four software environments – this unit was acknowledged as a measuring criterion for the Integrity quality dimension” (MANTRA, 2017).

The MANTRA units, as described above, offer a useful framework for directing researchers toward efficient data management techniques.

2.2.3.8 RDM Rose – RDM Training for Librarians

Librarians have always been regarded as the principal stewards of information and it seems to be a logical further step that they would provide researchers with research data management services as well. As this culture has developed, RDM tasks and responsibilities have been assigned to members of the data management team. This has led to RDM training for librarians being critical, as librarians are uniquely qualified to assist researchers with data management resources. The UK’s Joint Information Systems Committee (JISC) embarked on a project to develop librarians’ RDM skills and know-how (Guy, 2013). The RDMRose JISC Managing Research Data Project was funded by JISC “to produce taught and continuing professional development (CPD) learning materials in RDM tailored for information professionals. It is looking in particular at the specific needs of liaison librarians in university libraries, and deliverables” (Guy, 2013:1). The key success factors of this project included:

- Teaching about the method of analysis
- RDM needs a hands-on plan
- Encourage librarians to see training in RDM as career growth
- Teaching the best abilities to the right individuals
- Receive reviews (Guy, 2013).

Direct contact strategies are used in the RDM training curricula outlined above to educate librarians with basic to advanced research management abilities. There is no evidence in the literature that the issue of data quality is addressed as part of the training. As a result, the problem of a lack of skills is intensified by the lack of data quality as a training issue.

The preceding sections were written to provide a good understanding of RDM and data management training, as well as the benefits and drawbacks of RDM. From here onwards, the focus will be on data quality and how it may be incorporated into a social science-specific quality-focused RDM framework.

2.3 Data quality as a theoretical concept

The process of Research Data Management (RDM) is not given enough attention. Although there is discussion about quality at certain stages, there is disagreement about how to handle quality across the

whole range of RDM activities. To address this, RDM quality assurance initiatives should be implemented throughout the data lifecycle, especially during data collection and processing. Though various RDM tools, evaluation frameworks, guidelines, and best practices exist, they are not all applicable to the research data lifecycle and data quality. Even though enterprises all over the world have realised how important it is to improve RDM services, RDM is still relatively new in South Africa. Thus, it would be advantageous to implement a quality-focused RDM system (as a starting point) to help resolve the problems that researchers and institutions are currently facing. The purpose of this research is to develop a new RDM quality-based framework by applying guidelines. The foundation of this study is data quality as a concept. Curating data that has not been subjected to quality assurance procedures should be considered a waste of time and money. It is widely acknowledged that best practice guidelines must be produced to guide the research process and produce high-quality data. It is therefore critical to provide a thorough explanation and definition of the concept in the context of this research. Data quality is multifaceted. The way academic institutions use the term 'quality', and what it means in an institutional context, is frequently determined by past research project accomplishments. In line with this, Cooper and Kahn (2012) claim that an institution's ability to define the term 'quality' (data-based) depends on various factors. The authors explain that:

“Quality is to be described on the basis of data age, source and collection method; data representativeness (i.e. the percentage of total production represented by a sample), averaging methods, and approaches for protecting competition-sensitive company-specific information; methods used to estimate or justify missing data; and information for replicating key assumptions or methodological choices” (Cooper & Kahn, 2012:500). Gutema (2023) elaborated the scope of strengthening data quality through data management, stating that “data management focuses on maximizing the quality of data whereas data governance focuses on enhancing the value of organizational data assets” (Gutema, 2023).

Data representativeness is employed as a criterion for the Uniqueness quality dimension since the entire production represented by a sample must be accurately reflected in a data file (as the number of cases and variables). In the suggested quality process framework, the bulk of the elements listed above were employed as measuring criteria, including:

- “*Source and collection method*” (Cooper & Kahn, 2012) - For the data to be trustworthy, the source and collection technique must be credible; the statement "both reliability and validity are crucial considerations in research to ensure the trustworthiness and meaningfulness of the collected data and measurements" from Simplilearn (2023) reinforces this point of view. The Completeness quality dimension was measured using this component as a criterion.
- “*Data representativeness*” - Rudolph et al. (2023) define “a study sample to be representative of a well-defined target population if the results estimated in that sample are generalisable to

the target population”. Data representative was employed as a criterion for the Validity quality dimension.

- “*Methods used to estimate or justify missing data*” (Cooper & Kahn, 2012) – this factor was used as a measuring criterion for the Validity quality dimension. The phrase “missing completely at random” was used by Woods et al. (2023) and refers to the situation when “the likelihood of any given data point being missing is the same across all data points and unrelated to any other measured or unmeasured variables” (Woods et al., 2023:3).

In light of the diverse perspectives and definitions used by international RDM stakeholders, RDM programmes, major research funders, evaluation systems, and research methodology books, it appears that people who generate data (primarily academics) are the ones who decide on data quality. A researcher has to have the required “expertise to address technical data and metadata quality and to ensure that sufficient information is available for end users to make quality-related evaluations” (Doorn & Dillo, 2017:12). As a build-up on the expertise needed to address technical data and metadata quality Fischer et al. (2023:1510) explained that “readily accessible information on data quality and provenance is often missing or hardly accessible for users due to technical or perceptual barriers, for example, due to unstructured metadata information or missing references”. The researcher who understands data quality and how to maintain it not only specifies data quality but also delivers a high-quality prospectus for the secondary consumer (Andrews et al., 2012). According to Cai and Zhu (2015), the principles of data quality are:

- Data are a commodity with clients to whom both cost and value are available;
- Data are a commodity for clients that have both price and value for them;
- The quality of the data depends on many factors, including (at least) the reason for which the data is used, the user and the time of use.

Regarding data quality principles, Hassenstein and Vanella (2022) subsequently determined a technical set of data quality principles that are included in ISO/IEC 25012:2008(E) standards for software quality. Differentiating between inherent and system-dependent data quality, two broad kinds of data quality. As stated by the authors “inherent characteristics represent those with an intrinsic potential to meet the requirements, whereas the system-dependent characteristics suggest those conditional on the IT system used to manage the data” (Hassenstein & Vanella, 2022:502).

Data quality has been defined by a number of writers. The next section will provide a selection of such writers’ descriptions of data quality, as well as a definition of data quality as it relates to this research study.

2.3.1 Defining data quality

As previously stated, the term "data quality" has been broadly defined. Wang et al. (1995:623) state that "there are many approaches in the literature that can be applied to studying data quality". Scannapieco et al. (2005:7) define data quality as "a multi-faceted concept, to the definition of which different dimensions concur". Measurement criteria that inform data management operations performed by a researcher during research are included in the dimensions mentioned above. According to McKnight (2010:3) "data quality is not the absence of defects; it is the absence of intolerable defects". Data quality is a complicated notion with many characteristics that all contribute to a better understanding of the term. Cai and Zhu (2015) define 'data quality' as 'fitness for use'. Hassenstein and Vanella (2022) provided a more detailed definition of data quality, which is "Data Quality is, in essence, understood as the degree to which the data of interest satisfies the requirements, is free of flaws, and is suited for the intended purpose. Data Quality is usually measured utilising several criteria, which may differ in terms of assigned importance, depending on, e.g., the data at hand, stakeholders, or the intended use".

Federations such as the International Organization for Standardization have also provided comprehensive definitions of the data quality principle (ISO). The ISO 8000 standard is important for this study because "ISO 8000 provides a framework for improving data quality that can be used independently or in conjunction with the quality management system" (ISO, 2018:6). The ISO standards overlap and support one another, and the ISO 8000 structure can be utilised in conjunction with or independently of quality management system specifications such as ISO 9000, and ISO 9001 (ISO, 2018). The ISO 8000 standard is discussed in more depth in Section 2.7.1. The ISO 9000 definitions contain five primary reference terms that can be used to construct a data quality framework; some of them (as specified) were employed as measuring criteria in this study, and they are as follows:

- A data quality policy refers to an entity's overall goal and approach when it comes to concerns about the quality of data products – this reference term is a criterion for assessing the Methodological Soundness quality dimension.
- Data quality management is the management role that defines and implements the data quality policy – this term of reference is a criterion for the Consistency quality dimension.
- A data quality system includes the organisational structure, roles, policies, processes, and tools for implementing data quality management – this term of reference is employed as a measuring criterion for the Methodological Soundness quality dimension.
- The collection of organisational methods and operations used to achieve the required quality for a data product is known as data quality control – this term of reference is a measuring criterion for the Methodological Soundness quality dimension.
- Data quality assurance encompasses all of the planned and comprehensive steps required to ensure that a data product meets a specific set of quality specifications – thus, ensuring that

data meets certain quality specifications is employed as a measuring criterion for the Accuracy quality dimension.

The five terms of reference emphasise the importance of an organisational strategy for efficient data management in research. They serve as benchmarks for researchers to assess their approach to ISO's quality assurance areas. One of the terms of reference is the ability to create, gather, store, preserve, distribute, process, and present data in a timely and economical manner (ISO, 2018). The concept of data quality, which forms the basis for all the other definitions in this section, is the extent to which data is reliable, accurate, comprehensive, consistent, and timely. Other terminology that may be used in conjunction with the data quality debate include:

- **Quality Control:** When managing social sciences data, quality control refers to “methods and procedures implemented to ensure that data are collected, managed, and utilised with accuracy and precision” (Reisch & Webster, 2004:2). To a large extent, the manner in which data is managed is influenced by organisational procedures that are in place, hence this definition is used as a measuring criterion for the Methodological Soundness quality dimension. Quality Control (QC) “is a system of routine technical activities, to measure and control the quality of the inventory as it is being developed” (IPCC, 1997:6); Simplilearn (2023) defines quality control as “the process by which products/services are tested and measured to ensure they meet a standard. Through this process, a business can evaluate, maintain, and improve product quality. The primary objective of Quality Control is to identify and correct any deviations from the established quality standards”.
- **Data quality Measures:** If the data quality is appropriate for the purpose in question, a set of protocols that give a template and methods of measurement (Canadian Institute for Health Information, 2005); data quality is generally measured “against a set of criteria called 'data quality dimensions' that assess the health of the data, such as completeness, or uniqueness” (Curtis, 2021).
- **Data quality Rules (DQR):** adequate, organisation-specific definitions of data quality (Palmer, 2011); A Data quality Rule “is a constraint used to validate data values or records” (Black, and van Nederpelt, 2020). The definition of data quality rules is also given by Stedman and Vaughan (2023) as “specify required quality levels in data sets and detail what different data elements need to include so they can be checked for accuracy, consistency and other data quality attributes”. The DQR may have negative effects with respect to time, secondary users, and outcomes if they are not followed.

The term 'data quality' in this study refers to 'fitness for use,' as defined by various quality measurements in the form of best practices, guidelines, and standards that are correlated with the method used to measure or improve data quality. In essence, fitness for use means that "a high-quality product does

what its customers want in such a way that they use it" (Cohn, 2022). Data quality is a crucial aspect of research data, aiming to ensure high-quality and precise data for use. It is differentiated between scientific and technical data quality, providing context and quantifying the degree needed to aid in its development. Understanding the difference is essential for effective data quality management.

2.3.2 Scientific quality vs technical quality of data

The phrase 'data quality' has two distinct meanings: technical quality and scientific quality. Technical quality is linked to data management, but scientific quality judgments can be formed when quality indicators are reviewed, the researcher's competency is questioned, or additional scientific consequences are investigated. As a result, scientific quality judgments are made in the context of influencing factors, and the criteria for evaluation periodically alter, depending on which aspect is emphasised (Azimae, 2015). RDM interventions directly produce technical quality, as do the actions a researcher takes before commencing on the study journey – these processes include: “writing DMPs and support for various data management components such as creating metadata or choosing appropriate data repositories” (Rolando et al., 2015:43). The role of RDM was also explained by Zhou et al. (2023:4) “RDM requires awareness of technical, ethical, and legal issues as well as government policies regarding the sharing of research data”. Researchers’ experience, behaviour, and tacit knowledge significantly influence scientific knowledge discovery. The analytical quality of data is influenced by RDM techniques, while even poor technical data can still be scientifically significant. Table 2.2, below lists potential data errors that may occur during a research endeavour, along with the category (technical or scientific) to which the error belongs:

Table 2.2 - Data errors that may occur during a research project (Chapman, 2005)

Data error	Technical quality	Scientific quality
Data capture and recording at the time of gathering		X
Data manipulation prior to digitisation (label preparation, copying of data to a ledger, etc.)		X
Identification of the collection and its recording		X
Digitisation of the data	X	
Documentation of the data (capturing and recording the metadata)		X
Data storage and archiving		X

Data presentation and dissemination (paper and electronic publications, web-enabled databases, etc.)	X	
Using the data (analysis and manipulation)		X

A researcher makes the final choice on all data processing issues, which has an impact on the data's accuracy. The World Data System's (WDS) concept of data sharing supports this underlying information. The WDS data-sharing principle assigns responsibility for data quality to the individual who creates, manages, transfers, and uses data and metadata; this typically refers to a researcher working on a project (WDS, 2015). According to Jayawardene et al. (2016:2) “in light of the management axiom ‘what gets measured gets managed’ dimensions of data quality signify a crucial management element in the domain of data quality”. The University of California (2023) confirmed that “regular measurement and reporting keeps you focused, because you use that information to make decisions to improve your results”. As a result, it is critical to comprehend the impact of quality dimensions on data quality. The following section delves deeper into the various dimensions of quality.

2.3.3 Data quality dimensions

Wissik and Durco (2015:94) point out that “sharing and reuse of research data is not yet always an integral part of good research practice and that researchers are not familiar with data management plans etc.”. This effectively means that when it comes to producing and maintaining data, researchers are inexperienced with the concept of quality assurance. In the absence of a data quality system for social sciences, it was necessary to examine frameworks and literature from other domains in order to obtain some alignment (Anderson, 2014, Bui & Park, 2013; Chapman, 2005; Gassman et al., 2015; Hugo, 2012). The data management literature yielded a number of data quality dimensions (DAMA UK Working Group, 2013; Gassman et al., 2015; ICPSR, 2018; Jayawardene et al., 2016; Oracle, 2009; South African Statistical Quality Assessment Framework (SASQAF), 2010; UK Data Service, 2014). Data dimensions are used to describe “features of data that can be measured or assessed against defined standards in order to determine the quality of data” (DAMA UK Working Group, 2013:3). The next section discusses quality dimensions relevant to this study.

2.3.3.1 Quality dimensions as defined by DAMA UK

In an attempt to offer working data quality dimensions as general data quality dimensions, the Data Management Association of the United Kingdom (DAMA UK) has produced six best practice principles. These dimensions may aid in reducing uncertainty and misunderstandings that may arise while evaluating data quality (DAMA UK Working Group, 2013). The DAMA UK Data quality dimensions include completeness, uniqueness, timeliness, validity, accuracy and consistency. According to the DAMA UK Working Group (2013), these dimensions offer a realistic and trustworthy

basis for effective research data management across the research data lifecycle. Definitions for each of these dimensions are provided in table form in Appendix 16. Each of the DAMA UK Data quality dimensions is thoroughly discussed there as the final augmented quality dimensions. The DAMA UK Quality Dimensions are the key dimensions used in this study. Additional dimensions created by other organisations were added to the DAMA UK Quality Dimensions to confirm that they may be regarded as the fundamental quality dimensions. The table in Appendix 16 lists the DAMA UK quality dimensions and provides explanations of how other writers and organisations describe the same dimension. To contextualise the previously described quality dimensions (within the framework of this research study), the table presented in Appendix 4 includes all the measurements that are relevant for each quality dimension.

None of the quality dimensions discussed in this section exist in isolation. Because the aspects of quality are inextricably tied to standards, guidelines, and best practices, it is critical that institutions and researchers have access to them with an “awareness of the need to implement standards, and respect for certain best practices, in particular regarding a quality approach to managing and preserving data” (Rouchon et al., 2013:6). Subaveerapandiyani (2023:13) gives an account of how the University of St Andrews and the University of Glasgow “have employed integrated models, utilising standards and fostering an ongoing stakeholder dialogue. They have integrated research information systems with open access repositories, ensuring data is entered once and reused, enabling access to authorised individuals. The case studies highlight the benefits of these systems in improving research information quality and supporting compliance reporting”.

2.3.3.2 Quality dimensions as defined in the South African Statistical Quality Assessment Framework (SASQAF)

SASQAF refers to the South African Statistical Quality Assessment Framework. Statistics South Africa is South Africa's national statistical service, seeking to generate timely, reliable and official statistics to advance economic growth, progress and democracy (Statistics South Africa, 2018). Statistics South Africa (Stats SA) has specified preconditions in the form of eight quality dimensions known as the South African Statistical Quality Assessment Framework (SASQAF). Because statistical data management is similar to other types of data management, the SASQAF's eight quality dimensions overlap with the DAMA UK's quality dimensions. The eight SASQAF dimensions of quality include relevance, accuracy, timeliness, accessibility, interpretability, coherence, methodological soundness and integrity (Statistics South Africa, 2010). Two of these criteria (accuracy and timeliness) directly coincide with the DAMA UK Data quality dimensions, which are designed to make data as accessible as possible by precisely explaining the phenomenon they are measuring. The prerequisites of quality refer to “the institutional and organisational conditions that have an impact on data quality – these include the institutional and legal environment and availability of human, financial and technological

resources” (Statistics South Africa, 2010:4). Soelton et al. (2023:295) presented a different view on the conditions needed for an organisation to survive “an organization cannot survive without improving and learning new knowledge and coping with dynamic conditions”. The main objective of the SASQAF is “improving the quality of the data produced in Stats SA and all organs of the state” (Statistics South Africa, 2010:2). The SASQAF is discussed in more detail in Appendix 1.

The South African Statistical Quality Assessment Framework (SASQAF) and the DAMA UK Data Quality Dimensions have provided distinct and relevant quality dimensions that were considered when forming a comprehensive data quality assessment system for a research setting in the social sciences. The quality dimensions outlined above are sufficiently thorough to address areas where data quality may be jeopardised.

2.3.3.3 Victoria State Government – Data Quality Guideline

In both the private and public sectors, data management is a process; the Victoria State Government “is the executive administrative authority of the Australian state of Victoria” (Wikipedia, 2020). Coincidentally, it has created a data quality guideline that aligns particularly well with this study as it uses most of the data quality characteristics adopted here. The Data Quality Guideline for the Victorian State Government provides high-level direction on the evaluation, maintenance, and upgrading of the quality of essential and shared data assets to Victorian government departments (Victoria State Government, 2018). Data quality dimensions are considered the foundation of data asset evaluation and management, as well as how to improve research operations, according to the Victoria State Government – Data Quality Guideline.

The table presented in Appendix 15 (part A) outlines the seven data quality dimensions used by the Victorian Government to measure data quality and includes a summary of the properties included in the proposed quality framework.

The Victoria State Government's Data Quality Guideline is a valuable tool for assessing the quality of functional and administrative data. It uses attributes as measuring criteria, and the institution's data quality statement is sufficient for this study. The attributes provide a wide metric for assessing if researchers have addressed issues impacting data quality. Implementing data quality controls ensures the integrity, authenticity, and validity of research data, which are discussed in the following section.

2.3.4 Data ‘integrity’, ‘authenticity’ and ‘validity’

The purpose of implementing data quality controls is to ensure the integrity and authenticity of research data. These two characteristics are discussed in the following section. Data integrity refers to “maintaining and assuring the accuracy and consistency of the data over the entire data life-cycle” (Divecha, 2016:10). Authenticity, on the other hand, refers to:

“...the digital retention and preservation practice of verifying that a digital object has not changed. Authenticity attempts to identify that an object is currently the same genuine object that it was ‘originally’ and to verify that it has not changed over time unless that change is known and authorised” (Clarke, 2014:4).

Balogun (2023:428) subsequently stated that “ensuring the authenticity of digital records over time poses a major challenge. Authenticity in this context deals with maintaining provenance, rights management, technical environment, and preservation management, digital information requires a very high standard of guarantee of integrity and authenticity”.

The authenticity definition's latter half is utilised as a criterion for the Methodological Soundness quality dimension. Integrity and authenticity are complementary qualities since they both ensure the accuracy of data.

The two concepts (integrity and authenticity) can be considered as two sides of the same 'quality' coin, as they both necessitate safeguarding data and ensuring its accuracy (USAID, 2015). Reliable research data has a favourable impact on the quality of the results and all subsequent data interpretation and utilisation processes. Certain methods can be followed to ensure the quality of the research data. These methods include data quality control and data quality assurance; “data quality control is the set of operational techniques and activities that are used to attain the quality required for a data product” and “data quality assurance includes all those planned and systematic actions necessary to provide adequate confidence that a data product will satisfy a given set of quality requirements” (Wang et al., 1995:624). Shanmugam et al. (2023:79) explain that “maintaining data quality is often viewed as risky but essential for effective decision making. Various factors that can hinder data quality are identified in the literature, such as inadequate systems for ensuring timely and accurate reporting of data, insufficient standards and training for data collection, dysfunction and inconsistencies among organisations involved in data collection, and the need for new management strategies that utilise accurate and relevant data”.

If RDM and data quality are overlooked and not handled properly, this will have a negative impact on the data and usability. As a result, criteria must be in place to assist in the direction of research and the improvement of a suitable environment for data quality output. The next section discusses data quality guidelines.

2.3.5 Data quality guidelines

Data quality and usefulness will suffer if RDM and data management are neglected or improperly managed. Therefore, criteria must be in place to help guide research and establish a conducive environment for data quality outputs. The next section discusses the DataONE best practice guidelines.

2.3.5.1 *DataONE best practice guidelines*

Wiggins et al. (2013) embrace the Data Observation Network for Earth's (DataONE) best practice principles for data quality assurance, which include:

- Develop quality assurance and quality control

Quality assurance refers to the methods used to obtain the best possible data (Wiggins et al., 2013); The University of Wisconsin (2020) defines quality assurance as “strategies and policies for ensuring that data integrity, quality, and reliability are maintained at every stage of the project. This includes strategies for preventing errors from entering the datasets, taking precautions before data is collected, and establishing procedures while data is used in a study”. Multiple types of data collection errors can occur, necessitating quality control methods to prevent them. Data entry interfaces are a good example of this, since they may be set up to assist observers in providing the most accurate data, thereby reducing both errors and instances of missing data (Wiggins et al., 2013). The Completeness quality dimension is measured using this guideline as a criterion.

- Communicate data quality with research teams

It is necessary to “document quality control processes and any changes to research teams in as much detail as possible, as this benefits everyone who manages and uses data” (Wiggins et al., 2013:6). It’s also the viewpoint of Galli (2021) that “communication and coordination are the bonds that carry out management in all aspects, and they are the connection among people, thoughts, information, and resources. Hence, communication is not only crucial but is also essential to the success of a project”. When everyone on a research team understands the importance of data quality and what needs to be done to preserve it, they may work together to achieve a common goal of data quality. This guideline is utilised as a measuring criterion for the Consistency quality dimension.

- Mark data quality control flags

In a data file, reference points are important for indicating problematic areas. Frequently, data that has problems is 'marked' as potentially invalid (Wiggins et al., 2013). Because the data provided to the public, participants, project organisers, and researchers will differ, this is an important strategy (Wiggins et al., 2013). Secondary users perform different functions for different contexts, Abedjan et al. (2018:995) elaborate by explaining that “the ongoing process of adding new entries introduces duplicates and other data errors. A particular case of errors relates to inconsistent formatting”. A researcher can go back and repair possibly faulty areas in a data set after identifying them. This improves the quality of the data concerned; hence this guideline is utilised as a criterion for the Completeness quality dimension.

- Identify missing values and define missing value codes

It is crucial to address inconsistencies in order to preserve a data file's dependability and consistency. As a result, it is also crucial to give data file contents such parameter definitions and format clarifications, quality check notes (such as flags on untrustworthy data points), and missing values (Wiggins et al., 2013). The measuring criteria for the Validity quality dimension are thus the identification of missing data and the definition of missing codes.

- Ensure basic quality control

Quality control must be a part of the data collecting and processing procedures to ensure effective data management. Quality management and quality assurance processes “mitigate error input, and define and handle errors” (Wiggins et al., 2013:5), whilst Rohman et al. (2023:239) view quality assurance as “an activity to provide evidence to build confidence that quality or quality can function effectively”. This best practice is used as a measuring criterion for the Comparability and Coherence quality dimension.

- Double-check data entry

The process of data entry is prone to errors. Thus, to promote data analysis and reuse, “at a minimum, data collection protocols, measurement techniques, and QA/QC methods need to be included in documentation for all data sets” (Wiggins et al., 2013:9). This ensures that any deviations in the data collected, as well as established methods for recognising and correcting them in the data, are documented (Wiggins et al., 2013). This guideline serves as a measuring criterion for the Uniqueness quality dimension.

These instructions facilitate accurate data entry by providing basic practical methods that reduce the chances of a researcher making a mistake.

2.3.5.2 The ICPSR data management guideline

The Inter-University Consortium for Political and Social Research (ICPSR) “provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community” (ICPSR, 2018:4); this is also confirmed by the University of Michigan (2023) “ICPSR encourages and facilitates research and instruction in the social sciences and related areas by acquiring, developing, archiving, and disseminating data and documentation relevant to a wide spectrum of disciplines, and by conducting related instructional programs”. ICPSR provides data management assistance with quality aspects as part of its digital preservation and dissemination feature. The quality dimensions offer significant support for data quality even if they are not all-inclusive and differ from those employed by DAMA UK and SASQAF. The ICPSR data collecting and file creation section of the proposed quality framework was employed as measurement criteria for the Completeness

quality dimension in this study. The file creation section of the ICPSR data management guideline covers dataset creation, ‘integrity’, variable names, variable labels, variable groups, codes and coding, missing data and selecting missing data codes. The types of missing codes used in a data set can be problematic if inconsistent values are used to represent missing values. It is important to be consistent in the use of placeholders to represent missing codes in order to facilitate the usability of the data. The ICPSR ‘selecting missing codes’ guideline is used as a measuring criterion for the Interpretability quality dimension. The manner in which variable names and variable groups are handled for accuracy and consistency is used as a measuring criterion for the Uniqueness quality dimension. According to ICPSR (2018:21) “transcribing data from a questionnaire or interview schedule to an actual data record can introduce several types of errors”; Data integrity is preserved by following these measures, which help to address and minimise the occurrence of errors:

- “Separate the coding and data-entry tasks as much as possible. Coding should be performed in such a way that distractions to coding tasks are minimized” (ICPSR, 2018:21) – this step is used as a measuring criterion for the Validity quality dimension.
- “Use a data-entry program that is designed to catch typing errors, i.e., one that is pre-programmed to detect out-of-range values” (ICPSR, 2018:21) – This step serves as a criterion for measuring the Integrity quality dimension.
- “Perform double-entry of the data, in which each record is keyed in and then re-keyed against the original” (ICPSR, 2018:21) – This step serves as a criterion for determining the Accuracy quality dimension.
- “Carefully check the first 5 to 10 per cent of the data records created, and then choose random records for quality-control checks throughout the process” (ICPSR, 2018:21) – this step is used as a measuring criterion for the Completeness and Accuracy quality dimensions.

The goal of these steps is to ensure that the data obtained is of high quality, as following them will prevent discrepancies, errors, and duplications in a dataset. Because the integrity dimension of the ICPSR and the integrity dimension of the SASQAF overlap, the organisations' guidelines have been used in assessing parameters. Another module of the ICPSR that was recognised for this study is that concerning the code and coding guidelines. In quantitative research, researchers have to create codes and code categories to represent variables. The ICPSR code and coding guidelines appear in Table 2.3, below:

Table 2.3 - Recommended code and coding guidelines for researchers

Code and coding guideline	Description
Identification variables	Give each record the necessary fields at the outset to accommodate all identification factors. A distinct research number and respondent number to reflect each case are frequent identification factors.
Code categories	Code categories should be extensive, clearly defined, and mutually exclusive. Each interviewee should only offer one category of response. Ambiguity will make coding challenging and interfere with data interpretation.
Preserving original information	Code as much information as you can. It is more beneficial to record raw data rather than compressing or bracketing it, such as age and income. Secondary analysts are not limited to the brackets selected by others but can instead choose additional meaningful brackets with original or thorough data.

(ICPSR – University of Michigan, 2022)

If uniform codes and coding are developed, all statistical software packages will be able to handle the data, especially when translating data between different formats (e.g. SPSS to STATA). The utilisation of codes and coding facilitates the analysis of data files by representing survey replies through numerical codes that have specific meanings. Fieldworkers, or data capturers, have to shift from manual to automated data capture techniques when gathering data in order to promote automation, misinterpretations, and incorrect usage of the data collection instrument. Automating the data flow is the most efficient technique to minimise errors when gathering data (from field operations). Where applicable, mobile data collection apps should be used.

2.3.5.3 Data Archiving and Networked Services Guidelines

The Data Archiving and Network Service (DANS) is “the national organization in the Netherlands for storing and providing permanent access to research data from the humanities and social sciences” (DANS, 2010:1). Data must be handled and curated properly in order to advance new knowledge and benefit the scientific community. The purpose of the DANS Data Guide 8 is to ensure that data can be found and distributed in the future (DANS, 2010). The DANS data guidelines are intended to provide explicit direction for the handling of research data based on quality criteria. The summary guidelines include:

- “Research data that can be found on the Internet.
- The research data are accessible while considering ruling legislation with regard to personal information and intellectual property of the data.

- The research data are available in a usable data format.
- The research data are reliable.
- The research data can be cited” (DANS, 2010).

FAIRmat (2023) also developed guidelines for writing research data management plan, which include:

- “Organize your data: Create a clear structure for your data that makes it easy to find and access.
- Associate data to rich metadata: Metadata provides the contextualization and description of a data object, which allows for that data object to be properly interpreted and turned into information.
- Document everything: Keeping detailed documentation of your data using appropriate tools is crucial for ensuring its integrity and usability.
- Ensure data quality: To ensure that your data are reliable and accurate, you will need to develop a strategy for quality assurance. This may involve using automated data checks, conducting manual reviews, or establishing data validation protocols”.

These recommendations are intended to represent and support effective data management; DANS recognises the importance of a data management plan that covers the bulk of data quality areas. According to DANS (2010:11) “planning for the management and archiving of a data collection at the outset is critical to the project’s success. The cost of a project can be significantly reduced if careful planning takes place early in the project”. The DANS Data Guide 8 emphasises data management sections that should be considered by a researcher in order to develop and make high-quality data accessible. The following sections were considered data quality evaluation criteria for the proposed quality framework and are part of the DANS Data Guide 8 project plan:

Data and file structure: What is a data file going to look like and how will it be organised? What is the unit of analysis? Will there be one large data record or several shorter ones?

This section of the project plan is a starting point for good data collection management; answering these questions allows a researcher to create a structured plan before collecting data. This section serves as a criterion for the Accuracy quality dimension of the proposed quality framework.

Naming conventions: How will files and variables be named? What naming conventions will be used to achieve consistency? How will the study title be incorporated in file names?

In a data file, variables should be named in a consistent, exact, and accurate manner that reflects the aspect under examination. According to the University of Hertfordshire (2023) “putting time and effort into naming records consistently and logically will distinguish similar records from one another at a glance, and by doing so will facilitate the search and retrieval of records and enable users to browse file

names more effectively and efficiently.” File naming conventions are also important because files must be labelled precisely according to what they are designed to do (Creswell, 2014). The Integrity quality dimension is measured using this component of the project plan.

Data checks: With regard to the input of data, will the variable formats be numeric or character? What checks will be used to find invalid values, inconsistent responses, incomplete records, etc.? What checks will be used to manage the data versions?

The answers to the above questions are crucial for a researcher because the choices made about variable formats, invalid values, contradicting responses, and missing records can affect how well data can be read and analysed, as well as data usability. This section is recognised as a criterion for the proposed quality framework's Quality Check Processes quality dimension.

Preparing dataset documentation: What will the dataset documentation or metadata look like and how will it be produced? How much is necessary for future retrieval and archival processing?

Data documentation enables data to be used efficiently and accurately; the data documentation's design and content structure must be suitable for secondary use. Data usability is limited when data aspects are poorly documented and described (Brase, 2010). This component serves as a criterion for measuring the Integrity quality dimension; Salian (2023:4) used a different term to explain ‘data documentation’, according to Salian (2023:4) “choosing the correct format is essential so there is the highest usability of data. Portal managers should have open data in a standardised, machine-readable format that can be readily used by customers”.

Variable construction: What variables will be constructed following the collection of the original data? How will these be documented?

To ensure that data is clear and that secondary use is possible, a researcher must clearly explain and record all methodological decisions they make about variables. These questions are important because they enable a researcher to improve the correctness of variable formulation by removing any unnecessary information or feedback obtained during data collection. This section is recognised as a criterion for the Uniqueness quality dimension of the proposed quality framework.

Project documentation: What steps will be taken to document decisions that are made as the project unfolds? How will information be recorded on field procedures, coding decisions, variable construction, and the like?

The data collected (as part of field activities) must be timely and comply with agreed-upon methods and standards in order for the research questions in a study to be answered. For reporting purposes, any variations and choices made by a fieldworker must be agreed upon and reported by all parties. The section on documenting project decisions is recognised as a criterion for measuring the Completeness

quality dimension. Every data-related choice made by fieldworkers must be documented in order for the research team to keep track of any data modifications. The manner in which information is recorded on field procedures and coding decisions is used as a measuring criterion for the Flexibility of the data quality dimension.

The DANS Data Guide 8 project plan assists researchers in comprehending the practical aspects of managing their research data, particularly the matter of data quality. The portions of the project plan show that the researcher is solely responsible for the accuracy of the data; as a result, much of the section is included in the recommended quality framework. A subsequent section provides an integrated picture of the data quality characteristics that were defined, while the following section analyses the impact of data quality.

2.3.6 Impact of data quality

Institutions that strive to position themselves to manage data effectively and ensure high-quality data have an edge (Moreno, 2017); Sigmoid (2023:2) also supports this viewpoint in that “high-quality data is a valuable asset that enables businesses to make informed decisions and drive effective strategies. It ensures that information is reliable and trustworthy, reducing the risk of errors and misleading conclusions”. The efficiency of research-based work processes is determined by data quality. Data quality is crucial because it promotes ethical work practices that have a direct impact on an institution's reputation; if not correctly managed, it can result in reputational harm. Toni (2014:3) claims that “reputational risks is a top strategic business risk, being a key business challenge. A reputation risk that is not properly managed can quickly escalate into a major strategic crisis”.

To avoid reputational risk, a corporation that uses research as its primary business approach must ensure that its data asset is of high quality. Toni (2014:5) supports this statement by explaining that “how a company manages the expectations and performance related to its reputation determines whether value is created or destroyed”. Data quality has a positive influence on credibility since it has the following effects:

Decision-making

Decisions made in an organisation have a direct impact on the outcomes and achievement of goals. Data is a resource that is utilised to inform institutional decisions. Thus, the higher the data quality, the more trust users have in the outputs they generate, the lower the risk of outcomes, and the higher the efficiency (Moreno, 2017). Jeyadevi (2023:19) highlights the importance of data in decision-making by explaining “making decisions based on data enables the identification of fresh and alluring company chances. An in-depth insight into a company's essential activities can be obtained by examining publicly accessible visual data”. Because data quality ensures that outputs are correct, trustworthy, and consistent, and also that risks are avoided in decision-making, the phrase “garbage in, garbage out” is no longer applicable.

For the Usability of Data quality dimension, the confidence that users have in research results (depending on the quality of such outputs) is employed as a measurement criterion.

Compliance

Academic institutions operate in a legally regulated industry, which necessitates the ethical and appropriate use of data (Subaveerapandiyan, 2023). Data quality is a crucial link between law enforcement and the criminal justice system. Because regulations in different parts of the world are constantly changing, data quality should be a continuous process everywhere a firm does business (Moreno, 2017). When the data used for business is of poor quality, an organisation's reputation is jeopardised. According to Farha et al. (2017:3), “events which inflict reputational damage on a firm will ultimately result in losses that are far greater than the direct cost of the particular event”. The way consumers perceive and are able to use data has an impact on an institution’s reputation; In order to emphasise the significance of reputational harm, Perera et al. (2022:2) explain that “as organisations view success as financial and material gain, reputation is an important component to manage in ensuring this. Reputation is an intangible asset that affects all stakeholders of an organisation”. This is a criterion for determining the Usability of Data quality dimension. As noted previously, different authors and organisations assign diverse meanings to research data quality, yet these definitions help to arrive at a common ground of what the term signifies. The next section discusses how international RDM role-players perceive data quality. The reputational damage caused by poor-quality data (used in company operations) is utilised as a criterion for the Usability of Data quality dimension.

2.4 Data quality as documented by international RDM role-players

This section focuses on the important role-players who participate in RDM campaigns, services, and platforms. Appendix 15 (parts A - E), as noted below, discusses a variety of international RDM role-players. These role-players were selected because they have recently established RDM components that demonstrate their willingness to further improve RDM practice. It was necessary to carry out an extensive assessment process in order to select the most viable international role-players. To contextualise this selection process, 16 Appendices were created, each serving a different purpose, and titled as follows:

Appendix 1: Extended background information

Appendix 2: Quality Process Framework

Appendix 3: Data lifecycles with an emphasis on quality

Appendix 4: Measurements for each quality dimension

Appendix 5: Criteria used to identify relevant models

Appendix 6: Flow diagram symbol description – initial Quality Process Framework

Appendix 8: Interview schedule - Researchers

Appendix 9: Interview schedule – Data curators

Appendix 10: Research participant handout

Appendix 11: The initial quality process framework

Appendix 12: Results of the criteria used to identify relevant role-players/models

Appendix 13: Grouping of quality dimensions with stages of the research data lifecycle that form the quality process framework

Appendix 14: UK Data Quality Dimensions

Appendix 15: Data quality evaluations by different stakeholders

Appendix 16: Augmented UK DAMA data quality dimensions

After assessing all the potential models from the literature (Appendix 3) using the criteria specified in Appendix 5, a narrowed list of potential models was identified. These models deal with quality assurance and data quality to a certain level. The results of the selection criteria using Appendix 5 are represented in Appendix 12. This exercise resulted in seven models (role-players) that could be adopted for this study. The table presented as part B of Appendix 15 outlines the most significant information regarding international role-players that were selected and how they influence data quality

The data quality component of the multiple international RDM role-players contributes to a broader reason for developing RDM. It was required to assess each role-player's position on data quality in order to get a sound judgment about what international RDM role-players are doing and saying in terms of data quality. When data is collected and handled, most role-players have defined protocols and procedures in place. These procedures and suggestions are valuable because they provide multiple points of reference for dealing with data quality. The following section examines data quality as defined by major RDM programs.

2.5 Data quality as recognised by significant RDM initiatives

There are significant RDM projects, as well as international RDM role-players, that are actively interested in enhancing RDM and data quality. The key RDM initiatives that were reviewed are explained are represented in Appendix 15.

International role-players are making significant progress in transforming the way research is performed and managed. The concerns of data quality and quality assurance are dealt with in a variety of ways.

Working groups, culture of practice, and institutional techniques are used by the majority of international role-players to address consistency (WDS, 2015). As an additional means of putting quality assurance into practice, Daniel (2023) introduced the idea of 'consistency maintenance' which comprises "real-time monitoring ensures that data collected from various sources and across multiple time points adhere to predefined standards and data quality guidelines" (Daniel, 2023:37).

The source of funding for the research has a big impact on how researchers conduct their jobs. Research funders are an essential part of the research process since they give researchers the financial support they need. Additionally, applicants need to fulfil requirements imposed by funders related to the quality of their data. Some of the major funders of social science research are introduced in the next section, along with their views on data quality.

2.6 Data quality as an interest for large research funders

Researchers must adhere to certain conditions in order to receive funding for their research projects from research funding organisations. The most important "impetus for good data management practices might be compliance with funder requirements" (Bryant et al., 2017:8). There are a number of significant and large research funders that are well-known around the world for financing research projects in a variety of sectors. Researchers must understand the funders' requirements and the conditions under which they must comply in order to acquire research funding. According to Corti et al. (2014), researchers have many responsibilities regarding their data; The College of the Rockies (2023:9) elaborates by explaining that a researcher is responsible for identifying "who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible". Data management and sharing policies are being adopted by research funding agencies in the UK, the US, and Europe to increase data availability, openness, and accountability of the research they finance. Part D in Appendix 15 discusses how data quality is being addressed by researcher funders. Because these funders have already supported multiple research projects, it is vital to investigate the data quality assurance criteria included in their funding application requirements.

The research funders indicated in this section primarily ask applicants to submit a Data Management Plan. The DMP's strictness in terms of quality assurance and data quality varies, although it does not always preclude participants from applying. The majority of donors are concerned with how data will be disseminated and accessed. Participants must address data quality issues as part of their research DMPs. The following section looks at how data quality is viewed by evaluation systems.

2.7 Data quality as expressed in evaluation systems

An evaluation system refers to all metrics, methods and processes used to assess whether a programme has been implemented under a plan (monitoring) and has had the desired outcome (assessment)

(Latham, 2014). The evaluation systems that were used, as well as how the data's correctness and consistency are ensured, are discussed next.

2.7.1 International Standards Organization (ISO)

The International Standards Organization (ISO) has produced a set of quality requirements standards. Only those that help to improve data quality are discussed.

2.7.1.1 ISO 16363

ISO 16363 allows a repository to be evaluated and accredited as a trustworthy digital repository (The Consultative Committee for Space Data Systems, 2011).

The ISO 16363 standard includes evaluation categories that are relevant to this research. These categories are meant to serve as useful guides for the larger goal of improving the quality of social science research data quality assurance and RDM. The methods concerning how an organisation should handle its digital records are included in the ISO 16363 category below:

Section 3: Policies informing how a repository and institution at large should manage its digital information, including research data This portion of ISO 16363 is used as a criterion for measuring the Standards and Policies quality dimension.

2.7.1.2 ISO 8000

ISO 8000 is a data quality management process measurement standard that “specifies a structure for process measurement stacks that organisations can instantiate to measure the characteristics of processes for data quality management” (ISO, 2019). It was important to review this standard because ISO 8000 “defines characteristics of information and data that determine its quality, and provides methods to manage, measure and improve the quality of information and data” (ISO, 2019).

Although ISO 8000 is a comprehensive process measurement standard, the author has simplified some of its definitions for the purposes of this research study. Since the respondents are scientists in different domains and are not required to work with the intricate requirements of an ISO standard, the definitions have been made simpler. In the suggested quality process framework, it was crucial that the author recognised the ISO 8000 standard and included some of its recommendations as measuring criteria. The following section discusses sections of the ISO 8000 that were used as measuring criteria in the quality process framework.

- ISO 8000-8:2015: Data quality Part 8: Information and data quality: Concepts and measuring (ISO, 2011):

This category outlines fundamental concepts and data quality principles, as well as how these principles apply to quality control and management systems (ISO, 2011). The correct use of basic information and data quality principles is a good place to start when trying to improve how data is handled with quality-

based methodologies (ISO, 2011). This section serves as a criterion for assessing the Flexibility quality dimension.

- ISO 8000-62:2018: Data quality management: Organizational process maturity assessment: Application of standards relating to process assessment (ISO, 2011):

This category, according to ISO (2018), specifies certain elements of a maturity model. These aspects can be used by organisations to assess the maturity of data quality control systems in reference to their own evaluation metrics (this section serves as a criterion for assessing the Usability of Data quality dimension).

- ISO 8000-115:2018: Data quality – Part 115: Master data: Exchange of quality identifiers: Syntactic, semantic and resolution requirements (ISO, 2011):

The quality identifier specifications that are part of master data interchange are defined in this category (ISO, 2018). (Because the master data file is the final version that contains all of the results of a study, its access permissions and management should be handled with extreme caution). The management of the master data file is a measuring criterion for the Usability of Data quality dimension.

The author's interpretation of these concepts was intended to provide context for how the aforementioned definitions can enhance data quality and how they have been used inside the suggested quality framework.

2.7.2 Core Trustworthy Data Repositories Requirements (Core Trust Seal)

The World Data System (WDS) – Data Seal of Approval (DSA) and the International Council of Scientific Unions (ICSU) have established a repository certification process called Core Trustworthy Data Repository Criteria, which is focused on the catalogue and procedures of the two institutions. According to DSA-WDS (2016:1), “The Core Trustworthy Data Repositories Requirements are intended to reflect the characteristics of trustworthy repositories”. The Core Trust Seal “promotes trustworthy data repositories and practices by evaluating and certifying data repositories on sustainability and trustworthiness” (Stvilia, 2023:2). The requirements are mandatory (with some overlap), and each one is accompanied by guidelines that explain to participants what proof is necessary to verify that the repository concerned complies with the relevant standards. The requirements, all of which must be met by a repository, are organised into 15 categories. Three requirements (R7, R8, and R11) will be scrutinised for the objective and relevance of the study concerned since they helped to construct the suggested framework for RDM based on data quality. Appendix 1 examines these three requirements in greater depth.

For the purposes of this study, the following are the most essential data quality criteria extracted from the 16 Data Seal of Approval Guidelines (for producers and data repositories):

- “The data producer deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms” (DSA-WDS, 2016). This criterion is used as a measuring criterion for the Flexibility of the data quality dimension.
- “The data producer provides the research data in formats recommended by the data repository” (DSA-WDS, 2016). Using a data file format that is not supported by a repository limits a researcher's capacity to manage data on a secure platform (repository). A researcher's choice of acceptable formats is a criterion for the Flexibility of the data quality dimension.
- “The data repository uses due diligence to ensure compliance with legal regulations and contracts” (DSA-WDS, 2016);
- “Archiving takes place according to explicit workflows across the data lifecycle” (DSA-WDS, 2016);
- “The data repository ensures the authenticity of the digital objects and the metadata” (DSA-WDS, 2016). Discovering which mechanisms a repository employs to verify that the data's validity is not tampered with or altered is critical to ensuring data quality. The Accessibility quality dimension is measured using this guideline as a criterion.

Researchers and organisations who submit their data to CTS-DSA-certified repositories must resolve data concerns that could compromise data quality. Tsang (2014:14) explains that “the ICPSR was one of the first six data repositories to earn the Data Seal of Approval in 2011”. If a researcher meets the ICPSR self-assessment criteria, they will be eligible to deposit their data in a repository that “offers a number of specialized services designed to meet the needs of researchers collecting social science data” (ICPSR, 2018:4). The quality of data is primarily determined by the researcher, who conducts a study endeavour. ICPSR provides researchers with the chance to “conduct simple or complex analyses, recode and compute new variables, and subset variables or cases for downloading through our online data analysis system” (ICPSR, 2018:5). This method allows researchers to make favourable findings that could otherwise go unnoticed, thereby improving data quality.

Depositing data in the ICPSR repository benefits academics in two ways: it provides a qualified and constructive data management platform while also effectively encouraging data access and utilisation. With the ICPSR repository:

...the research data are available in a usable format (this benefit is used as a measuring criterion for the Accessibility quality dimension); the research data are reliable (this benefit is used as a measuring criterion for the Relevance quality dimension); the research data can be referred to” (Tsang, 2014:15) (this benefit is used as a measuring criterion for the Accessibility quality dimension).

Additionally, the University of Michigan (2023) clarified that “ICPSR makes quantitative data files available in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS,

SPSS, Stata, and R. Documentation is provided as PDF. For long-term preservation, data are stored in accordance with prevailing standards and practice”.

RDM is strongly linked to repository certification and how repository management is carried out. As a result, the general focus of assessment systems on repositories is reduced. At this level, assessment determines the quality of a repository by determining what functions it can and cannot perform. This is not, however, a quality assurance metric with a high level of relevance for the current investigation. The following section examines recommendations on social science methodology found in the literature.

2.8 Quality recommendations as expressed in books on social science methodology

Methodology is a system of methods applied in a given field (Oxford Pocket Dictionary, 2002). The approach a researcher takes when conducting research may be guided by established standards, ethics, and procedures. The application of a methodology should be methodologically sound. The research proposal is considered methodologically sound when the methods therein are able to answer the study question (Kaplan University, 2017). A research proposal should "help you to focus your thoughts on a topic, to define in broad terms the way in which you are going to approach and structure the work, and to familiarize you with the relevant literature," according to Walwyn and Chan (2023:11), who also list other purposes for research proposals. Because social science research focuses on human behaviour, it differs from other kinds of research. Social science research is written and reported by a vast number of academics. For the purposes of this study, it was essential to ascertain which methodological books are regularly used by researchers at the social science institution concerned. A list of the social sciences methodological books most frequently consulted at the social sciences institute was gathered and examined. Ten books were selected and their comments on data quality are discussed below. The goal was to highlight the points made concerning data quality in these significant texts on social science research methodologies.

The books reviewed covered a variety of data quality management topics that are relevant to the research process. Quality assurance and data quality issues are explored on a variety of scales and with varying degrees of focus. The most important element is that all of the methodological books analysed support the importance of the outcomes' validity and reliability. The authors of numerous methodological publications propose that researchers utilise statistical coding methods, error-checking procedures, and the creation of relevant documentation to clarify the details of data processing in order to answer the "how" question (of validation and dependability). Researchers will benefit from reading the methodological books reviewed since they do not overlook quality assurance and data quality. Other social science subjects, however, have been scrutinised (Frankfort-Nachmias & Nachmias, 1996;

Goertz & Mahoney, 2012; Kaplan, 2004; Maggetti et al., 2013; Scarbrough & Tanenbaum, 1998) to ascertain what is being said about quality data and quality assurance.

2.8.1 First book on social sciences methodological methods considered

Doing Qualitative Research – A Practical Handbook

Author: Silverman, D. (2017)

This book is considered an important guide both for researchers and for research students inexperienced in conducting qualitative research, and it is read by most social scientists before and during a study (Silverman, 2017). This book is also an excellent resource for seasoned researchers looking to improve their qualitative research skills and it provides practical advice on essential issues and methods, such as narrowing down a research subject, establishing 'originality', maintaining a study diary, drafting a research report, presenting the study to various audiences, and validating and reporting research findings.

On the subject of data quality, Silverman (2017) devotes only one chapter to validity and reliability. Silverman (2017:175) claims that “unless you can show your audience the procedures you used to ensure that your methods were reliable and your conclusions valid, there is little point in aiming to conclude a research dissertation”. The main point of the chapter is the author’s explanation of how to diagnose data difficulties and, in addition, he offers some practical validity remedies. Validating research findings is a quality assurance activity that increases data reliability and researchers should be aware of the many possible validation methods available to them. These include:

- Content Validity: this validation method “reflects how well the questions represent the possibilities of available questions” (Samuels, 2015:62); this method is used as a measuring criterion for the Validity quality dimension.
- Convergent validity: “is the extent to which scores on a measure share a high, medium or low relationship with scores obtained on a different measure intended to assess the similar construct” (Mohajan, 2017:18). This method is utilised as a criterion for the Interpretability quality dimension.
- Criterion-Related Validity: the validity of the metric is an approximation of the degree to which a measure corresponds with a gold standard (i.e. an objective metric of the measured phenomenon) (Bellamy, 2015). This method is used as a criterion for determining the Usability of Data quality dimension.
- Concurrent Validity: Estimates individual output at about the same time on various measures (Lin & Yao, 2020). This validation method, in essence, assesses how well a new test compares to a well-known test. This method is used as a measuring criterion for the Flexibility of the data quality dimension.

During the course of a research project, a qualitative researcher must overcome a biased urge. Overcoming this urge answers the question of how they are to “convince themselves and their audience that their findings are genuinely based on critical investigation of all their data and do not depend on a few well-chosen examples (known as anecdotalism)” (Silverman, 2017:385). In order for relevant research to be supported, data from a research study must be of good quality for secondary use. Silverman (2017) provides three interconnected techniques to think objectively about qualitative data analysis in order to arrive at "real" findings:

- The refutability principle

This involves a researcher “refuting their initial assumption about their data in order to achieve objectivity” (Silverman, 2017:388). Because it is abstractly centred on the assumption that a researcher is not prejudiced, this theory contributes only a small amount to data quality. In the suggested framework, this principle is employed as a criterion for determining the Validity quality dimension.

- The constant comparative method

This method entails that “a qualitative researcher should always attempt to find another case through which to compare” (Silverman, 2017:390). Using a different scenario to test a theory allows the researcher to double-check their data. Case testing aids data quality by ensuring data integrity. As a result, this method is employed as a criterion for measuring the Integrity quality dimension.

- Deviant-case analysis

This case analysis includes systematically searching for and solving deviant cases (Silverman, 2017). Because it is a hands-on technique for conducting data analysis, this approach is closely tied to improving data quality. Silverman (2017:394) provides a full overview of how to perform this procedure. This method begins with a small batch of data. A provisional analytic scheme is generated and the scheme is then compared to other data, and modifications are made in the scheme as necessary. This analysis method is used as a measuring criterion for the Interpretability quality dimension. The book also outlines the practical processes that researchers should follow when gathering qualitative data in a rich setting and supporting effective data management. Analytic induction, for example, is one of these processes, which is “to uncover casual relations through identification of the essential characteristics of the phenomenon studies” (Silverman, 2017:198). Using proper tabulations is another process. It “allows a researcher to participate in complete data treatment by evaluating and revising their generalizations” (Silverman, 2017:199). Simple tabulations are utilised as a measuring criterion for the Relevance quality dimension, and they are an effective technique to eliminate any question regarding the accuracy of data. In qualitative research, simple numbering systems can be employed as a validation tool (Silverman, 2017).

2.8.2 Second book on social sciences methodological methods considered

Data Quality Assessment

Author: Maydanchick, A. (2007)

The author of this book proposes methods for researchers and organisations to identify, correct, or avoid data quality issues within their organisations by ignoring theory and depending purely on what works and what is pragmatic (Maydanchik, 2007). The author investigates the topic of data quality, emphasising its importance in every advancement endeavour. The ability of firms and government organisations to acquire, organise, and use data on products, customers, competitors, and staff is critical to their success (Maydanchik, 2007). To be successful, the data must be of high quality. If the data is inaccurate, the organisation must work to improve the data quality, though it might be difficult to assess and determine whether an organisation's data is of good quality. It is critical to understand the theory underlying data quality, but it is also critical to have a realistic understanding of how to get the job done.

The author of this book seeks out researchers and anybody else involved in research to learn about the methodologies that support data quality measurement and good data quality. The book provides readers with vital information to assist them in managing their data in accordance with recognised standards. The following are some of the techniques covered:

- “*Data profiling and gathering metadata*” – this technique is used as a measuring criterion for the Completeness quality dimension.
- “*Organising rule and error catalogues*” – When secondary users are processing and attempting to build new extrapolations using another researcher's data file/s, mistakes may occur. An accompanying document (e.g. a user guide) that outlines how to deal with potential mistakes is beneficial. This technique is utilised as a criterion for determining the Quality Check Processes quality dimension.
- “*Ensuring accuracy and completeness of the data quality assessment*” – This technique is used as a criterion for determining the Accuracy quality dimension.
- “*Constructing the dimension data quality scorecard*” – this technique is used as a criterion for assessing the Quality Check Processes dimension.
- “*Executing a recurrent data quality assessment*” – this technique is used as a measuring criterion for the Accuracy quality dimension.

The book serves as a link between data quality theory and practical approaches for evaluating and improving data quality.

2.8.3 Third book on social sciences methodological methods considered

Approaches to Social Research

Author: Singleton, A., Straits, B.C., and Straits, M.M. (1993)

The theory of data quality is combined with practical approaches for analysing and improving data quality in this book. The authors look at two crucial factors in the planning or design of a study: the measuring of variables and the sampling of analytical units (Singleton et al., 1993). The authors emphasise the use of Reliability and Validity methods to contribute to the degree of measurement that should be used to evaluate the correctness of variables and what they represent, which helps to ensure that variables in a data file are consistent. According to Singleton et al. (1993:114), “the level of measurement provides a framework for interpreting the categories of a variable; it tells us what sort of inferences we can make about cases assigned to different categories”. Interpreting the categories of a variable and being able to make accurate extrapolations is used as a measuring criterion for the Relevance quality dimension. The author discusses the many grouping structures that variables can be a member of. This is a quality assurance of data consistency. A description of how reliability and validity are retained during research aids the reader in understanding how to apply these two principles to their own analyses. The next section looks at how measurement standards are determined by reliability and validity.

To further contextualise the concept of 'standard of measurement,' it is crucial to understand how the authors define the terms 'reliability' and 'validity.' These are crucial principles that a researcher must acquire and comprehend in order to work toward incorporating them into their research data. The authors define reliability as a phrase that refers to issues of consistency and accuracy. The operational meaning verifies that an aspect of a data file is truthful and trustworthy. Validity relates to the degree to which an organisational description is matched, consistent, or suited to the term to be measured, according to Singleton et al. (1993); Middleton (2023) echoes this notion, saying that “validity tells you how accurately a method measures something. If a method measures what it claims to measure, and the results closely correspond to real-world values, then it can be considered valid”. This indicates that a highly unstable measure cannot be genuine and that it will not be possible to measure what is required if the results vary too widely. Even where something can be calculated reliably by a researcher, if what is being calculated is not what it was intended to assess, an accurate measure can nonetheless be invalid (Singleton et al., 1993). The book discusses a variety of faults that can arise in data and lead to poor data quality. These problems include:

- Sources of error: “*Measurement results in the identification of differences in the cases observed different labels, ranks, ratings and scores*”. This suggests that the principles that a researcher wants to analyse are observable differences (presented in a data file). This variance

should account for the bulk of the observed changes if an operational measurement is correct (Singleton et al., 1993).

- Systematic measurement error: *“Results from factors that systematically influence either the process of measurement or the concept being measured”*. This implies that the results, which are customised to the researcher's preferences, are continually misrepresented in the form of ratings or assessments. Random measurement error: *“Is unrelated to the concept being measured; it is the result of temporary, chance factors, such as transitory upswings and downswings in the health and mood of subjects and respondents”*. This shows that if a person is bored, overwhelmed, or weary, they are more likely to give inaccurate replies. These kinds of errors are unpredictable since their occurrence, scope, and course differ from one person to the next. (Singleton et al., 1993).

Finding, recognising, and controlling potential data errors requires prior knowledge and expertise; otherwise, detecting and knowing how to correct such errors becomes difficult. Singleton et al. (1993) provide a number of reliability and validity measures to aid researchers in avoiding data quality errors:

Test-Retest Reliability: This is regarded as the simplest method for assessing the reliability of the data. This approach involves evaluating (i.e. measuring) the same individuals or units on two separate occasions. Thereafter, the statistical similarity between the sets of scores obtained from the two tests is measured and the resulting value serves as a reliability estimate.

Because it necessitates the use of statistical methods, statisticians and practitioners will be familiar with this procedure. A social sciences researcher may not be familiar with the Test-Retest Reliability technique because it falls outside of the usual technicalities of their research area. By using a statistical correlation approach to assess the validity of the measurements, this approach ensures the retention of data accuracy. Such a strategy is advantageous for social science researchers because it does not require a high degree of statistical knowledge and competence on the part of the researcher. This method is employed as a measuring criterion for the Relevance quality dimension.

Inter-coder Reliability: This method examines the extent to which different interviewers, observers, or coders, using the same instrument or measure, get equivalent results.

The data entry processes might present numerous obstacles during data collection, resulting in multiple inaccuracies in the data. Since fieldworkers frequently collect data using inaccurate codes and labels, it is critical to have a system in place to ensure that proper results are obtained. This method ensures the dependability of a data collection instrument (to arrive at identical results). It is a broad part of data quality improvement because it aids in determining the dependability and accuracy of the data source. This method gives readers (researchers) a practical way to improve data collection and serves as a criterion for determining the Consistency quality dimension.

Construct Validation: This measure emphasises the meaning of the responses to a measuring instrument; it queries whether the instrument is measuring the intended concept or whether it can be interpreted as measuring something else.

If there is no verification procedure in place, it is very easy to go from measuring the intended definition to measuring something different. The coherence between the participant's answer and the measuring instrument can be measured using this method. Responses are trustworthy and valid as a result of such initiatives, which contributes to data quality. This book offers a variety of strategies (for dealing with social science) that have been adapted from other fields of study. Singleton et al. (1993) demonstrate that handling high-quality research is dependent on the task at hand and not discipline-specific. Researchers in the social sciences should understand the concept of being able to employ realistic metrics to observe and perform error-free data processing processes.

2.8.4 Fourth book on social sciences methodological methods considered

Applied Social Research – Tool for the Human Services

Author: Monette, D.R., Sullivan, T.J. and DeJong, C.R. (2002)

This book focuses on the relationship between research and human services practice, with a focus on highlighting the rising variety of ways in which research is applied to practice and the increasingly extensive linkages that exist between research and practice (Monette et al., 2002). There are various aspects to research practice, which are described in terms of social research management. Previous research experiences, theoretical history, and institutional data management systems all inform a researcher's research activities (in the context of research operational procedures). Monette et al. (2002), as do the authors of the preceding books, use validity and reliability approaches to handle data consistency; however, statistical techniques are heavily emphasised to assure the accuracy of data acquisition stages. Statistics and the statistical data management elements of data processing and analysis are unavoidable aspects of the process. As a result, it is critical for researchers to have a basic understanding of how data might be linked and handled statistically. Monette et al. (2002:114) describe validity as “the accuracy of a measure: does a metric accurately measure the variable it is intended to measure?”; this definition of validity is used as a measuring criterion for the Relevance quality dimension. Monette et al. (2002) offer the following validity methods:

Face validity: “involves assessing whether a logical relationship exists between the variable and the proposed measure – it amounts to a rather common-sense comparison of what makes up the measure and the theoretical definition of the variable” (Monette et al., 2002:115). This validity method is used as a measuring criterion for the Relevance quality dimension.

Criterion validity: “refers to establishing validity by showing a correlation between a measurement device and some other criterion or standard that we know or believe accurately measures the variable

under consideration” (Monette et al., 2002:116). This method is used as a measuring criterion for the Validity quality dimension.

These two legitimate techniques offer readers two options for validating if a variable is valid: one based on intuition and the other on the use of existing criteria. Based on their skill set and preferences, social science researchers must select which approach works best for them. In the suggested quality framework, validity is a measurement criterion for the Validity quality dimension.

Monette et al. (2002) also recommend the methods of reliability that a researcher may use to check the accuracy of the measurements made in the data. According to Monette et al. (2002:117), reliability refers “to a measure’s ability to yield consistent results each time it is applied – reliability measures do not fluctuate except because of variations in the variable being measured”. The types of reliability methods proposed by the authors include:

Test-Retest: this technique involves “applying a measure to a sample of people and then, somewhat later, applying the same measure to the same people again. The two sets of scores are then correlated by using an appropriate statistical measure of association” (Monette et al., 2002:118). This approach is similar to the Test-Retest method of Singleton, et al. (1993) in that it calculates the same units in two distinct circumstances, then analyses the statistical correlation between the two sets of scores and uses the resulting number as an approximation of the reliability.

Test-Retest Reliability: This is the most basic way to measure the reliability of results. This method entails assessing (i.e. measuring) the same individuals or units twice, then computing the statistical similarity between the sets of scores acquired from the two tests, with the resulting value serving as an approximation of the reliability (Monette et al., 2002).

The precision with which the correlation is calculated indicates that the test utilised by the sampled group is a true association, and the data gathered from the sample should be identical. This brings consistency and completeness to the situation. Because the correlations between scores collected from two similar questionnaires should create the same correlation, this reliability technique keeps the results consistent. This technique is employed as a measuring criterion for the Relevance quality dimension.

Split-half reliability: The test group responds to the complete measuring instrument. The responses to the instruments are then randomly divided into two halves; each half is then treated as though it was a separate scale, and the two halves are correlated by using an appropriate measure of association” (Monette et al., 2002:118).

This method involves testing a single knowledge area by splitting it into two parts and then giving both parts to one group of participants at the same time. The scores from both parts of the test are then correlated. This technique is used as a measuring criterion for the Flexibility of the data quality dimension.

This is a statistical method for confirming that study findings are correct as regards the data collection tool (questionnaire). It enables a researcher to use the correct measure of association to equate two independent scales to a single scale using mathematics. The results would show that the scales are reliable when compared to the required measure of association. This technique is utilised as a criterion for assessing the Flexibility of the data quality dimension.

2.8.5 Fifth book on social sciences methodological methods considered

Research Design: Qualitative, Quantitative and Mixed Methods Approaches

Author: Creswell, J.W. (2014)

This book assists researchers who conduct a variety of research activities with how to identify the correct study design. The nature of qualitative, quantitative, and mixed-method research design is thoroughly compared in this work (Creswell, 2014).

Validity and Reliability are explored as techniques for dealing with data quality. According to Creswell (2014:199), “validity means that the researcher checks for the accuracy of the findings by employing certain procedures, whereas reliability indicates that the researcher’s approach is consistent across different researchers and among different projects”. Validity and reliability are two distinct approaches to research management that are used to verify findings at the dataset and project levels. These methods can help a researcher evaluate their work internally to ensure correctness before allowing data to be shared with others.

The validity approach provided by Creswell (2014) comprises the following:

Triangulation: the use of a number of methods to collect data on the same subject involving various types of samples as well as data collection methods (Creswell, 2014). This strategy ensures that the findings of the study are based on the validity of the chosen measures rather than the data collection tool and sample size. The Comparability and Coherence quality dimension uses this validity technique as a measuring criterion.

Authentication: involves “checking transcripts to make sure that they do not contain obvious mistakes made during transcription” (Creswell, 2014:199). Qualitative data is non-numeric data that reflects the perceptions, attitudes, and opinions of research participants, and therefore requires a different methodology than quantitative data (Johnston, 2014). Because transcripts are a direct-account translation of what the participant said, they are prone to inaccuracies such as misspellings, grammatical irregularities, and misrepresentation. This validation approach is a practical solution for researchers who want to confirm that their qualitative transcripts are free of errors. This validity method is utilised as a criterion for the Quality Check Processes quality dimension.

Definition documents: a data producer should:

...make sure that there is not a drift in the definition of codes, a shift in the meaning of the codes during the process of coding. This can be accomplished by continually comparing data with the codes and by writing memos about the codes and their definitions (qualitative codebook) (Creswell, 2014:203).

Researchers must understand that the usability of research data is contingent on the associated documentation, which provides descriptions and clarifies how to use the data. A quality assurance measure that will promote better secondary data utilisation is comparing data with codes and then writing notes (Creswell, 2014). This method is used as a criterion for assessing the Usability data quality dimension.

2.8.6 Sixth book on social sciences methodological methods considered

Research Methods in the Social Sciences

Author: Frankfort-Nachmias, C. and Nachmias, D. (1996)

The components of this book: research theory, architecture, data collecting, and data analysis, are all meticulously interwoven (Frankfort-Nachmias and Nachmias, 1996) and proper attention is paid to each. The book presents a non-technical approach to the key research techniques in the social sciences. The section on data collecting covers approaches for quality assurance and data quality that are routinely utilised in social science research.

2.8.7 Seventh book on social sciences methodological methods considered

The SAGE Handbook of Quantitative Methodology for the Social Sciences

Author: Kaplan, D. (2004)

The facts of social science research necessitate considerable investigation and explanation, hence this book concentrates on quantitative research methods. The goal of this book is to teach applied statistics to empirical researchers and graduate students alike, and to expose them to a wide range of cutting-edge quantitative approaches in the social sciences (Kaplan et al., 2004). Statistical strategies for dealing with quantitative data are presented in a way that leads readers towards best practices for dealing with quantitative data-based statistics, with the purpose of improving data quality.

2.8.8 Eighth book on social sciences methodological methods considered

Research Strategies in the Social Sciences: A Guide to New Approaches

Authors: Scarbrough, E. and Tanenbaum, E. (1998)

This book guides readers through a maze of specialised methodologies used across the social sciences spectrum (Scarbrough & Tanenbaum, 1998). The mathematics and statistics that guide social science

research are a dynamic that a researcher must understand and such a researcher must also be skilled at extrapolating credible facts here.

2.8.9 Ninth book on social sciences methodological methods considered

A Tale of two Cultures: Qualitative and Quantitative Research in the Social Sciences

Author: Goertz, G. and Mahoney, J. (2012)

In the social sciences, the justification for qualitative and quantitative research may have been understood similarly. The main contrasts between these two techniques, which influence practically every area of social science investigation, including design, purposes, causal effects and models, principles and measurements, data analysis, and case collecting, are described and explored in this book. (Mahoney & Goertz, 2012).

2.8.10 Tenth book on social sciences methodological methods considered

Designing Research in the Social Sciences

Author: Maggetti, M., Gilardi, F. and Radaelli, C., C. (2013)

A researcher's ability to make progressive research decisions might be hampered by a variety of research repercussions for the breadth of social science research. The goal of this book is to assist researchers in making decisions when conducting research. The authors explain what is at risk when choosing a research design and discuss the trade-offs that researchers must make when weighing factors, including causality, categories, categorisation, meaning, analytical levels, and time (Maggetti et al., 2013).

During the planning stage of data management, research funders and academic institutions expect researchers to provide specifics on how they will handle their research project. The following part examines research data management planning to see how planning tools might help researchers ensure data quality.

2.9 Quality as an element of data management planning

A Data Management Plan (DMP) “is a document that describes how you will collect, organise, manage, store, secure, back-up, preserve, and where applicable, share your data” (Pienaar, 2014:6). Bishop et al. (2023:1) go on to say that “the inclusion of data management plans (DMPs) for research proposals has become a near-global expectation for research funding. Well-executed DMPs add to research transparency, accountability, and reproducibility”. A research data management plan is essential for determining the first steps toward resolving data management issues that can have a negative impact on data quality. To help reduce the detrimental influence on data quality:

A data management plan should not be thought of as a simple administrative task for which standardised text can be pasted in from model templates, with little intention to implement the planned data

management measures early on, or without considering what is really needed to enable data sharing” (Van den Eynden, 2011:6).

RDM planning is essential when dealing with study findings. When planning how to handle research data management in conjunction with the Research Data Lifecycle, all connected concerns can be addressed (CESSDA/ERIC, 2017; Cooper & Kahn, 2012; ICPSR, 2018; Wissik & Durco, 2015; CEOS, 2012; Higgins, 2008; UK Data Archive, 2017).

Ogier et al. (2014:101) make an important argument that “at the heart of every institutional research-related service lies the need to understand the nature and context of the data produced by its researchers”. This includes the use of a framework to describe how research results are contextualised and developed. To encourage secondary use, researchers must remember that the data that will be shared and kept must be of high quality, therefore it is critical to obtain and document acceptable quality assurance options during the data preparation phases. According to Rouchon et al. (2013:2), “the quality approach to preserving digital documents covers all the procedures aiming at ensuring a high level of quality of the digital object itself”; Subaveerapandiyan (2023:12) identified the major drivers of RDM as “storage, security, quality, compliance, preservation, sharing and influencing factors such as acceptance, culture, demand, incentives, roles, governance, politics, resources, projects, skills, communications, and context”. Organisations have developed recommendations and plans to assist researchers and data teams in evaluating, analysing, and enhancing their RDM efforts as part of larger campaigns to advance RDM. The management and preservation of the data quality approach is used as a criterion for the Comparability and Coherence quality elements. Quality assurance is an essential factor that needs to be taken into account when it comes to data management. The most significant influence on data quality occurs at the data collection and processing stage of the data management cycle. A DMP must include explicit references to the quality assurance and quality control procedures that are to be carried out; this is covered in the next section on quality assurance and quality control.

2.9.1 Quality assurance (QA) and quality control (QC)

In the process of managing data over the research data lifecycle, the relevance of quality assurance, quality control, and data quality cannot be overstated. Domingus (2015:2) emphasises that “quality assurance and quality control should occur throughout the data lifecycle”; this has the advantage of making research methods more efficient, consistent, and accurate. Quality control refers to “methods and procedures implemented to ensure that data are collected, managed, and utilised with accuracy and precision” (Reisch & Webster, 2004:2). This definition of quality control is used as a criterion for the Interpretability quality dimension, with the goal of learning more about how academics think about quality control. Quality control of data is an integral part of all research and takes place at various stages e.g. during data collection, data entry or digitisation, and data checking, the Open University (2023) adds satisfactory practice at all phases of the data lifecycle is required for quality assurance and, more

specifically, data quality. In a DMP, the manner in which quality assurance and quality control are addressed must be specified (DANS, 2010). Quality assurance and quality control should be discussed in the DMP project documentation section; this section answers questions such as what measures will be taken to record decisions made as the project progresses (this question is used as a measuring criterion for the Completeness quality dimension), as well as how data on field operations, coding decisions, variable construction and the like will be documented (DANS, 2010).

2.9.2 File Formats

A file format is a file structure that instructs a program on how to display the contents of a file (Rolland, 2016); this is supported by Kent (2023) in that “digital file formats are structured data containers that store information in a specific way, making it accessible and usable through various software applications. These formats can encompass a wide range of data types, including text documents, images, videos, and audio files”. The file format type should be established in the DMP to guarantee that the chosen file format is compatible with data-gathering and processing software packages. The inclusion of the format type as a criterion for the Uniqueness quality dimension has been employed. A researcher should explain how they intend to assure future access to data files, considering the file formats they have chosen. Researchers should be aware of the software required to display and process data files, as well as the software's expected lifespan. It is recommended that a researcher utilise standard, secure, widely used file formats, including non-proprietary formats, especially for final versions (University of California, 2023). The file format definition in a DMP is a quality assurance measure that safeguards data quality against any technological issues. The file format specification is used as a measuring criterion for the Consistency quality dimension.

2.9.3 Documentation

A DMP must include any related documentation that defines the data features, how to use the data and workflows. Users can turn to the related documentation for a quick reference to the decision taken when the data files were created or changed (University of California, 2023). Micheli et al. (2023:2) also highlight that “documentation can indeed contribute towards increased data sharing between EU companies, public administration, universities, citizens and civic organisations, facilitating the discovery and use of quality datasets” The documentation serves the following goals in this regard:

- “Readme files and data dictionaries to provide digital documentation of data characteristics, workflows, progress, results, and software.
 - Document any database data field (variable/column) characteristics for later interpretation. Possible elements are field name, field description, and permitted range of values.
- (Documenting variable and column features aids data interpretation because the codes used in

the variables column imply something to the user. This factor is used as a criterion for assessing the Uniqueness quality attribute).

- Document data file and collection characteristics that are relevant for later interpretation. Possible elements are file name and path, relevant dates, creation method, and status.” (University of California, 2023).

This research study needed a DMP, so the DMPOnline tool was utilised to generate one and contextualise its usefulness. DMPOnline “is a web-based tool for building a data management plan. The tool offers templates for data management plans from the major UK funders, plus a generic template which can be used for any research project. Plans can be exported for inclusion in a funding bid, and shared with colleagues” (University of Oxford, 2023).

The DMPOnline tool includes sections that are comparable to those listed above, but that are not explicit, such as:

- “What documentation and metadata will accompany the data?” (Documentation and metadata make it easy for secondary users to use and interpret the data). This question is used as a measuring criterion for the Completeness quality dimension.
- “How will the data be collected or created?”
- “What data will you collect or create?”

2.9.4 Metadata

A DMP must establish the metadata standard and format in order for secondary users to understand and use the data properly. According to the University of Wollongong Australia (2017:6), “Metadata is a set of data that describes and gives information about other data. Metadata summarises the basic information, which can make finding and working with particular instances of data much easier”. Metadata incorporation improves data exploration and utilisation, as well as data analysis. David et al. (2023:4) highlight that “repository selection, licensing, formats and metadata recommendations, including use of identifiers for building linkages to external systems that provide data interoperability, as defined by the FAIR core principles, must be listed and updated during the project”. Researchers might consider employing discipline-specific metadata standards and schemas and storing the metadata file near the data (or embedded if possible) to ensure exploration (University of California, 2023).

A DMP must include detailed information on the metadata that will be used; the common fields of metadata should include the following: Identifier type; Data content; Structure; Permissions (so that it is discoverable for future use); Resource Type (with general type description sub-property); Format; Subject classification code; Version; and Date. The section of a DMP pertaining to data quality and quality assurance needs to be carefully considered and should contain information based on industry best practices and standards. The quality assurance standards that will be used in a DMP should also be

specified. This technique is applied as a standard to ascertain the quality aspect of Quality Check Processes.

2.9.5 Short-term storage and data management infrastructure

The technological infrastructure within an organisation is critical to the effective adoption and advancement of RDM. This portion of the DMP Online Tool allows a researcher to identify their data management technical infrastructure support system (for backup purposes) as well as the roles team members would play in ensuring sufficient data management and curation. Some support specifications must be developed, for example, if local support is offered, a researcher must demonstrate that they have addressed and authorised the specifications (DCC, 2021). The position parameters within which investigators must operate are another important aspect in this area. According to the DMPonline (i.e. an open source DMPRoadmap codebase, which is jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3)) The DCC and UC3 work closely with research funders and universities to produce a tool that generates active DMPs and caters for the whole lifecycle of a project, from bid-preparation stage through to completion” (Digital Curation Centre, 2023); the researcher must:

- “Describe the planned quality assurance and back-up procedures [security/storage].
- “Specify the responsibilities for data management and curation within research teams at all participating institutions” (DCC, 2021).

Such requirements guarantee that members of the research team are aware of their data management duties. This criterion serves as a benchmark for the Completeness quality dimension.

In addition, the following questions must be answered in a DMP:

- “Are the investigators co-located, or will you need infrastructure that accommodates secure remote access?
- How will data quality be monitored if you are working in a distributed network across several sites?” (DCC, 2021); this question is used as a measuring criterion for the Timeliness quality dimension.

A quality assurance management system for researchers is one of the answers to the above problems since it directs researchers to prepare appropriately and to have systems in place to deal with the difficulty of operating in multiple locations at the same time. This demonstrates that the DCC DMPOnline provides answers to queries about improving quality assurance in research.

2.10 Data quality evaluation in South African RDM initiatives

South African organisations are supposed to examine data quality using their own frameworks and procedures. The table presented as part E of Appendix 15 shows a number of academic and government institutions in South Africa that contribute to data quality management. The South African RDM initiatives mentioned in Appendix 15 show South African academic and government institutions' ability in and appreciation of data management. The level of emphasis varies from institution to institution, but the use of techniques for better management of research data is common to all. The RDM methods used by an institution have to be properly incorporated into RDM. The next section discusses how RDM can be refocused to integrate data quality.

2.11 RDM refocused to incorporate data quality

In order to achieve data quality improvement, it is critical to emphasise RDM areas that are both scientific and technical and the quality of data can be refocused by applying RDM. In this section, each segment should be linked to the key factors of behaviour change outlined in Sections 6.10 and 2.2.3.

2.11.1 Address the human side of change systematically

Change is unsettling, and individuals are resistant to it; any shift in meaning can bring problems for people (Burton, 2023). People can be hostile to the idea of changing their habits and routines, especially if they don't grasp the greater goal (Emerson, 2022). Thus, researchers and research teams must be informed about the benefits of RDM in addition to data quality. As part of the human aspect of transformation, it is also necessary to inform research teams about the type of change that is coming in order to reassure them that it is in their best interests. The usual adjustments applied as part of change management include that “new leaders will be asked to step up, jobs will be changed, new skills and capabilities must be developed, and people will be uncertain and will resist” (PWC, 2004:4). In order to be proactive, Lawton and Pratt (2023) suggest that “to be effective, the change management strategy must take into consideration how an adjustment or replacement will impact processes, systems and employees within the organisation. There must be a process for planning and testing change, communicating change, scheduling and implementing change, documenting change and evaluating its effects” For researchers and research teams aiming to achieve good data quality, change management entails a positive strategy that addresses issues on a case-by-case basis. For researchers and research teams aiming to achieve good data quality, change management entails a positive strategy that addresses issues on a case-by-case basis. Waddell and Sohal (1998:543) claim that “resistance is often viewed by managers as the enemy of change, the foe which must be overcome if a change effort is to be successful”. Resistance to change will derail the change management vision and cause unanticipated disruptions in the process. This means that it is critical to have a plan in place to deal with certain negative human elements. According to PWC (2004:4) “a formal approach for managing change –

beginning with the leadership team and then engaging key stakeholders and leaders – should be developed early but adapted often as change moves through the organization”. This technique will ensure that research teams in a social sciences organisation are up to date on the proposed changes in RDM and are implementing them correctly. One significant barrier is that change management is a job in and of itself. (Tushman et al., 2017). The respective leaders in an organisation must have common ground that specifies the expected improvements inside the organisation.

2.11.2 Change starts at the top and begins on day one

It is critical for senior management to demonstrate the shift they expect in order to create a data management culture that supports good data quality. This is accurate given that “our ability to manage change hinges on our ability to manage ourselves, including our fears and anxieties. These factors drive everything else, in business and in life” (Rinne, 2023). This can be accomplished by closely monitoring the key factors of behaviour change (Witte, 2015), as discussed in Section 2.2.3.1. Kotter (1997:19) emphasises that “an organisation’s ability to learn and adapt may also influence the success or failure of the change management program”. A suitable learning environment is crucial for researchers to understand the significance of high-quality data-aiming systems. Top management should provide both the learning environment and the intended change to help workers overcome resistance to change, promoting a more effective change management process. Hultman (1995:16) points out that “some employees resist because they believe the change will make it harder for them to meet their needs, as such they see the change as a threat, instead of a potential advantage”. To overcome this obstacle as suggested by Kumarasinghe and Dilan (2022:53) “getting an organizational change program off the ground necessitates the aggressive participation of many employees”.

In order for workers to fully embrace change management, top management must demonstrate effective implementation and roll-out of the intended change management initiatives. PWC explains (2004:4) that “the leadership must change first to challenge and motivate the rest of the institution, speaking with one voice and ‘walking the talk’ to model desired behaviour”. This is critical because, when senior management changes the way they do things, researchers and research teams can understand the level of priority necessary. Porras and Robertson (1992), on the other hand, make a significant point, that maintaining permanent change is a challenging endeavour because most businesses are made up of multiple communities, making it difficult to manage change. Hubbart (2023:5) highlights a crucial component of workers' acceptance of change in that “it is important to acknowledge that sometimes employees are change averse because they are experiencing organizational incompatibilities and conflicts that confuse proximate and ultimate productivity deficiencies”. Changing the actions, activities, and working principles of employees, like changing any other kind of management division in a firm, requires researchers to accept the principles of data quality.

2.11.3 Real change happens at the bottom

Data managers and the team on the 'ground' are the accountable custodians who work with the 'baton' (Hubbart, 2023). The effectiveness of any programme that directs researchers to deliver high-quality data “include[s] plans for identifying leaders and pushing responsibility for design and implementation down through the organization” (PWC, 2004:6). This will aid in the implementation of the desired change at all levels of an organisation. Researchers and research teams must understand the significance of data quality and the implications of poor data. When researchers are involved in this way, they can contribute to a broader organisational style of doing business as well as accomplish specific jobs better (Swedberg & Douglas, 2005). As a result, it is critical to identify the sorts of organisational change that will be undertaken and plan for keeping staff informed. Management can solicit feedback as you implement the change and then make changes to your change management strategy to ensure that your team has the resources it needs to maintain strong morale and assist the change from their end (Whatfix, 2022).

Finally, the research units of an institution will have a common vision and understand what is expected of them in terms of data management and high-quality data outputs.

2.11.4 Confront reality, demonstrate faith and craft a vision

Researchers have their own working preferences and research styles, and while research teams may not always adhere to the original vision, it is reasonable to expect worries about how changes may affect them. The creation of good data quality and research data management include a personal dimension that necessitates special statements to persuade people (researchers) of the importance of change. In this regard “a formal case for change and creating a written vision statement are invaluable opportunities to create (or force) leadership team alignment” (PWC, 2004:6). If research teams are matched with a structured case for change management, any realistic improvements that can be made would be easy for the team to understand. Rass et al. (2023:88) explain that leaders should “identify a need for change, creating and advertising a new vision and engaging their employees through inspirational influence, to establish the rationale for change that is needed to initiate the change process within the organization”. A key factor that is important for organisations that want to handle change is the awareness that “for organisations to survive in the 21st century, they must shift from traditional practices of management to contemporary practices characterised by attributes commonly described as vision; contemporary values; quality mindset; stakeholder focus; speed orientation; innovativeness” (Kotter, 1997:23). These traits are still vital in today's information and knowledge age.

2.11.5 Create ownership, not just buy-in

The intricacy of managing research data and the benefits of high-quality data present a step forward in terms of proving its feasibility and beneficial effects. In large part, it is practical to let researchers work

toward acquiring high-quality data. Therefore, change agents need to “over-perform during the transformation and be the zealots that create critical mass for change in the workforce” (PWC, 2004:7). This will help the organisation present a united front, and workers are more likely to feel comfortable making changes if they believe they can own the change they are being asked to make. Furthermore, Burton (2023:2) emphasises the concept of ‘Awareness of the need for the change’, “this represents a person’s understanding of the nature of the change, why the change is being made and the risk of not changing, and addresses the individual’s ‘What’s In It for Me’ (WIIFM) considerations”. Mabin et al. (2001:169) underline the importance of “prerequisites, in the context of change, should include amongst others, vision, mission, leadership, culture and communication because if these are not met, the change will not succeed due to what is often termed resistance to change”. Employee buy-in will not result in the necessary adjustments until employees believe they own and contribute to the change they require. Employees should demonstrate “ownership that’s willing to accept responsibility for making change happen in all of the areas they influence or control” (PWC, 2004:7). Clear ownership fosters a healthy working environment that researchers and research teams can proudly own and be responsible for, as well as a productive work culture. All change planning should involve the vertical and horizontal structures of the organisation concerned to allow for ownership of the intended change; this offers a sense of ownership of the change on the part of all employees (Edwards, 2020).

2.11.6 Practice targeted over-communication

Due to their unavailability, it may be difficult to get the 'change management' message across to researchers in a social science organisation, given their hectic schedules and work plans. According to Ford and Saren (2006:54), this undesirable situation can be stabilised when “organisations are required to foster good coordination, strong leadership, and clear communication in order to exploit and develop their resources”. Senior managers who are serious about change management “should develop a communication plan, a road map for change sponsors, integrated training programs and a plan for dealing with resistance” (The Society for Human Resource Management, 2023). These strategic leadership tactics are essential for persuading all workers (researchers) to agree on a fair deal that will improve work processes' efficiency. Targeted communication and effective leadership are critical in preparing the organisation for transition as they guide the organisation through difficult times (Handy, 1978).

It is critical to increase the interaction between researchers who provide high-quality data and change management, which might be accomplished by putting across “core messages through regular, timely advice that is both inspirational and actionable. Communication is both outbound and inbound. It should be targeted so as to provide employees with the right information at the right time” (PWC, 2004:8). Using this technique gives the necessary platform for researchers to ask questions and explore any potential roadblocks. Contact is vital in the daily work lives of those who are expected to drive the

suggested change management. This implies that “for successful change to occur, organisations are required to foster good coordination, strong leadership, and clear communication in order to exploit and develop their resources” (Ford & Saren, 2006:56).

2.11.7 Explicitly address culture and attack the cultural centre

Employees represent a mix of shared history, unique values and ideas, and shared attitudes and actions in a business (Nadler & Tushman, 2015). The goal of ensuring that researchers supply high-quality data must be accomplished in a way that does not compromise an organisation's basic objective. The most practical way to achieve this is to develop “a baseline through a cultural/organisational diagnostic, defining an explicit end-state or desired culture, and devising detailed plans to make the transition” (PWC, 2004:9). Moreover, “learning culture, inquiry and dialogue, and knowledge-sharing structure could also be alternative strategies for fostering change commitment” (Malik & Garg, 2017, as cited in Syafwan & Fathia, 2023). Culture may become a deeply embedded phenomenon that requires a great deal of effort and time to alter or eliminate. According to Beckhard and Pritchard (2012:37), “change management is most effective when learning and change processes interact positively with each other and when both employers and employees realise that change is a learning process and learning is a change process”. This awareness is critical because it pushes researchers to question every aspect of the larger change management approach when it comes to data quality improvement.

2.11.8 Assess the cultural landscape early

The way researchers treat their research data has its own culture, but it must be aligned with the company culture. The objective of a social science organisation is to conduct social science research and to produce viable research outputs that may be used to influence decision-making. In essence, the way research is conducted and managed creates unstated and ambiguous cultural expectations at various levels inside an organisation. PWC's (2004:9) interpretation is that successful change programmes pick up speed and intensity as they cascade down, making it critically important to understand and account for culture and behaviours at each level of the organisation. Mazikana (2023:9) further affirms this, stating that “Change efforts, including the adoption of new systems, change programmes, and change champions, are one of the ways that organisations work to improve their operations”. To accurately assess organisational readiness to change, bring important problems to the surface, uncover cultural elements that favour or hinder change, and target leadership and resistance sources, a full cultural diagnosis must be conducted early on (Dunphy, 2006). Armenakis and Harris (2009:137) believe that “there are many factors that can be used to categorise differences among change recipients that may affect their experience of the five beliefs. These differences can be social differentiation (their group memberships) and individual personality differences (openness)”. It is also vital to consider the cultural environment that affects an organisation in order to find potential options for resolving the change that is occurring in that institution. Recognising the reasons why adjustments are needed in an organisation

is an encompassing notion that necessitates the evaluation of the cultural environment. This will enable an organisation to consider the strategy it will employ to transition from its existing cultural environment to its desired one.

2.11.9 Prepare for the unexpected

The envisaged RDM organisational culture must be controlled and regulated in order to forecast unexpected events that may arise (Rass et al., 2022). People (i.e. researchers) are in charge of producing high-quality data. According to PWC (2004:11), “individuals will react in unexpected ways, areas of anticipated resistance will fall away, and the external environment will shift”. Working toward a researcher's vision of delivering high-quality data necessitates balancing change management with the removal of barriers that prevent an organisation from accomplishing this transformation. To keep the momentum of 'change management,' change management leaders must make necessary improvements. According to Robbins (1990), design change should not be haphazard; it should be planned with the expectation that employees' reactions will be opposite to what is expected. As a result, it is critical to ensure that all change activities are well coordinated with all key stakeholders (Dunphy, 2006). This is substantiated by Agnemyr & Bhargava (2023:3) who say that “for those that prepare well and in the right way – opportunities will most definitely present themselves”.

2.11.10 Speak to the individual as well as to the institution

Researchers' efforts result in the research that a social science organisation produces. This implies that in order to guarantee the production of high-quality data, the required adjustments must be made personally. This may influence how RDM initiatives aimed at improving data quality are perceived. One researcher and one research team at a time should be expected to make improvements. Mohebi (2023:30) emphasises a key aspect to remember: “unanticipated changes can have profound impacts on the organisation and its stakeholders, requiring rapid and effective responses. One of the common consequences of unanticipated changes is that the organisation may have to deal with new large-scale projects that were not part of its original plans or priorities”. It is important for researchers and research teams to know “how their work will change, what will be expected of them during and after the change program, how they will be measured, and what success or failure will mean for them” (PWC, 2004:11). In the same way that they interact with their work (data), researchers also need to be involved in and engaged in the process of change. Researchers and analytic teams are the primary guardians of improving data quality, with the exception of change management effectiveness. The range of formal and informal interactions influenced by those who operate them, namely the workers themselves, should be the focus (Nadler & Tushman, 2015). Employees require a transparent and consistent communication approach to prevent them from opposing the suggested RDM change management. Failure to discuss these matters with employees within an organisation can result in a poor trajectory of advancement. This is because “employee resistance has been identified as a critically important contributor to the

failure of many well-intended and well-conceived efforts to initiate change within the organization” (Bolognese, 2002:2). It is thus critical to emphasise that “leaders often spend a great deal of time communicating about the proposed change in order to gain buy-in before beginning the change initiative. However, change management strategies often fail or fizzle out when leaders don’t communicate enough after the initiative is announced” (Emerson, 2022).

The domains of change management listed above address negative issues that can hinder or discourage the necessary change management involved with social science researchers producing high-quality data. Change management is a critical component that necessitates careful planning and execution. The aforementioned areas promise to put a company on the correct track to effective change management in the generation of high-quality data by researchers. The following section covers all of the quality indicators defined and discussed earlier in this chapter.

2.12 A consolidated view of the data quality indicators

In light of the above discussion, a consolidated or integrated view of those factors regarded to be data quality proxies – both quality dimensions and related quality measures – was achievable. Some of the measures that may be used to provide an indication of the level of quality compliance have already been identified. Appendix 4 reflects both the dimensions and the potential measures. Appendix 3 lists the many authors and stakeholders who were evaluated for the data quality initiatives and definitions for each data quality dimension.

2.13 Summary

This chapter discussed various RDM definitions and outlined a strategy for developing a unique concept for this study. RDM can now be discussed in greater depth. Additionally, many areas of data quality and quality assurance were discussed in this chapter. To assist development, the present image of data quality from literature, issues and the philosophy of data quality management in terms of institutional, person (researcher), organisational, policy, technical, international role-players, guidelines, standards, and best practices were investigated. It can be inferred that the concept of data quality is usually addressed in one way or another but, in some cases, not at all.

The literature review revealed which aspects of RDM are best suited to act as measuring parameters and also sought to incorporate attempts to improve data quality. In the bulk of the RDM literature that was studied (i.e. guidelines, standards, best practices, and evaluation systems), data quality is not directly tied to the phases of the research data lifecycle, representing a shortcoming of this study.

It was critical to use the reviewed literature to define relevant measurement criteria that could be used when collecting and processing data to quantify crucial activities. The proposed quality framework is based on data quality dimensions established utilising the DAMA UK quality dimensions – as well as

the South African Statistical Quality Assessment Framework, due to its vast data quality coverage and unique data quality dimensions. While recommendations for data quality appear in the literature, there do not appear to be any working models offering this facility for research in the social sciences, particularly for organisations in developing countries. However, this chapter helped to illustrate all of the important components required for improving the quality of research data, as well as those of minor importance. The quality framework is intended to help researchers and data curators test the input needed to benchmark and deploy the organisation's best RDM solution. This chapter paved the way for the proposed quality framework to be developed on a solid foundation, and the literature can now be utilised. For the recommended quality framework to be implemented successfully, a well-considered and well-informed theoretical foundation that is essential, sustainable, and progressive must be applied. This chapter's expertise was crucial in determining which measurement criteria are the most satisfactory, and it will help detect discrepancies, propose improvements, and eventually alter the tide of data quality in the Social Sciences Research Organisation. The proposed quality framework will be built in the next chapter.

CHAPTER 3: A RELEVANT FRAMEWORK TO GUIDE DATA QUALITY

3.1 Introduction

The purpose of this chapter is to conceptualise the learning acquired in Chapter 2. The information gathered in Chapter 2 was utilised to create a theoretical framework for determining data quality in a social science organisation. To construct an effective measuring framework, the ideas gathered from the reviewed literature were merged to match the context in which the concepts will be used. A process framework was then developed to incorporate quality evaluation.

The chapter begins with an overview of frameworks and framework development. According to Carter and Rogers (2008:362), it is feasible to “characterize a theoretical framework as a collection of two or more interrelated suggestions that explain an event to provide understanding or propose testable hypotheses”. In essence, the backdrop for the development of the proposed quality evaluation framework is established. The principles that should be considered when developing a framework are also discussed in this chapter. The principles aided in the alignment of quality dimensions and phases in a data management cycle, as well as defining ideas behind the creation of the framework. This chapter identifies various examples of quality frameworks that were used to construct the suggested quality frameworks. The author's suggested frameworks for quality-focused RDM at a social sciences organisation are discussed at the end of the chapter (in Section 3.8). The next section discusses frameworks and framework development.

3.2 Defining frameworks and framework development

This section will go through the definition of a framework as well as the framework development process. There will be further discussion about the assumptions made for the proposed quality process framework and, thereafter, the definitions offered will be summarised.

3.2.1 Frameworks

Various development frameworks have been proposed in the literature (Holland et al., 2018; Jabareen, 2009; Ravitch & Riggan, 2012; Roberts & Khattri, 2012; Vetro et al. 2016). The development of a framework allows for the creation of individualised and well-informed ideas and principles that may be utilised to solve a problem or a set of problems in a single construct. (University of Southern California, 2020). According to Abraham et al. (2019:426), a framework “explains either graphically or in narrative form, the main things to be studied – the key factors, constructs or variables – and the presumed relationships among them”. In order to obtain ideal results, the relationships between variables in a framework must operate together. A framework must have a specified function that can be used in a variety of ways. A framework is a set of components that are linked together to serve as a guide for

improving their usefulness. The definition posited by the University of Southern California Libraries (USC Libraries, 2020) provides a precise description of what a theoretical framework is in the context of this study (i.e. the building of a data quality framework): “a framework consists of principles and the latest theory that is used for your specific research, along with their meanings and reference to applicable empirical literature” (USC Libraries, 2020). Within a research context, a framework must reflect the basic meanings of a theory or theories that are connected to a specific subject. Essentially, “framework outlines 5 categories of institutional carriers for data practices: (a) norms and symbols, (b) intermediaries, (c) routines, (d) standards, and (e) material objects” (Mayernik, 2015:2). To demonstrate the need for and importance of a framework, the following strengths of a framework were recognised for this study:

- The framework provides the researcher with established information. A basis for the conclusions is provided, driven by a relevant theory (USC Libraries, 2020).
- A framework outlines the main dimensions to be explored, the primary factors or variables and the assumed relationships, either graphically or in a narrative form (Anfara, 2008).
- By articulating the assumptions of a research study, a framework forces you to address questions about why and how. This helps you to move critically from simply explaining a phenomenon you have encountered to generalising about different aspects of that phenomenon (USC Libraries, 2020).
- A framework is a network of interconnected principles that, together, provide a detailed explanation of a phenomenon or phenomena (Jabareen, 2009).
- A framework determines which key variables affect a phenomenon of interest and illustrates the need to analyse how these key variables can vary and under what circumstances (USC Libraries, 2020).

These characteristics align well with the creation of a data quality framework. When dealing with framework literature, the strengths highlight significant features that must be considered. Creating a theoretical framework is not an easy task. According to USC Libraries (2020), a framework is made up of concepts that are gradually built together with their definitions and references to pertinent scholarly material. The author was able to identify RDM and organisational aspects that are assumed based on the literature review. Defining assumptions is an important part of establishing the context in which a framework will operate. The next section briefly describes framework development and the assumptions that were employed when building the proposed quality process framework in order to adequately enable the formation thereof.

3.2.2 Identifying assumptions

According to USAID (2022), “assumptions reflect our recognition that there are factors beyond our control that are necessary for successful achievement of objectives at all levels of the project”. Identifying assumptions that apply to this study was a crucial step in building the proposed quality framework. Identifying assumptions helps in assessing the accuracy and validity of a framework. The explicit stating of assumptions allows for scrutinising whether they align with reality, gathering data to validate them, and refining the framework as needed.

The assumptions recommended for this study are based on evidence discovered in the literature as well as broad assumptions that support the research goals and research questions of the study. The assumptions thus identified were broken down into four main assumptions and three minor assumptions.

- Accepted Assumptions
- The significance of RDM is understood by researchers.
- Organisational systems have an impact on a researcher's RDM methods.
- The importance of data curators in influencing how researchers manage their data cannot be overstated.
- RDM guidelines, which advocate for improved data quality, are implemented at the institutional level.

These assumptions are based on an organisation's implicit acceptance that the data it produces is of high quality. When reviewing the literature, the focus was not on case studies of organisations that lacked the quality component of RDM. All of the assumptions in this section help to draw clear conclusions regarding how data quality is maintained, the level at which the framework controls data quality, the capacity to conduct quality evaluations, and the potential RDM remedies once the quality framework is implemented. The assumptions also aid in determining the focus area on which the framework's variables should be based. Frameworks come in a number of different forms. The purpose of this discussion is to choose a framework that is appropriate for this research study. The following part goes over the many sorts of frameworks and then identifies the one that was chosen for this study, as well as the reasons for this choice.

3.3 Frameworks typology

The many framework types each have a distinct purpose. Theoretical and conceptual frameworks are the two basic types of frameworks (Anfara, 2008; Duranti et al., 2008; Ngulube & Mathipa, 2015). A third category, process models, will also be discussed. Adom and Hussein (2018:439) are of the view that “choosing a relevant framework type depends on the discipline or field of research, even within a particular discipline, a specific theory or theories that resonate with the area of inquiry must be

selected”. The full exploration of a topic is enabled by selecting a relevant framework. A theoretical framework differs from a conceptual framework. With this difference in mind, the following section delves into the differences between the two concepts in order to better understand how they are applied in this study.

3.3.1 Theoretical framework

A theoretical framework is made up of concepts and their definitions, as well as references to important academic literature (USC Libraries, 2021). A research challenge could be coming up with an entirely new topic or filling a knowledge gap. By integrating information from several sources, a theoretical framework can be used to examine a certain topic by synthesising “existing theories and related concepts and empirical research, to develop a foundation for new theory development” (Rocco & Plakhotnik, 2009:127). The creation of a new theory fundamentally establishes that there was a knowledge gap, and the theory explains the presence of the research topic under investigation.

At the operational level, Rocco and Plakhotnik (2009:127) explain, “a theoretical framework involves the presentation of a specific theory, such as systems theory or self-efficacy, and empirical and conceptual work about that theory”. All of the adopted elements that were brought together to form a single unit should be aligned and described in the presentation of a theory. In so doing, a researcher uses a “theory to build his/her own research inquiry – it serves as the foundation upon which a research is constructed” (Adom & Hussein, 2018:438). A research topic does not exist in a vacuum; there are theories and literature that are relevant to a particular research topic. A theoretical framework “provides a general or broader set of ideas within which a study belongs. It refers to specific or narrower ideas a researcher utilises in his/her study” (Adom & Hussein, 2018: 440). A theoretical framework uses already existing theory in an inquiry field that is reflected in the assumption of a research study. When a study is designed around a theoretical framework, it suggests that “the theory is the primary means by which the research problem is understood and investigated” (Walden University, 2024). A researcher then seeks to further explore variables that may influence or change the initial theory to arrive at a new theoretical understanding.

In conclusion, the properties of a theoretical framework enable a researcher to evaluate or embrace current ideas while also developing a new theory. This new theory can then be applied to a new set of research problems.

3.3.2 Conceptual framework

Grant and Osanloo (2014) suggest that a conceptual framework is used to guide a research project by drawing on concepts from various theories and research findings. According to Rocco and Plakhotnik (2009:126) “a conceptual framework grounds the study in the relevant knowledge bases that lay the foundation for the importance of the problem statement and research questions”. Concepts are brought

together and studied to establish key problem areas and identify those that are important enough to be investigated. A conceptual framework “is made up of theoretical and empirical work relevant to the manuscript’s purpose, where the purpose is not to further investigate a specific theory” (Rocco & Plakhotnik 2009:126). When a researcher employs a conceptual framework, their world-view is recognised, and their assumptions and preconceptions about the subject matter are outlined.

The knowledge base in a conceptual framework is knowledge that is created rather than discovered. With a conceptual framework, pieces of knowledge that originate from elsewhere are brought together, but the structure, the overall coherence, is something that is built from scratch, not something that exists ready-made (Ravitch & Riggan, 2012). Green (2014:3) emphasises that “where a framework is based on concepts, the framework should be called a conceptual framework and where it is based on theories it should be called a theoretical framework”.

A conceptual framework is made up of one or more well-known theories (in part or in whole), as well as other ideas and practical findings from the literature. Reviewing relevant literature is an important aspect of the conceptual framework since it seeks to show connections between ideas (concepts) and how those ideas apply to a specific research topic. Adom and Hussein (2018) created this table to demonstrate the fundamental differences between a conceptual and theoretical framework:

Table 3.1 - The differences between a theoretical framework and a conceptual framework

Theoretical Framework	Conceptual Framework
Well developed, designed and accepted.	Its design is not accepted, but it is a proposal of the researcher's answer to the research problem s/he has defined.
Offers a focal point for approaching the unknown research in a specific field of inquiry.	It is the framework that shows logically how the research inquiry is to be undertaken.
Consists of theories that seem interrelated with their propositions deduced.	It consists of concepts interconnected to explain the relationships between them and how the researcher seeks to answer the research problem defined.
Used to test theories, to predict and control the situations within the context of a research inquiry.	It is aimed at encouraging the development of a theory that would be useful to practitioners in the field.

In conclusion, conceptual frameworks connect concepts together to provide a clear picture of what already exists, whereas theoretical frameworks strive to look at existing theories and then generate new theories. The process model, described next, is the last of the frameworks to be discussed.

3.3.3 Process model as a framework

According to the American Productivity and Quality Center (2022), process models are tools that make abstract, complex work processes more tangible by providing a concrete illustration of how processes work and how they fit together. The Federation of American Scientists (2020:1) defines a process model as “graphical representations of existing or planned processes” and states that they are “excellent resources for understanding what functions are performed and how those functions interact”. A primary purpose of a process model is to facilitate the discovery of information needs within the enterprise. Researchers' research data management strategies are functions that this study tries to comprehend. The analysis is done in conjunction with the numerous activities of each step of the research data lifecycle to achieve this goal. The UK Data Archive data cycle depicts the lifecycle stages that data go through during the course of a research project. These stages are made up of smaller tasks that help with data management in general. The goals of three possible process models are represented in Table 3.2, below:

Table 3.2 - Typical process models

Goal	Description
Descriptive	Is concerned with producing an explicit and accurate representation of an organisation's actual processes, for purposes such as documentation, dissemination, analysis, and improvement (Munch et al., 2012).
Prescriptive	Prescribes the activities the inputs and outputs of the activities, how quality assurance is performed and how change is managed (University of Cape Town, 2018).
Explanatory	Allows “the construction of hypotheses about unobservable processes and structures that can be used to explain observable phenomena” (ScienceDirect, 2012).

According to Azimaee (2015), a process model is roughly an anticipation of how the process will look. With this in mind, it is possible that any of the process model types may be used here. However, a prescriptive process model best aligns with the goals of this study. It enables the charting of social sciences-related quality dimensions on the UK Data Archive data cycle model, resulting in a revised process model. All of the quality assurance processes that a researcher (as well as a curator) conducts could be plotted on the lifecycle because the quality of the data is altered across the different stages of a data lifecycle.

The goal of the quality framework is to establish new best practices and principles for documenting data management in a social science organisation. The quality dimensions introduce a refined and rigorous data management system, particularly for creating and processing data.

Only once the quality dimensions have been unpacked into quantifiable criteria and documented in a changed evaluation framework can the difference (that the extra quality dimensions will bring to the UK Data Archive data cycle) be quantified. It was necessary to identify and build both the dimensions and the evaluation elements into a descriptive process framework.

In conclusion, based on the discussion in this section, as well as the variations in framework types indicated above, the type of framework chosen for this study is a process model as a framework. The suggested quality framework incorporates elements from existing theories found in the literature (Chapter 2). The body of literature examined in Chapter 2 supplied resolutions that were assembled as interrelated concepts to aid in the development of an evaluation framework to support the process model. Furthermore, the theory on RDM, quality assurance, and data quality that was studied offered enough information to establish the initial building blocks (variables) of a quality framework. The framework's variables, as well as the measuring indicators, were identified through a review of the literature. The process model can be changed after data has been collected and insights from study participants have been obtained. The quality characteristics and measuring criteria are expected to be expanded and updated to incorporate data quality from the perspective of both the social sciences researchers and informed social sciences data curators.

To clearly visualise the link between the quality dimensions and the lifecycle phases, mapping the quality dimensions on a data cycle model, such as that of the UK Data Archive, is required. It is critical to understand the concepts that must be borne in mind when designing a process model before developing a framework and this is the purpose of the following section.

3.4 Principles for the development of a framework

Mandinach et al. (2006) 'Conceptual Framework for Data-Driven Decision-Making Skills' was utilised to give guiding principles for the construction of the descriptive process model for measuring data quality. Mandinach et al. (2006) developed a 'Conceptual Framework for Data-Driven Decision-Making Skills'. This conceptual framework is:

...founded on the notion of what it means for an educator to be data-driven. The assumption here is that individuals, regardless of where they are within an academic system, have questions, issues, or problems for which data must be collected, analysed, and examined in order to make informed decisions (Mandinach et al., 2006:6).

There is a continuous relationship between data, information and knowledge required for decision-making, which eventually informs how data can be translated into functional knowledge. The principles

behind Mandinach et al.'s (2006) framework are visualised in the figure below. The process of converting data into knowledge necessitates making informed decisions and it is critical to put these decisions into context:

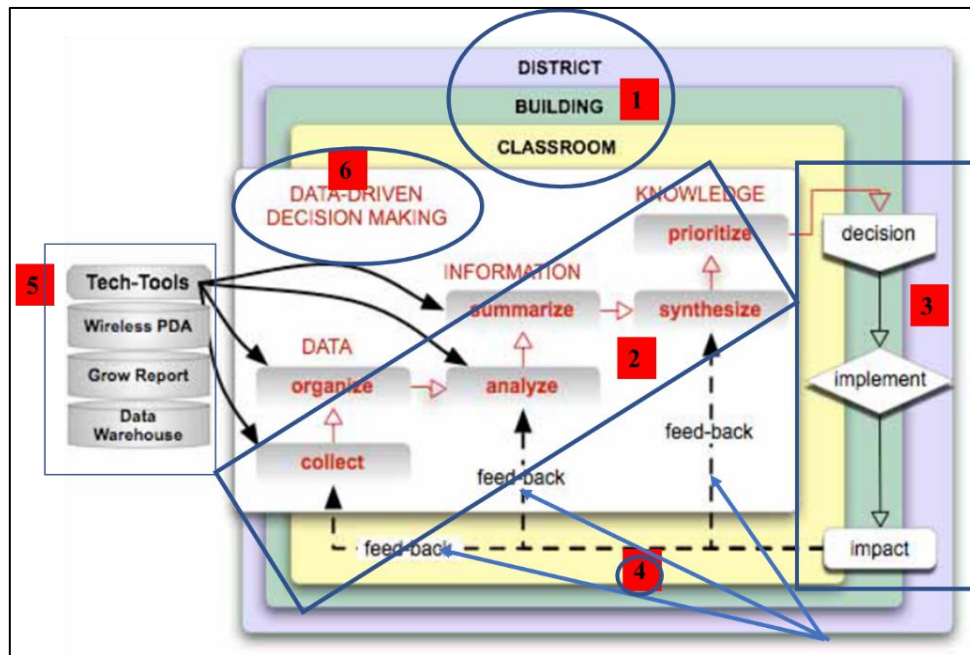


Figure 3.1 - Conceptual framework for data-driven decision-making skills
(Mandinach et al., 2006)

The framework of Mandinach et al. (2006) is based on the principles laid out below. While Mandinach et al. use a school context to illustrate the principles, they are also broadly applicable to a social science organisation.

Principle 1: Provide context.

For Mandinach et al. (2006:8), the context is provided by the district, the building and the classroom. The foundation of the conceptual framework is provided by the data-to-knowledge continuum, which is “grounded within the context of the classroom, the school, and the district, all of which will use different data in different ways to make decisions”.

There are specific reasons behind the development of a theoretical framework and these reasons are informed by a specific context. Rogers et al., (2020:2) expound that “to gain a greater understanding of implementation, it is essential to monitor these contextual conditions in research as the data generated is context-dependent. Literature suggests that theories, models and frameworks provide greater insight into the mechanisms of implementation”. The context in which a framework will exist and operate has a significant impact on its implementation success.

Principle 2: Identify all the crucial activities.

The attainment of desired objectives means that a theoretical framework must continue supporting existing processes or introduce new procedures. In the Data-Driven Decision-Making framework “the data-to-knowledge continuum is defined by the inclusion of six cognitive skills or actions that identify as crucial to the decision-making process” (Mandinach et al., 2006:8). These cognitive skills are linked to the data-information-knowledge continuum provided in the middle of the model and include the ability to organise, collect, summarise, analyse, prioritise, and synthesise (Mandinach et al., 2006:8).

It is important for an organisation to measure and identify all crucial activities and the current state of their processes when implementing a process framework to “identify process gaps or redundant processes, determine which processes provide value, and engage employees in process activities” (Lyke-Ho-Gland, 2015:3). Once critical activities have been identified, an organisation can benchmark its current state.

Principle 3: Show the correlations between actions.

There is a link between data, information, and knowledge, all of which have actions that are related. The management of these three resources necessitates unique actions.

Mandinach et al. (2006:78) explain that:

“...at the data level, users collect and organize data. Collecting data is typically the first step in a data-driven inquiry process. For example, teachers will collect work samples, classroom assignments, student portfolios, and other student performance data. These data may be supplemented with demographic data, health data, or behavioural data.”

After data are collected, they must be organised in a logical manner to make sense; an example of an action that may be used is triangulating among multiple sources of data and arranging them in a manner that will allow them to make sense (Mandinach et al., 2006). The information level of data entails users analysing and summarising information.

They use context in which to ground the data and transform them into information. The teachers will analyse the data, examining performance trends, drilling down to item levels, and looking at aggregated and disaggregated data to try and make sense of performance patterns” (Mandinach et al., 2006:78).

The final level is the knowledge level. At this level:

“Users synthesize and prioritize knowledge. The information then is synthesised in a way that helps the teachers form a knowledge base about student performance from which instructional decisions can be made. Finally, the teachers must prioritize among the information and knowledge to determine what courses of action to take” (Mandinach et al., 2006:78).

In order to use data in a meaningful way, the way a user navigates the analysis, interpretation, and management of data, information, and knowledge necessitates cognisant activities that are related.

Principle 4: Identify feedback loops.

De Lemos and Potena (2016:11) suggest that “in a feedback loop, uncertainty is associated to different sources (e.g. the environment), and appears in different forms (e.g. as noise in variables or imperfections in techniques being used)”. Identifying feedback loops in a process model must assist in knowing how well the identified actions are delivering the expected results. Mandinach et al (2006:74) explain that “there is a cyclical process of collecting and examining data, and applying the results in a feedback loop”. As illustrated in the Framework for Data-Driven Decision-Making (Figure 3.1, above), knowledge informs a decision that is implemented and produces an impact. The feedback loops (marked as dotted lines) are “diagnostic systems that provide almost real-time data that provide a tight feedback loop between assessment and prescriptive instruction” (Mandinach et al., 2006:78).

Principle 5: Identify any external influences that may have an impact on the system.

External influences can affect a system (also known as a process model) in either a positive or negative way by introducing factors that destabilise actions and processes. A researcher has the option of collecting new data or examining existing data sources. As a result, it is critical to recognise any external impacts so that any influences may be countered with creative solutions that allow the system to function properly. Data must be systematically formatted in some way in order to make sense of it and for it to be used efficiently. This is a useful concept within a framework because it is practically impossible to extract any meaning from raw data that has not been properly structured. This organisational scheme then enables the stakeholder to turn the raw data into information to enforce significance (Mandinach et al., 2006).

Principle 6: Understand the outcomes of the process.

The decision-making framework aids in understanding how to stabilise the components of a framework that will enable the process from effective data collecting to measuring the impact of decisions made and feeding back the learning acquired. The results of the process must be understood in order for them to make sense in the context of achieving goals. McDonald (2018) provides a comprehensive explanation relating to understanding the outcomes of the process:

You can use the process model you created to provide a ‘big picture’ view of your initiative and identify backlog items needed to support the process. This exercise can be done concurrently with or shortly after modelling the process, depending on how much time you have available” (McDonald, 2018:1).

Understanding the process outcomes aids in the creation of a big picture of the model and the making of any necessary changes.

Although the concepts are sourced from several domains, they provide a common understanding of framework development principles that can be applied to the social sciences. The framework development strategy that was principally employed to construct the quality framework model is discussed in the next section.

3.5 Framework development strategy deployed: University of Kansas process framework

The University of Kansas framework development strategy was chosen as the primary strategy for developing the suggested quality process model. This model was chosen because it allows for a systematic and orderly framework development process. Other resources that were considered did not include this methodical approach (Statistics South Africa, 2010; Roberts & Khattri, 2012; Mayernik, 2015; Vetrò et al., 2016; Holland et al., 2018; USC Libraries, 2021). The University of Kansas (2023) framework development strategy aims to assist patrons in coordinating their thinking and orienting programme design around intended outcomes rather than (often restricted) resources. The goal of this strategy is to provide a visual representation or roadmap (i.e. problem theory, action theory, logic model or framework) (Center for Community Health and Development, n.d.). Although the University of Kansas (2023) process framework does not have a formal diagram, the author was able to depict a basic framework figure after acquiring insight into the process framework. Figure 3.2, below, is the author's depiction of The Center for Community Health and Development framework:

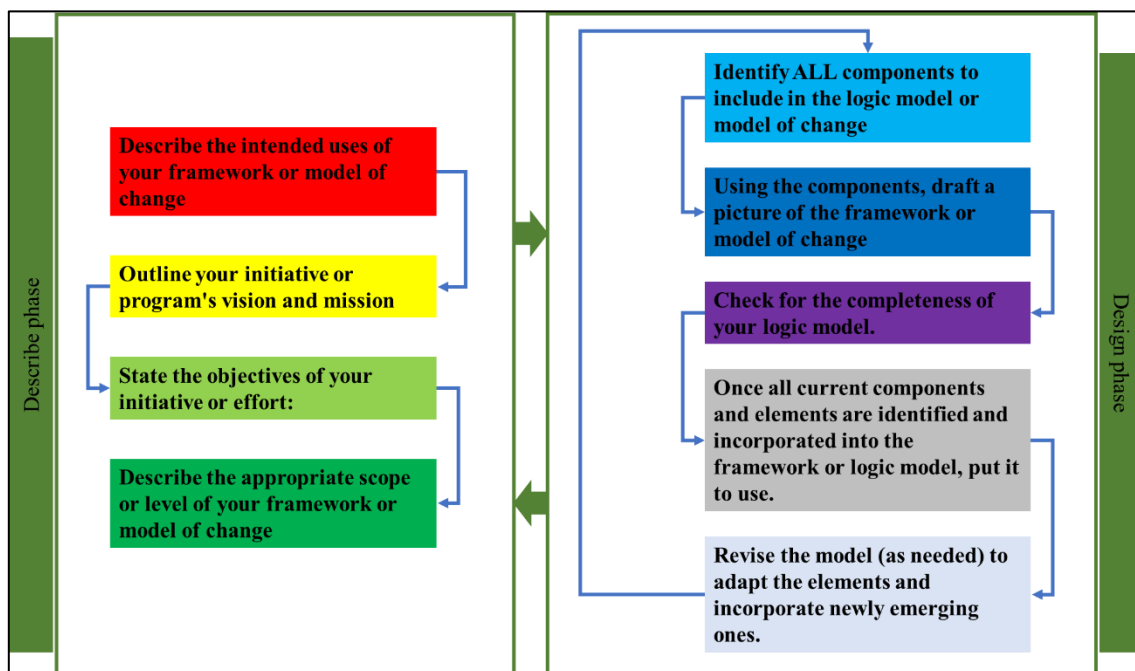


Figure 3.2 - Center for Community Health and Development framework (Author's depiction)

The complete framework of the Center for Community Health and Development contains eight phases. Each phase focuses on a distinct activity that is critical to framework development. The steps of the development framework are discussed in the next section. When creating the quality framework model, these steps were followed:

Step 1: Describe the intended uses of your framework or model of change

This step is primarily concerned with how the proposed quality framework would facilitate the desired improvements (Center for Community Health and Development, n.d.). The main objective is the intended usage, so every functional aspect of the framework, as well as how each of these aspects would interact with each other, must be perfectly evidenced. The intent is:

- To express the intent and direction of your effort or initiative
(Center for Community Health and Development, n.d.).
- Identifying more likely acts and strategies to contribute to the desired outcome.
- To explain how different factors combine to impact the issue or target
(Center for Community Health and Development, n.d.).

Step 2: State the objectives of your initiative or effort

This phase shows the importance of explicitly defining the purpose of the framework in order to identify the limitations of a method (Center for Community Health and Development, n.d.). Developing a framework aimed at improving the perceptions and behaviours of people in an organisation requires a clear understanding of potential barriers. These barriers can be created by environmental and personal factors. This step provides for the identification of key assumptions and hypotheses for observable outcomes, including:

- “Summarise all the particular observable outcomes you expect from your plan or program. These could include improvements in behaviour and related findings at the group level.”
- State your hypotheses and conclusions about the personal and environmental factors that lead to the issue or target” (Center for Community Health and Development, n.d.).

By specifying what a framework must create, the phases take a foresight approach.

Step 3: Describe the appropriate scope or level of your framework or model of change

This step includes identifying all the components that will function together in the effort to bring about development and alteration (Center for Community Health and Development, n.d.). Each component of a framework serves a purpose, hence the nature of every component in a framework must be defined

in order to make the essential improvements possible and clarification is required in the explanation of each:

- “The overall initiative can include all strategies and relationships used to influence change and to improve the overall problem or objective.
- A specific initiative or program may include only the components and elements of a particular component of the overall effort” (Center for Community Health and Development, n.d.).

Step 4: Identify all components to include in the logic model or model of change

This step refers to the functional component of a framework. It is important to define and account for all components that will move theory into practice (Center for Community Health and Development, n.d.). In the process of implementation of change, certain drawbacks may occur. Recognising drawbacks facilitates the degree of adaptation that may be required for the quality framework to operate successfully. In order to contextualise the final product, this method is often consistent with the identification of expected outcomes, including:

- “Open inputs, tools and assistance, as well as limitations or obstacles to achieving the goals of the initiative.
- What the plan or program is going to do to bring about progress practices or initiatives (e.g. improving help, improving access).
- Effects, outcomes or findings that are more generally assessed (may include immediate, intermediate and longer-term effects)” (Center for Community Health and Development, n.d.).

Step 5: Using the components, draft a picture of the framework or model of change

During this step, the relationships between different components have to make sense so that the specified representation of the framework or model of change can be created (Center for Community Health and Development, n.d.). Each piece of the framework is an empirical component of achieving the desired transformation. In this step, the following elements are clarified:

- “An estimated time sequence to organise the components and elements of the system or model (what happens before what).
- “Arrows or other strategies for transmitting paths of control and event sequences. Some arrows may point to demonstrate contact or reciprocal impact in both directions” (Center for Community Health and Development, n.d.).

Step 6: Check for the completeness of your logic model

This phase is regarded as a 'piloting stage' in which the assembled framework is tested and validated for practicality and completeness (Center for Community Health and Development, n.d.). External contributions, in the form of responses and comments from research participants, will be required to verify the completeness of the model. The major guideline for this stage is to revise and introduce new pieces/components to create a complete framework before putting the final version into a real-world application to test its usefulness. This provides a track record that demonstrates how a framework performs in a real-world environment. The tasks involved in this stage include:

- “Picking an actual or hypothetical case scenario from which input on the logic model can be obtained.
- Checking for the utility of model elements, (e.g. is it understandable?).
- Checking for model completeness (e.g. what is missing?).
- Making it more complete by revising it or adding new components”
(Center for Community Health and Development, n.d.).

Step 7: Put the model to use

This step illustrates how to handle the realistic use of a framework (Center for Community Health and Development, n.d.). Using the established framework in a real-life context can generate outcomes that may or may not be as anticipated (Center for Community Health and Development, n.d.). This step suggests that a management strategy should be established and that mid-course changes can be made. Changes in mid-course are beneficial because they place an author in a better position to note which elements of a framework require modification (Center for Community Health and Development, n.d.). Since any new initiative requires buy-in from its target population and stakeholders, advocacy and coordination are key parts of implementing the framework. According to the Center for Community Health and Development (n.d.), the practicalities of this move entail:

- “Orienting and supporting those doing the job, using it to explain how project or program elements work together, where contributors fit in, and what they need to be able to make it work.
- Using execution to evaluate what components you have in your project or program and those you do not have, create a management plan and make mid-course adjustments.
- Communication and lobbying are used to justify the progress of the plan/program and to demonstrate how investments can be used.”
(Center for Community Health and Development, n.d.).

Step 8: Revise the model (as needed) to adapt the elements and incorporate newly emerging ones

According to the Center for Community Health and Development (n.d.) if the desired outcomes have been achieved – that is, if a structure has been put to use – this must be reflected. This step focuses on how a framework can be altered and revised to better serve its intended purpose and objectives, and it contains the following activities for completion:

- “Link the direction of operations to the expected effects or results.
 - Planning to broaden activities to meet your objectives.
 - Comprehend the limitations of your program or initiative.
 - Adjust course to make unanticipated adjustments.
 - Develop a new system or new initiative for an extended effort”
- (Center for Community Health and Development, n.d.).

The University of Kansas (2023) framework development strategy requires compliance with eight defined steps (discussed above) that result in the development of a useful and reliable framework (Center for Community Health and Development, n.d.). The steps are structured in such a way that they create a realistic roadmap that can be revisited once the developed framework has been used to make changes and revisions. The steps of the University of Kansas (2023) framework development strategy enable a user to easily combine ideas and other constituents of their theoretical and functional components to construct a working framework based on theory and practice relevant to a specific environment. The steps of the framework development strategy will be used to discuss the actions taken throughout the creation of the proposed framework. (See Section 3.8 for details on how the proposed quality process model was developed).

3.6 Framework documentation

The management and storage of research data and digital records must be handled in a clear and concise manner. To do so, research data management should be carried out using:

...a number of different activities and processes associated with the data lifecycle, involving the design and creation of data, storage, security, preservation, retrieval, sharing, and reuse, all considering technical capabilities, ethical considerations, legal issues and governance frameworks” (Sanjeeva, 2018:2).

The numerous issues involved with data management must be addressed in a data management framework. Nwabugwu and Godwin (2020) identify the following challenges related to digital data management: technology obsolescence; technology fragility; lack of guidelines on good practice; inadequate financial and human resources to manage data well; and the lack of evidence about best infrastructures.

The issues mentioned above should be addressed in data management policies to guarantee that all potential causes of such issues can be addressed as soon as possible. This viewpoint is stated in Section 2 of the Australian Code for the Responsible Conduct of Research (National Health and Medical Research Council, 2018): “Policies are required that address the ownership of research materials and data, their storage, their retention beyond the end of the project, and appropriate access to them by the research community” (Australian Research Council, 2018). To attain this goal, principles for implementing policies, strategies and standards for the long-term preservation of digital data must be adopted. In line with this, it is critical to pay attention to the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) research project, which explored the development, maintenance and preservation of digital documents.

Fundamentally:

“the InterPARES project was established with the aim to develop the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form and provide the basis for standards, policies, strategies and plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity” (Duranti et al., 2008).

The key finding of the research is that such a system of principles can be created to promote the formation, maintenance and preservation of documents, irrespective of jurisdiction (Duranti et al., 2008). These principles share features that are also important to the development of an RDM quality framework. Aside from a researcher's data management skills, the ever-increasing volume of new data has an impact on the generation of high-quality documents and research data. Ng'eno and Mutula (2018:29) carefully explain that “the availability of huge amounts of new data, often referred to as a data deluge, impacts the method in which research is carried out for societal benefit”. RDM must therefore be conducted in the most efficient manner possible because it benefits researchers and research institutes in a variety of ways (Lewis, 2010; Dora & Kumar, 2015). Some of the advantages are stated by Ng'eno and Mutula (2018:29) as follows:

- “Ability to share research data, minimizing the need to repeat work in the field or laboratory.
- Research data gathered at considerable cost is not lost or inadvertently destroyed.
- Retrieval, comparison, and co-analysis of data from multiple sources can lead to powerful insights.
- New research themes can emerge from re-analysis of existing data or comparisons with new data.
- Long-term preservation of data provides for validation check of the data and this enhances the credibility and transparency of the research data used”.

The management and storage of research data allows for research data to be well maintained, allowing for later use. Research data management, storage, and sharing hold a number of advantages, including increasing the return on major investments, reducing costly data duplication, promoting economic development, and furthering human knowledge (Ray, 2014).

Principles for Record Creation and Preservation

As part of the InterPARES research initiative, the policy team created two complementary sets of principles. The first set of principles was created for record creators and the second for record preservers, all with the goal of promoting the establishment of the relationship between creators and preservers by describing the nature of that connection (Duranti et al., 2008). Data from studies that can be used for systematic evaluation come in a variety of formats (Bowen, 2009). Researchers must recognise that managing research data necessitates a high level of expertise in order to provide data that is suitable and dependable for secondary use. The principles for record production and preservation are discussed in the next section.

Principles for the record creator

The principles that apply to record creators are designed to promote (1) reliable, (2) accurate, and (3) authentic digital record handling. These three traits should be at the heart of a system in order to create an efficient RDM quality framework, as they will allow for exceptional professionalism, outputs, and working culture (Abiteboul et al. 2016; Chapman, 2005). These qualities, according to Duranti et al. (2008), should be presented explicitly and individually and promoted within an organisation's records management guidelines. It is critical to have unique procedures in place that can assess the extent to which record creators (i.e. researchers) handle and store digital records/research data. Organisational policies, practices, and standards should contain these measures. For this study, it is critical to include compliance indicators in policies and standards. The key pillars on which an organisation's RDM efforts are built are policies and standards (Domingus, 2015). In order to systematically pursue the quality aim, the various tasks played by project team members (through the research data lifecycle) must be clearly differentiated and defined. The principles for record creators "[are] directed to the persons responsible for developing policies and strategies for the creation, maintenance and use of digital records within any kind of organization, and to national and international standards bodies" (Duranti et al., 2008:2). This ensures that data creators retain the high level of accuracy and consistency required for research data development. In a data-driven framework, this principle is critical (as is the case for this study).

Principles for record preservers

The principles for record preservers:

"...are directed to the persons responsible for developing policies and strategies for the long-term preservation of digital records within administrative units or institutions that have as their core

mandate the preservation of the bodies of records created by persons, administrative units or organizations external to them” (Duranti et al., 2008:2).

In contrast to the principles of record producers, data preservers can be thought of as working in the background in the managerial task of providing high-quality outputs (data).

Definition summaries of framework development were provided in this section. It was necessary to clarify the notions that a framework could encompass, such as data decision-making skills and principles for the execution of policies, plans, and standards. The knowledge gained in this section has opened the basis for a much more in-depth examination of the framework's development. The principles of framework development are a starting point for creating a framework that is based on policies, strategies, and standards. The next section goes over some of the existing frameworks and principles that have been identified as being relevant to this research.

3.7 Relevant frameworks

The management of research data is guided by a variety of RDM frameworks and models. It was critical to identify the most widely used RDM frameworks in order to understand not only why they are seen to be the most beneficial, but also how they specifically handle data quality.

Various RDM frameworks can be used to inform a research data quality management framework, as the literature review chapter illustrated. The goal of this study is to identify the most helpful quality assurance and data quality characteristics of the frameworks that have been studied. The reviewed literature provided many options that can be used as variables for the data quality model. The following section goes over a carefully chosen group of frameworks. Appendix 3 contains a list of potential frameworks derived from the literature that might be used in the quality framework. Following a thorough examination of the frameworks, the following sections focus solely on those that were chosen and are significant. The chosen frameworks either provided appropriate concepts that could be applied to the suggested quality frameworks and contributed to RDM, or they did not.

3.7.1 UK Data Archive model as a framework

The proposed quality process model works well with the UK Data Archive model as a base model. Ball (2012:9) explains, “The UK Data Archive provides a data lifecycle model as an aid to researchers considering how data management relates to the lifecycle of a research project”. The UK Data Archive model is a well-known, comprehensive data lifecycle to help researchers with data management (Van den Eynden, 2011; Bishop, 2012; Pryor, 2012; Beedham et al. 2015; Nhendodzashe & Pasipamire, 2017). RDM activities are associated with each of the stages represented in the model depicted below:

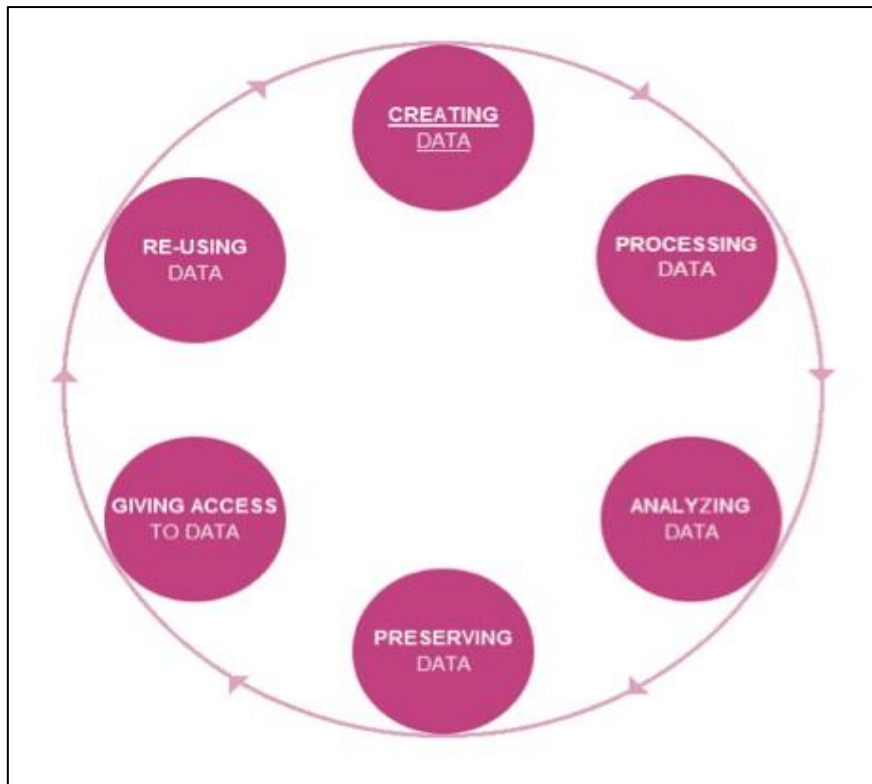


Figure 3.3 - UK Data Archive research data lifecycle (UK Data Archive, 2017)

The creating and processing data stages of the UK Data Archive model are the most significant in the context of this study. The quality of data is at risk throughout the research data lifecycle, but the most vulnerable phases are when data is created and processed, as human errors are most common during these two stages. It was also critical to comprehend the procedures involved in the creation and processing of data in the UK Data Archive model. This knowledge was required to determine whether additional quality dimensions were required to assess quality compliance during the data creation and processing stages. The next section delves into the theoretical underpinnings of the data creation and processing stages.

Creating Data

Pryor (2012:6) highlights that “in each and every phase of the data lifecycle the researcher will gather and use data and/or generate new data”. This entails creating new data; the first phase of the data lifecycle includes the compilation or extension of data holdings (UK Data Archive, 2017). A researcher's data collection activities are used to obtain information needed to answer research questions. Furthermore, the development of research data necessitates the physical collection of data using traditional methods. Other data generation activities listed by Etheridge (2016) include design, consent preparation, collection and management, and metadata acquisition and creation. These operations must be carried out appropriately for the sake of data quality, requiring the inclusion of

quality assurance methods in the process. Ball (2012) outlines the steps that should be accomplished when creating data. These tasks, according to Ball (2012:9) are:

- “*Design research*” – This task forms part of the data management plan, which is used as a measuring criterion for the Accessibility data quality dimension.
- “*Plan data management (formats, storage etc.)*” – This task serves as a criterion for assessing the Accuracy quality dimension.
- “*Plan consent for sharing*” – This task is used as a measuring criterion for the Usability of Data quality dimension.
- “*Locate existing data*” – This task serves as a criterion for assessing the data quality of the Accessibility dimension.
- “*Collect data (experiment, observe, measure, simulate)*” – This task serves as a criterion for assessing the Validity quality dimension.
- “*Capture and create metadata*” – This task is used as a measuring criterion for the Accuracy quality dimension.

The literature studied concerning this stage (producing data) does not specifically include advice on quality assurance processes to be applied when data are created (UK Data Archive, 2017; Etheridge, 2016; Pryor, 2012; Van den Eynden, V. 2011). Researchers are free to employ whichever methods and preferences they prefer when collecting data. Given that this is the case, it is critical to have quality assurance mechanisms in place. The procedures in the UK Data Archive's Creating data stage are rigid, but they provide a good overview of the duties a researcher must do when collecting data. Quality assurance will be strengthened by including a quality dimension in the data creation stage.

Processing Data

Data processing is a sequence of data actions or steps to validate, organise, transform, incorporate, and extract information in an appropriate output form for subsequent use (UK Data Archive, 2017). Data processing might take place in the field or in an office setting, depending on how the data is obtained. The methods used by a researcher or fieldworkers (data collectors) on the data in both environments should not jeopardise the integrity or quality of the data. The procedure used must be carefully recorded when the data are processed to ensure the usability and credibility of the data (UK Data Archive, 2017). It was vital for this study to understand what processes the UK Data Archive recommended for determining the inclusive level of quality dimensions, just as it was with the data creation stage. The following activities must take place during the processing stage, as per the UK Data Archive model:

- “*Enter data, digitise, transcribe, translate*” – This activity is used as a measuring criterion for the Validity quality dimension.

- “*Check, validate, clean data*” – This activity serves as a criterion for assessing the Validity quality dimension.
- “*Anonymise data where necessary*” – This activity is used as a measuring criterion for the Completeness quality dimension.
- “*Describe data*” – This activity is used as a measuring criterion for the Interpretability quality dimension.

The processing activities of the UK Data Archive are viable and beneficial to the data quality problems. The level to which the activities answer the 'how' question, on the other hand, is unclear. All of the quality dimensions that are customised for the processing phase are informed by important questions that will disclose how quality is achieved when data is processed. Researchers are required to participate in data processing activities on a hands-on basis. This necessitates a researcher's implementation and upgrading of data management during research in accordance with best practices and standards (Van den Eynden, 2011).

The stages of creating and processing data are critical in the research data lifecycle; however, data quality can be impacted by the other phases of the research data lifecycle, which are covered in the following section. Because it provides acknowledged and best practice activities for when data is created and handled, the UK Data Archive is pertinent to the proposed quality framework. The data quality dimensions will be used to evaluate and consolidate the administrative procedures of the UK Data Archive. Furthermore, including data quality dimensions provides quality assurance standards, lowering the danger of errors and of jeopardising data quality and integrity. The relevance and applicability of the remaining data lifecycle stages are informed by activities related to the first two stages of the research data lifecycle. If data has undesirable characteristics or is of poor quality, it is fruitless to move it through the next stages of the data lifecycle. This viewpoint is depicted in Figure 3.4, below:

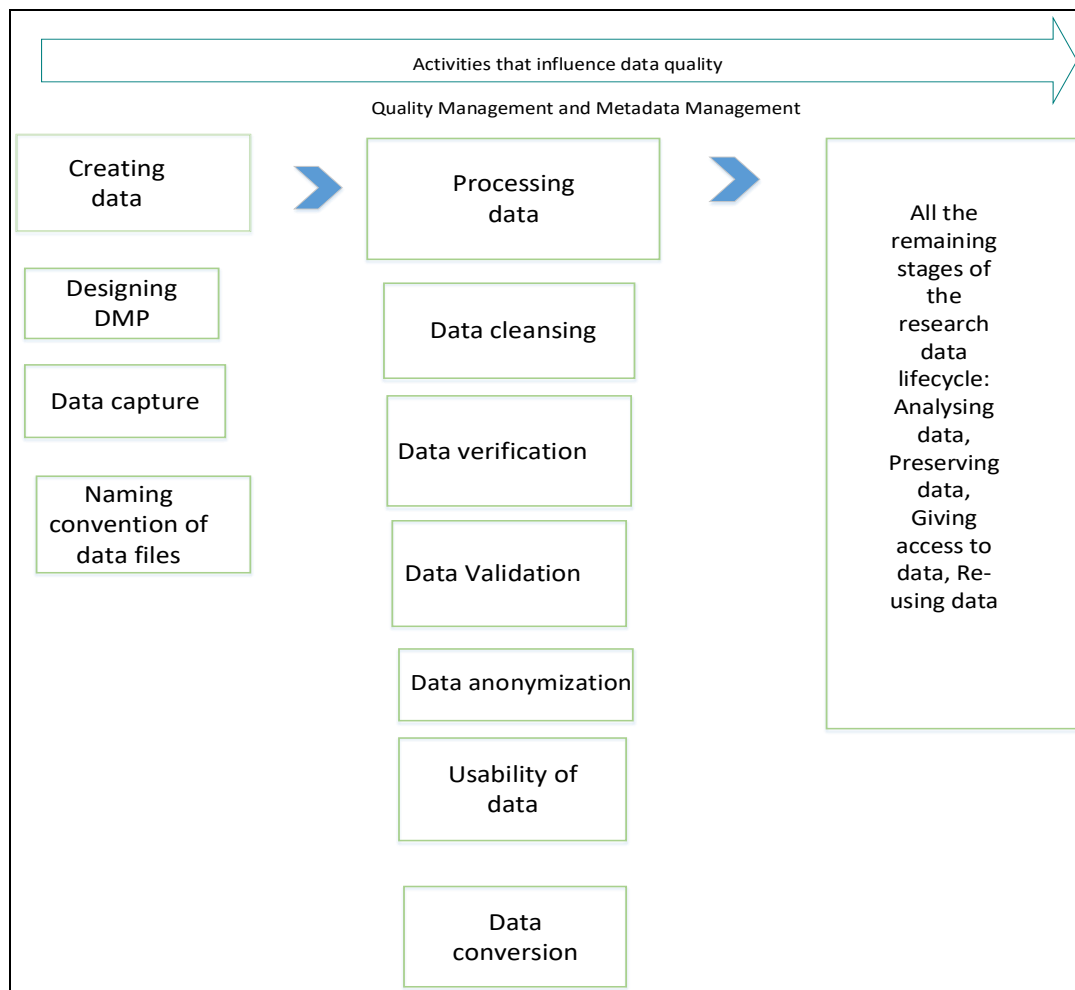


Figure 3.4 - Activities influencing data quality in the first two stages of the research data lifecycle (UK Data Archive, 2012; Rolland, 2016; International Organization for Standardization, 2018)

The following section considers all the other stages of the research data lifecycle, as well as the activities that go along with these, such as data quality and quality assurance.

Analysing Data

The next stage in the UK Data Archive research data lifecycle after processing data is to analyse it. Data analysis “involves actions and methods performed on data that help describe facts, detect patterns, develop explanations and test hypotheses” (UK Data Archive, 2017:1). Analysing data is done using techniques such as statistical software, and modelling. These techniques require a researcher to interpret, derive, produce outputs, publish and prepare the data for sharing (Etheridge, 2016). The usefulness and dependability of data are determined by how well it is processed; thus, it is critical to demonstrate and maintain a high degree of quality assurance during this stage. This stage includes (but is not limited to) the following activities (UK Data Archive, 2017):

- Interpret data
- Derive data
- Produce research outputs
- Author publications
- Prepare data for preservation

These activities include both theoretical and practical endeavours. Vosloo (2014:355) points out that “the analysis and interpretation of data represent the application of deductive and inductive logic to the research”. Individuals interpret and view things differently, which can lead to different extrapolations of meaning from the data studied. As a result, quality assurance is dependent on the individual. A gap that needs to be filled is the lack of guidance on how to conduct data analysis while maintaining quality assurance. Additionally, quality assurance is required since data analysis (whether quantitative or qualitative) can be used to characterise and summarise data, discover relationships between variables, compare variables, identify differences between variables, and anticipate outcomes (Azimae, 2015).

Preserving Data

The actions taken to preserve data include “procedures to keep data for future use” and “data archiving and/or data submission to a data repository” (UK Data Archive, 2017). These actions require a researcher to develop data descriptions, documentation and metadata (UK Data Archive, 2017). The development of these documents ultimately makes research data logical, findable and easy to use. When doing preservation actions, a researcher should be able to ensure that quality is maintained. Barsky (2015:8) provides a valuable description of quality assurance that is relevant to preservation by stating that, even if the decision is taken not to follow a metadata standard, it is still important to properly document the data. At the very least, the following should be documented in a readme.txt file stored alongside the data:

- “Context of data collection (the goal of your research)
- Data collection methods (protocols, sampling, instruments, coverage)
- Structure of files
- Sources used
- Quality assurance (data validation, checking)
- Data modifications
- Confidentiality and permissions
- Names of labels and variables
- Explanations of codes and classifications” (Barsky, 2015:8).

If what is handled and preserved using quality assurance techniques is to be accessed efficiently well into the future, quality assurance is vital as part of the preservation effort.

Giving access to data

The UK Data Archive (2017) provides a detailed explanation, stating that “the transparency and accountability plan, driven by a range of communities, including the general public, governments, funding bodies and the research community, pushes for data to be made available to provide the evidence that research has been conducted properly”. Providing access to research data involves the creation of descriptive documents that make data accessible, such as the following:

- Related documentation: provides an overview of the research context and design, data collection methods, data preparation and results or findings enabling the secondary user to make informed use of the data (UK Data Archive, 2017). Documentation of this nature may include user guides, a Codebook and a readme file.
- Metadata: providing standardised structured information explaining the purpose, origin, time references, geographic location, creator, access conditions and terms of use of data (UK Data Archive, 2017).

These documents make data more accessible by allowing it to be explained and transferred and by encouraging secondary data usage. Because the material must be understandable, relevant, current, and beneficial, quality assurance is inextricably tied to the development of the above-mentioned documents.

Reusing data

- The UK Data Archive research data lifecycle concludes with efforts that promote data reuse. These activities include:
 - Follow-up research
 - New research
 - Undertaking research reviews
 - Scrutinising findings
 - Teaching and learning

Without an appreciation for quality assurance, these processes may be implemented reactively as opposed to pro-actively. Surkis and Read (2015:155) explain that “the last stage of the data lifecycle involves reusing the data to conduct new studies or to test the reproducibility of the original results”. Follow-ups, research reviews, new research, teaching, and learning are all made easier by reusing data (Etheridge, 2016). However, quality assurance techniques on how the activities of reusing data should be carried out are not documented.

The UK Data Archive research data lifecycle provides an overarching management structure that guarantees researchers do not neglect or skip critical activities while doing research. Planning is an important aspect of the research data lifecycle that must be considered. The UK Data Archive research data lifecycle was chosen for its comprehensive approach and then customised with data quality dimensions.

3.7.2 Inter-university Consortium for Political and Social Research (ICPSR) model

The ICPSR Guide to Social Science Data Preparation and Archiving is intended to help researchers deal with challenges that may occur during the research process. These issues include applying for a research grant, the data collection phase, and the preparation of the data for deposit in a public archive (ICPSR – University of Michigan, 2018). The ICPSR Guide is a compilation of best practices gathered from the experience of many archivists and researchers (ICPSR – University of Michigan, 2018). The ICPSR uses a lifecycle (process framework) illustrating key considerations relevant to archiving at each step in the data creation process (ICPSR, 2018). Below, in Figure 3.5, is a graphical representation of the ICPSR Data Lifecycle:

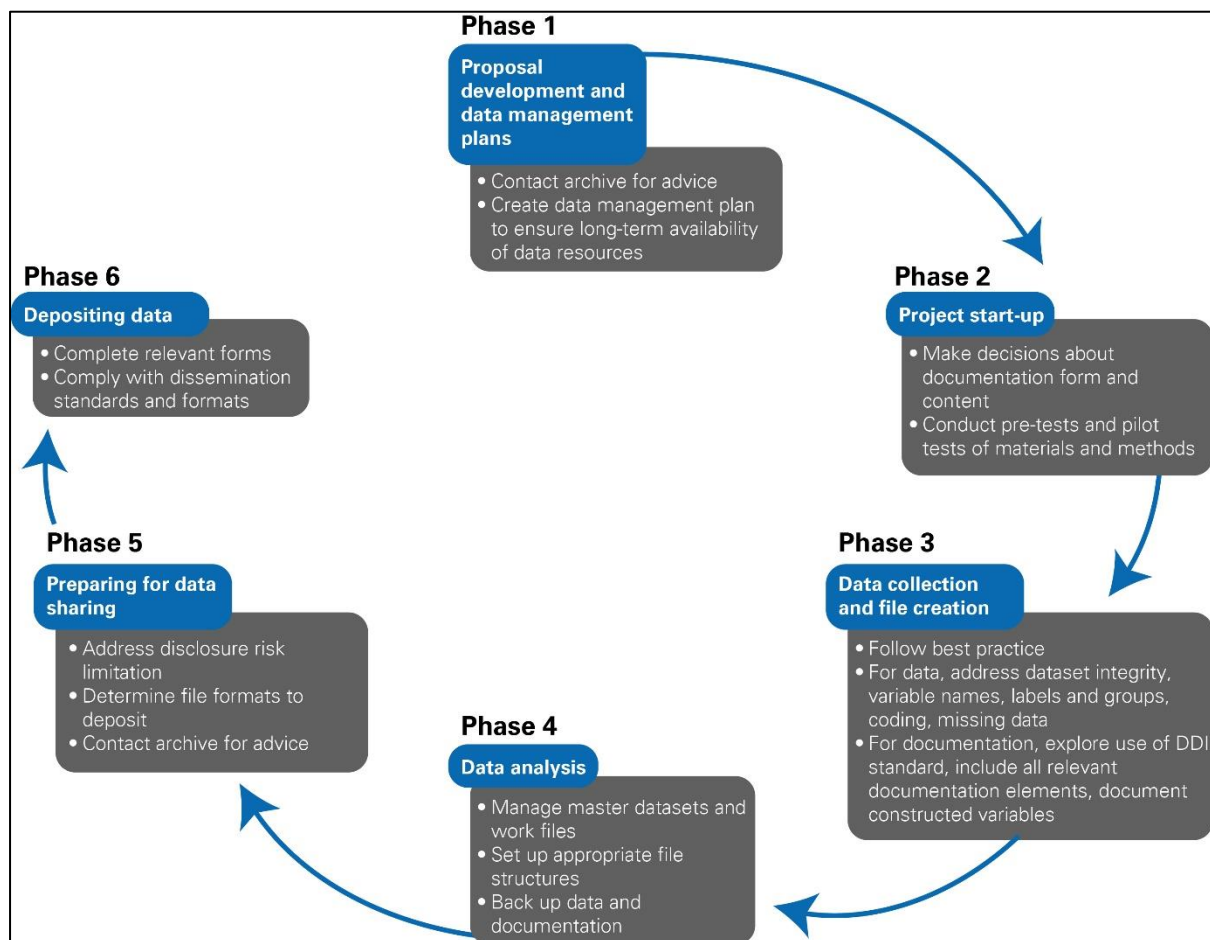


Figure 3.5 - ICPSR Data Lifecycle (ICPSR – University of Michigan, 2018)

The ICPSR Guide to Social Science Data Preparation and Archiving “provides a detailed overview of what an institutional repository can do to appraise data, prepare them for storage, and ensure that the preserved data are independently understandable” (ICPSR, 2018:3). Researchers and custodians should be able to appraise material in an institutional repository so that they can know which data should be prepared for storage and preservation. The ICPSR Guide “provides a detailed overview of what an institutional repository can do to appraise data, prepare them for storage, and ensure that the preserved data are independently understandable” (ICPSR, 2018:11).

Phase 1 (Project Development and Data Management Plans) of the ICPSR Data Lifecycle is a relevant component of the ICPSR Guide that was utilised in this study. This phase outlines clear data management procedures that address critical data quality issues and guides a researcher through the data management and data collecting planning process. These contributions are regarded as functionality components for the quality assurance of the quality framework. This phase is also used as a measuring criterion for the Completeness, Uniqueness, Validity, Accuracy, Consistency, Interpretability, Integrity and Relevance quality dimensions in the proposed quality process framework. Appendix 1 covers all of the other phases (i.e. from Phase 2) of the ICPSR Data Lifecycle.

Phase 1: Project Development and Data Management Plans

Planning for data management is a crucial step because it establishes a research project's RDM direction and techniques. As a result, the ICPSR Project Development and Data Management Plan was used as a variable to evaluate the established quality dimensions of the study. The 'Designing DMP' activity, which is part of the Creating Data phase of the quality process framework, is worth noting as a measurement criterion. The DMP in the ICPSR Guide to Data Preparation uses extensive questions that should be included in a data management plan to ensure adequate data management and reduce errors.

3.7.3 The South African Statistical Quality Assessment Framework (SASQAF)

The importance of SASQAF was described in Chapter 2 (Section 2.3.3.2) and in Appendix 1. The national statistics service of South Africa, Statistics South Africa (Stats SA) has defined eight dimensions of quality known as the South African Statistical Quality Assessment Framework (SASQAF) (Statistics South Africa, 2010). The dimensions are “relevance, accuracy, timeliness, accessibility, interpretability, coherence, methodological soundness, and integrity” (Statistics South Africa, 2010:4). These dimensions are intended to make statistical data as useful as possible by characterizing the processes they measure as precisely as possible. SASQAF has quality prerequisites in addition to the eight data quality dimensions, which include “the institutional and organisational conditions that have an impact on data quality – these include the institutional and legal environment, and availability of human, financial and technological resources” (Statistics South Africa, 2010:4).

The SASQAF contains important measurement components that are considered critical to the formation of the proposed quality framework. The SASQAF's eight quality dimensions and prerequisites contribute by providing a baseline for data quality in all government agencies. The suggested quality measurement framework considers both quality dimensions and data quality measuring indicators. The SASQAF served as the foundation for the proposed framework in Section 3.8 and was effectively expanded to incorporate the technical quality dimensions. The quality dimension description is a useful place to start when evaluating RDM scientific quality compliance. The SASQAF was used to assess the various stages of the lifecycle and is part of the data quality dimensions of the proposed theoretical quality framework. Table 3.3, below, depicts an example of SASQAF dimensions used in the quality theoretical framework.

Table 3.3 - Example of the SASQAF dimension (Statistics South Africa, 2010)

Dimension/Prerequisite	Description	Meaning of quality dimension in the proposed quality process framework
Integrity	“The integrity of statistical information refers to values and related practices that maintain users’ confidence in the agency producing statistical and ultimately in the statistical product” (Statistics South Africa, 2010:4).	Such firmness will resemble a high level of good data quality if correct and best practice quality assurance procedures were followed when data was created and processed. This dimension aims to investigate a researcher's quality assurance approaches used during the Creating data and Processing data stages.

Since its main focus area is to increase the 'quality' of statistical/research data, the SASQAF quality dimensions and prerequisites are part of the quality process framework. The SASQAF quality dimensions and prerequisites are expected to offer the quality process framework a boost in terms of examining elements that negatively impact data quality. Identifying such flaws will aid in the development of reasonable recommendations for enhancing data quality. The following section brings all of the previously examined material together to create a process model and quality assessment framework for social science data (Statistics South Africa, 2010).

3.8 A process model and quality assessment framework for Social Sciences data

To design the proposed framework for quality evaluation of social sciences data, the framework development strategy (see Section 3.5) of the Center for Community Health and Development at the University of Kansas is used in conjunction with the framework development principles identified by Mandinach et al. (2006) (see Section 3.4).

3.8.1 A quality-focused RDM for Social Sciences data

RDM is based on a number of principles. These principles apply to a wide range of RDM topics and can be linked to the quality level of the data file concerned. The RDM literature that was examined for the current research (Domingus, 2015; Flores et al., 2015; Knight, 2015; Tripathi et al., 2017; University of Wollongong Australia, 2017; Borghi et al., 2018; Sanjeeva, 2018), along with the literature on data quality (Scannapieco et al., 2005; McKnight, 2010; Andrews et al., 2012; Cooper & Kahn, 2012; Cai & Zhu, 2015; Doorn & Dillo, 2017) presented specific ideas that can be used to build a theoretical process framework.

The proposed quality process model must explore important quality assurance and data quality areas in order to change the data quality planning process, data collection, and data processing in the Social Sciences Research Organisation. To accomplish this goal, quality-based standards and procedures to direct data management in all phases of the research data lifecycle must be created and executed. Various data quality dimensions and RDM quality metrics identified in the literature are presented in the suggested quality process model. The following section implements the University of Kansas framework development strategy, as discussed in Section 3.5.

3.8.2 The intended uses of the model and the framework

Downs et al. (2021:3) state that “uncorrected errors, missing data, and undocumented corrections and modifications could influence findings resulting from the analysis of data”. To reduce the likelihood of such issues, the suggested quality measurement framework will be utilised to analyse researchers' research data collecting and processing techniques, as well as data curators' curation practices. These procedures have an impact on the quality and reusability of the data. The quality measurement framework will assess participants' RDM and quality assurance methods in order to attain this goal. The quality measurement framework has borrowed a variety of guidelines, standards, and best practices from the literature consulted. In line with this, the quality measuring model will be utilised to see if research participants are employing effective quality process approaches.

3.8.3 The objectives of the model and the framework

This process allows essential assumptions and hypotheses for observed results to be identified. The proposed data quality framework aims to analyse and direct the process of data management planning in a Social Sciences Research Organisation by using specific aspects of data quality. The elements of quality were assessed using metrics specific to social science research data. The quality dimensions are designed to see how researchers handle issues that could jeopardise data integrity. In summary, the following are the expected observable outcomes:

- Researchers recognise the importance of adding quality assurance into their research practices in order to deliver high-quality results.
- Curators take a more active role in encouraging researchers to work toward the creation of high-quality data.
- Data quality and quality assurance become a major institutional priority when current data management policies and processes are followed.

Researchers (as data producers) should generate data that is trustworthy, accurate, and genuine. The proposed quality process model's primary goals are to first comprehend the RDM environment at a social sciences research institution. This involves researchers and curators providing their RDM experiences. The second goal of the process model is to put in place a theoretical agenda based on unambiguous quality characteristics and indicators in order to stimulate and lead RDM planning. The third purpose of the model is to provide a means for data curators to play a clear, active role in influencing researchers' data management decisions to produce high-quality data. Finally, the model aims to investigate the details of RDM at each stage of the research data lifecycle.

3.8.4 Developing the process model

Principle 2 of the Conceptual Framework for Data-Driven Decision Making Skills (Mandinach et al., 2006) requires the “identification of all the crucial activities” associated with the process. The data quality dimensions and measuring criteria (i.e. the first block on the left) that are used to manage data in the different stages of the data lifecycle make up the overview of the quality review process (Figure 3.6). In order to prepare the measuring framework, a quality review process is implemented based on the quality measurement dimensions and associated measuring criteria. All of the quality assurance steps to be employed during all phases of the data cycle are covered by the quality review process. Once the quality review is in place, its findings can be combined to generate a quality rating (i.e. Quality Review). The measuring criteria in this regard are unpacked in Appendix 4. The data quality evaluation results are depicted in Figures 3.9 and 3.10 as Spirogram results of the quality assessment. The quality review process that was created for this study employing quality dimensions and measuring criteria is shown in Figure 3.6 below:

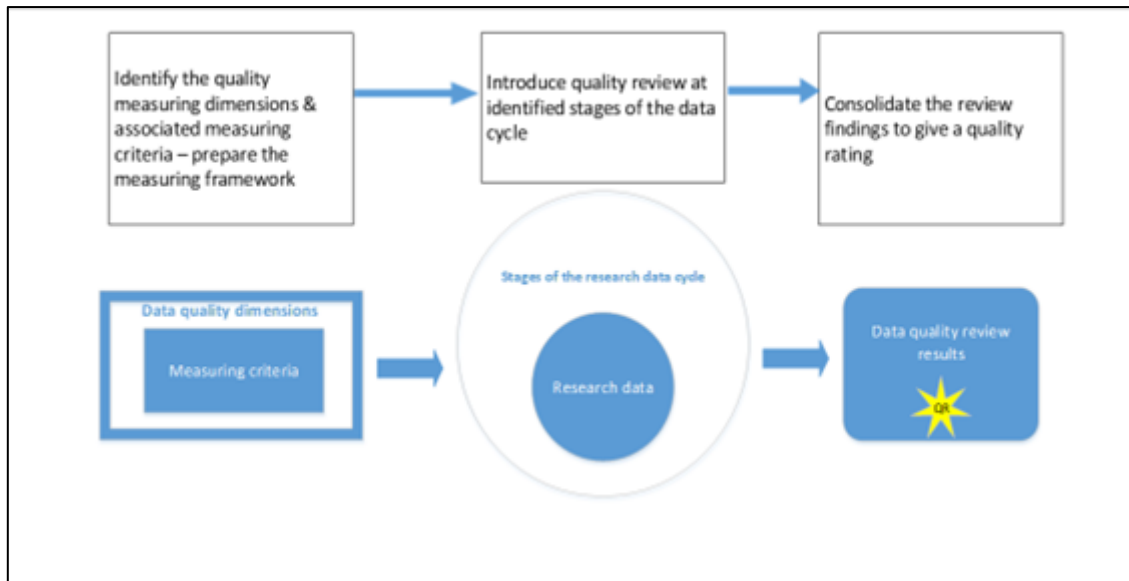


Figure 3.6 - Overview of the quality review process

The data quality dimensions and measuring criteria are utilised to create the data quality measuring framework, as mentioned previously. The process of developing the measuring framework is discussed in the following section.

3.8.5 Developing the quality measuring framework

The procedure for establishing the quality review framework is depicted in Figure 3.7. The literature review enabled the identification of RDM best practices and standards that encourage improved data management. RDM is used in a variety of disciplines, with each area having its own set of rules and best practices. Figure 3.7 begins by acknowledging RDM, guidelines, standards and best practices that are unique to a particular discipline - in the case of this study the social sciences discipline. Among the key findings from the reviewed literature are the data quality dimensions (discussed in Section 2.3.3), which are used to measure technical and scientific quality. All of the data quality dimensions are informed by various measuring criteria taken from RDM guidelines, standards and best practices. Each of the scientific and technical quality dimensions has data quality measuring criteria that assess how data is managed throughout the research data lifecycle. The assessment done according to the data quality measuring criteria results in a quality review framework, which goes into the quality process model (Figure 3.8). The principles for the development of a framework (discussed in Section 3.4) feed into the framework development strategy (discussed in Section 3.5), which, in essence, is the foundation of the quality review framework. The blueprint for developing a quality review framework is depicted in the diagram below. To determine adequate measuring criteria for the selected data quality dimensions, RDM guidelines, evaluation systems, and best practices (as discussed in Chapter 2, Sections 2.3.3, 2.3.5, 2.7) were employed. The scientific and technical quality dimensions of data were used (as discussed in Chapter 2 Section 2.3.2).

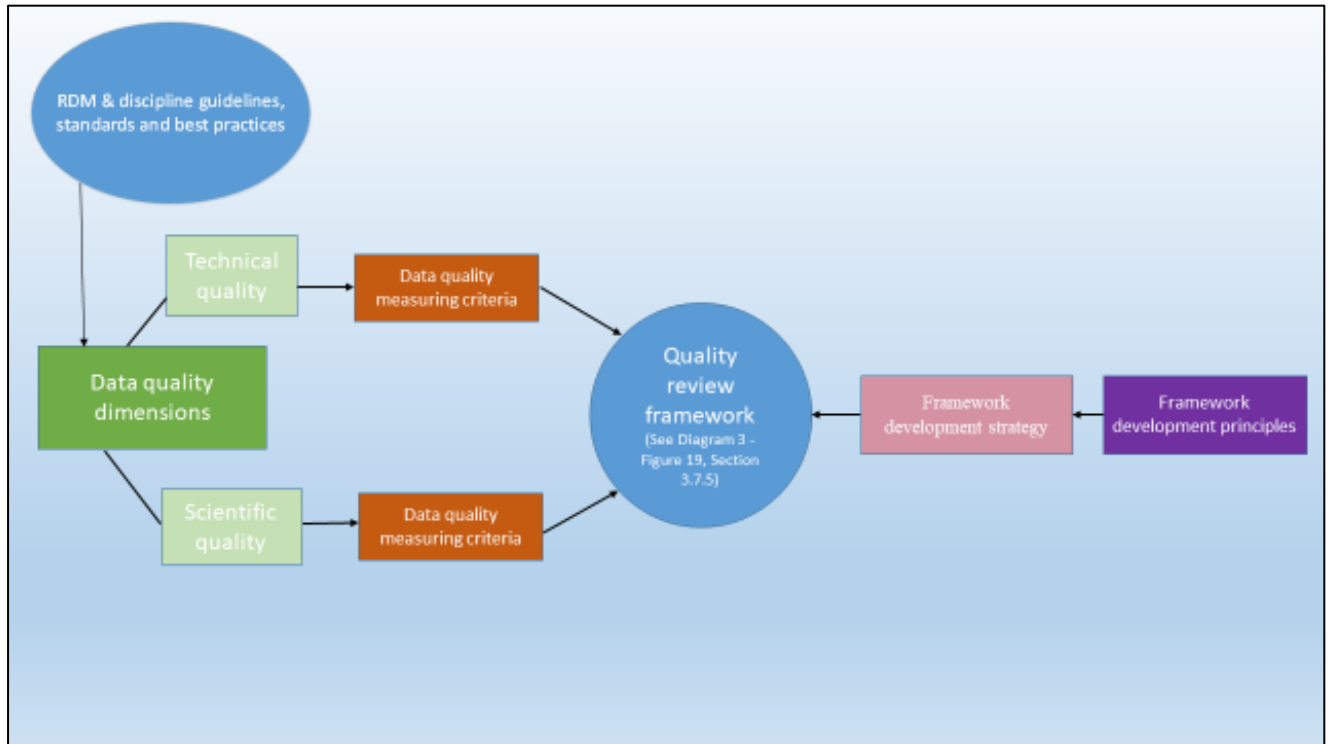


Figure 3.7 - Establishing the quality review framework

The blueprint for developing a quality review framework is depicted in the diagram above (Figure 3.7). To determine adequate measuring criteria for the selected data quality characteristics, the RDM guidelines, standards, and best practices that were discussed (in Chapter 2 Section 2.3.3, 2.3.5, 2.7) were employed. Scientific and technical quality dimensions are utilised to create a quality review framework. Appendix 2 shows the potential metrics for the quality dimensions – the Quality Process Framework for determining quality.

The preceding social sciences data measuring criteria, which were based on the literature consulted, formed the foundation for an initial quality measurement framework. A potential complete Quality Process Framework is presented in Appendix 2.

3.8.6 Components included in the quality review framework

The data quality dimensions are the most important elements of the quality review framework. The quality review framework is broken down into scientific and technical measurements, totalling 16 scientific and 16 technical measures. A total of 16 measures related to each research participant group (i.e. researchers and curators) are included in the quality review framework.

To analyse a given quality dimension, the quality review framework employs a best practice or quality assurance metric (derived from the literature). The scale is divided into four levels: Level 1, Level 2,

Level 3, and Level 4. The tables presented in Appendix 4 summarise what each level entails (using the quality dimension 'Completeness' as an example for both Scientific and Technical measures):

Specific best practices, standards, and evaluations were studied to inform the above levels. For each of the quality dimensions, the levels apply to both scientific and technical measures. The Quality Process Framework (Appendix 2), which is the primary measuring and quality assurance component of the quality process model, contains the assessment outcomes of all the quality aspects (Section 4.1 and Section 4.2 in Appendix 4). The components of the quality process model are discussed in the next section.

3.8.7 The components included in the quality process model

The proposed quality process model is based on the research data lifecycle, which was adapted from the UK Data Archive. Throughout the many stages of the research data lifecycle, data exists and is managed. Managing research data should be done using entrusting quality-based techniques and standards. To achieve this, a researcher's steps should be measured using adequate data management processes, guidelines, and best practices.

The quality process model requires a researcher to perform data management planning by detailing all of the actions that will be utilised to gather, organise, manage, store, back-up, preserve, and communicate data as part of data management planning. The factors should be documented in a data management plan, in which a researcher outlines the data quality methods and dimensions that will be applicable to their research study. A researcher can identify and use only applicable data quality dimensions by completing a 'data quality preparation' section in a data management plan. A researcher must account for the relevant quality dimensions and data quality methods in the DMP when doing research. The DMP does not currently require researchers at the social sciences research institution to declare the data quality procedures they will apply in their research projects. This phase is the essential rallying point of the Quality process model. If the relevant quality dimensions and data quality procedures have not been defined (as part of Data quality and management planning), the researcher must describe the quality assurance procedure to be followed before proceeding to the Creating Data phase – this course is represented by the turquoise arrow in the quality evaluation framework (Figure 3.8). Once a researcher has a plan in place, he or she must make informed quality assurance judgments, which will have an impact on data quality. The research project should begin by moving to the Creating Data phase when data management planning and quality assurance decisions have been made.

The quality process framework, as detailed in Section 6.8, serves as a best practice and quality assurance reference point for data management at all phases of the data lifecycle. The quality process framework includes all applicable measuring criteria as well as data quality dimensions. The goal is for researchers and data curators to use the quality dimensions as a reference point to review RDM discipline

guidelines, standards, and best practices (by assessing their present RDM procedures and making required improvements).

Various interruptions might occur throughout the course of a research endeavour, which can derail or severely affect data management. In the suggested quality process model, these disruptions are represented as 'External influences,' and are located at the bottom of the process structure. External influences are continual and affect all phases of the research data lifecycle because they have the potential to affect any RDM action as data is handled throughout the data lifecycle. To overcome external effects, a researcher must develop management and quality assurance plans (that are informed by best practices).

After collecting data, a researcher must ensure that it is accurate and reliable by performing frequency and spell checks to find and correct errors and inconsistencies. Here, a researcher must accurately and logically characterise variables and, should problems in the data be discovered, a researcher must consult best practices and guidelines (i.e. the quality review framework) to effectively resolve the errors concerned. This course is depicted by the orange arrow in the quality evaluation framework. It is possible to process data only if it is free of errors. The reliability and precision of frequencies and syntax are crucial aspects that a researcher must examine throughout data processing. In the event that discrepancies are detected during data processing, a researcher must clean the data using best practices to ensure that it is correct and error-free. This course is represented by the purple arrow in the quality process model. The data can then be analysed once it is free of discrepancies. All value labels, codes, variable names, grammar, anonymisation, and speaker tags must be correct during data analysis for the data to make sense and answer research questions and the quality review is documented at this point. The quality process model mandates that errors and inconsistencies must be handled if the data does not make sense. This signifies that a researcher must consult best practices and rules (such as the quality review framework) in order to correct errors and inconsistencies; the red arrow represents this path. All mistakes and inconsistencies should be handled at this point, thereafter the data can proceed to the next stages of the data lifecycle (i.e. preserving data, giving access to data and reusing data). Data that can be clearly analysed can be preserved and the quality evaluation system involves the addition of preservation metadata as part of the preservation. If preservation metadata is not added, a researcher must consult the quality review framework and utilise the appropriate preservation metadata, the blue arrow in the quality process model depicts this path. Access to data can be granted when it has been adequately preserved (including the addition of preservation metadata). A researcher must ensure that the repository documents the level of data quality at this stage (i.e. allowing data access). If a repository does not guarantee that the level of quality is documented, a researcher must reference the quality review framework, which is represented by the green arrow in the quality process model. Data can be reused if the repository can provide assurance that the degree of quality has been documented. Preservation metadata is included, the level of quality assured by the repository is documented, and the level of

quality is validated in the remaining stages of the data lifecycle. In the same way that the first three stages of the data lifecycle (Creating data, Processing data, and Analysing data) require a researcher to be informed by quality assurance best practices and guidelines, so too do the last three stages of the data lifecycle.

3.8.8 Depicting the quality process model

During this step, the relationships between different components have to make sense in order to achieve the specified goal at the beginning by creating a representation (Center for Community Health and Development, n.d). All of the framework's building pieces are an empirical component in achieving the desired transformation. In this step, the following elements are clarified:

- *An estimated time sequence to organize the components and elements of the system or model (what happens before what) (Center for Community Health and Development, n.d);*
- *Arrows or other strategies for transmitting paths of control and event sequences. Some arrows may point to demonstrate contact or reciprocal impact in both directions (Center for Community Health and Development, n.d).*

According to Principle 6 of the ‘Conceptual Framework for Data-Driven Decision-Making Skills’, in order to attain the aforementioned domains, “it is necessary to understand the outcomes of the process” (Mandinach et al., 2006), which are to verify whether the defined standards, guidelines, and best practices actually assess the selected quality dimensions and if the quality process framework is approved by the participants. A thorough knowledge of what the proposed framework will do is required at this level.

While the initial picture of the quality process model (as presented in Chapter 2) was mostly centred on theoretical principles and the reviewed literature, a thorough literature review was necessary in order to identify relevant quality metrics. This guaranteed that the concepts of 'quality,' 'quality assurance,' and 'data quality' would be addressed from every RDM perspective. Figure 3.8, below, depicts the initial quality process model (for flowchart legend descriptions, see Appendix 6).

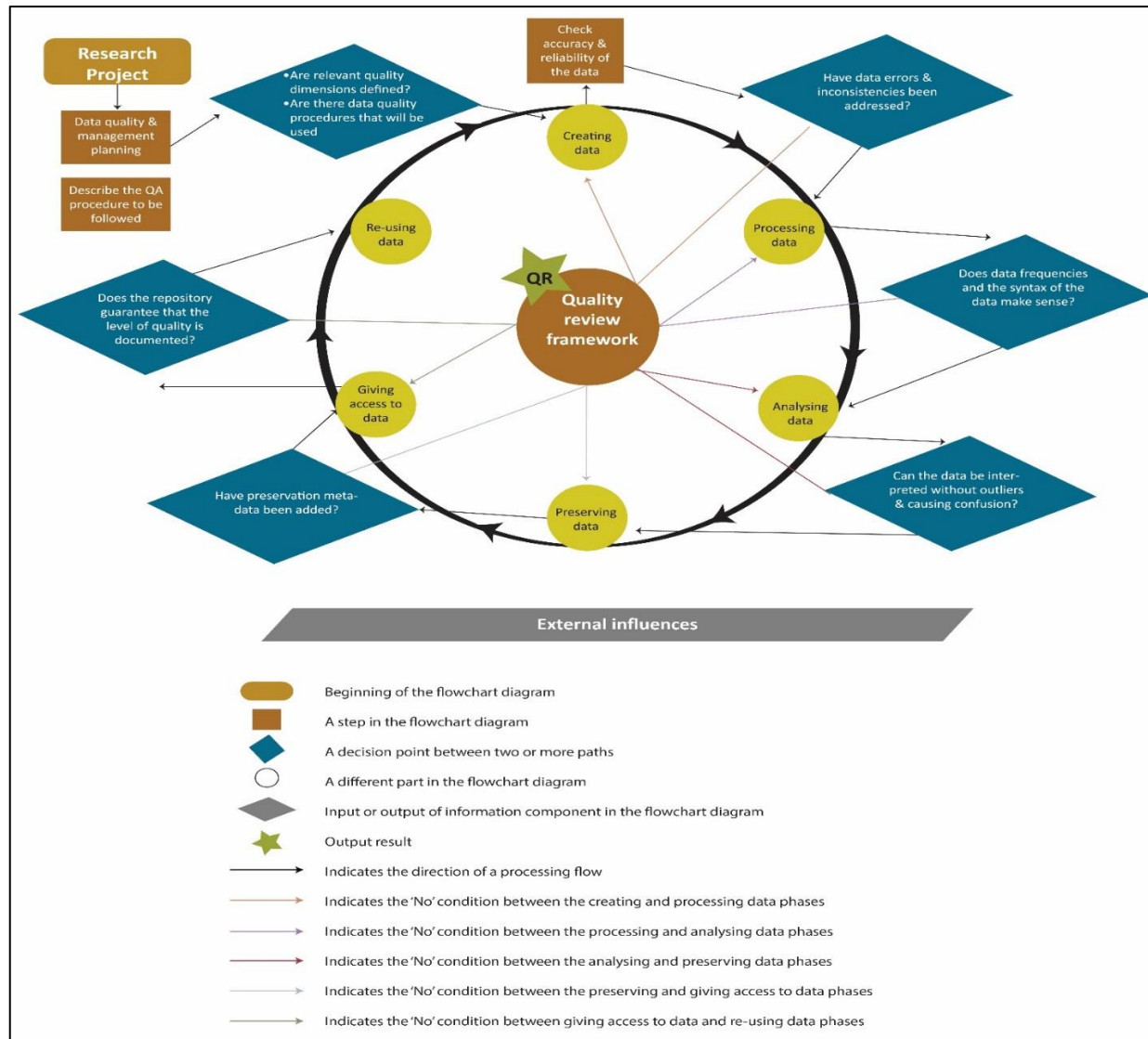


Figure 3.8 - Proposed quality process model

An explanation of the quality process ‘Decision’ feedback loops

The quality dimensions in the quality evaluation framework do not apply to all the stages of the research data lifecycle and different quality dimensions apply to certain phases. In the proposed (initial) quality process model, should the answer to a specific question be ‘No’, and should the model also make reference to the quality review framework, this means that the quality dimensions for that specific phase of the research data lifecycle will apply. Appendix 13 shows each data phase of the data lifecycle and its associated quality dimension. The following description is an example of this process using the Creating Data phase:

- If the answer in respect of the first ‘Decision’ (i.e. Have data errors and inconsistencies been addressed) is ‘No’, the framework takes you through the quality review process. Since this ‘Decision’ condition comes from the Creating Data phase, the quality dimensions (as part of the quality review framework) of this phase will apply (as indicated in Appendix 13). In the event that the response is in the affirmative, a researcher must explain and adhere to the necessary measuring criteria as outlined in the Quality Process Framework (Appendix 2):

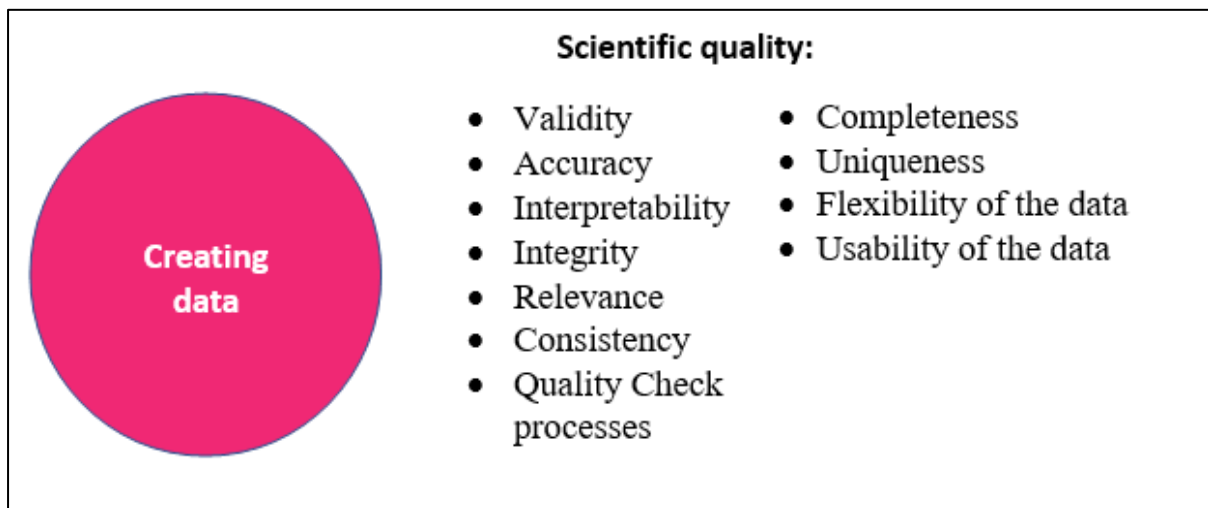


Figure 3.11 - Creating Data phase quality dimensions as part of the quality review framework

This means that a researcher has to answer all the questions and meet the criteria of the quality dimensions assigned to the Creating Data phase. Once this is done successfully, the quality process framework progresses to the next phase (i.e. ‘Processing data’). This approach applies to all the ‘Decision’ conditions associated with every phase of the data lifecycle as these appear on the quality process framework. The quality process framework should be used together with Appendix 13 (i.e. Grouping of quality dimensions with stages of the research data lifecycle).

When data is effectively managed throughout the research data lifecycle, a 'quality review' (represented as QR) is generated by the quality process framework. Because the review is influenced by data quality dimensions, which ensures a quality-based assessment, the quality review is closely associated with the

quality review framework. The graphics in the 'Reviewing data quality' section show the relationship between the data quality dimensions, the type of quality to which they belong, the review procedure, and how it all comes together to produce the quality review results:

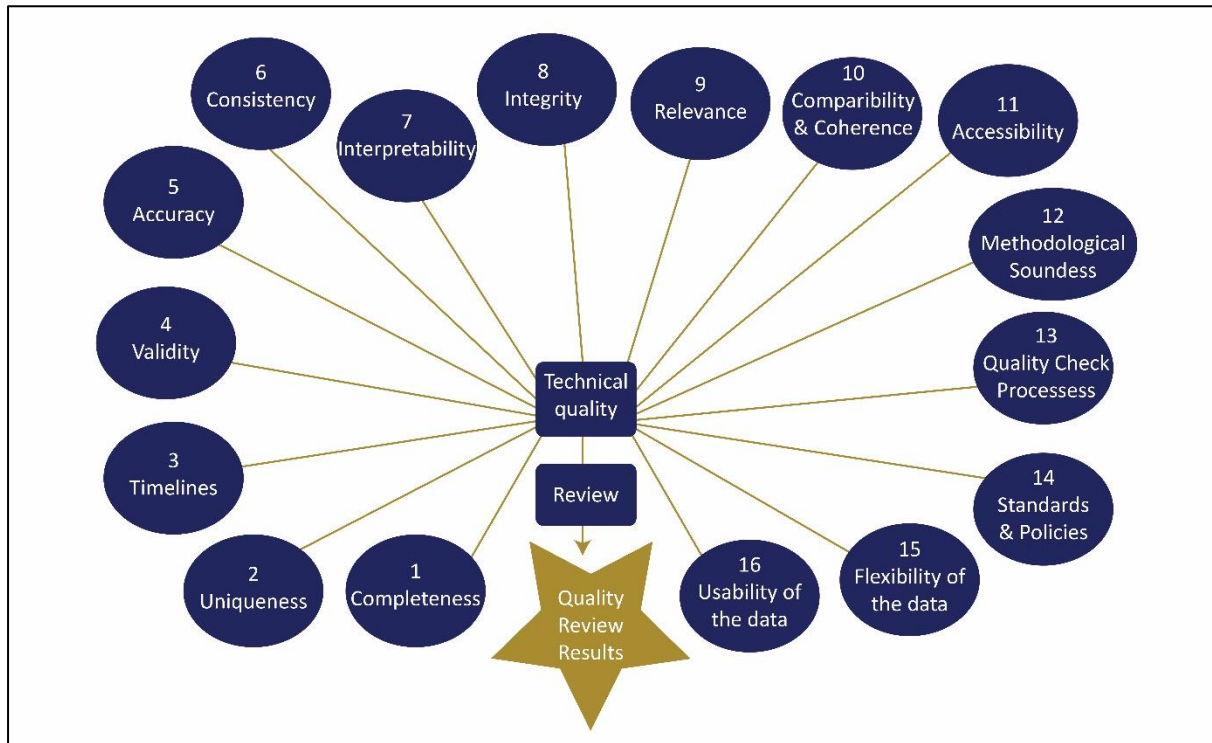


Figure 3.9- Reviewing data quality (Technical measures)

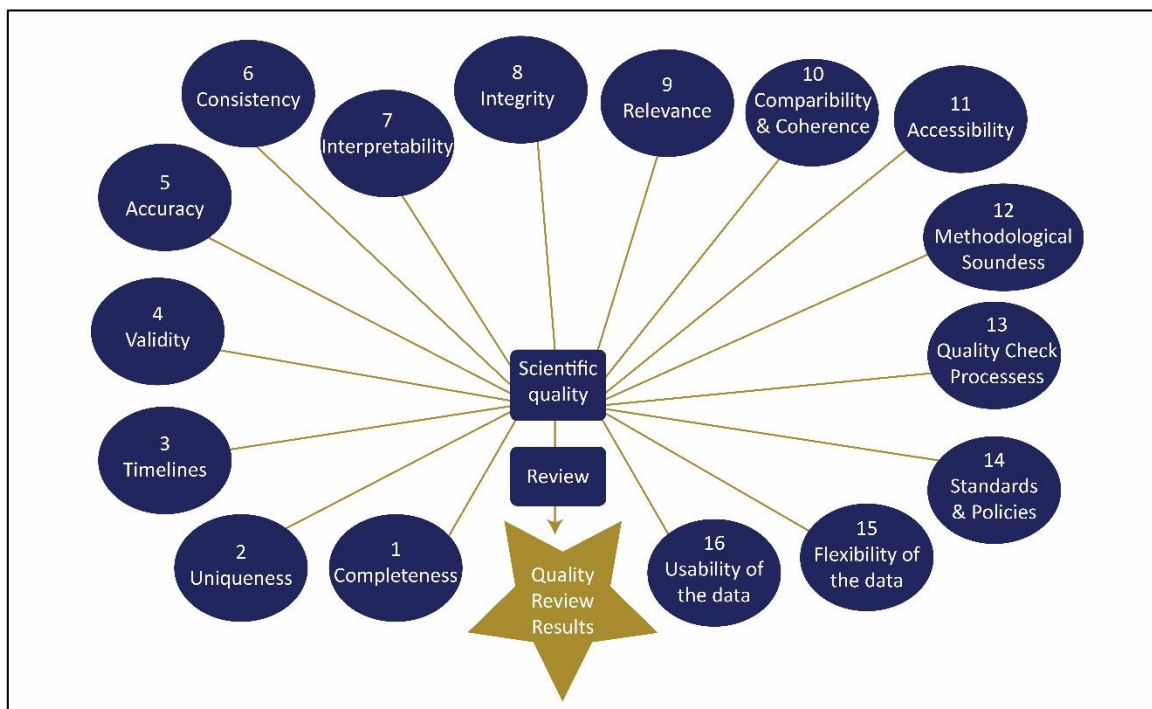


Figure 3.10 - Reviewing data quality (Scientific measures)

The figures above depict the relationship between technical and scientific criteria and all data quality parameters. Quality review findings are generated after data is assessed using the Technical and Scientific quality dimensions. Figures 3.10 and 3.11 show that the results of quality reviews are evaluated using both technical and scientific data quality standards.

3.8.9 Checking the completeness of the quality measuring framework

This phase is regarded as a 'piloting stage' in which the assembled framework is tested and validated for practicality and completeness (Center for Community Health and Development, n.d). Before implementing the quality framework with the research participants, the author felt it was necessary to pilot it first. In the words of Kim (2010:191) "a pilot study is referred to as a feasibility study that comprises 'small-scale versions' of the planned study". The purpose of the pilot study was to ensure that the quality framework was complete. The data quality dimensions needed to be confirmed based on the inputs from the pilot study to ensure that they assessed the relevant aspects of data quality and quality assurance. Checking for completeness provides a track record of how a framework has performed in both the piloting stage and in a real-world situation. According to the Center for Community Health and Development (n.d.), the following tasks are engaged in this stage:

- *Picking an actual or hypothetical case scenario where you can get input on your logic model*
- *Checking for the utility of model elements (e.g. was it understandable?)*
- *Checking for model completeness (e.g. what was missing?)*
- *Revising, making the model more complete, and adding new components.*

The quality framework developed required testing using study participants with similar backgrounds to the real research participants. Pilot studies are also defined as "mini versions of a full-scale study (also called 'feasibility' studies), as well as the specific pre-testing of a particular research instrument such as a questionnaire or interview schedule" (Van Teijlingen & Hundley, 2001). The pilot study was carried out with the help of a senior data manager with extensive RDM and curation experience. The proposed quality framework was used as a reference guide during the piloting stage in order to ensure that the quality framework was both useful and complete. The pilot participant was presumed to be aware of RDM and its significance when conducting a research study. The pilot participant was also expected to be familiar with the majority of the terms used in the quality framework. During the interview session, any confusing sections were discussed to ensure that the various components of the quality system were understandable to the participant.

The primary target audience for the suggested framework is researchers and data curators and it is critical that the framework be explained to them. In addition, researchers and curators are expected to be familiar with all of the technical parts, features, and context of the framework. The input obtained from

the pilot participant regarding how complete the proposed quality framework was revealed that the framework would be easily understood by researchers and curators.

3.8.10 Putting the framework to use

Crawford (2020:44) is of the view that “an essential source of a framework is the published research literature on [one’s] topic”. Principle 5 of the ‘Conceptual Framework for Data-Driven Decision-Making Skills’ (Mandinach et al., 2006) requires the identification of any external influences that may have an impact on the system. Since the quality framework aims to improve institutional RDM, it was vital to utilise it when interviewing researchers (from diverse research programmes) and data curators. Prior to their interviews, researchers and data curators were asked to interact with the quality framework and provide feedback on their interpretation of the framework, quality assurance, and data quality problems.

Using a developed framework in a real-life context has the potential to generate outcomes that may or may not be as anticipated (Center for Community Health and Development, n.d). This reality necessitates the development of a management strategy as well as allowances for mid-course adjustments. An author is thus able to present his/her asserted remedies to the problem s/he has defined (Adom & Hussein, 2018). Changes in the middle of a project are advantageous since they allow an author to better identify which aspects of a framework require revision. (Center for Community Health and Development, n.d). Since any new project requires buy-in from its target population and stakeholders, advocacy and coordination are key parts of putting the framework into action. In order for this shift to be realistic, it had to include:

- “Orientation and support for those doing the job, using it to explain how project or program elements work together, where contributors fit in, and what they need to be able to make it work.
- Using execution to evaluate what components you have in your project or program and those you do not have, create a management plan and make mid-course adjustments.
- Communication and lobbying were used to justify the progress of the plan/program and to demonstrate how investments can be used” (Center for Community Health and Development, n.d).

Participants in the study must have a thorough comprehension of the entire quality framework at this point. The theoretical quality aspects included in the framework, on the other hand, were thought not to be well comprehended by research participants. Kivunja (2018:45) supports this viewpoint in that “theory, research, and practice are part of a continuum for understanding the determinants of behaviours, testing strategies for change, and disseminating effective interventions”. By putting the framework (theories) to use, the main results pertaining to the notion of data quality and how research

data should be managed as a result will be disclosed. A theory, according to Gabriel (2008), is a broad statement of abstractions or ideas that asserts, explains, or anticipates correlations or connections among occurrences. The theories that were used to create the process framework and quality model were applied in reality at this point. The proposed quality model thus served its intended aim of determining the extent to which researchers and curators adhered to the data quality characteristics described in this study.

3.8.11 Revising the model

It was expected that the research participants' experiences and understanding would provide significant insight. Their inputs influenced how the framework was modified and expanded to address the study's research questions and it became feasible to identify the missing components in the researcher's RDM behaviour at this time. Researchers were aware of the necessity to manage their research data, and it is noted that the Data Curation Unit provides RDM and curation support to the organisation's numerous research units. The degree to which researchers and curators comprehend what the data quality dimensions imply and require is a key study topic investigated by the quality framework.

The quality framework will be modified based on research participants' interpretations and input. At this point, the aim and compatibility of the expansions were discussed with study participants.

3.9 Summary

While RDM is in progress, the development of a quality process model acts as a tool to guide quality. Clarification of various concepts, as well as quality assurance and the generation of high-quality data, necessitated consulting a variety of resources. The framework development strategy of the University of Kansas (2023) provided a guiding outline that aided in the creation of the suggested quality process model. Reviewing numerous development frameworks in the literature was necessary for developing a measurement framework. To solve data quality challenges, the use of quality assurance and RDM best practices were indicated here. The links between variables in a framework need to work together to allow for beneficial outcomes and they do so by integrating similar components, thereby serving as a guide for boosting usefulness. The construction of the measuring framework, as addressed in this chapter, served as a starting point for the development of the quality process model.

The development process was not achieved in a vacuum and various framework development principles, data quality dimensions, frameworks and models were consulted to achieve this objective. Possible measuring criteria were also established and these may be utilised to determine which standards, best practices, and guidelines are available to serve as proxies for determining conformity with the chosen quality dimensions. The chapter concluded with a discussion of the proposed framework for quality-focused RDM. A research approach was employed to contextualise the

methodology of this study with the proposed frameworks. The methodology approach employed in this investigation is discussed in the following section.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 Introduction

The goal of this chapter is to explain the research methods and design that were used to look into how researchers and data curators think about data quality assurance and data quality. The chapter begins with a summary of the study's objectives and goals, followed by a description of the research paradigm and research design approach. A discussion of the research design used for the study, as well as the target population, data collection methods, and the pilot study, is also included. The chapter concludes with an explanation of data collection tools, data sampling for the study, research location selection, data analysis and interpretation, data management plan, validity and reliability.

The method of investigation behind a piece of research consists of a number of elements that determine how that study will be carried out. The approach to research and the strategies utilised to complete the thesis are documented in the chapter on research methodology (Onwuegbuzie and Collins, 2007). The chapter also covers data collecting, data processing, and interpretation methodologies. Research activities are well-organised and planned in order to add to the body of knowledge already available (Fellows & Liu, 2015). Getting started on a research endeavour is not a one-dimensional process that takes place in a vacuum; rather, it is structured within a framework of inquiry (Schmerse et al., 2024). Because the current research set out to construct and examine the utility of a developed framework based on data quality for a social science organisation, it is classified as a qualitative rather than a quantitative study.

4.1.1 Short overview and goal of the study

The purpose of the study was to develop a framework for data quality-focused RDM that could be used to guide the data management planning process in a social science organisation. The study attempted to solve the research questions (as stated in Chapter 1) and analyse RDM, quality assurance, and researchers' perceptions of data quality, as well as the role of data curators in persuading researchers to provide high-quality data. The social sciences organisation had eight research units as of January 2022. The research units are made of research teams that comprise junior researchers, administrators, principal investigators and senior researchers with various responsibilities. For the purposes of this research study, principal investigators, senior researchers and data curators were included in the sample. Data capturing errors are common when data is collected (during fieldwork) and this has a detrimental impact on the dependability and correctness of such data. The current situation at the social sciences research institution is that, after a researcher has completed his or her study, the data is submitted for curation (to the data curation unit). Only during this step do curators notice errors and flaws in the data that have a negative impact on the quality. In addition to the errors that occur during data collection, errors also frequently arise throughout the data processing stage due to mismanagement and technical issues.

Curators can help by providing data processing instructions that reduce error production while data is collected and processed.

In order to develop a framework for data quality-focused RDM that is able to guide the data management planning process in a social sciences organisation, the main research question is:

“What would an effective, quality-focused framework, to guide the data management activities of a Social Sciences Research Organisation, look like?”

Additional research questions must be answered in order to establish a framework for data quality-focused RDM, including:

1. What is meant by the term ‘data quality’?
2. How does the network of stakeholders, involved in research data management, address data quality issues?
3. What will be the characteristics of a framework for quality-focused data management?
4. Considering the data quality dimensions, how is data quality measured?
5. What are the unique remedies that the proposed quality-focused framework could incorporate?
6. What recommendations regarding change management would be applicable to the changing of RDM practices at South African social science research organisations so as to include data quality?

To guide the data management planning process, the proposed frameworks for data quality-focused RDM strive to understand researchers' expectations and views of quality assurance as data is managed across the research data lifecycle. The main problem is that data curators are not involved in the data planning, gathering, or processing stages. However, when data sets are shared, the same curators must attest to the quality of the data. Researchers plan, acquire, and process data using their own methodologies. In order to answer the above research questions, a suitable research methodology had to be developed. The methodology is embedded in a research paradigm. The following section discusses research paradigms and then explains why a specific paradigm was chosen for this study. A proper research approach had to be created in order to answer the above research questions. A research paradigm is embedded in the methodology. The next section addresses research paradigms before explaining why the particular paradigm adopted was chosen for this study.

4.2 Research paradigm

A research paradigm is defined as “a collection of overarching and interconnected assumptions about the essence of reality” (Chand, 2017). Human constructs are types of research paradigm that deal with the fundamentals of showing where the researcher is coming from in order to generate meaning from data (Denzin & Lincoln, 2000, as cited in Kivuja & Kuyini 2017:26). Creswell (2014) views a research paradigm as epistemology or ontology, or even research methodology. The most appropriate research paradigm must be chosen since it provides the framework for a research study and permits research questions to be answered. The next section examines the most common research paradigms in social science research, as well as the paradigm ultimately selected for this study.

4.2.1 Interpretivist paradigm

According to Willis (2007), the interpretivist paradigm typically attempts to explain a specific context, and the central conviction of this paradigm is that truth is socially constructed. This means that a researcher investigates an issue from the subject's point of view. It allows a researcher “to view the world through the perceptions and experiences of the participants” (Thanh & Thanh, 2015:24). In order to accomplish this, a researcher must have a thorough awareness of the context and organisational setting of the subject (i.e. academic, technical, or social). A researcher should embrace and seek out the interpretivist paradigm, which includes multiple perspectives, openness to change, the use of iterative and dynamic data collection procedures, the promotion of participative and holistic analysis, and the use of inductive and deductive methods (Willis, 2007). The interpretive paradigm is centred on subjectivity rather than objectivity as the main component. The major goal is to consider the participants' own perceptions and perspectives. Thanh and Thanh (2015:25) state that “in order to explore understandings of participants, an interpretive methodology provides a context that allows a researcher to examine what the participants have to say about their experiences”. This attribute is ideal for this study since there is no better way to understand data management issues that affect the quality of data in a social scientific organisation than for participants to elaborate on them.

The interpretivist paradigm considers a variety of perspectives and helps to provide a more accurate understanding of the event (Morehouse, 2012). All of the characteristics highlighted about the interpretivist paradigm imply that it would be a useful paradigm to explore with regard to researchers' and data curators' perspectives on data quality. The interpretivist paradigm is expected to aid in obtaining data curators' perspectives on their role in influencing researchers to generate high-quality data. The level and type of support that curators require during the stages of data planning, data collection, and data processing are consistent with the findings of the current research and it is for this reason that the interpretivist paradigm was chosen for the study.

4.2.2 Positivist paradigm

The positivist paradigm differs from the interpretivist paradigm in that it is based on the belief that science and scientific endeavour are the only ways to learn about facts. Park et al. (2019:690) explain that “positivism relies on the hypothetic deductive method to verify a priori hypotheses that are often stated quantitatively, where functional relationships can be derived between causal and explanatory factors (independent variables) and outcomes (dependent variables)”. The positivist paradigm is controlled to a certain degree considering that a researcher is objectively restricted to records collection and evaluation (Creswell, 2014). The study outcomes under the positivist paradigm are usually quantifiable to a significant extent and derived from direct observations. A key factor of this paradigm is that “real events can be observed empirically and explained with logical analysis” (Kaboub, 2008:343). In the positivist research paradigm, real occurrences are recreated in a laboratory-like setting by conducting micro-level research studies that remove the external world's ambiguity (Kaboub, 2008). The positivist research paradigm is used to draw conclusions based on findings from controlled experimental methodologies, which are generally used to inform policies and decision-making. When a researcher employs the positivist paradigm, it is important to be aware of some of its major shortcomings. Experience is a dependable source of understanding under the positivist paradigm. A broad range of fundamental and essential principles, such as cause, time and space, are, however, not based on experience (Business Research Methodology, 2018).

This section considered the two basic types of research paradigms that are widely employed in social science research and justified the choice of the interpretivist paradigm to be applied in this investigation. The next section discusses various types of research approaches before describing the one used in this review.

4.3 Research approach

In the field of academic research, there are two main research methodologies that can be utilised to collect data for research objectives. These research methods offer a variety of options for data collection and information exploration. There is a clear distinction between quantitative and qualitative research approaches. Mixed-method research refers to the combination of these two approaches. Each of these is discussed in further depth below.

4.3.1 Qualitative research

Qualitative research is a method of investigation that entails looking at a social or human problem from various angles. It is conducted in a natural setting and entails the creation of a dynamic and complete depiction of the topic under investigation (Creswell, 2014). When qualitative research is used, the goal is to fully comprehend an issue from several perspectives, allowing the researcher to create a nuanced and holistic picture of the phenomenon of interest (Creswell, 2014). As a definition, data quality is a

subjective phenomenon that necessitates extra attention in order to perceive it from a personal perspective. Different interpretations of qualitative research seek to better convey the subject in its entirety. The following is a detailed explanation of qualitative research, as given by Mason:

“...grounded in a philosophical position which is broadly ‘interpretivist’ in the sense that it is concerned with how the social world is interpreted, understood, experienced, produced or constituted. While different versions of qualitative research might understand or approach these elements in different ways (for example, focusing on social meanings, or interpretations, or practices, or discourses, or processes, or constructions), all will see at least some of these as meaningful elements in a complex – possibly multi-layered and textured – social world” (Mayson, 2002:3).

This description outlines all of the key features that make qualitative research a valuable tool for determining how study participants perceive and experience a phenomenon. Qualitative research has the distinct advantage of being “good at simplifying and managing data without destroying complexity and context” (Atieno, 2009:16). The responses of research participants are presented without being tempered, which means that finer input data of research participants is obtained. As a result, the qualitative technique was perfect for this study, which focused on identifying researchers' and data curators' perspectives on data quality and quality assurance. Qualitative research establishes a formal framework that allows all participants in the study to use the same set of questions. Atieno (2009:16) adds to this by saying that “qualitative methods are highly appropriate for questions where pre-emptive reduction of the data will prevent discovery”. This aspect of qualitative research provides the author with an ideal opportunity to learn more about RDM methods and the role data curators play in encouraging researchers to provide high-quality data. Another feature of qualitative research that lends credence to this study is its ability to adjust to the nature of the knowledge acquired. If the available data does not appear to yield any results, the research will shift gears right away and attempt to collect data in a different direction. This helps a researcher to guide the interview questions deliberately to areas with more useful references and meanings. According to Austin and Sutton (2014:436), “qualitative research involves asking participants about their experiences of things that happen in their lives. It enables researchers to obtain insights into what it feels like to be another person and to understand the world as another experiences it”. The following are examples of qualitative research methods:

Phenomenological research

Via explanations given by the individuals involved, a phenomenological study explores human experiences; these experiences are called lived experiences (Donalek, 2004). It is tough to assess and identify a research participant's current perspectives and experiences, and it is even more challenging to do so by identifying earlier experiences of research participants. Because the purpose of

phenomenological research is to define a lived experience rather than clarify or measure it in any way, it differs slightly from other types of qualitative research. Neubauer et al. (2019:92) claim that “the goal of phenomenology is to describe the meaning of this experience – both in terms of what was experienced and how it was experienced”. Live experiences provide rich, raw, and thorough insights that can be given only through first-hand experience. The typical steps in a phenomenological research study are detailed in the table below:

Table 4.1 - Steps of a phenomenological research study

Step	Description
<i>Bracketing</i>	This is the process by which any preconceived ideas, views or notions about the phenomena being studied are detected and held in check (Neubauer et al., 2019). This implies that a researcher detects potential perceptions and prejudices that may cause those biases against the studied phenomenon.
<i>Intuition</i>	The phenomenon is completely understood by researchers. In this case, the researchers avoid criticising, evaluating, or expressing an opinion on items supplied by participants, instead focusing on the phenomenon under investigation, resulting in a more accurate image (Umanailo, 2019).
<i>Analysis</i>	The researchers determined the significance of the event that was discovered and investigated the similarities and connections between the data and other phenomena. Important data is thoroughly studied (Umanailo, 2019). This is when a researcher uses his analytical and understanding ability to create patterns that explain the experience accurately from the viewpoint of those who have encountered it.
<i>Description</i>	The important parts that are based on the categorisation and grouping phenomenon are communicated and described in writing by researchers. The goal is to explain the experience's meaning and relevance (Umanailo, 2019). A researcher may have thoroughly analysed the perspectives of a phenomenon during this process and will formulate independent conclusions.

Phenomenological research aids in the development of a comprehensive, precise, straightforward, and articulate explanation and interpretation of a single human event or experience. This method of inquiry

is utilised to investigate areas where there is little literary expertise. The key factors in phenomenological studies are the way a person experiences responses, emotions and perceptions (Donalek, 2004). This is not an acceptable strategy for this study because the subject matter (RDM, data quality) is well-known and discussed in the literature.

Ethnographic research

Ethnographic studies are part of the data collection and analysis for cultural groups. Ethnographic research is a research method using “naturalistic, open discovery, and largely inductively derived emic modes and processes with diverse strategies, techniques, and enabling guides to document, describe, understand, and interpret meanings, experiences, symbols, and other related aspects bearing on actual or potential nursing phenomena” (Farren, 2015:419). Direct observation and definition construction are the major goals of ethnographic investigations. The aim is to become part of a culture they want to learn and then, through the 'eyes' of its members, show an image of that culture (Farren, 2015).

The data gathered through interviewing study participants may be limited in its ability to generate diverse types of outcomes and a better knowledge of social action. Reeves et al. (2013:1365) elaborates, stating that by way of collecting “observations, interviews and documentary data, which are triangulated (i.e. compared and contrasted with one another) ethnographic research offers a qualitative approach with the potential to yield detailed and comprehensive accounts of different social phenomenon (actions, behaviour, interactions, beliefs)”. The ability to evaluate a phenomenon and gather opinions that are generally 'hidden' from the public gaze is at the heart of ethnographic research. Social phenomena may share characteristics, despite not being connected because of their overlapping relationships. The primary goal of ethnographic research is to develop a comprehensive view of social accounts. Researchers and curators must be asked thorough questions about their RDM activities in order to determine what data quality means to them. This does not include investigating and watching their cultural lives in order to infer significance. As a result, ethnographic research is insufficient for this topic.

The qualitative research approach enables a researcher to comprehend the relevance of the activities, conditions, and behaviours of interest, as well as the descriptions of life and experiences provided by research participants. This feature of qualitative research is one of the key reasons why it is the best research method for this research study. The qualitative technique is an excellent way to investigate and reveal the complexities of a complicated subject like data quality, quality assurance, and the difficulties and practices that research participants experience in a social sciences organisation. The next section looks at the quantitative research approach.

4.3.2 Quantitative research

Contrary to qualitative research, quantitative research is a type of research “in which the researcher decides what to study; asks specific, narrow questions; collects quantifiable data from participants; analyses these numbers using statistics; and conducts the inquiry in an unbiased, objective manner” (Terrell & Edmonds, 2017:4). The primary focus of quantitative research is numbers. This is in line with Muijs' definition of quantitative research as being “an explanatory phenomenon” because it collects numerical data that are analysed using mathematically based methods (in particular statistics) (Muijs, 2004). Quantitative research differs from qualitative research in that it tests a hypothesis, analyses cause and effect, and predicts outcomes, whereas qualitative research tries to comprehend and interpret social experiences (Creswell, 2014). When a researcher employs a quantitative research approach, they simply calculate and evaluate data using statistical procedures to determine whether or not a hypothesis is valid. Because the expectations and behaviours of a social sciences organisation's data quality are not evaluated as a hypothesis, but rather objectively examined, this approach is unsuited for this study. According to Creswell (2014), the aim of quantitative methods is to determine whether a theory's predictive generalisations are valid. Qualitative research attempts to fully analyse a human or social problem from various points of view and viewpoints, while quantitative research seeks to clarify a phenomenon by gathering numerical data (Muijs, 2004). Since it aims to analyse and understand the perceptions, objectives, and behaviours of data quality in a social sciences organisation, and establish a framework that focuses on data quality that helps direct the organisation's development of good data quality, this study can be classified as qualitative rather than quantitative research.

In quantitative research “data collection is highly managed and organized, resulting in a researcher having a profound influence on the outcomes of the investigations under research” (Finlay, 2006:3). Quantitative research focuses on data in the form of numbers and figures that may be used to compute or quantify a wide range of things. Adu (2015) used a descriptive figure to describe the key differences between quantitative and qualitative studies:

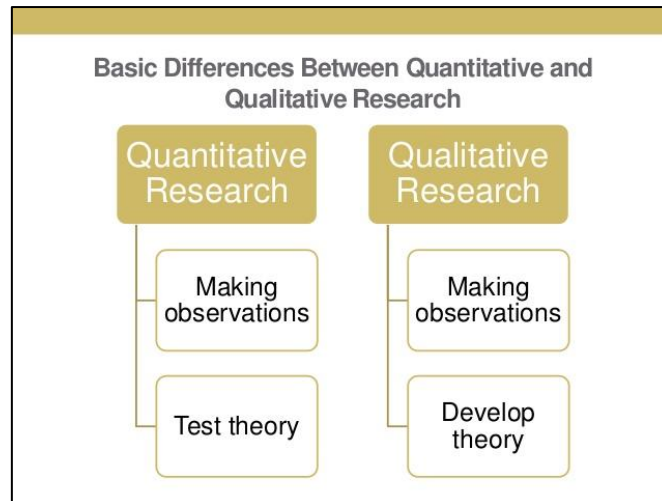


Figure 4.1 - Basic differences between quantitative and qualitative research (Adu, 2015)

Quantitative research, like qualitative research, uses a variety of approaches, including correlation research and survey research:

Correlation research

In this type of quantitative research, the differences in features or variables of two or more entities are investigated. According to Kumar (2011), a correlation exists when one variable increases or decreases in line with the other variable. In a specific group, a researcher will collect data on two or more variables. A correlation study is the best research strategy for a researcher whose research goal is to explore the relationship between two variables. This is because correlational research is a type of non-experimental study in which the researcher calculates two variables and evaluates the statistical relationship (i.e. the correlation) between them with little to no regard for foreign factors (Price et al., 2013). A researcher may consider employing correlation research after creating and studying statistical correlations (which can be deduced between two variables). However, correlation research implies that a researcher cannot influence the independent variable because it is difficult, impractical, or unethical (Price et al., 2013). The degree of control that a researcher must be able to exert during a corrective study is reduced. Cross-tabulation and correlations are typically the methods used in the correlation study (Simon & Goes, 2011). Because of the open nature (lack of control) of correlation research, it lacks the flexibility needed to completely uncover perceptions and practices of data quality, making it unsuitable for this study.

Survey research

Survey research is a method of gathering data in a standardised manner, which “focuses on people, the vital facts of people, and their beliefs, opinions, attitudes, motivations and behaviour” (Mathiyazhagan & Nandan, 2010:34). When specific features within a population are described, they can be beneficial for research reasons. These qualities are also easily recognised within a group or feature as a whole.

Glasow (2005:1) states that “survey research is used to quantitatively describe specific aspects of a given population [and that these] aspects often involve examining the relationships among variables”. A researcher's capacity to acquire personal data from a large group hinges on time and requires a lot of resources. In such a case, surveys can be used to conduct research on a subset of the population, with the results then being generalised to the entire population (Glasow, 2005). This element of the survey is especially valuable in census surveys because of the large sample size covered by such a survey. Because a researcher may utilise other elaborative ways to obtain data, surveys on a smaller scale do not have the same impact in gathering the needed data. Even though a survey is a method of gathering information on a large group of people's characteristics, behaviour, or attitudes (Pinsonneault & Kraemer, 1993), it is not appropriate for a small population. The number of research participants in the current study rendered survey research unfeasible. One of the purposes of this study is to evaluate the attitudes towards and behaviours of researchers and curators in respect of data quality. Furthermore, the author's primary goal was to comprehend the specific RDM issues that arise during the research data lifecycle.

4.3.3 Mixed-method research

Mixed-method research is a type of methodology in which a researcher uses a combination of qualitative and quantitative research methodologies (Terrell & Edmonds, 2017). Creswell (2009:9) defines mixed-method research as “both a method and methodology for conducting research that involves collecting, analyzing, and integrating quantitative and qualitative research in a single study or a longitudinal program of inquiry.” Creswell (2009:9) continues that the purpose of mixed-method research is that, when they are combined, qualitative and quantitative research deliver a better understanding of a research problem or issue either of the approaches alone – and that this is also one of the key advantages of the approach.

When a researcher uses mixed-method research, he or she has the knowledge and experience to benefit fully from both methodologies. It is important to understand the advantages and disadvantages of mixed-method research, which include:

Advantages:

- A researcher has the advantage of using numbers to make phrases, images, and narratives more accurate (UK Essays, 2017).
- Provides versatility in methods and is adaptable to multiple designs of studies, such as observational studies and randomised trials (Wisdom & Creswell, 2013).
- The use of the mixed research approach helps the researcher to answer a wider and more complete range of research questions (UK Essays, 2017).

Weaknesses:

- A researcher must learn about different strategies and methods and understand how to combine them properly (UK Essays, 2017).
- Studies with mixed approaches are difficult to prepare and execute, raising the difficulty of assessments. To explain all aspects of the study, careful preparation is needed (Wisdom & Creswell, 2013).
- The mixed research approach is more expensive and time-consuming than any other research method due to its duplicity of content (UK Essays, 2017).

Depending on what a researcher wants to accomplish, the research methodologies discussed in this portion of the chapter can be utilised to answer research questions. The quest to comprehend participants' RDM actions involves a thorough examination of their practices and perspectives. In order to investigate RDM, quality assurance, and data quality, the author used a qualitative technique. To help achieve this goal, the next section discusses the specific research design of the study.

4.4 Research design

A plan is required to address research questions in a research investigation. A research design is a method for accomplishing this goal that includes a strategy for gathering and analysing data. A research design involves the structure and environment that make up the scientific and academic ideas of a certain study field (Creswell, 2014). The research designs used in this study are discussed in this section.

4.4.1 Case study

In a case study, the researcher investigates a particular entity or phenomenon and explains the interactions and/or actions of a person maintained by an outside observer (Creswell, 2009). A case study allows a researcher to focus on a single problem and discuss it thoroughly. Burns and Grove (1999:23) explain that “the word ‘case’ means ‘an’ and the central feature of case study research design is the investigation of the one or more specific ‘instances of’ something that comprise the cases in the study”. A study subject is chosen by a researcher based on characteristics that can be researched and give meaningful results. Because there are many facets of what comprises a case study that can be explained at many levels, a case study broadens the meaning of a topic. A case can be anything physical, like an organisation, a group, or a person, or something more abstract, like an event, a management decision, or a change programme (Burns & Grove, 1999). As the chosen case study to be researched, the social sciences organisation that has been identified is a part of this research.

The subject's day-to-day environment, what they do, how they interpret the world around them, and how to represent these scenarios can be difficult to complete. A case study is a solution to such a problem because “a case study is an empirical research method used to investigate a contemporary

phenomenon, focusing on the dynamics of the case, within its real-life context; case studies are most suitable for ‘how’ and ‘why’ types of research questions” (Teegavarapu et al., 2007:4). A researcher's data quality prospects are a 'case' that can be explored in and of themselves, but doing a case study requires extensive planning and development of the case study's angle. A crucial aspect of a case study is the questions that will be used. Teegavarapu et al. (2007:5) add to this by stating that “before getting ready to conduct a case study, the observer should have the questions articulated. This is the most time-consuming part of case study design upon which the success depends, and thus should be done meticulously”. Yin (2003) lists some of the most typical aspects of a case study, including:

- “A comprehensive, often longitudinal, analysis of a limited number of cases (prospectively or retrospectively).
- Data about a large number of characteristics of each case is collected and analysed.
- In their real-life context, cases are studied; understanding how the case affects and is affected by its context is also of central interest to case researchers.
- Cases naturally arise in the sense that, as in an experiment, they are not manipulated.
- The use of numerous data sources, including interviews, observations, archival records and even physical artefacts, to enable findings to be triangulated” (Yin, 2003).

The goal of this study is to create a framework that focuses on data quality and then to use that framework to evaluate what data quality and quality assurance mean to researchers and curators at various points of the data lifecycle. Once input from research participants has been gathered, it is expected that the framework will be adjusted to provide an accurate picture of how data quality can be handled. Accordingly, this study used a case study research methodology to better understand the perspectives of data quality among researchers and curators in a Social Sciences Research Organisation.

According to Yin (2003:13), “a case study investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident”. These characteristics of a case study enable a researcher to quickly grasp the gist of the primary themes under investigation. Zainal (2007:1) explains that “case study research, through reports of past studies, allows the exploration and understanding of complex issues - it can be considered a robust research method particularly when a holistic, in-depth investigation is required”. Since study participants are the primary source of findings, seeing subjects in their natural environments is the most effective technique to determine the true existence of factors that influence subjects in their daily activities. This is because “the detailed qualitative accounts often produced in case studies not only help to explore or describe the data

in a real-life environment, but also help to explain the complexities of real-life situations which may not be captured through experimental or survey research” (Zainal, 2007:4).

In the literature, much attention is paid to deciding the sort of case study to utilise when conducting research. The overarching purpose of the study will determine which type of case study design is used (Yin, 2003). The responses that a case study should provide are a crucial factor in deciding which case study to use - i.e. “are you looking to describe a case, explore a case, or compare between cases?” (Yin, 2003:15). After these criteria have been discussed, the sort of case study design to be used should be chosen. The three types of case study research designs include exploratory, explanatory, and descriptive (Hancock & Algozzine, 2006). Merriam (2001) clarifies that doing case study research involves deciding what we know about a research issue in order to assess its relevance and the need for more research on it, to recognise previous research strengths and limitations, and to identify areas of adequate and inadequate research as well as methods used to study it. The type of case study research design to adopt is determined by the study's final goal. The table below lists the various case study research designs available:

Table 4.2 - Case study typology

Case Study Type	Definition
Explanatory	This kind of case study would be used if you were trying to answer a question that sought to explain the supposed causal associations in real-life interventions that are too complex for the survey or experimental techniques. The explanations would connect programme execution in the evaluation language with programme effects (Yin, 2003).
Descriptive	This type of case study explains the intervention or occurrence and the real-life context in which it has occurred (Yin, 2003).
Exploratory	In a specifically formulated theory that can be tested and/or a particular research context that restricts the choice of technique, the exploratory case study explores various phenomena characterised by a lack of detailed preliminary research (Mills et al., 2010).
Collective	In order to have a general understanding, the collective case study is carried out using a variety of instrumental case studies that either take place on the same site or come from several places (Harling, 2012).

When deciding which case study design to use, it is important to keep in mind that “the focus of the study is to answer “how” and “why” questions; you cannot manipulate the behaviour of those involved in the study and the boundaries are not clear between the phenomenon and context” (Yin, 2003:13). The explanatory case study research approach was used for this study because it allowed researchers to comprehensively investigate the phenomena of RDM practices in a social science organisation. Authors such as Stake (2005); Honebein (1996) and Yin (2003) are in agreement when they explain that a case study research design is based upon a constructivist paradigm. The constructivist paradigm is “an approach that asserts that people construct their own understanding and knowledge of the world through experiencing things and reflecting on those experiences” (Adom et al., 2016:2). A case study research design portrays the idea that learning happens only when the learner discovers knowledge in the spirit of exploration and success (Kalender, 2007).

Applications of case study design

The case study research design has traditionally been viewed “as lacking rigour and objectivity when compared with other social research methods” (Rowley, 2002:16). As a result, a researcher must exercise extreme caution while using and conducting a case study as a research strategy. When employing a case study, if the data-gathering strategy is not validated, the study will be derailed because irrelevant data will be obtained. Eisenhardt (2009) explains that case studies have often been seen as a useful tool for a research project's preliminary, exploratory stage, as a basis for the creation of the 'more formal' tools required in surveys and experiments. When it comes to the type of research to be undertaken, case studies should be carefully evaluated. Rowley (2002) established three characteristics to assess whether the optimum research design is a case study research design:

- “The types of questions to be answered,
 - The extent of control over behavioural events, and
 - The degree of focus on contemporary as opposed to historical events”
- (Rowley, 2002:17).

Selecting cases

When an in-depth understanding of an issue, event, or phenomenon of interest in its natural real-life setting is required, the case study approach is particularly effective (Crowe et al., 2011). It should also be established whether the usage of a certain case is appropriate in answering the research questions when choosing a case to study. This is important because:

...objections to single case research can arise from concerns about the representativeness of the chosen case, the extent to which generalisability is possible and the vulnerability to confirmation

bias, the tendency for the collection or analysis of data to be biased so as to confirm the researcher's preconceived opinion (Rose et al., 2015:5).

When a researcher undertakes a study on a certain organisation, these elements must be taken into account. Using a single case study raises the potential that comparative research will not be conducted; yet, there are valid reasons to use a single case study. A researcher must determine what type of question they want to answer when picking a case to study. Rowley (2002) describes the categories of research questions that a research strategy can address:

Table 4.3 - Types of research questions that a research strategy should answer

Choosing an appropriate research strategy	
Strategy	Form of the research question
Experiment	How, why
Survey	Who, what, where, how many, how much
Archival analysis	Who, what, where, how many, how much
History	How, why
Case study	How, why

The how and why questions that a case study is supposed to address are answered by analysing a phenomenon in its real-life setting (Rowley, 2002). When looking closely at the context in which research participants live and understanding the finer nuances, a case study is a realistic and appropriate design to apply. The following section explains why the case study was chosen as the research design for the current research.

4.4.2 Justification for using a case study design

The explanatory case study was chosen as the ideal research design for this study because it provides for an in-depth focus and inquiry on a single social sciences organisation and its present, RDM, quality assurance, and data quality procedures.

According to Chigwada et al. (2017:2), “research data management is important because data are a valuable resource whose production requires time and money”. RDM is a multifaceted ideology that requires thorough examination in order to completely comprehend it and how it is performed. The most efficient way to explain such a complicated phenomenon is to analyse it using methodologies capable of capturing its complexities. Thus, qualitative research is ideal for this study because “qualitative research produces holistic understandings of rich, contextual, and generally unstructured, non-numeric data” (Ponelis, 2015:537). The major goals of the study are to gain a comprehensive understanding of researchers' present RDM procedures and how curators might help influence data quality. Participants

in the study come from a variety of backgrounds (in terms of research management skills and disciplines), therefore their contributions and points of view are diverse. Case study research offers a direct and effective platform for researching, communicating, and assessing participants' RDM behaviours. A case study encourages participants to participate and interact with research participants in their natural environment (Creswell, 2009). Direct inquiry is used to assess RDM processes and establish a quality-focused RDM framework. In line with this, Ponelis (2015:537) claims that “case study research is appropriate for descriptive and exploratory studies that ask ‘why’ and ‘how’ questions”.

Stake (2005) notes that, in its general setting, a case study will concentrate on defining processes, individual or group behaviours, and/or the sequence of events in which the behaviour happens. The contexts in which researchers operate and manage research data are diverse. Different data management procedures take place on the ground (applied by fieldworkers) and in the office (applied in a controlled environment). The case study method is appropriate for studying data management methods in various environments because it provides support for theory development, which is especially useful in areas where present theoretical and conceptual structures are insufficient. The case study research design was deemed well suited for this study since it would allow for the discovery of new RDM, quality assurance, and data quality concepts and discoveries. These results are specific to social scientific organisations – that is, the findings are applicable to other social science organisations.

4.5 Target population

In order for a research study to be successful, research participants must be part of the target group defined for the study. Alvi (2016:10) describes a target population as “all the members who meet the particular criterion specified for a research investigation”. It is critical to carefully select a target population so as to avoid data being collected from inattentive and impracticable people. The target population is such an important entity in research because “it is the researchers’ quest to contribute to academic debate and knowledge they gather data or information from participants” (Asiamah et al., 2017:1607). The participants represent the study's target demographic and they have characteristics that will provide viable study results. All researchers from the numerous research programmes of the Social Science Research Organisation, as well as all data curators employed by the same organisation, were considered members of the target population in this study. The target population totals 16 people. It was thus decided to employ a sampling technique in order to choose a representative but manageable number of persons to interview. The following section goes into further detail concerning sampling.

4.6 Sampling

A research study generally analyses a certain question based on a subject, which must be chosen from a specific population; this procedure is known as sampling. Alvi (2016:11) defines sampling explicitly

as “the process through which a sample is extracted from a population”. A sample is made up of a limited number of participants drawn from a population under investigation. Sampling is an important step because “in an investigation, it is impossible to assess every single element of a population so a group of people (smaller in number than the population) is selected for the assessment” (Alvi, 2016:11). Using a sample enables a researcher to draw informed extrapolations from a larger population, which would not have been possible if the complete population was employed. Asiamah et al. (2017) provide a noteworthy conclusion on sampling by stating that the more representative the sample population is, the greater the precision of the inferences and the better the generalisability of the results. However, the term “generalisable” should be considered in the context of a population that shares similar traits. The larger the sample, the more certain the investigator can be that the results can be extrapolated to the target population. There are various sorts of sampling procedures; the next section describes each of them, as well as the one chosen for this study.

4.6.1 Probability sampling

Probability sampling, often known as representative sampling or random sampling, is the first type of sampling method. Every member of the population has a known (non-zero) probability of being included in the sample in probability sampling; some kind of random selection is used (Alvi, 2016). Using this sampling method requires a researcher to precisely know and define a population whereby probabilities may be allocated to each unit of the population objectively (Barreiro & Albandoz, 2001). When employing this sampling strategy, it is vital to remember that sampling should not be used when the population is too broad. That is, the population must lack distinctive features that distinguish and identify people in some way. A researcher can benefit from the practical advantages of probability sampling, especially when dealing with a known population. Alvi (2016:12) highlights the following as the most common benefits of probability sampling:

- “The methods minimize the chance of sampling biases.
 - A better representative sample is produced using probability sampling techniques.
 - Inferences drawn from the sample are generalizable to the population”
- (Alvi, 2016:12).

Participants in qualitative research are recruited because they possess traits that will provide valuable findings and answer research questions. Because of the random selection of participants when utilising probability sampling, this sampling method was deemed inappropriate for this study.

4.6.2 Non-probability sampling

Non-probability sampling, often known as non-random sampling, is an alternative to probability sampling. In qualitative research, non-probability sampling is used to select the population to study using non-random methods (Phrasisombath, 2009). A researcher can use this sampling strategy to select

a population sample depending on specified features. The sample in non-probability sampling is “gathered in a process that does not give all the individuals in the population equal chances of being selected” (Explorable, 2009). Non-probability sampling gives a researcher complete freedom and control over selecting individuals based on subjectivity and desire. According to Explorable (2009), most researchers are bound by time, money, and labour, and it is nearly impossible to randomly sample the entire population due to these constraints. This was the situation with this study, in which research participants were chosen based on prior knowledge of their significance. The non-probability sampling method was utilised in this study because the researcher wanted to explicitly locate and interview research participants in a social sciences organisation.

Purposive sampling was used in this study as a type of non-probability sampling. Purposive sampling “is a sampling technique in which researcher relies on his or her own judgment when choosing members of the population to participate in the study” (Quinn & Cochran, 2002:10). This study's sample was identified and chosen based on their practical experience in data organisation, data management planning, technical skills, policy compliance, producing data related documentation and data sharing and publishing skills. The criteria that were used for the purposive sampling in this research study have been explored by Marton (2013); Palinkas et al. (2015):

- Relevance to research objectives: Participants are selected based on how well their traits, backgrounds, or viewpoints relate to the goals or study topics.
- Diversity of perspectives: involving people with different viewpoints, experiences, or backgrounds in order to collect a variety of information and ideas.
- Experience: people who have been exposed to particular circumstances, treatments, or settings that are pertinent to the study.
- Theoretical contribution: selecting participants according to how well they may be able to advance or validate theoretical ideas.

Purposive sampling: making sure that the study contains respondents who can provide the most pertinent, varied, and informative data for the research aims by carefully selecting participants based on the aforementioned criteria.

RDM in a social sciences organisation. The Elite Institute (2019) supports this standpoint by explaining that “researchers often believe that they can obtain a representative sample by using a sound judgment, which will result in saving time and money”. Once the target population and sampling methods have been chosen, data can be collected from the specific sample. Collecting data requires using the use of a data collection method. The next section discusses data collection methods and identifies the one used in this study.

4.7 Data collection methods

The data collection procedure is an important step in acquiring data of considerable value and utility and this section explains the data collection method used for this study. Choosing the optimum data collection method enables a researcher to analyse the relationship between research design and data collection methods in order to address the research topic at hand (Terrell, 2012). A researcher must fully recognise that the research questions to be addressed are directly dependent on the research design and the manner in which the data obtained will be examined and analysed (Wisdom & Creswell, 2013).

Prior to planning and data collection, an extensive literature review was done. This provided a better understanding of the literature on RDM, data quality, and quality assurance. Examining literature enables a researcher to benchmark and compare secondary data collection method outcomes (Creswell, 2014). The reviewed literature offered an overview that allowed for the gaining of a comprehensive understanding of the concepts of data quality and quality assurance, as well as of how they are generated and executed. A review of data quality evaluation in South African RDM projects was also conducted to evaluate how South African academic institutions implement RDM initiatives and deal with data quality challenges. The University of Cape Town (UCT), the University of Pretoria (UP), the South African Medical Research Council (MRC), the Council for Scientific and Industrial Research (CSIR), the Human Sciences Research Council (HSRC), and DataFirst were among the academic institutions included in the literature review. The selection of these academic institutions and initiatives was based on their recognition of the need for and importance of RDM. Each institution has undertaken research on how to improve RDM and all of them have established data management programmes to improve RDM. Academic institutions were excluded if no evidence of RDM or quality assurance systems was discovered in the reviewed literature.

In addition to studying the literature, a second primary data-collecting strategy was applied and is explained in the next section. This strategy enabled the researcher to undertake a complete review of the data management procedures of the research participants. An evaluation of data management procedures was conducted to identify where researchers and curators are lacking in their application of certain data quality features. The process of selecting an effective data collection method is crucial and should not be disregarded. Once data has been acquired, it must be adequately and precisely recorded in order for precise data analysis to take place. Terrell (2012) states that, ultimately, the kind of research that needs to be done will decide the kind of data collection techniques a researcher uses. However, it remains true that the same data collection method can be utilised for many sorts of studies. There are various qualitative data collection methods, such as interviews, past records, observations, documents and audio-visual materials (Kaplan, 2004; Crowe et al., 2011; Leedy & Ormrod, 2013; Creswell, 2014). The explanatory case study research design, as stated in a previous section of this chapter, was employed to investigate a phenomenon characterised by a lack of extensive preliminary research – specifically,

formulated hypotheses (Mills et al., 2010). Semi-structured interviews were utilised as a data collection method for this qualitative study. The section that follows discusses the various types of qualitative data collection methods, as well as the one chosen for this study.

4.7.1 Qualitative data collection methods

Interviewing

Interviews are a personal and direct data collection strategy. According to Kvale and Brinkmann (2009), an interview is a conversation aimed at gathering details of the interviewee's [real-world] in terms of understanding the meanings of the identified phenomena. When attempting to uncover new knowledge about a hard topic, interviews are an excellent data collection approach to employ. It is much easier for a person to express their thoughts and feelings in a casual atmosphere than it is in a formal and controlled setting (Creswell, 2014). According to Alshenqeti (2014:40) an interview “is an extendable conversation between partners that aims at having an ‘in-depth discussion about a certain topic or subject, and through which a phenomenon could be interpreted in terms of the meanings interviewees bring to it”.

Researchers' RDM practices are best stated and explained by the researchers themselves. Direct discussion and conversation are the finest ways to communicate what RDM entails and how it is carried out. Alshenqeti (2014:39) is of the opinion that “the value of interviewing is not only because it builds a holistic snapshot, analyses words, reports detailed views of informants; but also because it enables interviewees to speak in their own voice and express their own thoughts and feelings”. Interviewing is thus appropriate for this study since it produces a viable and functional communication environment that will discover critical RDM flaws and areas that require development. A number of authors (Berg, 1998; Kvale and Brinkman (2009); Dörnyei, 2007) advocate for the use of interviewing as a data collection tool for social research because it allows for the discovery of 'direct' explanations for human activities through extensive speech interaction. A number of phrases commonly associated with RDM might be perceived differently by different people; as a result, RDM jargon can act as a barrier to common understanding between study participants and a researcher. The next section looks at the types of interviews that can be used, as well as at the type selected for this study.

Structured interviews

This type of interview allows for a more controlled setting whereby the questions are fixed and are answered in a particular order (Fowler, 1995). When using structured interviews, the interviewer cannot deviate from the original questions and the interviewer is “generally unable to provide any explanation beyond repeating the question” (Harrell & Bradley, 2009:28). When deciding whether to use structured interviews, it is critical to consider the population size. When working with a big population sample, structured interviews are ideal, since they may be generalised to a broad population. The format of

questions used in structured interviews is carefully structured, similar to that of a questionnaire, and they are “planned in advance, sometimes with the help of a pilot study to refine the questions” (Mathers et al. 2002:2).

Unstructured interviews

The structure of this type of interview is simple. Because of this “the interviewer approaches the interview with the aim of discussing a limited number of topics, sometimes as few as one or two, and frames successive questions according to the interviewee's previous response” (Mathers et al., 2002:3). Unstructured interviews rely exclusively on a build-up strategy based on earlier answer feedback. This method, however, has the potential to divert discussions in a variety of areas as “the interviewer does not exert much control over the course of the discussion. He or she might follow explanations with additional questions based on the topics that the respondent brings up” (Harrell & Bradley, 2009). The process of conducting unstructured interviews is relatively lengthy and time-consuming. The data collected through unstructured interviews may be rich and complex, but it may take a long time to collect. This type of interview is suitable when a researcher has ample time to spend with the community being studied (Creswell, 2014). Given the researchers' limited availability and busy schedule, time was a critical aspect to consider in this study. The utilisation of unstructured interviews was thus deemed unsuitable due to time constraints and the limited availability of researchers and data curators.

Semi-structured interviews

This type of interview consists of a series of open-ended questions about the topics the researcher wants to explore. Mathers et al. (2002:2) explain that “the open-ended nature of the question defines the topic under investigation but provides opportunities for both interviewer and interviewee to discuss some topics in more detail”. In semi-structured interviews, both the researcher and the interviewee have the opportunity to elaborate and seek clarification as needed. A researcher can employ a standard set of questions with all research participants in a semi-structured interview. According to Van Teijlingen and Hundley (2001:17), semi-structured interviews use predetermined questions, though their order can be modified depending on the interviewer's perception of what seems most appropriate. In addition, question wording may be altered and explanations provided, while questions that might be inappropriate for a particular interviewee can be omitted and others added.

The format of a semi-structured interview includes an interview guide containing questions and all the topics that must be covered (Creswell, 2014). A feature of semi-structured interviews that offers the researcher complete control over the line that the interview will take is that “the interviewer has some discretion about the order in which questions are asked, but the questions are standardised, and probes may be provided to ensure that the researcher covers the correct material” (Harrell & Bradley, 2009:27). This aspect allows for gathering comments that are to be utilised as the foundation for the proposed

quality framework. Mathers et al. (2002:2) are of the opinion that “if the interviewee has difficulty answering a question or provides only a brief response, the interviewer can use cues or prompts to encourage the interviewee to consider the question further”. Semi-structured interviews were employed in this study based on the characteristics described above. A further reason for this choice was that semi-structured interviews would allow for the collection of detailed information as the researcher sought to delve deeply into the RDM and quality assurance methods of researchers and that they would allow for a full understanding of the responses supplied. The following are further reasons why semi-structured interviews were used:

- Everyone is asked the same main questions, but there is flexibility in how they are asked and what questions to use for follow-up or probing.
- They are especially useful for exploring a person's views of something.
- Well suited to investigating habits, attitudes, beliefs, and motivations, which may be useful in sensitive areas.
- In assessing truthfulness/validity and urgency, non-verbal indicators support making evaluations.
- The interviewer can encourage the process so that it answers any question.
- The method guarantees that the respondent acts on his or her own (i.e. provides personal views and opinions) (Van Teijlingen & Hundley, 2001).

The semi-structured interviews were conducted using an interview schedule comprised of open-ended questions. The interview schedule for researchers and curators was made up of 17 questions over three sections. Before using the interview schedule on the research participants, it was important to test it by performing a pilot study. The pilot study is discussed in the next section.

4.8 Pilot study

The ability to test concepts and research approaches can be an effective way to find flaws and identify opportunities for improvement. The term "pilot study" has been interpreted differently by various authors. A pilot study is traditionally defined as “a small-scale methodological test conducted to prepare for a main study and is intended to ensure that methods or ideas would work in practice” (Kim, 2010:191).

A pilot study was explored for this study since the author needed to determine how well the data collection instrument would cover all of the questions in the interview schedule and whether the interview schedule would provide useful responses that answered the research questions. The pilot study was also utilised to establish how long the interviews would be so that sufficient time could be scheduled. Tashakkori and Teddlie (2003) suggest that a pilot study, regardless of methodology, is useful as preparation for a full-scale investigation. To make the research process as effective as possible,

a customisable roadmap that allows a researcher to improve their research approach is required. An advantage of a pilot study is that interview methods can be strengthened by piloting them. Kvale and Brinkman (2009) emphasises the need to determine whether there are flaws or deficiencies in the interview design that require any aspect of the research to be revised. In addition, the data collection tools utilised were examined to identify drawbacks that the researcher would need to analyse and remark upon. It was also feasible that the structures and questions of the data collection tool could be updated and improved through a pilot study. Harding (2013) highlights that piloting the interview questions and changing the interview guide accordingly before embarking on major research is obviously beneficial. Based on this perspective, the researcher believed that it would be beneficial to pilot the interview schedule. The pilot study was conducted in collaboration with a senior data manager from the social sciences organisation being studied. This was prompted by Kim's (2010:191) explanation that a pilot study is also a feasibility study “that comprises small-scale versions of the planned study, trial runs of planned methods, or miniature versions of the anticipated research in order to answer a methodological question(s) and to guide the development of the research plan”.

The data manager was chosen by the researcher based on her in-depth understanding of data management, data curation, and information systems. The interview schedule was trialled and, in practice, it was revealed that several of the questions were difficult to understand and needed to be rephrased in simpler words. Many of the questions were revised following the pilot study. The pilot study also found that participants would require more time than originally thought to complete the interview schedule. The pilot study greatly aided the researcher in estimating the time it would take respondents to answer all of the questions. The following part examines data collection tools in general as well as the data collection tool utilised for this study.

4.9 Data collection tools

In addition to using a data collection strategy, a data collection tool must be described and chosen. The section that follows describes the various types of data-collecting tools and which tool was used for this study.

4.9.1 Questionnaires

A questionnaire is a research instrument that consists of a set of questions or other types of prompts and that aims to collect information from a respondent (QuestionPro Survey Software, 2021). Data collection (developing the data collection instrument, facilitating its distribution, and collecting the data) has financial repercussions that a researcher must consider. Questionnaires, as opposed to interviews, are likely to be a less expensive approach because they are “simply mailed to the respondents with a minimum of explanation - further, the questionnaire can often be sent to a large number of individuals simultaneously” (Gangrade, 2012: 356). This feature of questionnaires is useful when

gathering information from a large and divergent group. Another advantage of questionnaires is that they use a standardised question sequence, standardised phrasing, and standardised instructions for recording responses (Abawi, 2013). This type of standardisation promotes consistency across multiple measurement scenarios. The identity of study participants should be protected and treated with care (anonymised) at all times (Creswell, 2009) and questionnaires thus give participants direct control over whether or not to reveal their identity. Gangrade (2012:357) is of the opinion that, with questionnaires, “respondents may have greater confidence because of their anonymity, and thus feel more free to express views which they fear might be disapproved of or might land them in trouble”. Participants answer questionnaires on their own, which means they form their own opinions about what the questions signify. When using questionnaires as a data collection tool, the response rate is not always assured; respondents may postpone answering the questions or skip the majority of them. Interviews and direct engagement impose pressure on respondents to react to a question right away, which may cause them to not think deeply about their comments. A questionnaire allows respondents enough time to consider their responses. This is advantageous because “when the subject is given ample time for filling in the questionnaire, he can consider each point carefully” (Gangrade, 2012:357). When creating a questionnaire, a researcher should attempt to utilise basic, straightforward language that is easy to understand by displaying clarity, conciseness, unambiguity, consistency, and directness (Abawi, 2013).

4.9.2 Interview schedule

Data collection is a procedure that should be thoroughly planned before it is carried out. An interview schedule allows a researcher to plan and structure appropriate questions. An interview schedule is a set of questions formulated and presented with the specific purpose of testing an assumption or hypothesis (Farooq, 2013). Based on these reasons, the method of data collection that was used in this research study are qualitative interview. The qualitative interview questions provided in-depth insight and opinions on RDM, quality assurance, and data quality. Individual interviews, as opposed to focus group conversations, are more intimate and allow for more conversation-like interaction. Automated transcription was used, and the transcriptions were subsequently edited because automated transcriptions create language and grammar problems. The section that follows provides an overview of the interview schedule's categorisation.

The interview schedule is divided into sections derived from the quality process framework, which served as the basis for the Quality Process Framework (Appendix 2). Before delving into the sections, the logic of the evaluation framework' is explained thus:

- Each quality dimension was graded on a scale ranging from 'No quality control' [Level 1] to 'Quality compliance' [Level 4].

- The assessment levels were determined by questions and statements that examine RDM, quality assurance, and data quality. Each question/statement is a criterion for measurement derived from the reviewed literature.
- The quality measurement framework is a compilation of measuring criteria drawn from the evaluated literature; hence, it is a synthesis of all the main components of the measuring criteria informing a quality dimension.

The questions in the interview schedules were divided into three sections:

Section A (researchers and data curators): of the interview schedule looked at the demographics of the research participants, such as their educational background and number of years in their professional positions.

Section B (researchers and data curators): of the interview schedule included questions about assessing data quality. The questions were based on how respondents viewed the evaluation framework's data quality dimensions (both scientific and technological).

Section C (researchers and data curators): of the interview schedule asked respondents questions based on the quality process model. The questions were designed to gain an understanding of the applicability of the process model based on respondents' data management experience.

The interview schedules were created in such a way that the deeper significance of respondents' judgments of data quality and quality assurance could be discovered. The following section outlines the study's research location.

4.10 Selection of research location

The study participants were researchers and data curators from a social sciences organisation. Due to the outbreak of the COVID-19 pandemic, face-to-face interviews could not take place at the original scheduled location: the offices of the social sciences organisation in Pretoria, Gauteng province, South Africa. The study interviews were thus conducted via online meetings using a virtual meeting platform (Zoom). This platform was chosen based on the researchers' and data curators' convenience and their access to the platform, also considering their busy schedules. The choice of the Zoom platform thus allowed researchers and curators to be accessed via a virtual platform.

Once the research data had been collected, it had to be analysed and interpreted and the following part details how this was done.

4.11 Data analysis and interpretation

Data analysis is defined as the methods and procedures required to get from raw qualitative data acquired to a point where the data is explained and a knowledge or perception of the individuals and

circumstances researched is gained (Sunday, 2016). The diversity component of qualitative data requires comprehensive and organised analysis in order to yield significant insight. The primary goal of data analysis is to extract meaningful content from the data collected, which allows for:

- Describing the details and summarising them,
- Identifying the variables' relationships,
- Comparing the variables,
- Identifying the distinctions between the variables (Flick, 2013).

This section includes details of how the data was analysed and evaluated, and inferences were drawn from semi-structured interviews. The data in this study were analysed using content analysis. Using a grouping method, the acquired data were organised into 'Themes' and 'Measures' that represent categorised thinking about specific topics. Grouping responses into themes was employed in question number six (Measuring Data quality Dimensions) of the interview schedule. Bowen (2009) states that organising replies into "Themes" and "Measures" facilitates the structured organisation, interpretation, and analysis of qualitative data. The following procedures were used to analyse the data that was gathered for this research study:

Transcription: Automated text transcriptions of audio recordings were used to compile the data.

Familiarisation: Several readings of the data were made in order to gain a general idea and begin to see some basic trends.

Identifying themes: The following steps were undertaken:

Group responses into themes: Group similar responses together to form broader themes. Themes represent overarching topics or patterns that emerge from the data.

Refinement: Review and refine themes to ensure they accurately reflect the data. This may involve merging, splitting, or redefining themes.

Thematic analysis: Analyse the themes to understand the patterns, relationships, and meanings within the data.

Synthesise findings: Combine the insights from the thematic and measure analyses to draw comprehensive conclusions about the data.

Contextualisation: Relate the findings to the broader context of the research question or literature (Caulfield, 2023).

The aforementioned method of data analysis enabled the author to comprehend and evaluate the research participants' responses. In an attempt to raise the calibre of data at the social sciences research organisation, the analysed data offered insightful RDM information (for example data management planning, and data collection processing) that was crucial to understand.

The data analysis methods described in this section aid in the examination of data collected using a qualitative research strategy. The data coding and content analysis methodologies mentioned provide an external validity perspective that considers the specifics of data quality. Content analysis provides a highly practical inspection viewpoint of the social science organisation's specified methods. The following section examines the data management plan employed for this study while the results of data coding and content analysis are presented in Chapter 5.

4.12 Data management plan (DMP)

The research process requires the use of an organised roadmap, which necessitates a researcher considering how they will address essential data management topics. A data management plan is a written document that describes the information you aim to gather or produce during a research project, how you intend to treat, define, assess, and store the data, and what techniques you want to use to share and maintain your data at the end of your study (Stanford Libraries, 2020). Considering the large amount of literature available on RDM, quality assurance, and data quality, as well as on the data collection methods used, a DMP was a significant and necessary planning document for this study. Various institutions have developed DMP templates and guidelines to help their researchers (and other users) plan their studies. These DMPs are designed to provide direct responses to exact funder requirements, with extra prompts that institutions can modify (DCC, 2021). The DCC DMPOnline Tool was used to create the data management plan for this study; a copy of the DMP can be seen in Appendix 18. The DCC DMPOnline Tool employs the DCC Lifecycle Model as an underpinning structure to improve its comprehensiveness; this model is intended to assist researchers in defining their data roles and responsibilities, recognising risks that occur at transition points, and ensuring an effective and stable digital data custody chain (Donnelly et al., 2010). The following sections of the DCC DMPOnline Tool enabled the author to consider essential data management areas that are critical for this research:

What data (type, format) will be generated and how?

- Plans for related metadata and paperwork,
- How data can be accessed and exchanged, justifying any restrictions (e.g. embargoes),
- Intellectual Property Management and Ethics,
- The plan for long-term archiving and data sharing (Donnelly et al. 2010).

The DCC DMPOnline Tool enabled the author to make informed data management decisions and plan ahead. This aided in arranging the data in a more cohesive manner, as well as handling data accessibility and long-term preservation. Refer to Appendix 7 for the DMP used for this research.

4.13 Validity and reliability

The investigation and measurement of an entity (as part of the research process) must be consistent and accurate. Consequently, ensuring validity and reliability in research is critical. Mohajan (2017:1) agrees with this point of view, stating that “reliability and validity are the two most important and fundamental features in the evaluation of any measurement instrument or tool for a good research”. This section outlines how the study's validity and reliability were ensured. A brief overview of ‘rigour’ and ‘trustworthiness’ will also be discussed to contextualise the qualitative nature of this study.

According to Surbhi (2017) reliability is used to mean the degree to which, if the calculation is repeatedly carried out, the measurement method produces reliable results. Validity is required to establish how well an instrument does what it is designed to do, while holistic reliability refers to how well any measuring tool compensates for random error (Surbhi, 2017). An organised strategy is utilised to measure reliability, which includes test-retest, internal consistency methods, and alternate forms. (Bellamy, 2015). The types of questions utilised in the data collecting instrument (Appendix 8 and Appendix 9) had to be validated as adequate and coherent in order to fully comprehend the RDM dynamics in a social sciences organisation, hence the ability to assess reliability was critical in this study.

It makes sense for a data collection device to be both reliable and valid. The degree to which an instrument tests what it seeks to measure is referred to as validity; validity also involves an instrument's reliability (Surbhi, 2017). Validity is also vital for this study since the legitimacy of the various RDM perspectives/feedback of research participants must be assessed, depending on how the participants were chosen. The following table summarises the distinction between validity and reliability:

Table 4.4 - Validity and Reliability comparison (Surbhi, 2017)

Basis for comparison	Validity	Reliability
Meaning	Validity implies the extent to which the research instrument measures what it is intended to measure.	Reliability refers to the degree to which scale produces consistent results when repeated measurements are made.
Instrument	A valid instrument is always reliable.	A reliable instrument need not be a valid instrument.
Related to	Accuracy	Precision

Value	More	Comparatively less.
Assessment	Difficult	Easy

Though validity and reliability are two fundamental elements of every piece of research, the reliability of research findings is also influenced by the ideas of rigour and trustworthiness. Researchers claim that the notions of validity and dependability, which are all essential elements of quality, are equivalent to the rigour of qualitative research (Cypress, 2017:254). The characteristic of being thorough and accurate, or the state of being extremely exact, meticulous, or with exacting precision, is the simplest definition of rigour (Finlay, 2006). Given that qualitative research is a process of explanation and discovery that does not lend itself to rigid boundaries, the phrase "qualitative rigour" itself is oxymoronic (Cypress, 2017:254); and should be carefully considered.

The basic strategy to ensure rigour in qualitative research is systematic and self-conscious research design, data collection, interpretation, and communication (Crawford, 2020); which requires developing a comprehensive research plan outlining data collection and analysis methods. Beyond this, there are two goals that qualitative researchers should seek to achieve: to create an account of method and data which can stand independently so that another trained researcher could analyse the same data in the same way and come to essentially the same conclusions; and to produce a plausible and coherent explanation of the phenomenon under scrutiny (May & Pope, 1995:110)

Trustworthiness is described in different ways by researchers. Trustworthiness refers to quality, authenticity, and truthfulness of findings of qualitative research (Cypress, 2017). The degree of confidence or trust that readers have in the results to be able to apply them in a meaningful and justified way is what is referred to as trustworthiness. Yin¹⁷ (1994) describes trustworthiness as a criterion to judge the quality of a research design. Trustworthiness addresses methods that can ensure one has carried out the research process correctly and ensuring trustworthiness is crucial in establishing the credibility and reliability of qualitative findings, given their subjective nature (Cypress, 2017:254). The concept of trustworthiness in qualitative research comprises various essential elements, such as credibility, transferability, dependability, and confirmability (Ahmed, 2024:1). According to Ahmed (2024:2) "establishing reliability in qualitative research is crucial for influencing future research paths. Researchers can utilize reliable discoveries as a basis for information, which can then inform the development of hypotheses and the planning of future investigations". Reliable data is also the basis in seeking feedback from fellow scholars to identify potential biases and improve data analysis. According to Cypress (2017:256), "reliability and validity should be taken into consideration by qualitative inquirers while designing a study, analyzing results, and judging the quality of the study".

In this study, reliability is employed to ensure that no question (from the interview schedule) is answered twice by the same research participant and that each person has answered most of the questions on the interview schedule. This allowed for a complete portrayal of the findings. Validity is used in this study to ensure that relevant questions (in the interview schedule) are used to answer the research questions and contextualise the literature review. Reliability and validity were determined during the pilot phase before the interviews with the research participants were conducted.

4.14 Ethics approval by the university

The University of Pretoria's ethics approval procedure includes ensuring that all research conducted under its affiliation conforms to ethical standards and guidelines (University of Pretoria, 2023). Due to the university's conformity to research practices, ethical approval for this study was required. Additionally, agreement from research participants was required, as was transparency with the social sciences organisation under investigation. The author was required to submit an ethics application form to the EBIT (Faculty of Engineering, Built Environment and Information Technology) Ethics Committee in order to receive ethical approval. The ethics application form consists of the following sections:

1. RESEARCHER DETAILS: *(Please include your Supervisor details in this section if you are a student)*

2. RESEARCH PROJECT TITLE *(use a descriptive title)*

3. RESEARCH PROJECT DETAILS:

3.1 Provide a complete but concise description

4. RESEARCH PARTICIPANTS

Does the project involve people as participants, either individually or in groups?

If Yes, please answer questions 4.1 to 4.7. If No, continue to section 5.

5. ENVIRONMENTAL IMPACT and HAZARDOUS MATERIALS

5.1 Does the project have a potentially detrimental environmental impact, or are hazardous materials used in the project?

6. DISSEMINATION OF DATA

6.1 How and where will your results be published and/or applied?

7. DECLARATION *(Tick the relevant boxes)*

I accept and will adhere to all stipulations pertaining to ethically sound research as locally, nationally and internationally established.

8. SUBMISSION CHECKLIST

Each item to be submitted should be submitted as a separate pdf file, using the naming convention given earlier in this document or below.

8.1 Have you submitted confirmation that the research proposal has been approved?

Please submit as a pdf file with this filename format: Confirmation.pdf

(University of Pretoria, 2023)

Following the completion of the application for ethics and submission to the Committee for Research Ethics and Integrity (EBIT), a double-blind examination of the research study application is conducted before a decision is made. An applicant could get one of four answers:

- *Approved: there were no significant ethics-related issues in your application.*
- *Approved with conditions: Minor mistakes or ethics-related challenges were found.*
- *Revise Resubmit: significant ethics-related challenges were found in your application. Revision of specified sections must be addressed and resubmitted for an additional round of reviews*
- *Rejected: The ethics challenges present in the application go against University policies, laws and regulations, and you may not continue with the suggested research” (University of Pretoria, 2023).*

The ethics application of this research study was Approved with conditions; the condition that had to be met was “When permission letter(s) is obtained these need to be uploaded again under Post Approval submissions”. This condition was met and ethical clearance was approved (Reference number **EBIT/230/2021**). The study was also approved by the organisation’s CEO (Gatekeeper approval for Protocol No 3-09-2021).

Two data collection instruments were developed – one for researchers and the other for data curators – both of them semi-structured interview schedules. Owing to the Covid-19 issue. Microsoft Teams was used for all virtual interviews. The respondents' identities were anonymised and the interviews were transcribed verbatim. To give participants background information before the interview, a participant handout (Appendix 10) was developed and shared with respondents before the interview.

4.15 In summary

This chapter explained the process of selecting an interpretivist paradigm to be used in conjunction with a qualitative approach. The research design was an explanatory case study, and the data-gathering methods were semi-structured interviews. This chapter further described why it was required to test the data-gathering instrument using a pilot study to determine the tool's usefulness. Purposive sampling was used for the research and Pretoria was chosen as the research location. The chapter concluded by detailing the data analysis and interpretation procedures utilised (categorising the data and conducting content analysis). The results of the analysis procedure are examined in greater detail in the following chapter.

CHAPTER 5: RESEARCH FINDINGS AND ANALYSIS

5.1 Introduction

The findings of this research are presented in this section, along with an analysis of the data and the interpretation of the results. Together, these serve as the foundation for the study and aid in comprehending the implications and results of the stated research questions.

The purpose of this chapter is to document the findings drawn from the semi-structured interviews. This chapter also seeks to interpret the data collected concerning the study questions posed in Chapter 1. The chapter gives a thorough account of how the research participants responded to questions about quality and quality assurance across a variety of quality dimensions. To illustrate the idea of quality and how the participants address it, the chapter examines and evaluates the consequences of these findings. The personal preferences and experiences of the research participants are covered in the final section, along with the handling of data quality and quality assurance in the social sciences.

5.2 Background of the case study and data collection

The research study employed purposive sampling, with participants selected based on their actual RDM experience in a social sciences organisation (as discussed in Chapter 4, Section 4.6.2). The research participants, 24 researchers and four data curators. The selection of data curators was based on their proficiency in upholding guidelines for data quality and integrity, which guarantees the accuracy and dependability of the data utilized in research. Data curators were chosen in part because they should be adept at solving problems in order to handle data-related difficulties. These difficulties can include missing data, discrepancies, and integration problems. In order to maximize the long-term impact of research data, data curators were included for another major reason: their engagement can help ensure that data remains relevant and usable for future study. Because of their subject expertise, researchers were included in the study; they have in-depth knowledge and experience in their particular disciplines and may offer insightful opinions and significant insights. The selection of researchers was also based on their background in research and their capacity to adjust to various study designs and methodologies, which makes them excellent subjects for a range of investigations. In order to gauge their level of expertise in handling data, from gathering and organizing it to analyzing and interpreting it, researchers were invited to participate.

5.3 Structure of the data analysis

The findings are documented in accordance with the interview schedule structure. The discussion of the findings is linked to the research sub-questions.

Section A, (refer to Section 5.4) reports on demographic and educational data. The intention is to determine whether a participant's educational level and number of years as a researcher/data curator influence data management, quality assurance, and perspectives about data quality.

Section B, (refer to Section 5.5) which focuses on data quality, links interviews to a quality process framework. There is a unique set of questions for each quality dimension. In the six sub-sections that make up Section B, the participants were asked to (i) name the quality dimensions that they think should be included in the quality process model, (ii) name their own dimensions that reflect the quality of scientific data, and (iii) add and discuss non-framework dimensions that describe the technical quality of data. The final question (iv) uses sixteen dimensions to evaluate different aspects of data quality.

Section C, (refer to Section 5.6) aimed to gather opinions on the preliminary quality process model, eliciting perspectives, criticisms, and suggestions for improvement. The researchers and curators reflected on the model from their own perspectives. The section also addresses data management issues and the potential contribution of RDM to the research project.

5.4 Findings: Section A - Demographics of the research participants

This section presents interview schedule results, analysed thematically. The research findings collected from all the research participants are discussed in the next section.

The sections below outline the educational backgrounds and duration of professional experience of the research participants. Four had Master's degrees while 20 of the 24 researchers held PhDs. Three of the four data curators had Master's degrees, while the fourth had a Bachelor's degree. Information about the research participants' levels of formal education is provided in the sections that follow.

5.4.1 Educational background of the research participants

The table below represents the education qualifications of the research participants (researchers and data curators). For the sake of anonymity, the specific code representing the participant is not indicated.

Table 5.1 - Academic qualifications of participants

	PhD	Master's	Hons	Bachelor's	Total
Researchers	20	4	0	0	24
Curators	0	3		1	4

The participants were mostly (83%) researchers with PhD degrees, while 75% of the curators participating held Master's degrees. One of the curators (with a bachelor's degree) was deemed an expert participant and was included in the participant group due to her extensive experience as a data curator. It was expected that the participants' answers to questions about RDM would be influenced by the specifics of their educational backgrounds and the length of time they had been charged with their various professional responsibilities. The research participants' educational backgrounds suggest that they were well-educated.

5.4.2 Expertise as a researcher and data curator across time

The table below indicates that the majority 21 (87.5%) of the researchers who took part in this survey had more than six years of experience as researchers, while three (12%) had 0–5 years of experience. Three of the curators (75%) had 0–5 years of experience and one curator (25%) had 6–15 years of experience.

Table 5.2 - Years of experience

	0–5 years	6–15 years	More than 15 years	Total
Researchers	3 (R1), (R6), (R19)	8 (R3), (R8), (R10), (R12), (R14), (R20), (R22), (R23)	13 (R2), (R4), (R5), (R7), (R9), (R11), (R13), (R15), R16), (R17), (R18), (R21), (R24)	24
Curators	3 (C1), (C2), (C4)	1 (C3)		4

Researchers require a comprehensive understanding of the research process and a solid understanding of data management. As indicated, three of the four data curators had less than six years of experience, demonstrating that the three of the four curators have limited (less than five years) professional experience.

Discussion: Section A indicated the make-up of the participant group. These participants were purposively selected (as explained in Chapter 4, Section 4.6.2) as being representative of the organisation's researchers and data curators in the data curation unit. The researchers participating had more experience in research than the curators had in terms of data curation. It was not clear from the literature whether this was also the case in other organisations.

The next section discusses the findings from Section B of the interview schedule, i.e. measuring data quality.

5.5 Findings: Section B - Measuring data quality

The goal of Section B of the interview schedule was to establish how participants felt about the technical and scientific quality of research data in general. The replies from the respondents painted a picture of how they individually manage data and how institutional policies and standards provide specific guidance for this process. This section also discovered how the quality dimensions are perceived by respondents in terms of yielding data that is of good quality.

5.5.1 Findings: Quality dimensions to add (Question 3) – Researchers

Each researcher seemed to favour certain qualities that they felt were essential for effective data management. This section reports the quality attributes that were relevant to the researchers but that were not included in the list that was used in this study.

Q3. What quality dimensions should have been listed but were not?

Of the 24 researchers interviewed, 14 (58%) were able to provide additional data quality dimensions. The dimensions that the researchers identified vary and are dependent on how they manage data in their respective research units. The remaining ten researchers (R1, R5, R8, R10, R12, R17, R18, R20, R22, R23) indicated that the set was complete. An example of what one researcher said appears below:

“I can't think of one dimension that is not here.” (R17)

Listed next is a selection of quotes (text in bold is the author's interpretation) from research participants about quality dimensions that they felt should have been listed:

*“**Rigour** – has it been collected with rigour? One way of looking at it is to involve some **triangulation** where you use different methods, for example, mixed methods, or you use observations and interviews to triangulate your data.” (R2)*

*“Trustworthiness. [This] basically involves **triangulation** from **different sources of data**.”*
(R14)

*“I think that you need a **revision policy** there. It will depend on what data you are working with. We work with data that involves GDP (Gross Domestic Product), [which gets rebased more or less] every year. So, there's a **revision policy** put in place there. The policy is also there to identify errors that were maybe curated or preserved...”* (R4)

*“**Internal consistency**, which is a different dimension to consistency, **repeatability**, being able to redo/repeat and **cultural context** – if data is culturally irrelevant then it's of no use.”* (R5)

*“The **correctness of data** [that is often categorised as **validity**] is a quality dimension that should have been listed.”* (R11)

*“So, **analysis** and **sample size**. Those are the two that I would like to have in my own dimensions of quality table.”* (R13)

*“...**credibility**, **transferability**, **dependability** and **confirmability**. In some cases, these are given different terms in the handbook.”* (R16)

*“I guess **validity** can be defined in different ways. Maybe in my type of research, we define validity in terms of, you know, the data actually representing the truth. The **truth** of the subject matter that you are dealing with. There is then also **external validity** that is representative of the validity that you are talking about.”* (R18)

*“**Reliability** is not here and [neither is] **veracity**. I think those are the two that I usually rely on, besides some of the ones that you've listed here.”* (R21)

*“... the issue of **how easy it is to share** it. How easy is it to find the data that has become available to the end user.”* (R15)

The comments in the next group speak more to methodology than to the quality of the data and these comments were not considered to be new or additional quality dimensions (refer to Table 5-3).

*“So, **analysis** and **sample size**. Those are the two that I would like to have in my own dimensions of quality table.”* (R13)

*“I want to know [...] **if your methodology** is actually sound. Because, when it comes to quality control or the quality of the data, it starts with the writing of the proposal. If the methodology is not sound, I would likely reject the entire paper.”* (R24)

*“I suppose two from my perspective. The one quality dimension relates to **ethics or human rights**. The dimension speaks to anonymisation or privacy. As captured by POPIA [or in its sub-protocols] so that's one. The other is potentially one related to automation or **interoperability** – so, whether the data or the metadata is sufficiently described to support some kind of automated harvesting and reuse. So, **machine readability** is part of the core of it.” (R19)*

The quality dimensions identified by the curators are discussed in the next section.

5.5.2 Findings: Quality dimensions to add (Question 3) – Curators

Q3. What quality dimensions should have been listed but were not?

Two curators (C1, C3) stated that all quality dimensions had been listed. They explained:

“Well, I think you’ve already listed them all. I cannot think of anything you have not already listed.” (C1)

“I would say that this is a comprehensive list of the dimensions for the quality of the data.” (C3)

A suggestion for further additions was:

*“I think, the **assignment of a DOI**. So, when a data set receives the DOI, that can be an indication that some sort of quality assurance process has actually been followed. **Format compliance** could possibly also fit in with standards or with consistency.” (C2)*

5.5.3 Analysis and discussion of the findings

Of the 24 researchers interviewed for this research study, 14 (58%) (R2, R3, R4, R5, R7, R11, R12, R13, R14, R15, R16, R19, R21, R24) provided additional data quality dimensions. Twelve (85%) of the 14 respondents had more than 15 years of research experience. It appears that experienced researchers can evaluate their work and identify suitable quality dimensions with more confidence than less experienced researchers. Nevertheless, their familiarity with managing data and comprehension of it also leads to the usage of synonyms for formal quality dimension words. The table below, Table 5-3, shows that they seem to be unaware that their phrases are synonyms, which leads them to use synonyms for formal quality dimension terms in addition to those terms, giving rise to redundancies.

The curators were less vocal (suggested fewer items) than the researchers when suggesting additional dimensions. Among the items suggested by the curators were format compliance and DOIs but, since these do not guarantee quality, they were integrated with existing quality dimensions.

The table below presents a summary of the extra dimensions that the research participants (researchers and curators) recommended, along with the potential integration of each of these new dimensions with a dimension used in this study.

Table 5.3 - Alignment between suggested and available quality dimensions

Additional dimensions as suggested by Participants	Respondent(s) suggesting the addition	Dimension(s) with which the suggested dimension can be integrated
Ease of sharing data	R15	Accessibility
Assignment of a DOI	C2	Accessibility
Internal consistency	R5	Accuracy
Veracity	R21	Accuracy
Rigour	R2	Accuracy
Trustworthiness	R2, R14	Accuracy
Confirmability	R16	Completeness
Analysis	R14	Completeness
Dependability	R16	Consistency
Interoperability	R19	Flexibility of the data
Transferability	R16	Flexibility of the data
Reproducibility	R5, R9	Flexibility of the data
Reliability	R2, R7, R21	Integrity
Credibility	R16	Integrity
Ethics	R19	Methodological Soundness
Human rights	R19	Methodological Soundness
Triangulation	R2, R14	Methodological Soundness
Revision policy	R4	Standards and Policies
Machine readability	R19	Usability of the data
Format compliance	C2	Usability of the data
Validation	R11, R18	Usability of the data

The responses from the research participants suggest that the researchers are acquainted with terms relevant to RDM quality, which, if used consistently, may close the gap in understanding data quality. The dimensions used in this study were closely matched by the dimensions identified by the research participants (researchers and curators), with the names of the dimensions differing only slightly. Nevertheless, the descriptions of the original dimensions must be improved by incorporating the descriptions so that it is clear that each of the aspects is reflected in the identified dimensions. The

definitions of the quality dimensions (see Appendix 14) were the original definitions; after data collection and receipt of new dimensions (such as those in Table 5.3 above), the definitions were amended by incorporating the extra meanings (dimensions) that the research participants specified. The updated definitions of the dimensions appear in Appendix 16. The dimensions that research participants believed could be used to measure the scientific quality of the data are discussed in the next section.

5.5.4 Interview schedule Question 4: Scientific quality – Research participants

It is necessary to share here the level of uncertainty that the researchers had regarding the quality dimensions presented to them in their interviews, an impression confirmed when they were asked to specify which of these dimensions (and any other dimensions) they believed could be used to measure the scientific quality of the data.

Q4. Which of the dimensions mentioned (your own included) would determine the scientific quality of the data?

The dimensions that respondents (both researchers and curators) believed would determine the **scientific quality** of data are listed in the following table:

Table 5.4 - Quality dimensions that can determine scientific quality

Dimension	Respondents	Total researchers (n=24)	Total Curators (n=4)
Accessibility	R1, R2, R3, R4, R5, R6, R11, R12, R13, R14, R15, R17, R19, R21, R22 C1, C2, C3	15	3
Accuracy	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R20, R21, R22, R24 C1, C3, C4	23	3
Comparability	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R21, R22, R23, R24 C1, C3, C2	21	3
Completeness	R1, R2, R3, R4, R6, R7, R8, R11, R13, R14, R15, R16, R17, R18, R21, R22, R23, R24 C1, C2, C3, C4	18	4

Dimension	Respondents	Total researchers (n=24)	Total Curators (n=4)
Consistency	R2, R3, R4, R5, R6, R7, R8, R9, R11, R12, R13, R14, R15, R16, R17, R18, R19, R22, R23, R24 C1, C2, C3, C4	20	4
Flexibility of the data	R2, R3, R4, R6, R7, R14, R15, R17, R18, R19 C1, C2, C3, C4	10	4
Integrity	R2, R3, R4, R5, R6, R7, R8, R9, R11, R12, R13, R14, R15, R16, R17, R18, R19, R21, R22, R23, R24 C1, C2, C3, C4	21	4
Interpretability	R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R13, R14, R15, R17, R19, R21, R22, R23, R24 C1, C2, C3, C4	19	4
Methodological Soundness	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R22, R23, R24 C1, C2, C3, C4	22	4
Quality Check Processes	R1, R2, R3, R4, R5, R6, R7, R8, R10, R11, R12, R13, R14, R15, R16, R17, R18, R21, R22, R23, R24 C1, C2, C3, C4	21	4
Relevance	R1, R2, R4, R5, R6, R7, R10, R13, R15, R17, R18, R19, R21 C1, C2	13	2
Standards and Policies	R2, R3, R4, R5, R6, R7, R11, R12, R13, R14, R15, R17, R18, R19, R21, R22, R23, R24 C1, C2, C3	18	3
Timeliness	R1, R2, R4, R6, R9, R11, R12, R13, R14, R15, R19, R21, R22, R24 C2, C4, C3	14	3

Dimension	Respondents	Total researchers (n=24)	Total Curators (n=4)
Uniqueness	R2, R3, R4, R5, R6, R7, R8, R13, R14, R17, R18, R19, R22, R23 C1, C2, C3, C4	14	4
Usability of the data	R2, R3, R4, R5, R6, R7, R8, R9, R12, R13, R14, R15, R16, R17, R18, R21, R22, R23, R24 C1, C2, C3, C4	19	4
Validity	R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R21, R22, R23, R24 C1, C2, C4	22	3

The researchers were not unanimous in selecting all the quality dimensions that should be used. It is, however, the ten scientific dimensions that they put forward most frequently were: Accuracy, Comparability, Consistency, Interpretability, Integrity, Methodological soundness, Quality Check Processes, Standards and Policies, and Usability of the data and Validity.

The ten most essential quality dimensions that the curators identified were: Comparability, Completeness, Consistency, Flexibility of the data, Integrity, Interpretability, Methodological Soundness, Quality check processes, Uniqueness, and Usability of the data. The only distinction was that the curators recognised "Uniqueness" – possibly because they understood that each data file is distinct and varied (i.e., has its own traits and proportions). The thoughts of respondents regarding measuring technical quality are discussed in the section that follows.

5.5.5 Interview schedule Question 5: Technical quality – Research participants

The technical quality of the data is just as important as its scientific quality. In order to discover the dimensions that, in their opinion, would define the technical quality of data, the research participants were asked to list the dimensions.

Q5. Which of the dimensions mentioned (your own included) would determine the technical quality of the data?

The dimensions that the research participants (both researchers and curators) said will affect the technical quality of data are included in the table below, along with any extra dimensions not included in the interview schedule.

Table 5.5 - Quality dimensions that can determine technical quality

Dimension	Number of respondents	Researchers	Curators
Accessibility	R4, R6, R7, R9, R11, R13, R16, R17, R18, R19, R21, R22 C2	12	1
Accuracy	R1, R2, R3, R4, R5, R6, R7, R8, R10, R11, R13, R14, R16, R17 R18, R20, R22, R23, R24 C1, C2, C3, C4	19	4
Comparability	R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R13, R14, R16, R17, R18, R20 C1, C2, C3, C4	16	4
Completeness	R1, R2, R3, R4, R6, R7, R8, R9, R11, R13, R16, R17 R18, R22, R23, R24 C2, C3	16	2
Consistency	R1, R2, R3, R4, R5, R6, R7, R8, R9, R11, R12, R13, R14, R16, R17, R18, R19, R22, R23, R24 C1, C2, C3, C4	20	4
Flexibility of the data	R2, R3, R4, R7, R10, R12, R16, R17, R18 C2, C3	9	2
Integrity	R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R13, R14, R16, R17, R18, R23, R24 C1, C2, C4	17	3

Dimension	Number of respondents	Researchers	Curators
Interpretability	R1, R2, R3, R4, R7, R8, R9, R10, R11, R12, R13, R14, R16, R17, R18, R19, R21, R23 C1, C2, C3	18	3
Methodological Soundness	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R16, R17, R18, R19, R22, R23, R24 C1, C2, C3, C4	21	4
Quality Check Processes	R2, R3, R4, R5, R7, R8, R9, R10, R11, R12, R13, R14, R16, R17, R18, R19, R22, R23 C1, C2, C3	18	3
Relevance	R2, R4, R5, R6, R7, R8, R10, R13, R16, R17, R18, R19, R22 C1, C2, C3	13	3
Standards and Policies	R1, R2, R3, R4, R5, R6, R7, R9, R10, R12, R13, R14, R16, R17, R18, R19, R22, R23 C1, C2, C4, C3	18	4
Timeliness	R1, R2, R4, R5, R6, R7, R8, R11, R12, R13, R14, R16, R19 C1, C2, C4	13	3
Uniqueness	R2, R3, R4, R5, R6, R7, R8, R12, R13, R14, R17, R18, R19, R22, R23 C3	15	1
Usability of the data	R1, R2, R3, R4, R5, R6, R7, R8, R9, R11, R12, R13, R16, R17, R18, R19, R23	17	3

Dimension	Number of respondents	Researchers	Curators
	C1, C2, C3		
Validity	R1, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R17, R18, R22, R23, R24 C1, C2, C3, C4	18	4

The researchers consider a variety of dimensions when evaluating the technical quality of a study. They identified the following ten dimensions as the most important to evaluate technical quality: Accuracy, Comparability, Completeness, Consistency, Integrity, Interpretability, Methodological Soundness, Quality Check Processes, Standards and Policies, and Validity.

The data curators identified the following eleven dimensions as the most important: Accuracy, Comparability, Consistency, Completeness, Integrity, Interpretability, Methodological Soundness, Quality Check Processes, Standards and Policies, Timeliness, and Validity. The curators identified Timeliness while the researchers did not, which is the only difference. This can be explained by the fact that the curators – given the nature of their task – are aware of how crucial it is to make data accessible as quickly as possible so that it can address relevant and current issues.

After the quality dimensions were established, attention turned to how the dimensions could be broken down individually to produce metrics or levels of compliance specific to each of them. The next section discusses the participants' reactions to the measurement of data quality dimension compliance.

5.6 Findings: Section C - Measuring the Data Quality Dimension compliance - Research participants

According to the researchers, numerous criteria can be used to evaluate different aspects of data quality compliance. To determine how comfortable the researchers were in utilising and comprehending the dimensions employed in this study, it was crucial to have discussions about how they would measure these dimensions.

The section that follows provides a summary of how the research participants felt about how the data quality dimensions can be measured and identifies key areas that contribute to the development of the quality process model. Items that were mentioned but that are not related to the measurement have been noted.

Q6. Considering the data quality dimension(s) you identified above, please indicate how we should measure [compliance to] the dimension(s).

5.6.1 Accessibility

When asked how to measure accessibility, 24 (85%) of the respondents interviewed provided methods for measuring accessibility. The relevant measures that emerged from the responses given by all the research participants include:

- **Measure 1:** The *ease with which data can be made accessible* (R1, R3, R4, R5, R7, R12, R16, C3), e.g.: “Accessibility, for me [would mean] how easy or difficult it is to access the data.” (R3)
- **Measure 2:** The duration between the **request submission** and **data access** (R1, R15, R18), e.g.: “I need to know how quickly or how soon can data be accessed by internal and external users.” (R15)
- **Measure 3:** Data access **platforms** (R9, R14), e.g.: “What platforms are you using to share the data?” (R14)
- **Measure 4:** Complying with **funder requirements** (R17), e.g.: “The guidelines on how data can be accessed come from the funder, because the funder, most of the time, dictates data access.” (R17)
- **Measure 6:** Availability of data **in various formats** (C2), e.g.: “When you make data accessible, you must make it accessible in various formats.” (C2)
- **Measure 9:** *Number of downloads* (R22), e.g.: “Looking at the number of downloads to see that, in fact, people are accessing the data.” (R22)
- **Measure 10:** *Citation index* (R13), e.g.: “You measure accessibility through the citations, [using] the citation record. I mean, if data is accessible, it is obvious that people [researchers] will make use of the data.” (R13)

The researchers have access to a wide range of platforms for disseminating their findings, as seen by the preference of two Master's degree holders (R9 and R14) for disseminating their findings across many platforms.

Although they were able to convey their ideas, the researchers did not do so in terms of the definitions employed in this study. Researchers R2, R11, and R21 appeared not to comprehend the term ‘Accessibility’ and thus provided measures that were invalid for this dimension. Measuring items that were mentioned but not relevant include:

- Factors that impact **methodological conformance** (R19, R21, R23).
- Determining the **integrity** of the data source (R2).
- The **usefulness** of data to the user community (R21).

5.6.2 Accuracy

Of the 24 researchers interviewed, 23 (95%) explained how accuracy can be measured. All the curators (100%) provided suggestions as to how data accuracy can be measured.

Measures that emerged:

- **Measure 1: Verification of methodology** elements (R1, R7, R18), e.g.:
“You have to check what tools were used, because in research, sometimes if you are using certain tools they are not 100% accurate themselves.” (R7)
- **Measure 2: The inclusion of external expertise** to assist with **RDM** (R2, R22), e.g.:
“Appointing maybe two or three data managers or statisticians who will go through that data.” (R22)
- **Measure 3: Ensuring the relevance and accuracy** of the collected data’ (R3, R8, R15, R16, R17, R20), e.g.:
“The accuracy of the documents would be looking at how it reflects what is actually happening and what is actually being investigated.” (R16)
- **Measure 4: Documenting research processes** to facilitate secondary usage (R4, R5, R10, R14, C1, C3), e.g.:
“During data collection in the field something sometimes happens that goes against the protocol being used. I document the process because then it means that the protocol was not complete enough.” (R10)
- **Measure 5: Strategies for confirming the accuracy** of data documented (R6, R9, R12, R14, R21, C2, C4), e.g.:
“I think accuracy is when we can use specific statistical software like SPSS, we can run frequencies to identify any missing information or values or outliers or empty fields or missing cases.” (C2)
- **Measure 6: Data collection authentication methods** (R11, R13, R24), e.g.:
“Here I would do scientific tests [...] I run my normal scientific tests [...] and that will actually tell me the accuracy of the data scientific tests.” (R24)

All but three of the researchers (R8, R14, R15) held PhD degrees and reported using various methods to ensure data accuracy. One experienced researcher (R24) suggested measuring data accuracy through scientific tests, emphasising the importance of understanding different types of tests for accurate research data measurement. The concept – that of using tests to verify accuracy – is also reported in Chapter 2 (Section 2.8.3).

The response of one curator (C2), the holder of a Master's degree, demonstrates that the person concerned had statistical knowledge and the ability to analyse technical data. This exemplifies how useful analytical and statistical skills are for curators.

5.6.3 Comparability

Of the 24 researchers interviewed, 20 (83%) provided responses on measuring comparability, while three of the curators (C1, C2, C3) (75%) explained how Comparability can be measured:

Measures that emerged:

- **Measure 1:** The ability to **assess and analyse** data in order to **make comparisons and identify patterns** (R2, R3, R4, R6, R7, R13, R14, R16, R18, R20, R21, R24, C3), e.g.:
“You need more than one survey, you need more than one dataset, whether data are collected every year or quarterly, even biannually. I think you can probably use the data depending on which methodology you use to collect that data. You can compare that with other countries.” (R4)
- **Measure 2: Quality control measures and ensuring consistency** in processes (R1, R8, R9, R10, R11, R12, R14, R17, R23), e.g.:
Cross-checking, and I will ensure that there is standardised uniformity that serves as proof of how things are carried out.” (R8)
- **Measure 3:** Data interoperability (C1, C2), e.g.:
“Ensure that data files are available in various formats when we do the curation. So, we've got the ASCII files, we've got an ASCII CSV file, the SAS program file, the do file for Stata, a dictionary file and an SPSS syntax file.” (C2)

A PhD-educated researcher (R4) discussed comparing data sets with similar data from other countries, demonstrating that he was of the opinion that a level of comparison/benchmarking is important.

Two curators (C1, C2) with Master's degrees listed data format as a metric, implying exposure to different formats being used by researchers when submitting data for curation.

5.6.4 Completeness

Of the 24 researchers interviewed, 22 (91%) selected measuring data completeness, while three of the curators interviewed (C2, C3, C4) (75%) explained how Completeness can be measured.

Measures that emerged:

- **Measure 1:** Following predefined guidelines set by industry experts to ensure completeness (R1, R6, R7, R8) e.g.:

“The word benchmark comes to mind [and] conforming to those standards should be how this dimension is measured.” (R1)

- **Measure 2:** Eliminating inaccurate information to make data understandable (R2, R3, R11, R15, C3), e.g.:

“Having fewer errors. [I]f the person has responded to something they shouldn’t have responded to, then you need to remove that because it’s not relevant to your data set.” (R2)

- **Measure 3:** Documented **quality assurance management** (R4, R5, R9, R14, R18, R21, C2, C4), e.g.:

“For completeness, it is important to check whether your questions from the questionnaire do tie up with what you have in the data set and that all the answers are included in the data file or in the transcriptions [...] I think we should remove any notes or markings from the researchers before we disseminate the files [...] but there might be cases where these markings are very important for secondary users and then we might want to include those notes either into the code book or the user guide.” (C2)

- **Measure 4: Managing** data according to the **research design** (R16, R17, R20), e.g.:

“The completeness would be measured by what we set to look for at the start of the project. When we set out to look for 20% of a type of population, we must reach that 20% for qualitative saturation.” (R17)

- **Measure 5:** Data cleaning and processing strategies (R22, R23, R24), e.g.:

“Accounting for the degree of missing values, accounting for non-responses, or [refusal] to answer, [as well as for] invalid responses.” (R23)

One researcher (R1), with 0-5 years of experience, referred to conforming to benchmarked standards and meeting research requirements. The researcher was familiar with using and conforming to RDM standards, regardless of their experience. This viewpoint is similar to what was reported in Chapter 2, Section 2.13 on benchmarking crucial RDM activities.

One curator (C2) with 0-5 years of experience provided practical data-cleaning steps. Some steps (e.g. removing notes or markings in a data file) should be the responsibility of data managers or the Principal Investigator. This responsibility, which should be allocated to data managers, is similar to USAID’s Conducting Data Quality Assessment Framework (Appendix 15, part D).

5.6.5 Consistency

Of the 24 researchers interviewed, 22 (91%) discussed measuring consistency in data, while of the four curators interviewed, 3 (75%) (C1, C2, C3) provided different responses on measuring consistency.

Measures that emerged:

- **Measure 1:** Data **validation** using **logical and quality checks** (R2, R3, R6, R8, R11, R12, R17, R23, C1, C2), e.g.:
“I think variables and terminology need to be used clearly. So, I think you would probably have to go back to your questions in the questionnaire and compare [the data gathered with those].” (R2)
- **Measure 2: Benchmarking** research processes for **comparative analysis** (R2, R4, R5, R7, R9, R19, R23), e.g.:
“The extent to which there are common data definitions across multiple data sets, as well as standards for data definitions.” (R19)
- **Measure 3:** Checking **consistency** in research **methods** (R14, R15, R16, R18, R22), e.g.:
“One aspect of consistency is being able to apply the same research methods to achieve the same or similar sets of results.” (R16)
- **Measure 4:** Data accuracy quality control methods (R22, R24, C3), e.g.:
“The first thing in curation is the frequency [and] that reveals a lot of answers about the accuracy of the data.” (C3)

An experienced researcher holding a PhD (R2) emphasised the importance of distinguishable variables and clear terminology. This person also discussed the application of international standards to micro-level data. An example of quality requirement standards, as discussed by the International Standards Organization (ISO), was mentioned in Chapter 2, Section 2.7.1. The researchers of the social sciences institution know the importance of variable constructions and of referring to international standards as a reference guide for data management.

Two curators (C2, C3) with different levels of education placed a strong focus on helping researchers, suggesting that a curator’s educational background does not influence how they view the importance of providing research support.

5.6.6 Flexibility of the data

Of the 24 researchers interviewed 14 (R2, R3, R4, R5, R7, R10, R12, R13, R14, R15, R16, R17, R18, R19) (58%) responded on how the Flexibility of data can be measured. Two curators (C2, C3) (50%) explained how the Flexibility of data can be measured.

Measures that emerged:

- **Measure 1:** The **application of data** in various **environments** (R2, R3, R12, R13, R18), e.g.:
“Look at [whether] the data set [can] be used in different ways.” (R3)
- **Measure 2:** Ensuring that proper procedures and guidelines are followed when handling data (R4, R10, R19), e.g.:

“Look at what guidance is provided on how the research data can be used or reused.” (R19)

- **Measure 3:** Data transformation and management strategies (R7, R14, R16, C2), e.g.:
“We run frequencies to see if there’s any data missing or strange variables in the data set that might not tie in with the questionnaire.” (C2)
- **Measure 4:** Understanding fundamental methodological approaches guiding the research process (R6, R15), e.g.:
“Research data management should at least be discussed at the [time of] data collection or prior to data collection. [It is preferable to discuss this matter at the] proposal stage so that we know exactly what the philosophical assumptions of the researchers are, as well as their approach to the particular research problem.” (R15)
- **Measure 5:** Preparing researchers to conduct research effectively (C3), e.g.:
“Data integrity, data quality, [be it] scientific or technical, all comes from the researchers. That’s why they must be so well trained when it comes to data collection.” (C3)

Four of the researchers (R20, R21, R22, R24), all holders of PhD degrees, felt that the Flexibility of data cannot be used to measure the scientific and technical quality of data – which runs counter to the statements presented in Section 6.6.5 – ‘Flexibility of data’. In this regard, despite the fact that the researchers’ perspectives differed from that of the author, it was chosen to accept this viewpoint because it illustrates how the levels of comprehension among academics may differ.

5.6.7 Integrity

Of the 24 researchers interviewed, 20 (R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R13, R14, R15, R16, R17, R18, R21, R22, R23) (83%) explained how Integrity can be measured. All of the curators (C1, C2, C3, C4) (100%) interviewed responded to this question.

Measures that emerged:

- **Measure 1:** Ensuring that researchers follow ethical guidelines in their research practice (R4, R15, R23), e.g.:
“We have to get a formal statement from the research organisation in general; somebody must take responsibility for the integrity of all the data that we put out there.” (R15)
- **Measure 2:** How RDM processes align with or differ from existing research or entities (R5, R7), e.g.:
“How it (data) compares with similar work that others are doing.” (R5)
- **Measure 3: Steps of verifying the reliability** of a data source (R9, R10, R17, R21, C1), e.g.:
“The integrity would be based on the level of errors in the data, which speaks to data quality. And, to an extent, it’s about the quality of people that participate in the studies.” (R9)

- **Measure 4: Documenting the principles and guidelines** researchers follow when **conducting research** (R11, R13, R14, R16, R18, R22, C4), e.g.:
“Somebody will look at your documentation now to see how you did your research, what’s the methodology that you followed. [They will also] look at your ethical clearance forms.” (R11)
- **Measure 5: Managing research data** to ensure data **integrity** (R2, R8, R24, C1, C2, C3), e.g.:
“That’s basically the integrity of the provider. Do they provide good quality data so that the integrity of the programme or institution is maintained? I mean, is the quality of the data good? Has it been collected soundly?” (R2)
“The integrity of data, I actually have to run basic recurrences and then also [check for] consistency in the responses.” (C3)
A researcher (R2) with over 15 years of experience emphasised the importance of producing quality data for measuring integrity. Data integrity relies on practical management methods used by researchers. Quality data is not random, as R2 suggests. This feedback is aligned with ensuring data integrity, as discussed in Chapter 2, Section 2.3.4.

5.6.8 Interpretability

Of the 24 researchers interviewed, 21 (R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R21, R23) (87%) provided a way in which Interpretability can be measured. Three (C1, C2, C3) (75%) of the four curators interviewed also responded to this question.

Measures that emerged:

- **Measure 1: The ability of secondary users** to effectively **interpret** and **make sense** of data (R16, R12, R17, R18, R19, R21, C1), e.g.:
“Interpretability is the ability of anybody to be able to use the data in a way that makes sense.” (R16)
- **Measure 2: Providing descriptive information** that describes the **characteristics** and **meaning** of data (R1, R2, R7, R23, C2, C3), e.g.:
“Publishing metadata that helps the user to understand each variable and, maybe, how to analyse that variable.” (R23)
“I think to measure interpretability you will have to use monitoring checklists or data checklists as well as completing the standard operating procedures.” (C2)
- **Measure 3: How well data aligns with research objectives** (R3, R4, R6, R8, R11), e.g.:
“Interpretability can only be determined based on what we currently know about the thing that we are trying to investigate.” (R11)

- **Measure 4:** Statistical interpretation to understand the significance and implications of data (R5, R9), e.g.:
“We can analyse and interpret the statistics coming out of the data.” (R9)
- **Measure 5:** Data governance focusing on **ethical and effective use of data** (R10, R14, R15), e.g.:
“I document it in the definition in the methodology section of the report, and how it was written in the proposal. It starts from the methodology section of the proposal.” (R10)

Creating data documentation that clarifies the data for secondary users was mentioned by some of the researchers holding Master’s degrees (e.g., R23). The discussion that appears in Chapter 2, Section 2.3.5.3 lends credence to this understanding.

5.6.9 Methodological Soundness

Of the 24 researchers interviewed, 23 (95%) (R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R20, R22, R23, R24) provided a way in which Methodological Soundness can be measured. All of the curators (C1, C2, C3, C4) (100%) responded to this question.

Measures that emerged:

- **Measure 1:** Evaluation of the research methodology to determine if it addresses the research question (R1, R3, R4, R5, R6), e.g.:
“Methodological soundness can be assessed through various indicators, but all depend on the context as well as the objective of the research.” (R6)
- **Measure 2: Quality management methods** to ensure **consistent processes** (R7, R8, R10, R11, R12, R14, R18, R22, R23, C1), e.g.:
“Methodological soundness. I think, Glenn, that your data should probably be weighted.” (C1)
- **Measure 3:** Using the same methods and processes consistently to address research objectives (R13, R15, C4), e.g.:
“Whether the methods speak to the issues they are out to address.” (R13)
- **Measure 4: Reproducibility** in research (R9, R16, R17, R19), e.g.:
“I would say it’s the level at which we can [reproduce the results] or the data using the method of data collection.” (R9)
- **Measure 5:** Using language that is clear, concise and easily understandable by participants (R3, R24), e.g.:
“If people have understood your questions, then your results will tell you that.” (R24)
- **Measure 7:** Applying standards in the research approach (C2, C3), e.g.:

“Which international and national peer review standards do you use when curating and managing research data and how do organisations that manage data do it?” (C3)

Higher education enables researchers to recognise important aspects of RDM. A PhD-level researcher (R6) emphasised the importance of following indicators that assist in meeting research objectives. This indicates an understanding that complying with such indicators will guide research to meet its objectives.

In their remarks, two curators (C2, C3) with varying educational backgrounds (Master's vs. undergrad) emphasised standards. This illustrates how experience enables one to have a broader understanding of critical RDM elements, such adhering to standards. Quality Check Processes

Of the 24 researchers interviewed, 23 (95%) (R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R20, R22, R23, R24) responded how accessibility can be measured. Three of the curators (C1, C2, C3) (75%) explained how Quality Check Processes can be measured.

Measures that emerged:

- **Measure 1: Processes and techniques** used to ensure **data meets quality standards** (R1, R2, R3, R5, R8, R9, R11, R12, R14, R16, R19, R21, R22, R24), e.g.:
“I suppose checking if there are quality check processes available. Asking what people are using to do the quality checks and ensuring that it is a complex system that looks at the kind of data that’s being collected.” (R3)
- **Measure 2: Verifying the correctness of data** when conducting research (R6, R15, C1, C2), e.g.:
“The researcher would have to hire a data manager. Someone who will observe that whatever that is being collected is accurate, or [at least] consistent with the questionnaire, before they can even send it for curation.” (C1)
- **Measure 3: Creating related documents** that allow for meaningful interpretation and analysis of data (R7, R10, R18, R23, C3), e.g.:
“The training manual provides a project overview, the team structure, the interview schedules, instructions on how to do the interviews and instructions for community entry as well.” (R10)

The doctoral-level researchers (R1, R2, R3, R4, R5, R6, R7, R9, R10, R11, R12, R13, R16, R17, R18, R19, R20, R22, R24) provided responses that evaluate processes, standards, repeatability, protocols, documentation, methodology, paradigm, and best practices in research projects. This suggests that more educated researchers link the methodological soundness element to the techniques for efficiently handling research data. Two curators (C1, C2) holding Master’s degrees expressed the need to know about the research process and data management assistance in a research team. An example of the type

of data management assistance is discussed in Chapter 2, Section 2.3.5.2, as provided by the ICPSR data management guideline.

5.6.10 Relevance

Of the 24 researchers, 23 (95%) (R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R21, R22, R23, R24) offered a method for measuring quality check processes. Two of the curators (C1, C2) responded to this quality dimension.

Measures that emerged:

- **Measure 1: The extent to which research results are useful** in addressing the **research question** (R2, R5, R6, R7, R11, R15, R16, R18, R19, C1), e.g.:
“I think you measure relevance in terms of [whether it applies] to the intended question. Does it address any of the issues that you think you have targeted?” (R5)
- **Measure 2: Identifying data outliers** to ensure **data accuracy** (R4, R10, R13, R13, R17, R21, C2), e.g.:
“Checking the relationship between variables and whether it is relevant or not.” (C2)

One of the researchers (R5) expressed the sentiment that, for research data to effectively address the questions at hand, data quality is crucial. According to one curator (C2), variables and their definitions must be accurate if data is to respond to the intended study questions. The answers provided above show how closely related RDM ideas are in generating high-quality data.

5.6.11 Standards and Policies

Of the 24 researchers interviewed, 20 (R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R12, R13, R14, R15, R17, R18, R19, R21, R22, R23, R24) (83%) explained how Standards and Policies can be measured. Three of the curators (75%) (C1, C2, C4) explained how Standards and Policies can be measured.

Measures that emerged:

- **Measure 1: Quality assurance** based on **organisational standards and policies** (R1, R2, R3, R4, R5, R6, R7, R9, R10, R12, R13, R14, R15, R16, R17, R18, R19, R21, R22, R23, R24, C1, C2, C3, C4), e.g.:
“Do you have a standards and definitions document? Do you have a procedures document used at various phases for collecting the data? And for the processes [such as the] fieldwork phase and the data processing phase, what were those procedures? Were the policies in-house? Was a consistent policy used in every survey, or is it a different policy or methodology?” (R4)
“We’ve got to use the policy and we’ve got to comply with it. So, we’ve got organisational policies in terms of authorship guidance. We’ve got one for data management and sharing.

We've got a research output policy, so there are various policies that can inform how data management should be done.” (C2)

The researchers, though they had varying levels of experience, recognised the importance of adhering to different standards and policies in their work. Even those with less expertise were aware of protocol adherence. One curator (C2) mentioned complying with policies and adhering to them, which should be a standard practice across the organisation.

5.6.12 Timeliness

Eighteen (75%) researchers (R1, R2, R4, R5, R6, R7, R8, R9, R11, R12, R13, R14, R15, R16, R18, R19, R21, R24) gave responses on how the Timeliness quality dimension can be measured. Answers related to measuring timeliness were given by two curators (C1, C2).

Measures that emerged:

- **Measure 1: Timely** and **reliable** access to data (R1, R2, R5, R11, R12, R14, R15, R16, R18, R19, R21, R24, C1, C2, C3), e.g.:
“You should look at when a study was completed, enrolment or follow up, and when they are made available online or when they're curated. How long is it between when the data is collected and when it is made available.” (R12)
“Always try to ensure that you do your whole project and also that everything is done according to the timeframe that you set.” (C1)
- **Measure 2: Disseminating** data while the **issue is still relevant** (R6, R7, R8, R9, R13), e.g.:
“Does it respond to the topical issues at the time? If the data misses that point, then it's not timely.” (R9)

The majority of the researchers with PhDs (R1, R2, R5, R11, R12, R16, R18, R19, R21, R24) believed that data should be made available promptly – this includes three curators (C1, C2, C3) with different education backgrounds (i.e. Master's and Undergraduate). Another group of PhD researchers (R6, R7, R9, and R13) felt that making data available when the topic the data should address is still relevant qualifies as being timely. The comments of the research participants show that academics from a variety of educational backgrounds think that data should be made available as quickly as feasible so as to address pressing contemporary concerns.

5.6.13 Uniqueness

The Uniqueness dimensions received responses from 15 (62%) of the researchers (R2, R3, R4, R5, R6, R7, R8, R12, R14, R16, R17, R18, R19, R22, R23). Of the four curators interviewed, two (C2, C3) (50%) outlined a method for determining uniqueness.

Measures that emerged:

- **Measure 1:** The **uniqueness** and **reliability** of data (R2, R3, R5, R16, R17, R22, C2), e.g.:
“Uniqueness applies to something that is found in only one instance. It's something that is not repeated elsewhere.” (R16)
- **Measure 2: Data description** and **consistency** (R6, R7, R8, R12, R14, R18, R19, R23, C3, C4), e.g.:
“You measure uniqueness by having no duplication and by being able to define each variable properly and clearly, and [also by making] sure that there are no overlaps” (R18).
“Are the questions properly phrased? Do you get unique responses from your questions?” (C3)

The researchers who did not respond (R1, R9, R10, R11, R15, R20, R21, R24) hold advanced degrees. This indicates that academically literate researchers do not consider Uniqueness as a quality dimension for assessing technical and scientific quality.

Two curators (C2, C3) agreed that each record in a data file should be distinct, ensuring accuracy and consistency, which is a technique that was explained in Chapter 2, Section 2.8.2 and Chapter 6, Section 6.6.5. The two curators (C1, C4) who chose not to reply share a similar educational background (Master's degrees), suggesting that they have similar thought processes.

5.6.14 Usability of the data

Twenty (83%) of the 24 researchers (R1, R2, R3, R4, R5, R6, R7, R8, R9, R11, R12, R13, R14, R15, R16, R17, R18, R21, R23, and R24) questioned provided a response outlining a procedure for figuring out whether or not data is usable. Three curators (C1, C2, C3) (75%) of the four interviewed explained how the Usability of data can be measured.

Measures that emerged:

- **Measure 1: Descriptive** information of a **data file** (R2, R3), e.g.:
“I think it's got to be done with clarity. So again, with the metadata you need to explain about data.” (R2)
- **Measure 2: Data format versions** (R4, R14), e.g.:
“Is the data in a format that is easy to understand?” (R4)
- **Measure 3:** The **usefulness** of a **data file** (R1, R6, R7, R8, R24, C3), e.g.:
“The master data must contain all the original variables as well as be coded and have aggregated variables, which is quite important, and it depends where the recorded and aggregated variables are because it's very important that the data curation team checks that these recoded and aggregated variables are there.” (C3)

- **Measure 4: How usable** is a **data file** (R5, R9, R16, R23), e.g.:
“I would say, firstly, if the data is considered to be very useful by many people. So that's one aspect of usability. In other words, you don't put up a dataset that nobody ever looks at.” (R16)
- **Measure 5: Data accessibility restrictions** (R15, C2), e.g.:
“Determine if there are going to be restrictions on the use of the database by, say, a funder, because of the sensitivity of the nature of the project in an organisation.” (R15)
- **Measure 6: Increasing data usage** (R11, R12, R13, R18, R21, C1), e.g.:
“It would be outputs that people generate from the data set. That could be the only way that you can determine the usability of data.” (R11)

The most seasoned curator (C3, 6-15 years of experience) said that the master data file needed to have all of the original variables, be coded, and have aggregated variables. This feedback relates to the discussion that was reported in Chapter 2, Section 2.7.1.2 – i.e. access permissions and management of the master data file should be handled with extreme care.

5.6.15 Validity

Twenty-two (91%) of the twenty-four researchers (R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R21, R22, R23, R24) interviewed provided a response. Three of the curators (C2, C3, C4) explained how Validity can be measured. The Measures that emerged from the responses include:

- **Measure 1: Quality control verification methods** to ensure data accuracy (R2, R3, R4, R6, R7, R8, R10, R12, R14, R15, R16, R18, R22, R23, C3, C2, C4), e.g.:
“Validity all depends on the different statistical tests. Like significant tests you're doing. Things like regression are statistical tests that you can use to measure the validity of your findings.” (R6)
“Missing data comes out in your frequencies, of course. If there is missing data, you will first look at whether it was a skip question.” (C3)
- **Measure 2: Data relevance in solving research objectives** (R5, R9, R11, R13, R17, R24), e.g.:
“Validity will also have to do with the issues that the data was out to address. From the very beginning, does it speak to the question that it was out to address?” (R13)
- **Measure 3: Data referencing and citation** (R21), e.g.:
“So, usability and validity, you know, it's only when people start citing you, quoting your work, that you realise that you've probably done a good job.” (R21)

According to researcher R6, it should be possible to assess the validity of (any) data if statistical tests are applied. With the exception of two (R19, R20), all of the researchers holding PhDs responded to this question, demonstrating that educationally driven researchers are capable of statistical analysis. These researchers demonstrated outstanding analytical abilities, regardless of their areas of specialisation. The use of statistical tests, as discussed in Chapter 2, Section 2.8.3, is comparable to the position taken by R6 concerning the use of statistical tests as a technique for assessing the validity of the data.

5.6.16 In conclusion and after considering all the measures collected

The study found that data quality compliance measures are essential and that certain data management areas are crucial for both curators and researchers. It was also established that more experience leads to the identification of better measuring techniques, but no definitive measures could be recommended. Significant further research is needed to find measures that could be applied across the organisation. In addition, further research is needed to understand the relationship between education, experience, and data management, as well as to consider human characteristics that could positively (or negatively) affect strategies to improve data quality.

5.7 Personal experience and preferences in data quality management planning

The interview schedule revealed the participants' data quality management planning preferences, revealing various methodologies and strategies. The participants also shared their experiences with data quality and management planning, emphasising the importance of understanding and implementing effective data quality management strategies. The quality dimensions chosen for this research study, such as accessibility, accuracy, consistency, integrity, interpretability methodological soundness, relevance, timeliness, uniqueness and validity were already familiar to some of the researchers. Others, however, identified quality dimensions of their own, similar to those chosen for this research study. The new dimensions that were reviewed included understanding the data (all the data points), having data backup and documentation, internal validity, accuracy, having standards and Standard Operating Procedures SOPs in place, and data cleaning. These quality dimensions combine to create a comprehensive technique for assessing data quality. Researchers focus on these issues to ensure the authenticity, trustworthiness, and usability of data in research projects. In this section, the participants' responses are not only grouped (as in the previous section) but direct quotations are also used to represent the responses that were given. This was done to preserve the contextual and practical meaning of what a person said.

5.7.1 Quality dimensions and measuring criteria that researchers address

The goal of this section was to determine whether the quality dimensions of the study were in line with the techniques and approaches already employed by the researchers to manage the quality of their research data. The researchers were required to mention the quality dimensions and measuring criteria they address when working with their research data.

Q7. What quality dimension and measuring criteria do you usually address when working with your own research data?

All the related dimensions and measuring criteria identified by the researchers are grouped into the following Criteria:

- **Criterion 1: Practical data quality control measures** (R7, R8, R16, R17, R24), e.g.:
“I’ll go back to the data capturers and inform them that I’m seeing a lot of errors [...] So for me that’s the best way.” (R8)
- **Criterion 2: SASQAF-based dimensions** (R1, R4, R6), e.g.:
“Those eight quality dimensions in SASQAF are obviously the things that I conform to when I look at my work.” (R1)

The author did not ask again for the measuring aspects for all of the dimensions because they are the same as those used for SASQAF evaluation. The SASQAF framework is discussed in detail in Chapter 3, Section 3.7.3.

Some of the researchers (R2, R5, R9, R10, R11, R12, R13, R14, R15, R18, R19, R20, R21, R23) mentioned dimensions that were already used in this study and that also form part of SASQAF and, for this reason, those dimensions are excluded from the discussion. The dimensions are tabulated below:

Table 5.6 - Dimensions excluded from the discussion

Dimension	Number of respondents	Researchers
Accessibility	R10	1
Accuracy	R12, R21, R23	3
Completeness	R12, R23	2
Consistency	R5, R9, R13	3
Flexibility of the data	R13	1
Integrity	R9, R12, R18	3

Interpretability	R5, R10, R11, R23	4
Methodological Soundness	R12, R13, R14, R15, R18, R19, R23	7
Quality check processes	R12, R19	2
Relevance	R2, R9, R11, R12	4
Standards and policies	R23	1
Timeliness	R2, R9	2
Uniqueness	R2, R12, R23	3
Usability	R2, R12, R23	3
Validity	R12, R15, R18, R21, R23	5

A selection of quotes from research participants about the quality dimensions and measuring criteria appear below:

- “Those eight quality dimensions in SASQAF are obviously the things that I conform to when I look at my work.” (R1, R6)
- “We make use of the statistical value chain outlined by the statistical quality assessment framework [SASQAF].” (R4)
- “Well, for me, the key dimension is consistency. Also, the issue of accuracy is very important for me, and another dimension that is important is integrity.” (R5)
- “Methodological soundness is important as are our quality check processes. Methodological dimensions are important for promoting research to other audiences.” (R19)

When asked what quality characteristics and measurement criteria they consider while working with research data, the researchers gave a variety of different answers. Some of these were notable, such as the quality dimensions used in the Statistical Value Chain (SASQAF). The quality process model of this research study was developed using the framework that these researchers discussed, making it possible for some of them to relate to and better grasp the setting of the study. The quality dimensions, measuring criteria and evaluation standards that curators use when curating research data are the subject of the following section.

5.7.2 Quality dimensions, measuring criteria and evaluation standards curators use when curating research data

Understanding the quality dimensions that the curators consider when curating data is essential since it allows one to gauge the extent of best-practice RDM support that they are providing to the researchers.

Q7. What quality dimension(s) and measuring criteria do you usually use when curating research data?

The curators interviewed all gave different responses to this question. Quotations from their responses are presented below:

- “We have a monitoring checklist. That helps us to assure that we've looked at all the data quality issues and that these are addressed during the curation process.” (C2)
- “Do your data quality checks, logical checks, and frequencies.” (C3)
- “I use validity and reliability. I also [check for] consistency and uniqueness.” (C4)

The comments above suggest that each curator employs a unique procedure for curating research data. To handle data efficiently and offer researchers appropriate RDM support, the curators must embrace standardised best practices, such as those explained in Chapter 2, Section 2.3.5 and Chapter 6, Section 6.7.3. The quality practices that research participants utilise or are aware of are covered in the next section.

5.7.3 Quality procedures that researchers use/are aware of

The approaches to quality procedures that researchers use throughout a research study directly determine the level of quality exhibited in a data file. Understanding the quality controls that respondents use is crucial.

Q8. What quality procedures do you make use of/are you aware of?

All related quality procedures are grouped to form the following three broad approaches:

- **Approach 1:** Compliance with methodological frameworks as part of RDM (R1, R2, R4, R18, R19, R21), e.g.:
“We make full use of the methodology. We use the Frascati manual. That's for one project and the Oslo manual for another project. Obviously, there are quality procedures and [so on for] each article [and those are in] the data collection manual.” R4)
- **Approach 2:** Methodological soundness based on quality assurance processes (R5, R6), e.g.:
“I think, first and foremost, my quality procedures ensure that you pilot your work. You test your systems even if they're electronic or they're paper-based. Piloting your work is very

important. You can miss a comma and find that everything stops working. [It's really a matter of] testing your systems before you go out there and try to collect data." (R5)

- **Approach 3:** Quality control processes used for data cleaning (R3, R7, R8, R9, R11, R12, R14, R15, R17, R20, R22, R23, R24), e.g.:

"We have teams that go back and do re-visits. Our statisticians will run quality checks [on the data that comes in] to identify inaccuracies in the data set." (R11)

Piloting is a key factor in determining the effectiveness of a certain component (such as a data collection instrument) in a research study, as described in Chapter 4, Section 4.8. Finding out that some of the researchers mentioned piloting was encouraging.

It was interesting to establish that some researchers were not aware of any quality procedures. This was the case with one researcher (R10) who stated:

"I'm not aware of any quality procedures, apart from checking data and accuracy." (R10)

This comment highlights the need for researchers to be provided with training and information on quality controls in order to improve RDM across the research data lifecycle.

Another researcher (R12) described the order in which the quality procedures in his research are applied. The researcher (R12) explained:

"We have people immediately review surveys and make sure that they are completed fully. Then we have a second-level check where we look for any inconsistencies or inaccuracies in the data. We produce reports where we check things like randomisation and making sure that there aren't any issues with those processes." (R12)

Reviewing and checking must be done repeatedly as a research project is being conducted. Hands-on quality control should be applied as soon as possible, and this should be supported and promoted throughout the entire institution as a "best practice" procedure.

The following description includes the quality controls that one researcher (R22) mentioned:

"Yes, it's, for instance, checking for duplicates. Checking for missing data and then making sure that those are clear. [You must also] consider the issues of your data, your variable types [and whether they have]' merit. You also need to be aware of those [as well as of] the issue of completeness." (R22)

The Principal Investigator's RDM procedures should be used regularly by data managers where one or more data managers is/are employed in order to prevent the errors raised above. The data should be

provided to the Principal Investigator in a “cleaned version,” and he or she should run tests to verify consistency and accuracy.

To make sure that the data they utilise for a study are accurate and trustworthy, researchers employ a variety of data quality procedures. Data cleaning, data validation, data transformation, and data analysis are a few examples of these processes. By using these processes, researchers can find and fix any data mistakes or anomalies, ensuring that their findings are reliable, error-free and informative, which ultimately improves the quality of data. In order to ensure the validity and integrity of research findings, it is crucial to apply reliable data quality processes, as demonstrated in this section. The quality control procedures that the curators employ are covered in the next section.

5.7.4 Quality control processes that curators use

As part of its oversight function, the data curation unit ensures that deposited data files are correctly handled so they can be separated from other data files. This matter was investigated as part of this study and the curators’ responses are considered below.

Q8. What quality control processes do you use to ensure that deposited data files are accurately processed to make them distinguishable from other data files/versions?

This question brought only two Processes to light:

- **Process 1: In-house quality assurance** methods are used (C1, C2, C4), e.g.:
“If a data file has different versions, each data file has its own questionnaire and receives a unique/different name.” (C1)
- **Process 2:** Standardised data curation verification methods are implemented (C3), e.g.:
“The quality checks [are] sometimes logical checks that I [do] on the data.
That [takes care of the quality checks] and then [I look at the] variables [within] the interview schedule.” (C3)

The use of different file names is a quality assurance technique that makes it easy for secondary users to differentiate between data files. However, the data curation unit’s naming convention must be based on best practices in order for it to make sense and be understandable for secondary users. Curators must be able to respond to the naming convention requirements listed in Section 2.3.5.3 of Chapter 2 when constructing a naming convention.

Data cleaning, data storage, data validation, conversion, and passive analysis are a few of the quality control techniques that the data curators interviewed use to manage a data file. By identifying errors and inconsistencies in the data, these procedures serve to ensure its accuracy and integrity. The handling and upkeep of high-quality data for use in research and analysis requires the application of specific

methods. The quality process model is discussed in the next section, based on Section C of the interview schedule.

5.7.5 Planning data quality at the same time as data management planning

To increase the possibility of producing data that is of high quality, researchers should pay equal attention to both managing data quality and managing data (in general). The opinions of the researchers on the management and planning of data quality are examined in this section.

Q9. What is your opinion – will it be acceptable to plan data quality at the same time as when data management is planned? Please elaborate.

All related proposed modifications are grouped to form the following two options:

- **Option 1:** Planning for data management and data quality together is acceptable (R3, R5, R6, R10, R11, R12, R13, R14, R16, R18, R22, R23, R24), e.g.:
“Putting a quality process in place and data management planning should happen at the same time.” (R5)
- **Option 2:** Planning data management and data quality at the same time is undesirable (R8, R15, R17, R19, R20, R21), e.g.:
“I think they [data quality and data management] should be separated. The data quality [...] checking process, in and of itself, is quite long and, sometimes, data management seems to mess with the data quality process.” (R8)

More researchers agreed that combining data management and quality planning is essential than did not believe this to be a good idea. Developing data quality measures is crucial in the initial stages of data management planning to ensure precise data collection and useful analysis and decision-making. Prioritising data quality improves the accuracy of findings and influences the field of study, making it a more appropriate approach for researchers to take.

5.7.6 Procedure in place to confirm data quality preparation completed - Curators

Understanding the processes and techniques the data curation unit uses to verify that researchers have completed the production of high-quality data was seen by the curators as crucial. The necessary level of support interventions is determined by this understanding.

Q9. Is there a procedure in place that confirms that data quality preparation was completed by researchers? Please elaborate on your answer.

Each curator mentioned a different procedure and some of their insights are presented below:

- “At [the social sciences research organisation] we have a particular **process flow diagram** that is based on the activities from [the organisation’s] Digital Curation Centre lifecycle model and that is also linked to our data preservation and sharing policies. [The diagram specifies] what the researcher should do before submitting data for curation.” (C1)
- “Yes, there is **a data checklist** for quantitative and qualitative data available to researchers and they can use that checklist to assure that the data quality is good before they deposit the data for curation.” (C2)
- “It’s the basic things that you’re looking for, [such as **outliers** and **frequencies**]. If you start off with basic steps of validation when the data comes into the curation centre, then you will see the quality of the data.” (C3)

In light of these comments, it becomes clear that the data curation unit applies quality control techniques informed by best practices (such as The Digital Curation Centre lifecycle). However, it appears that the researchers are not obliged to prepare for data quality and that the existing checklists are not mandatory documents. This requirement should be addressed in the data management plan that is created at the beginning of the research process and that describes the procedures for data collection, organisation, storage, and analysis, as discussed in Chapter 6, Section 6.7. The plan must be kept relevant and effective in achieving the objectives of the research and can be achieved by making changes as frequently as necessary.

5.7.7 When to measure the scientific quality of data - Researchers

Researchers perform different measurement and assurance tasks during the course of a research project. The period during which researchers measure scientific quality will be discussed in this section.

Q10. When do you propose we measure the scientific quality of the data?

Five options were mentioned:

1. At the beginning of a research project: Five researchers (R2, R8, R15, R16, R22) mentioned that the scientific quality of data can be measured from the beginning of a research project.
2. Throughout the research process: Ten researchers (R1, R3, R5, R6, R7, R12, R17, R18, R20, R23) mentioned that the scientific quality of data can be measured throughout the research process.
3. At the end of the research project: Two researchers (R14, R19) stated that the scientific quality of data should be measured closer to the end of the research project.
4. When the data instrument is being designed: Another researcher (R4) was of the view that the scientific quality of data should be measured when the data instrument is being designed.
5. During the data collection phase: Five researchers (R10, R11, R14, R21, R24) said that the scientific quality of data should be measured during data collection – i.e. as data comes in.

The relevant comments that were used to reflect on when the scientific quality should be measured are as follows:

- “*We manage the scientific quality of the data from the beginning.*” (R2)
- “*You should start looking at quality reading we do when you’re designing your instruments to collect the data.*” (R4)
- “*Throughout the research process because then if you wait till the end, you might not be able to fix anything that is wrong.*” (R17)
- “*Closer to the end of the research project. As a kind of retrospective evaluation.*” (R19)

Considering the comments provided by the researchers, it is evident that a comprehensive data management plan should be utilised to evaluate the scientific quality of data, which is an essential element. This plan should be created at the start of the research project and periodically evaluated as the project develops. By placing more emphasis on data quality, researchers can increase the reliability of their findings and have a greater influence on their respective disciplines. The next section discusses the responses that the curators gave regarding how the scientific quality of data can be measured.

5.7.8 When to measure the scientific quality of the data - Curators

Various quality control duties are carried out by the curators throughout a research project. The time frame for the measuring by the curators of scientific quality is covered in this section.

Q10. When do you propose we measure the scientific quality of the data?

Only two options could be drawn from the responses that the curators provided. These are:

- **Option 1:** Before data is submitted for curation (C2), e.g.:
“I think it should be measured before [the data] is sent for curation because, if it's disseminated and it's not scientifically correct or qualified then we're in for big problems.”
(C2)
- **Option 2:** Throughout the data lifecycle (C1, C4), e.g.:
“[I think it takes place] throughout the lifecycle of the data.” (C4)

One curator (C3) did not understand when the scientific quality of data should be measured. In addition, this curator did not name any measuring items that would be of use to scientific quality.

Based on their remarks, the curators offered an assessment of the scientific quality of data in various phases of the data management process. Examples of these phases include the design of the questionnaire, preparation for data collection, data format, data cleaning, handling of ethical issues related to data-gathering, and making sure the data complies with relevant standards and guidelines.

The ultimate goal of the curators is to provide trustworthy, excellent data management support to enable researchers to produce excellent research outcomes. The next section focuses on the feedback from the researchers concerning when the technical quality of data should be measured.

5.7.9 When to measure the technical quality of data – Researchers

Researchers were asked to explain when should the technical quality of data be measured; their responses are provided in the section below.

Q11. When should the technical quality of the data be measured?

As with measuring the scientific quality of data, the responses provided by the curators concerning when the technical quality of data should be measured are grouped as options:

- **Option 1:** At the beginning of a research project (R2, R5, R10, R12, R24), e.g.:
“[At] the beginning of data collection.” (R10)
- **Option 2:** Throughout the research project (R3, R6, R7, R8, R11, R15, R16, R17, R18, R20, R21), e.g.:
“I would say also [that] it's a continuous process right from the beginning and right up to the end.” (R16)
- **Option 3:** At the end of a research project (R1, R13, R19, R22, R23), e.g.:
“At the completion phase.” (R13)

The researchers believed that several stages of a research project, including data collection, processing, analysis, and dissemination, should be employed in order to evaluate the technical quality of the data. This is similar to how the researchers assessed the quality of their own work. Making sure the data is reliable, accurate, consistent, and in accordance with all relevant guidelines and standards is crucial in scientific research since it makes it possible to check the veracity and precision of the results. The opinions of curators regarding the appropriate times to measure the technical quality of data are covered in the next section.

5.7.10 When to measure the technical quality of data - Curators

The curators had their own viewpoints concerning when technical quality should be measured and this section looks at the responses that they gave.

Q11. When should the technical quality of the data be measured?

Again, each curator identified a different option:

- **Option 1: During** data curation (C1), e.g.:
“I think, during the curation part.” (C1)

- **Option 2: After** the curation is complete (C2), e.g.:
“I think, maybe, when the curation is completed, because then you can get a really good idea if the cleaning [and the anonymisation were] done well.” (C2)
- **Option 3: Throughout** data collection (C3), e.g.:
“Throughout data collection [and] data capturing.” (C3)
- **Option 4: When** data is collected (C4), e.g.:
“You start when the data [is] collected, [with the] procedures that we use to collect the data.” (C4)

The interviewees' perspectives on the technical quality of data were influenced by different curation experiences. To ensure the validity and reliability of research findings, curators should assess the technical quality of data early on, as well as regularly throughout the research data lifecycle. This includes the initial data input, ingestion, processing, conversion, cleaning, and analysis. Detecting and rectifying any data quality concerns as early as possible is crucial. A robust quality control system for data curation is essential for conducting high-quality, reproducible scientific research.

5.8 Findings: Section C - Opinions about the proposed quality process model

After learning about the personal experiences and preferences of the research participants in respect of data quality, it was important to understand their opinions about the proposed quality process model. The data quality process model was seen in a variety of ways by the research participants, who also provided feedback on how to simplify the model. The participants in the study valued the development and refinement of the models since such models raise the overall integrity and quality of the data. The participants thought that the data quality process model may promote usefulness, transparency, and the production of high-quality data if it were simplified, identified the accountable parties, and included some input (in the form of feedback loops) on activities.

The participants in the study were required to ask any questions they might have had regarding the quality process model before responding to questions about it in this section.

The questions and comments that researchers provided in response included:

- “I’ll be quite honest with you, I do find some of these become too busy.” (R2)
- “Your formative research is missing, what I’d say is basically missing here, is your formative research...” (R24)

(Formative research refers to research carried out during programme/research production to identify and describe the target audience).

In light of the aforementioned remarks, researchers might be worried about specific data management-related challenges that are relevant to their line of research. It is vital to address these questions in order to make sure that researchers have a clear understanding of the model and how it might help their work.

This section is significant because it discusses the viewpoints of the suggested quality process model, which is the primary subject of this research study.

5.8.1 Opinions expressed by Researchers

It was anticipated that the research participants would perceive the quality process model from various angles. A candid and fruitful discussion with them was thus necessary to assess the usefulness of the model. A question about the model's practicality was posed to the research participants.

Q12. Is the model too complex/too simple for practical use? Please explain your viewpoint.

The responses provided by the research participants are encapsulated in one of three measures:

- **Opinion 1:** The model is too complex (R1, R2, R3, R5, R6, R7, R8, R13, R17, R20, R23), e.g.:
“Well, for me, I think it's a bit complex. How usable can it be? People technically more advanced than others, can understand it better, so we probably can simplify it, but [this] doesn't mean that the bigger model needs to be discarded.” (R5)
- **Opinion 2:** The model is simple (, R4, , R9, R10, R12, R14, R15, R16, R18, R19, R21, R22, R24), e.g.:
“I actually found it very straightforward and simple.” (R16, R21)

Considering the above explanation, it is critical that researchers and curators alike thoroughly comprehend the benefits of modifying their RDM methods by utilising the data quality process model as a guiding framework, because it will help them address RDM concerns that, previously, they would have neglected to address (such as how to deal with anonymisation or utilising consistent file names). Researcher (R5) highlighted that more technically advanced people would easily understand the model. For this reason, it was important to modify the data quality process model and use a simple and understandable layered approach. This modification is recorded in Chapter 6, Section 6.7.

The model was changed, adopting a layered approach that looks at each stage of the data lifecycle, one at a time (as illustrated in Figure 6.9, Section 6.7.10). By using this method, a user can read the phase's accompanying explanation and put what they learn into practice. In order to tackle any potential difficulties that may develop during the data management process, it is crucial to make sure that the right quality control measures are in place. The proposed model can be a helpful tool for researchers across a range of areas when used in the right context. The next section focuses on the opinions of the curators regarding the proposed model.

5.8.2 Opinions about the proposed model - Curators

As with the researchers, the viewpoints of the curators regarding the practical use of the data quality process model were examined. This section discusses the responses that the curators gave.

Q12. Is the model too complex/too simple for practical use? Please explain your viewpoint.

It was fascinating to learn that the model gave rise to a 50% split in opinion: two curators (C2, C4) thought that the model was overly complex, while two curators (C1, C3) found the model to be straightforward to comprehend and to have been well done. This implies that, while some curators will be able to use the model with ease, others may find it difficult. The comments that follow show that the curators were aware of the challenges associated with data curation, indicating that certain curators value the help and standardisation that the proposed model brings to the table.

One of the curators stressed that the model was complicated:

- “It is a complicated model. But I think [that the] curation process is already a complicated process, so I don't think it's incomprehensible for people who do curation.” (C2)

The following statement was made by another curator, who felt that the model was not complex:

- “I think that the model is not complex. It's simple to understand”. (C1)

The curators' responses indicate that the model should be fairly straightforward. The significance of quality control processes and how they solve any potential issues that can arise during data curation must be explained in order for research participants to understand the model and for it to be fully implemented. This is accurate, as explained in the discussion in Chapter 6 of how each data quality dimension contributes to data quality (Section 6.6). As long as adequate quality control procedures are employed, the proposed model can be a useful tool for both curators and researchers. The opinions of research participants on how the model should be modified to meet their needs are the subject of the following section.

5.8.3 Required changes to the model

The final version of the data quality process model was created through iteration and reflection on research participants' input. The practical approaches that researchers and curators employ to handle data impact how the model should be altered. The next section addresses the proposed improvements identified by the research participants.

Q13. Based on your answer in Q12, how should the model be changed to comply with your requirements?

All related proposed modifications are grouped to form themes. Three distinct themes could be identified:

- **Theme 1:** The importance of a clear **interactive approach** to quality management (R1, R4, R6, R10, R11, R12, R14, R15, R17, C2), e.g.:
“In the bottom right-hand corner, where you have ‘analysing data’, I would suggest you have some kind of feedback loop. [Perhaps] some kind of dashed line that illustrates that, when you’re analysing data, you encounter problems and [and that these] could be big problems, such as encountering anomalies in the data. [You could] then have a yes/no kind of branch the way you do with the others.” (R1)
- **Theme 2:** Clarity in quality management to ensure that **the structure minimises complexity** (R5, R9, R16, R18, C1), e.g.:
“I don't know whether it's reporting or whether the arrows going back into the quality process framework are [some kind of interim reporting] about what's happening at each and every step.” (R18)
- **Theme 3: User-friendly design** in quality management to facilitate effective use (R2, R3, R7, R13, R19, R22, R24, C4), e.g.:
“You may want to shift things around; the model has to have checks on processes in every stage of the research data lifecycle. As a researcher progresses in the research data lifecycle, [he or she] has to go back and look to verify if everything is OK.” (R2)

The suggestions proposed by the researchers and curators call for an interactive, simplified, and user-friendly data quality process model. Accordingly, it was critical to adjust the model in a straightforward, interactive, and intuitive manner.

During data collection, the technical aspect of the first draft of the data quality process model became a concern after the author realised that participants were not fully understanding and engaging with the model. Thus, the above remarks confirmed that the model needed to be simplified, and the changes are recorded and explained in more detail in Chapter 6, Section 6.7. A Feedback Loop (FBL) component that feeds into a Check Progress (CP) component was added to every stage of the process. The information from the FBL and CP components then triggers the generation of an Interim Report on whether the activities of that stage have been done properly. The interim report is submitted to the responsible person (either the Data Manager, Principle Investigator or the Data Curator) for reviewing and checking of the results. If everything has been done correctly and is of good quality, the process may move to the next stage.

The areas in which the study participants believed the model would be useful are discussed in the next section.

5.8.4 In which areas will the model assist? - Researchers

The researchers were asked to identify any areas within research work or projects where they thought the model could be useful. The responses provided were analysed and are represented as themes. Where unique responses were provided, these have been quoted.

Q14. Taking your research work/projects into consideration, which data quality concerns do you think the model will assist with?

With the ever-changing RDM landscape of technology and processes, researchers face a variety of issues related to data quality during a research project. It was, thus, critical to identify concerns that the model is expected to address. All related proposed modifications are grouped to form the following areas:

- **Concern area 1: Data management** aspects (R1, R3, R12, R16, R21, R23, R24).
- **Concern area 2: Complying** with and using related **SASQAF dimensions** (R4, R11, R19, R20).
- **Concern area 3:** Effective data planning, collection and analysis (R5, R8, R9, R14),
- **Concern area 4:** Ensuring the reliability and validity of research results (R6, R7, R17, R22)

The following quotes are explanations that researchers provided regarding the data quality concerns researchers thought the model would assist with:

One researcher (R2) thought that the model – because of the terminology that it uses – was primarily **focused on quantitative data**. The researcher (R2) gave the following explanation:

“I think your model has more to do with quantitative data, data sets and their curation. That is my sort of general feeling, and that's largely because of the terminology used.” (R2)

The preceding remark is no longer a true reflection of the model because the redesigned data quality process model includes data management planning (Chapter 6, Section 6.7.3), which guides both qualitative and quantitative data.

Respondent R14 felt that the model would assist with certain dimensions, and the following quotes are provided to substantiate this:

“Data capturing will assist. The methodological approaches will also assist, [as will] the analysis. I think that those will be key.” (R14)

The preceding description illustrates the relationship between a research approach, data collection, and analysis. The methodological approach used by a researcher in a research project determines the type of data collection strategy and, to a large extent, how data will be analysed. Each approach (i.e.

qualitative and quantitative) has different data analysis methods (as discussed in Chapter 4, Section 4.3). A researcher must comprehend all of the implications of the methodological approach chosen so that data management within that approach does not become difficult and compromise data quality.

One researcher (R5) mentioned that the model would assist with a key aspect of this research study, which is **data management planning**. The researcher (R5) explained:

“It's really the issue of data management planning. It's a very important thing and it's also a requirement. Even to get ethics approval, you need a data management plan.” (R5)

A data management plan is the essence of good RDM, its importance cannot be overstated. The comprehensive activities around how data management planning should be done are presented in Chapter 6, Section 6.7.3. Before starting a research project, a researcher must define how he or she will handle data quality and management planning.

Researchers can recognise quality dimensions that they have previously encountered. The following response provides proof that the researcher concerned connected the quality dimensions utilised in this study with those of other quality-based frameworks:

“In my own research work, in the work that we do, we use all these dimensions. But, obviously, once you've done your study and you've published, I can guarantee you there are many students out there –maybe honours students or [those]going to do their Master's or even a PhD [who will not be] aware of these models. SASQAF is a complicated model if you look at it [and] you must read the whole manual. But your model is simple and it's relevant and it's easy to understand. So, I think it will [be of quite a lot of] future academics and even people who want to [conduct] research projects where data would be a huge point or outcome of that project.” (R4)

Given that the SASQAF dimensions were used as the key data quality dimensions in this research study, when a researcher says, “we use all these dimensions” it may be assumed that he or she has seen and is aware of the SASQAF data quality dimensions. This shows that, because they have already been exposed to SASQAF, some researchers will be more willing to implement the idea of using quality dimensions.

According to the comments provided by the researchers, the model will assist in addressing a number of challenges (such as organising a research study and data analysis, issues with data quality, content analysis, evaluation of the research's quality, and data preservation). The use of quality control techniques and the adoption of best practices enable this. The model will also allow researchers to actively monitor and preserve their data over time and this will reduce the likelihood of errors and increase the overall accuracy of the data. The next section focuses on the areas that curators thought the model would assist with.

5.8.5 In which areas will the model assist? - Curators

The areas in which the curators thought that the model would help were sought. The author focussed on the degree to which curators assist researchers with managing data. The responses are discussed in this section.

Q14. Taking your curation work into consideration, which data quality concerns do you think the model will assist with?

All related proposed areas that curators identified are grouped in the following concerns:

- **Concern 1:** Navigating and managing research data throughout its lifecycle (C1, C2), e.g.:
“I think [that]the entire model will be helpful because we are aiming to be part of the entire research lifecycle process and this [...] model can be very helpful in the total process of the research lifecycle, together with the data curation process.” (C2).
- **Concern 2:** Compliance will result in quality data (C3), e.g.:
“If people comply with all these steps, then we should have quality data.” (C3)
- **Concern 3:** Data collection planning (C4), e.g.:
“The data collection planning, because I think if people have knowledge of the model, data quality [will] be better because of data planning.” (C4)

One curator (C2) linked the future prospects of the data curation unit with how the model can assist in achieving the desired results. This response demonstrates that altering and upgrading RDM services is a work in progress – one that necessitates fresh and quality-based inputs (such as the data quality process model). The data curation unit must strive to offer a consistent (through the entire research data lifecycle) support structure for researchers, and the quality process model will be the tool to assist in doing so. The data quality process model was well received, according to curator comments, and it is anticipated that it will be able to influence data management through data curation and data collection planning. The model will also be able to assist researchers in lowering the likelihood of errors through ongoing monitoring and maintenance of data by researchers, with the help of curators, as discussed in Chapter 6, Section 6.6. Better research findings, enhanced data management choices, and assured secondary data users (who depend on high-quality data) requesting access to the institution's research data are all possible results. The data management areas that researchers felt that the model covered well are discussed in the next section.

5.8.6 The data management areas covered well by the model - Researchers

It was vital to determine which data quality management areas researchers believed the model adequately covered because researchers work in a variety of data management fields. Knowing which

data management areas researchers thought the model covered well was necessary in order to modify the model accordingly and these modifications are reported on in Chapter 6, Section 6.7.

Q15. Which data quality management areas do you think the model covers well?

The demographic information, specifically the ‘Expertise as a researcher’ (as discussed in Section 5.4.2), indicates that the majority of the researchers had more than 15 years of experience as researchers at the time of their responses. This suggests that the responses provided by the researchers to this question are provided by experts with a wealth of knowledge. All the related areas that the researchers identified are grouped in the following opinions:

- **Opinion 1: Overarching** objective of **preserving high-quality data in research** (R1, R3, R4, R8, R9, R11, R12, R14, R15, R18, R21), e.g.:
- “It gives one a helicopter view. It looks at the process of creating, processing and analysing data, [as well as at] that processing step...” (R1)
- **Opinion 2: Strategies and techniques** for managing data quality that are **effective** in helping to achieve the **necessary quality criteria** (R5, R7, R13, R16, R17, R19, R20, R22, R23), e.g.:
- “Guaranteeing the level of quality of the data that is preserved.” (R16)

The comments above show that researchers were able to see the model's overarching quality-focused potential concerning RDM in general. Their understanding here emphasises the advantages (described in Chapter 2, Section 2.2.2) that may be associated with good RDM and the application of the proposed data quality process model. The issue of change management, which is covered in Chapter 6, Section 6.10, is relevant here. With the right vision and understanding in place, researchers will be able to implement the necessary RDM changes, enabling improved data quality. Researchers should be aware, however, that improving RDM necessitates changing processes and the way things are done. The likelihood of errors will be reduced and the accuracy and readiness of the data for secondary use will be ensured by actively monitoring and maintaining all quality control processes during the data collection and processing phases. A proactive strategy for preserving data integrity is provided by the quality process framework (Chapter 6, Section 6.7.10). The next section looks at the data management areas that curators identified as being covered well.

5.8.7 The data management areas covered well by the model - Curators

It was vital to determine which data quality management areas curators believed the model adequately covered since they deal with and examine a variety of data management issues in their curation work.

Q15. Which data quality management areas do you think the model covers well?

This question uncovered only three opinions:

- **Opinion 1: Data management activities** (data collection, data processing) (C4), e.g.:
“The issue of data collection or processing.” (C4)
- **Opinion 2: All data preparation steps** in the data lifecycle (C2, C3), e.g.:
“If the researchers understand that they can use it throughout their data collection, data preparation, data processing and data analysis...” (C3)
- **Opinion 3: Accuracy** and completing the curation activities **on time** (C1), e.g.:
“I think that this one will assist with accuracy, as well as [with] being able to complete the curation on time.” (C1)

Curators appear to have slightly diverse perspectives on the areas of data quality management that the model covers well. The curators' identification of several features' points to the various RDM components, such as accuracy, a quality dimension, data collection and data processing, which are part of managing data in the different stages of the data lifecycle. that the model can help with. The opinions listed above might be used in the instruction process for curators as all of the aforementioned RDM components would be included in the training materials for the data quality process model. The next section looks at the areas of concern that the researchers thought the model would not address.

5.8.8 Areas of concern that the model will NOT address – Researchers

This section examines the issues that researchers brought up regarding the areas (data quality concerns) that they felt that the model would not address

Q16. Taking your research work/projects into consideration, which data quality concerns do you think the model will NOT address?

All the areas of concern that researchers thought the model would not address are grouped to form the following themes:

- **Theme 1:** Data quality planning activities (R1, R12, R15, R16, R18), e.g.:
- **Theme 2:** The **usefulness of research data** (R2, R3, R11, R24), e.g.:
“The design of things is what's going to influence everything else, especially in terms of quality. That interpersonal space around how questions are designed and the relevance of the questions [will be important].” (R3)
- **Theme 3:** Identifying unethical methods of data acquisition (R5, R19), e.g.:
“I think it's dishonesty, the human factor. A person might try to cheat, they might try to be dishonest. The human element of when someone decides to be dishonest and break the rules – and that can, unfortunately, be controlled only by proper supervision and by ensuring that you put systems [in place] to avoid the human elements.” (R5)
- **Theme 4:** Management of research methodology components (R3, R7, R17, R23), e.g.:

“Interpersonal space around how things [are] – how questions are designed and the relevance of questions.” (R3)

The issue raised by (R3) under Theme 2, above, is significant and requires a proactive and thoughtful approach if it is to be managed, but it is outside the scope of this research. The quality control procedures for data checking presented in Chapter 6, Section 6.7.4 are intended to address the areas of concern that researchers identified.

In summary: In areas including timeliness, interpersonal data management, data collection, processing, and analysis, permission, and human factors, the researchers felt that the current model may be deficient. Methodological difficulties, study design components, and inadequacies in data reporting and dissemination could all be issues in this regard. These topics ought to be discussed during the data planning stage, where a researcher outlines how each of these topics will be addressed once the research project gets underway. As research procedures evolve and data curators are given the opportunity to help researchers identify answers to data management issues, it will be imperative that researchers continue to address these constraints.

5.8.9 Areas of concern that the model will NOT address - Curators

It was necessary to know the areas (concerns about data quality) that the curators believed the model would not cover. The issues raised by the curators are examined in this section.

Q16. Taking your research work into consideration, which data quality concerns do you think the model will NOT address?

All the related areas of concern identified by the curators and proposed below are grouped in the following theme:

- **Theme 1:** Accuracy of the data (C3, C4), e.g.:

“Accuracy. I’m curious in terms of processing or collecting data.” (C4)

Two curators (C1 and C2) didn’t have any data quality concerns to mention.

The preceding comments demonstrate how seriously at least two of the curators considered the data collection and processing phase of research to be. These curators were of the opinion that the majority of data errors occur during these two stages (data collection and processing). The recommendations provided in Chapter 6, Sections 6.7.4 and 6.7.5 to avoid problems that can arise during data collection and processing were developed to address these concerns. The following section examines any further discussion topics the researchers raised.

5.8.10 Anything else regarding the quality of research data - Researchers

In this section, the researchers expressed their RDM concerns and raised other areas that they felt deserved to be addressed. The comments that researchers provided concerned not only data quality but other institutional and unit-level matters as well. The following comments represent a selection of the lecturers' reflections on the model:

- “The model allows for breaking research and data management areas up into small parts.” (R20)
- “A more general model is needed to help researchers change their project designs.” (R2)
- “It would be best to have two models – one for qualitative research and one for quantitative research.” (R3)
- “Data collectors (field workers) do not have the skills and expertise [needed to] approach respondents and, ultimately, collect relevant data.” (R17)
- “Advice should be provided to researchers on anonymising data and dealing with confidentiality processes.” (R22)
- “Some processes are in place to ensure that the data that we gather can meet international scrutiny.” (R21)

The above responses represent a mixture of viewpoints and are based on data collection training, anonymisation issues, research methodology areas, and compliance with international and local standards.

When the model was updated (refer to Section 6.7.10), these suggestions were considered and, where feasible, the suggestions were incorporated. The curators too were requested to offer final comments about the quality of the research data.

5.8.11 Anything else, regarding the quality of research data - Curators

Unlike the researchers, the curators didn't have much to add. One curator (C1) raised the matter of data collection moving from paper-based to online processes in the future:

“We tend to move to the space of doing everything online. I think, probably in two to three years, that this thing of collecting the data physically [will have] come to an end.” (C1)

As RDM processes advance, data collection and processing via online and electronic systems and platforms are starting to become the standard. Curators and researchers need to be more skilled and collaborate from the beginning when handling electronic data. The next section serves as a summary of this chapter.

5.9 Chapter summary

According to the demographic and educational information provided in Section A, the majority of the research participants interviewed – 20 (87.5%) of the researchers and one (25%) of the curators – had extensive expertise. Considering the useful feedback that was received, it can be seen that the more field experience a person has, the more knowledgeable he or she is likely to be in matters related to data quality and the less complex it is likely to be for him or her to understand why data quality is important to the RDM function. Illustrating this, the respondents were able to grasp unknown concepts and were also willing to consider potential suggestions that were shared with them.

The research participants gave responses in the *Measuring Data quality* component of the interview schedule (Section B) that were based on their actual experience. They were able to identify measures of data quality for all 16 of the quality dimensions considered. Nonetheless, these measures will need considerably more investigation than was possible during this research.

The Quality Process Model section (Section C) offered research participants the chance to interact with the suggested model. It is clear from the data findings and analysis in this chapter that the model can help researchers plan their data management more accurately, even though opinions on the data quality process model varied. There was consensus that the general quality of the data, as well as the interpretation and the quality of the decisions based on the data, can all be improved when researchers adhere to certain predefined stages of quality control. The stages mentioned were data preparation, data cleaning, and data validation. The conclusions and suggestions drawn from the results of the study are discussed in the next chapter. The model was also updated to incorporate the feedback and suggestions provided by participants. The updated model is provided in Figure 6.9 in Section 6.7.10.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

It was possible to generate recommendations for the social sciences organisation in light of the discussion of the data findings in the previous chapter. These recommendations should influence the way that curators and researchers see and manage the overlap between RDM and data quality. The chapter first responds to the main research question and its sub-research questions and, thereafter, presents conclusions and discusses recommendations.

6.1 Introduction

In light of the literature review chapter (Chapter 2) and the research findings, this chapter aims to reflect on the study's objectives and research questions (described in Chapter 1, Sections 1.1 – 1.4). While many of the problems related to the topic have been resolved by this study, there is still more work to be done, as detailed in Section 6.11. The conclusions and recommendations are based on the data that were obtained, analysed, and compared to what was established from the literature and are also covered in this chapter. A summary of the research study concludes the chapter.

6.2 Overview of the study

The objective of this study, as stated in Chapter 1 (Section 1.3), was to provide an effective, quality-focused framework to direct the management of quality-assured data across the stages of the research data lifecycle. The main research question of the study is: “What would an effective, quality-focused framework, to guide the data management activities of a Social Sciences Research Organisation, look like?” Six sub-research questions have been used, as described in Chapter 1 (Section 1.4), to help answer the main research topic. The following sub-questions were used to realise this goal:

1. What is meant by the term ‘data quality’?
2. How does the network of stakeholders, involved in research data management, address data quality issues?
3. What will be the characteristics of a framework for quality-focused data management?
4. Considering the data quality dimensions, how is data quality measured?
5. What are the unique remedies that the proposed quality-focused framework could incorporate?
6. What recommendations regarding change management would be applicable to the changing of RDM practices at South African Social Sciences Research Organisations so as to include data quality?

The next section discusses the findings of the sub-research questions first to provide a firm foundation before the main research question is answered.

6.3 Defining the term ‘data quality’

The study reviewed various definitions for data quality, revealing little consensus on data quality or its place in the research data lifecycle. The concept is based on the ability to produce, gather, store, preserve, distribute, process, and present data in a timely and cost-effective manner to support business activities (Azimae, 2015; Black and van Nederpelt, 2020; Cai and Zhu, 2015; Charles et al., 2017). Various perspectives and definitions from international RDM stakeholders, RDM programmes, major research funders, evaluation systems, and research methodology books (as reported in Chapter 2, Sections 2.4, 2.5, 2.6, 2.7, and 2.8) are influencing the concept. It is expected that RDM quality assurance measures should be performed throughout the research data lifecycle (Torana Inc, 2023).

The following sections look at the concepts found in literature and what was learned empirically:

As was reported in Chapter 5, Section 5.7.3, research participants associate the phrase "data quality" with how accurate and clean the data is. Considering what was stated in the chapter on the literature review, data quality is related to other terms, such as quality control, data quality measures and data quality rules that are discussed in Chapter 2, Section 2.3.1. The definitions derived from the literature have the drawback that they define "data quality" from a more general organisational and management perspective in terms of:

- Organisational methods and operations used to achieve the required quality for a data product.
- Organisational structure, roles, policies, processes, and tools for implementing data quality management.
- The management role that defines and implements the data quality policy.
- The research participants associated data quality with the practical ways that are used to manage data, not necessarily on a policy level, in contrast to the definitions provided above (from the literature), which do not match what was discovered through empirical research – in other words, where researchers utilise their own established and personal data management techniques.

The statements that follow are examples of how research participants perceive and characterise data quality:

- “Research teams go back and verify that everything was done appropriately. Additionally, statisticians and a data manager will do quality checks on the received data, identifying any flaws in the data set so that the other programmes will be aware of the issues that need to be fixed.” (R11)
- “Prior to data collection, one has to know and set the targets for the quality of the data that you would want to collect.” (R15)

- “There is a data checklist for quantitative and qualitative data available to researchers [and] they can use that checklist to assure that the data quality is good before they deposit the data for curation.” (C2)
- ‘The fundamental elements (outliers and frequencies) are what to look out for. When data enters the curation centre, fundamental processes of validation are performed to determine the data's quality.’ (C3)

The empirical research in this study showed that when working with research data, researchers reflect on a variety of quality dimensions and metrics (Chapter 5, Sections 5.7.1 and 5.7.2). Researchers are determined to give secondary users data with high-quality prospects because they are aware of data quality. The definition of data quality is strengthened by additional concepts such as quality control, data quality measures, and data quality rules (addressed in Chapter 2, Section 2.3.1). In research projects and institutional settings, stakeholders define data quality, so it is crucial to review and consider the definitions offered in the literature because these stakeholders are the main stewards and contributors to the definition of data quality. Organisations that define RDM policies and practices are referred to as custodians in the literature review (see Sections 2.4, 2.5, 2.6, and 2.7), which list the stakeholders engaged in determining data quality. Selected international role-players offer a variety of alternatives for implementing RDM policies and practices, while international role-players in general offer resources to solve data quality issues.

The many research units in the organisation for social sciences are composed of research teams under the direction of the Principal Investigator (PI). Research participants believe that the Principle Investigator determines the quality of the data by verifying that the data manager handled and processed the data correctly (Chapter 5, Sections 5.6.4, and 5.7.3). Experiences with data management and dealing with data management documents have shaped this perspective. Research teams employ standardised templates as part of the data collection process to guide the actions to take. These templates are viewed by researchers in large part as indicators of the quality of the data. The usage and incorporation of a DMP template were reported in Chapter 4 (Section 4.12).

Conclusion: Taking into consideration all the reviewed literature, the empirical findings, the term ‘data quality’ could be defined as: *‘fitness for use’* as determined by a variety of quality metrics in the form of best practices, recommendations, and standards that are connected to the technique used to assess or enhance data quality.

Recommendation: Using standardised definitions and terminology is advisable across the organisation and in all related documentation. A standard quality assurance strategy must be developed by the organisation for social sciences to guide quality assurance and quality control (described in Chapter 2, Section 2.9.1) throughout a research project. This will make it easier to produce research results that

are well-managed and of high quality. The terms "data quality" and "quality assurance" should be used in accordance with accepted standardised terminology and guidelines.

It is also recommended that a deviation mechanism be developed, which would be applied when a researcher cannot adhere to the quality assurance requirements. How stakeholders address issues with data quality is covered in the following section.

6.4 Stakeholders addressing data quality issues

This section aims to demonstrate how different formal RDM initiatives at major international RDM players, large research funders, evaluation systems, a South African RDM initiative, and South African academic and governmental institutions recommended that data quality issues be addressed.

6.4.1 Large international RDM role-players

As noted in Chapter 2 (see Section 2.4), a number of international role-players have developed approaches, protocols, strategies, and practices to deal with problems associated with data quality. The social sciences institution's data curation unit has made an effort to include some of these methods as part of its quality control procedures (see Section 5.7.4). The RDM and quality control strategies that international role-players are using and which have been identified as recommendations in this study include:

- The application of FAIR data principles to advance data correctness, consistency, and interoperability (the FAIR data principles are mentioned in Section 2.2.3.2 as well as by various stakeholders in their RDM initiatives – see Appendix 15 part C);
- Employing automated techniques to identify and report some of the most prevalent problems in survey or numerical data, such as missing values, repetition, outliers, and direct identifiers (see Appendix 1 – ICPSR Data Life Cycle);
- Advancing a cyber infrastructure that supports the data lifecycle through a multidisciplinary, multi-institutional, and collaborative relationship (see Appendix 15 part B);
- The establishment of various working groups and communities of practice that are focused on tackling concerns related to data quality and quality assurance (see Section 6.4.4);
- Providing researchers with data management training that involves teaching researchers about the importance of data quality, quality assurance and conducting rigorous and reliable research (see Section 2.2.3.6).

Conclusion: Large international RDM role-players are actively tackling data quality challenges through a variety of strategies and activities, as noted in Chapter 2, Section 2.4. As research procedures move toward being digital, these processes must become integrated with best practices in order to retain the validity and reliability of research findings. Researchers and organisations are better able to produce

data that is of good quality when they are knowledgeable and skilled in using RDM best practice strategies.

Recommendation: Individualised data management training for researchers is necessary, with a focus on quality control methods to use when collecting and processing data; including digital data management training. The social sciences organisation should also provide specialised training to researchers and curators on all of the quality control and RDM methods it wants to deploy. The social sciences organisation should also take part in RDM developments, or at the very least, monitor them.

6.4.2 Large research funders

According to the discussion of key research funders in Section 2.6, funders have begun adopting requirements and programmes that demand applicants to produce a Data Management Plan that specifies data management obligations and at what level of detail data quality will be checked. This includes describing the applicants' RDM experience and skills. At the social sciences organisation, researchers indicated that as part of receiving ethical clearance, a DMP needs to be submitted. A key activity that large research funders do is to require applicants to outline all the [quality] measures that will be taken while carrying out their study (referred to in Chapter 2 - see Appendix 15, part D). As a suggestion, researchers should prioritise submitting funding requests from funders who need applicants to offer a basic explanation of how the quality of the data will be assured; it is recommended that this should be the first option for researchers. Another strategy that funders are employing is a merit system (referred to in Chapter 2 - see Appendix 15, part D) that mandates applicants to carefully craft their proposals in order to ascertain how quality assurance may be ensured throughout the study process. To be eligible for financing, applicants must fulfil specified data management requirements. Another tactic employed by research funders is requiring applicants to adhere to the FAIR data principles as a condition of receiving funding (Chapter 2, Section 2.6).

Conclusion: As discussed in Chapter 2 (Section 2.6 - see Appendix 15, part D), research funders require grant applicants to include a plan (e.g. Project Plan, Conducting Data Quality Assessment, Data Management Plan) as part of their grant proposals. These plans outline how researchers will collect, document, store, and manage their data, including data quality control measures. Research funders are increasingly recognising the importance of data quality in research projects and are putting various measures in place to address data quality issues. Research funders play a critical role in promoting data quality in research by having requirements, providing resources, and enforcing policies and standards that encourage researchers to maintain and enhance the quality of their data.

Recommendation: Researchers should be compelled to provide a thorough Data Management Plan (DMP) as part of the data curation deposit procedure, just as they did for their ethics clearance applications. The DMP should include a declaration from the researcher outlining the precise procedures they will follow to produce high-quality data. The data curation unit should actively take

part in the review of funding proposals before they are submitted by researchers, working with the research ethics office. Additionally, the data curation unit should monitor the RDM recommendations made by funders and assist researchers in deciding which sponsors to approach for financing. This will ensure that all aspects of quality assurance have been addressed by the researchers.

6.4.3 Evaluation systems

In Chapter 2 (Section 2.7) the discussion on evaluation systems demonstrated how organisations (i.e. role-players) have started and are working on projects with quality standards that users (researchers) can take advantage of while managing their research data. Evaluation categories, a repository certification procedure, and risk assessment audit techniques are some examples of the quality requirements.

Conclusion: In order to address problems with data quality in research, evaluation systems are essential. These systems are designed to make sure that data utilised for analysis, decision-making, and reporting are accurate and up to par. Overall, evaluation systems play a crucial role in maintaining data quality standards, advancing the legitimacy of research, and guaranteeing the accuracy and validity of research results. These systems offer the checks and balances required to successfully resolve problems with data quality. The evaluation systems that were discussed in this research study (Chapter 2, Section 2.7) offer a balanced coverage of data management areas, ranging from trustworthy repositories and auditing repositories to organisational processes and managing a master data file. These topics were brought up even though they have no direct bearing on data quality; yet, they are significant since they are a part of the broader picture of data management and should not be disregarded.

Recommendation: It is advised that the data curation unit evaluate the institution repository policy with the help of the IT department and ensure that RDM and quality assurance audit practices are addressed in the policy. Researchers should review their RDM process and quality assurance audit techniques that evaluate data quality during the course of a research project. The evaluation systems examined in this study concentrate on the evaluation of a data repository, the laws governing data repositories, data formats that repositories can manage, legal requirements for data repositories, or repository operations. It is advised that researchers and the data curation unit become familiar with the specifications of evaluation systems. They will gain a broader perspective as a result, and the organisation will be able to meet specific criteria (such as becoming a trusted repository), thereby creating an environment that is suitable for handling research data.

6.4.4 South African RDM initiatives

According to the literature study in Chapter 2 (see Appendix 15, part E), the various academic and governmental organisations in South Africa show a significant understanding of and dedication to the improvement of RDM. Academic institutions have created and tailored their own systems and processes that have different maturity levels. It should be noted that South African institutions are still in the

planning stages of starting their RDM projects; this in turn affects the coverage of data quality and quality assurance because the institutions have no policies that inform and promote RDM. The RDM-related documents and processes that were reviewed show that the social sciences institution is moving in the correct direction with regard to RDM and quality assurance.

Conclusion: South African (RDM) initiatives are actively trying to address data quality issues to improve the reliability and usability of research data. Organisations in South Africa are recognising the importance of data quality in research and are implementing various initiatives (as discussed in Chapter 2 – see Appendix 15 part E). These initiatives, however, are not of a general kind that speaks to data quality on a broader level. They are, instead, in line with organisational missions. By implementing these initiatives, the efforts will assist researchers (in the respective organisations) to produce data of high quality, facilitating impactful research and supporting data-driven decision-making in various fields.

Recommendation: The social sciences institution should conduct an RDM policy workshop, reviewing all institutional policies in line with business goals, mission, and vision. This should include internationally recognised best practice standards. Collaborating with academic institutions can create a community of practice groups and forums, allowing all the institutions to benefit from the knowledge of other institutions and providing helpful RDM growth among all.

6.4.5 South African academic and government institutions

According to the examined literature, the majority of RDM programmes offered by South African universities are primarily concerned with integrating RDM and ICT projects. The responsibility for quality assurance is associated with researchers and research teams. Even though this strategy can be useful in enabling different organisations to collaborate on research initiatives. South African organisations are also focusing on the issue of preserving and making data accessible. The institution of social sciences is advised to broaden the application of RDM and incorporate quality assurance modules into its data management programmes. Nonetheless, this will need organisational adoption and ongoing RDM training and awareness campaigns will help to raise everyone's understanding of the value of RDM, data quality, and quality assurance.

According to the discussion of South African academic and governmental institutions in Chapter 2 (see Appendix 15, part E), institutions have different perspectives on RDM and its significance. Government agencies are offering definitions and attempting to create an environment that promotes data accessibility and interoperability. According to the literature review, government departments are concentrating more on managing archiving facilities and performing a monitoring function. Although these actions are a part of managing data in accordance with the research data lifecycle, they have no direct bearing on quality control and the generation of high-quality data.

Government departments have placed a high priority on developing methods that will guarantee the exchange and sharing of research data and documentation among research groups on a regional, national, and international level. In South Africa, government departments have implemented policies to promote the generation of new information in a variety of research sectors; however, these initiatives do not address problems with data quality and quality assurance. Sections 6.4.5 and 6.6.5 covered the significance of establishing a policy that directs and addresses the RDM practices of an institution.

Conclusions: South African academic institutions are actively trying to address data quality issues to ensure the reliability and credibility of research outcomes, policy decisions, and public services. The disparity in institutions' RDM maturity levels, however, has led to an imbalance in the country's overall RDM environment. Government departments should extend their data management reach and collaborate more with organisations that have already begun to advance their RDM initiatives. Through a change in perspective, governmental departments will be better able to manage data quality as well as accessibility and utilisation. Approaching the institutions that were covered in Chapter 2, Section 2.10 (see Appendix 15 part E) and beginning RDM partnership activities is an ideal place for government departments to start. Organisations such as DIRISA, DataFirst, UCT, and the University of Pretoria may be able to offer insightful guidance on RDM and how to get started with it.

Recommendations: Government departments should request the assistance of academic institutions that have already created RDM processes and data management policies and the social sciences organisation could facilitate this collaboration. This support should not be limited to academic institutions only; government agencies should compare their data management procedures to those of major international actors, such as those mentioned in Chapter 2, Section 2.4 (see Appendix 15). Workshops, symposiums, and conferences may be used for this kind of collaboration, following which, government departments will be responsible for putting RDM policies into effect and making improvements to RDM procedures.

6.5 The characteristics of a framework for data quality-focussed RDM

The first version of the framework was created using a framework development strategy (as discussed in Chapter 3, Section 3.5) as a quality process model (see Chapter 3, Section 3.8). This model was presented to the research participants, who were then asked what they thought of the model. Researchers and data curators had different opinions on aspects of the model. Some claimed that the model is difficult and needs to be simplified, while others were of the opinion that the framework is simple to understand and utilise. The author chose to modify the model in response to the feedback that was received from the research participants. In doing so the author developed a simple and easy to use, yet complex framework to improve data quality. Section 6.7 goes into further detail about the adjustments that were made. For the sake of easy reference, the initial proposed quality process model (that was used during data collection) is repeated here:

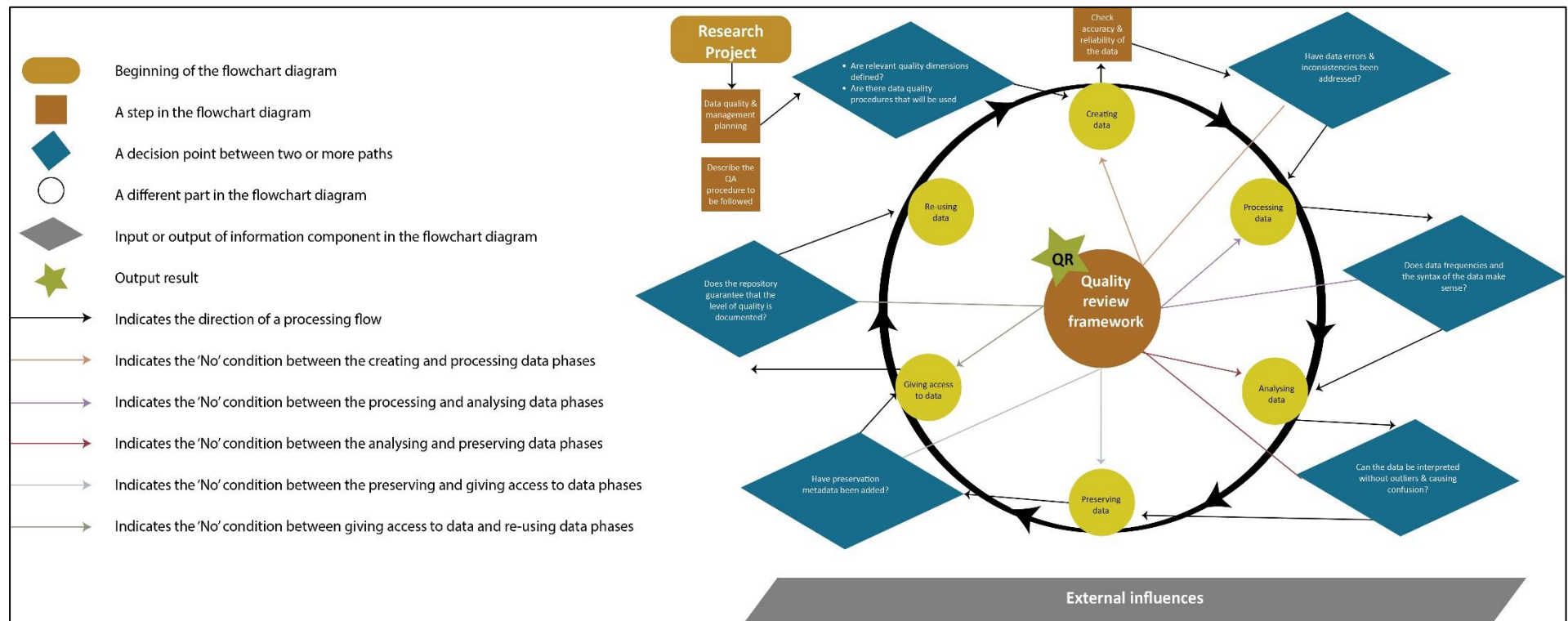


Figure 6.1 – The initial quality process framework

(Key in Appendix 11)

Conclusion: The characteristics of a framework for data quality-focussed RDM are based on the research data lifecycle (adapted from the UK Data Archive), while the stages of the data lifecycle are informed by measuring criteria of the Quality process framework (Appendix 2) that is made up of a mixture of theoretical frameworks, best practices and guidelines, as discussed in Chapter 2, Section 2.3. Considering the extensive literature review that was conducted, the characteristics of a framework for data quality-focussed RDM include:

- A detailed lifecycle, such as that provided by the UK Data Archive research data lifecycle (discussed in Chapter 3, Section 3.7.1)
- Quality assessment criteria, such as those provided by the South African Statistical Quality Assessment Framework (SASQAF) (discussed in Chapter 3, Section 3.7.3).
- Both scientific and technical quality dimension measures – measuring criteria adopted from best practices, models, guidelines and standards (Appendix 4).

When the characteristics of a framework for data quality-focussed RDM are brought together, the framework aims to (as discussed in Chapter 3, Section 3.8.3) examine and guide the process of data management planning in a social sciences research organisation. Utilising criteria tailored for social science research data. The quality dimensions (that are based on the various measuring criteria) are intended to evaluate how researchers address problems that could compromise the integrity of their data.

After receiving feedback (Chapter 5, Section 5.5, 5.6) from research participants about the first version of the quality process model, it was necessary to change the model and make it simpler and more user-friendly. The changes will be discussed in relation to the different stages of the research data lifecycle (in Section 6.7).

Chapter 3 offers a description of the aforementioned model. The majority of participants in the study believed that the model was too complicated and difficult to understand (see Sections 5.7.9 and 5.8.2). The author consequently opted to remove the green blocks representing the decision points between two or more paths. A layered approach was followed and, as a result, each stage of the data lifecycle was represented as a framework layer. Because of the layered approach, each stage of the research data lifecycle has its own best practice standards that must be fully met before moving on to the next stage (or layer). Section 6.7 offers a focused view of each phase. The changes made to the first version of the quality process model include:

- Removing the different arrows that create loops in the model
- Removing the different decision points between two or more paths

The features that were kept from the first version of the quality process model include:

- The data management phases
- All the different stages of the data lifecycle, and
- The external influences component.

The full discussion of the revised quality process model is provided in Chapter 6, Section 6.9. The findings of the data quality dimensions (i.e. sub-question 4) are covered in the next section.

6.6 Measuring data quality using the data quality dimensions

Data dimensions are “features of data that can be measured or assessed against defined standards in order to determine the quality of data” (DAMA UK Working Group, 2013:3). Data quality dimensions are crucial for determining and enhancing data quality. The quality process framework was built on the basis of the quality dimensions that were employed in this research study. The data quality dimensions were not used in isolation but rather were seen on a higher level where they were integrated with other best practices, standards, measuring criteria and guidelines (see Appendices 15 and 16). It is crucial to highlight what the literature asserted and what was found in the empirical research and compare the similarities or differences before explaining the findings (response) to sub-research 4. The ideas offered in literature and what was learned empirically are examined in the following section:

6.6.1 Data quality assurance

Data quality assurance is a notion that is closely related to data quality dimensions. Data quality assurance was defined as: "all of the planned and comprehensive steps required to ensure that a data product meets a specific set of quality specifications"; and "all those planned and systematic actions necessary to provide adequate confidence that a data product will satisfy a given set of quality requirements" in the literature (Chapter 2, Section 2.3) that was reviewed. Putting systems in place to ensure that data satisfies specific quality requirements is mentioned in the literature review. The participants in the empirical research fell short since they did not follow a specific set of standards for quality. Internal quality control methods were used instead. Curators are therefore encouraged (as a remedy) to suggest established quality standards for researchers to follow.

6.6.2 Quality control

In the literature (Chapter 2, Section 2.9.1) quality control is defined as “the set of operational techniques and activities that are used to attain the quality required for a data product”; and data quality is defined as “data quality assurance includes all those planned and systematic actions necessary to provide adequate confidence that a data product that will satisfy a given set of quality requirements; and methods and procedures implemented to ensure that data are collected, managed, and utilised with accuracy and precision. The definition of quality control techniques in the literature review chapter did not include any reference to employing methods that are informed by known best practices, which was a minor

shortcoming. The fundamental definition of quality control in the literature omitted any reference to the benchmarking element. Participants in the study were able to name an international role-player to whom their quality control practices are compared (for example, the lifecycle model of the Digital Curation Centre).

Researchers should begin utilising data quality indicators as a remedy when managing data throughout all phases of the data lifecycle. Section 6.7.3.5 provides specifics on how the data quality indicators function.

6.6.3 Data ‘integrity’ and ‘authenticity’

The accuracy and integrity of the data determine its quality. Understanding the meanings of these two phrases was essential because they provided the foundation for understanding the bigger picture of data quality dimensions. In Chapter 2 (Section 2.3.4) integrity is defined as “maintaining and assuring the accuracy and consistency of the data over the entire data life-cycle”; and authenticity refers to “the digital retention and preservation practice of verifying that a digital object has not changed. Authenticity attempts to identify that an object is currently the same genuine object that it was ‘originally’ and to verify that it has not changed over time unless that change is known and authorised”. There was a limitation in the approach to how research participants defined these two concepts (integrity and authenticity). According to empirical studies, preserving correctness and consistency across the lifecycle of research data is not something that researchers strongly link with data integrity. Integrity is a concept that depends on how effectively researchers take care of their data. By highlighting errors and inconsistencies in the data, these strategies aid in ensuring the accuracy and integrity of data.

As a remedy, the Producing Data phase of the quality process model offers useful best practice steps to apply when data is being collected, enabling researchers to produce data that is authentic and has integrity. These procedures (both qualitative and quantitative) will ensure that any errors that might occur during data collection are prevented and corrected. As a result, the data gathered will be reliable and trustworthy.

6.6.4 Data validation methods

Different data validation methods were mentioned as part of quality control procedures in the literature review chapter (Section 2.8.1). The literature indicated a wide range of efficient data validation methods, including concurrent validity, convergent validity, content validity, and criterion-related validity. Regarding data validation techniques, there was a balance between what was found in the literature and what the empirical research showed. Logical and quality checks are used by researchers and curators as data validation techniques (described in Chapter 5, Section 5.6.5). The validation procedures help assure the accuracy of the data, even though they are not as thorough as those listed in the literature. The quality process model's Processing phase (Section 6.7.5) offers further validation recommendations; these best-practice methods are practical and clear.

Answer to sub-question 4: data quality using the data quality dimensions

It was established (in Chapter 2, Section 2.3) that measuring data quality requires having certain measuring indicators to be in place. These measuring indicators are data quality dimensions and help to measure certain aspects of data. The next section discusses how data quality dimensions assist with measuring data quality.

6.6.5 How each data quality dimension assists with data quality

Each data quality dimension serves a particular purpose by addressing various aspects of data management, such as reliability, precision, and usability, and contributes to assessing and improving data quality overall. The answer to sub-question 4 is provided in the section that follows and discusses how each dimension enhances the quality of data. After the data was gathered, research participants identified additional quality dimensions (Table 5.3), which were combined with the ones already in use (in the definition). The revised definitions of the dimensions are given below:

Accessibility: can assist in data quality by ensuring that the data is readily available and easily accessible to the users who need it; this viewpoint is similar to the description that was used in Chapter 2, Section 2.7.2. By making data readily and easily accessible, organisations can reduce the chances of data being lost or forgotten, and ensure that it is used promptly. This helps to prevent data from becoming outdated or inaccurate. Accessibility can also promote data transparency and accountability. When data is easily accessible, it can be reviewed and verified by multiple stakeholders, increasing the chances of identifying and correcting any errors or inconsistencies. Overall, accessibility plays a crucial role in maintaining data quality by enabling efficient data retrieval and ensuring its reliability and accuracy. Accessibility also refers to "Ease of sharing data" and "Assignment of a DOI" as per the extra dimensions proposed by research participants.

Accuracy: plays a key role in maintaining data quality. When data is accurate, it means that it is free from errors, inconsistencies, or inaccuracies. Accurate data provides a reliable and trustworthy foundation for decision-making and analysis. Accuracy ensures that the data correctly reflects the real-world entities and events it is meant to capture. This is essential for any business or organisation to make informed decisions and take appropriate actions based on reliable information. Accurate data helps maintain data integrity. When data is accurate, it reduces the risk of duplicate records, incorrect calculations, or misleading insights. This, in turn, improves the overall quality and reliability of the data. This justification fits the descriptions presented in Sections 2.8.5 and 2.9.1. Accuracy is vital for data quality as it ensures that the data is reliable, reflective of reality, maintains integrity, improves decision-making, and increases user confidence in the data. According to the additional characteristics provided by research participants, accuracy also relates to "Internal consistency", "Veracity", "Rigour" and "Trustworthiness".

Comparability: is a key factor that can assist in data quality. Comparability refers to the ability to compare data across different periods, geographical locations, or other dimensions; this description aligns with what was reported in Section 2.2.1. In order to ensure data comparability, organisations need to establish consistent definitions, classifications, and measurement methods. This helps in ensuring that the data collected and reported is consistent and can be compared accurately over time, across different regions, or between different datasets.

Completeness: Complete data contains all the necessary information required for analysis and decision-making. By ensuring data completeness, organisations can avoid making decisions based on incomplete or missing information, leading to more informed and accurate insights. Completeness is also an essential aspect of data quality because it allows a researcher to document gaps in the data allowing for effective secondary usage – see Sections 3.5 and 3.8.9. When data is complete, it means that all the necessary and relevant information is present, without any missing or incomplete entries. This helps in ensuring that the data is comprehensive and provides a holistic view. Completeness plays a significant role in data quality as it ensures that all the required information is present and accurate, leading to reliable analysis, decision-making, and stakeholder trust. Completeness includes “Confirmability” and “Analysis” among the additional qualities identified by research participants.

Consistency: Consistency in data helps to eliminate duplication, discrepancies, and conflicts that may arise due to variations in data formats or data entry practices. It also enables efficient data integration and reduces the likelihood of errors during data processing. By maintaining consistency in data, organisations can ensure that their data is reliable, accurate, up-to-date and structured in a standardised manner; this description aligns with what was reported in Section 2.3.3. This, in turn, enhances the accuracy of analysis and decision-making processes – overall, consistency contributes to data quality by promoting accuracy, reliability, and efficiency in data management and usage. Consistency also refers to "Dependability" in accordance with the additional dimensions provided by research participants.

Flexibility of the data: assists in data quality by allowing for easier updates and modifications. With flexible data, organisations can easily adapt to changes in their data needs and requirements. This means that they can make necessary adjustments to their data structures, formats, and definitions as needed, ensuring that the data remains accurate and relevant. Additionally, flexible data allows for easy integration with other data sources, enabling organisations to combine and analyse data from different systems and sources, leading to more comprehensive and reliable insights. Ultimately, the flexibility of data enables organisations to continuously improve and refine their data quality processes, leading to better overall data quality. This perspective of refining data quality processes links with the description used in Section 6.7.3.3. The term Flexibility of the data also refers to "Interoperability", "Transferability" and “Reproducibility”, according to the extra dimensions that research participants offered.

Integrity: Integrity plays a crucial role in ensuring data quality. When data has integrity, it means that it is accurate, reliable, and consistent. This helps organisations trust their data and make informed decisions based on it. By maintaining data integrity, organisations can avoid inaccuracies, inconsistencies, and errors in their data across the data lifecycle (linked with description in Section 2.3.4). Data integrity ensures that data is complete, valid, and free from any corruption or unauthorised changes. This helps in improving the overall quality of data and enhances its usefulness for analysis, reporting, and decision-making purposes. Furthermore, data integrity helps in preventing data duplication, ensuring that each piece of data is unique and free from any redundancies. This eliminates the risk of duplicate records and ensures that data is not unnecessarily duplicated or overwritten, leading to improved data accuracy and reliability. Integrity also includes "Reliability" and "Credibility" according to the other aspects offered by research participants.

Interpretability: Interpretability can assist in data quality by providing a clear understanding of how data is obtained, processed, and transformed. It allows data analysts and stakeholders to interpret and explain the meaning and context of the data, ensuring transparency and reliability. With interpretability, organisations can identify potential issues or biases in the data collection process, such as data incompleteness or errors. This understanding helps in detecting and rectifying data quality issues, ensuring that the data used for analysis or decision-making is accurate and reliable. Additionally, interpretability enables better communication and collaboration among different teams and stakeholders involved in data analysis. By explaining the data (in related documents and metadata) in a clear and understandable way, it becomes easier for team members to validate and verify the data, leading to improved data quality; the usage of metadata to explain data links is described in Section 2.9.4. Overall, interpretability plays a crucial role in enhancing data quality by promoting transparency, accountability, and effective communication in data-related processes.

Methodological Soundness: may significantly help to ensure the quality of the data by offering a methodical approach to data collection and analysis. In producing data that is of good quality, Methodological Soundness can assist with:

- Clear research design: A well-designed research plan helps define the objectives, target population, and data collection methods. This ensures that the right data is collected from the right sources.
- Standardised data collection: Methodologically sound approaches employ standardised tools and protocols for data collection, reducing the chances of errors or bias. This enhances the reliability and validity of the data.
- Quality control measures: Sound methodologies include the application of internal quality control processes/guidelines to identify and rectify errors in data collection or measurement; the usage of international guidelines aligns with what was reported in Section 2.12.

Overall, methodological soundness improves data quality by minimizing errors, reducing bias, and enhancing the overall credibility and reliability of the data. The term Methodological Soundness also refers to "Ethics" and "Human rights", in accordance with the extra dimensions that the research participants suggested.

Quality Check Processes: play a crucial role in ensuring data quality by identifying and resolving any errors or inconsistencies in the data. These processes involve various steps such as data cleaning, validation, verification, and reconciliation. By implementing these quality check processes, organisations can identify and rectify any issues or errors in the data, leading to enhanced data quality. This, in turn, improves the reliability and validity of the data analysis and interpretation, ultimately supporting better decision-making and problem-solving. As discussed in Section 2.12, in the course of managing research data, a research institution must have a quality management system in place.

Relevance: The degree to which the data is acceptable and valuable for a specific analysis or decision-making process is referred to as relevance in data quality. Relevant data is that which directly addresses the study topic or issue at hand. This is crucial because unnecessary information might skew the analysis and add noise and confusion. By ensuring that the data used in an analysis is relevant, organisations can improve data quality by:

- **Filtering out unnecessary or unrelated data:** By removing data that is not relevant to the analysis, organisations can reduce the amount of noise and increase the signal-to-noise ratio, leading to more accurate and reliable results.
- **Increasing the accuracy of analysis and interpretation:** Relevant data provides a clearer picture of the phenomenon being studied, allowing for more accurate analysis and interpretation of the results. This leads to more reliable insights and better decision-making.

Relevance plays a crucial role in improving data quality by ensuring that the data used in the analysis is directly applicable and useful for the specific purpose at hand.

Standards and Policies: Standards and policies play a significant role in ensuring data quality. They provide guidelines and rules for collecting, storing, and using data, which helps establish consistency and accuracy in how data is managed. Standards set specific criteria for data collection and storage. By following these standards, organisations can ensure that data is collected in a consistent and standardised manner, reducing the chances of errors or inconsistencies. For example, standards may dictate the format in which data should be recorded, ensuring uniformity across different sources. Standards and policies can also help in data integration and interoperability. When different systems or databases need to share or exchange data, having common standards ensures that data can be seamlessly integrated without loss of information or discrepancies. This promotes better data quality and consistency across systems. Standards and policies assist in data quality by promoting consistency, accuracy, and adherence to best practices in data management. They provide a framework for maintaining data integrity, security, and privacy, ultimately improving decision-making and enabling

organisations to leverage their data effectively. "Revision Policy" is one of the new characteristics suggested by research participants, and it is included in Standards and Policies.

Timeliness: Timely data ensures that information is up-to-date and relevant, enabling decision-makers to base their decisions on the most recent information available. Outdated data can lead to inaccurate analysis and ineffective decision-making. Overall, timeliness plays a critical role in maintaining data quality by ensuring that information is current, reliable, and useful for decision-making. It allows organisations to stay agile, responsive, and competitive in today's fast-paced business environment.

Uniqueness: Uniqueness is a crucial factor in ensuring data quality. When data is unique, it means that each record or piece of information is distinct and not duplicated. This helps prevent redundancy and inconsistency in the dataset. Uniqueness also plays a significant role in data integration and data deduplication efforts. When merging datasets from different sources, having unique identifiers allows for seamless merging and avoids creating duplicate entries. Uniqueness in data promotes efficiency, accuracy, and integrity in various data-related activities, leading to improved data quality and better insights.

Usability of the data: The usability of data can greatly assist in improving data quality. When data is easily accessible, organised, and presented in a user-friendly manner, it becomes easier for individuals to understand and work with the data; this viewpoint is aligned with the description presented in Appendix 14. By providing clear and intuitive interfaces, data usability allows users to quickly navigate through datasets, locate the information they need, and make informed decisions based on the data. This reduces the likelihood of errors or misunderstandings that can lead to poor data quality. If the data is presented visually or in a structured format, it becomes easier to spot any discrepancies, outliers, or missing values. This allows data analysts or data quality professionals to address these issues and improve the accuracy and completeness of the data. Data usability improves data quality by facilitating easy access and understanding and enabling users to identify and address any inaccuracies or inconsistencies more effectively. Usability of the data also refers to "Machine readability", "Format compliance" and "Validation", according to the extra characteristics given by research participants.

Validity: is a crucial aspect when it comes to data quality. It ensures that the data collected accurately represents the phenomenon or concept being studied. By focusing on validity, researchers can enhance the accuracy and reliability of their data. Validity assesses whether the data collected can accurately predict or correlate with other variables or outcomes; this is in alignment with the definition in Section 4.13.

Conclusions: The data quality characteristics that were chosen for this research study can aid in data quality by providing a framework to assess and improve the overall quality of the data. Researchers can improve data quality by finding and correcting errors in data collection and processing by using the data quality dimensions and their measuring criteria (Appendix 4). The data quality dimensions are crucial

metrics that make sure researchers can provide useful findings that can be utilised to address the current research questions and are unambiguous, and the results may be employed in multiple scientific contexts.

Recommendation: Given the extensive nature of the data quality dimensions, it is advised that the data curation unit include a section on data quality dimensions (as stated above) in their Standard Operating Procedure document. Because of this inclusion, researchers will need to make plans and keep records of how they will guarantee that relevant dimensions are chosen at the start of the project and that they plan how each quality dimension will be met. The next section discusses the remedies proposed for the revised quality-focused process model.

6.7 Remedies for the proposed quality-focused process model

The advice given by research participants was rather general in scope. It might be argued that researchers have varying levels of knowledge regarding data management and quality control. Nonetheless, research participants were able to provide valuable feedback that was used as remedies for the quality process model, which is the answer to the main research question.

Research participants understand that data must be efficiently managed for it to have real value. It is evident from the data gathered that "data quality" is an important concern for data curators. The various RDM documents they employ to manage research data, including the monitoring checklist, qualitative and quantitative data appraisal checklists, data deposit form, user guide, Readme file, and a standard operating procedure document, are a testament to this. It was found (refer to Chapter 5) that most data quality problems start while data is being collected and processed.

The discussion in this section responds to the main research question, i.e. *"What would an effective, quality-focussed, framework, to guide the data management activities of a Social Sciences Research Organisation, look like?"* The high-level characteristics of a model for data quality (covered in Chapter 3, Section 3.8) will be presented in order to lay the foundation of how the main research question was answered, which is followed by a description of the final quality process model. This contrast is necessary to show how literature and empirical research come together to answer the main research question. The next section discusses the high-level characteristics of a model for data quality.

6.7.1 The characteristics of a model for data quality

The high-level characteristics of a model for data quality, as extracted from the literature, appear below:

- Frameworks: These explain the essential subjects of the study, such as the important variables, constructs, or components, and the assumed connections between them, either visually or narratively. A discussion on Frameworks is provided in Chapter 3 (Section 3.2.1).

- Identifying assumptions: identifying assumptions enables the evaluation of the validity and accuracy of the framework. By declaring assumptions up front, it becomes possible to examine whether they correspond to reality, collect evidence to support them and adjust the framework as necessary. The assumptions for this study are discussed in Chapter 3 (Section 3.2.2).
- Frameworks typology: a Theoretical framework, a Conceptual framework and a Process model as a framework are the three framework types that were considered in this study. Each type has its unique characteristics which are discussed in Chapter 3 (Section 3.3). The type of framework selected for this research study is the Process model as a framework.
- Principles for the development of a framework: there are different principles that can be used to develop a framework. The principles of the 'Conceptual Framework for Data-Driven Decision-Making Skills' (Mandinach et al., 2006) were used to develop the quality process model. These principles are discussed in Chapter 3 (Section 3.4).
- Framework development strategy: The process of creating a framework must be organised and systematic. The quality process model was developed using the University of Kansas (2023) framework development strategy. This method of developing a framework is covered in Chapter 3 (Section 3.5).

6.7.2 Features of an effective, quality-focused framework

This section answers the main research question – What are the features of an effective, quality-focused framework to guide the data creation process of a Social Sciences Research Organisation?

The main research question was addressed in the context of the research data lifecycle applicable to the social sciences organisation's data quality assurance, and RDM. The author defined data quality as "fitness for use", as determined by a variety of quality metrics in the form of best practices, recommendations, and standards that are connected to the technique used to assess or enhance data quality. Since the quality process framework (as discussed in Section 6.7.2) does not alter pre-existing models, the initial iteration of the quality process model required revisions. Thus, the suggestions provided in the preceding sub-question (i.e. change management) represent a justifiable advancement in the application of the quality process model's final version. The study's assessment of the literature demonstrated that the idea of "data quality" has several sides and a variety of characteristics, all of which contribute to the definition of the term. As was also suggested by the literature review in Section 2.3, data quality incorporates principles that convert data into a commodity with clients who give a price and value to it. Different universities, RDM initiatives, evaluation systems, and large research funders have been proven to approach RDM in different ways. RDM was further demonstrated to include quality dimensions that are primarily focused on evaluating and enhancing data quality. The study also demonstrated the various frameworks and a framework development process used in Section 3.2 to create the quality process and evaluation frameworks (Sections 3.8.4 and 3.8.5). The

primary measurement basis (i.e. quality dimensions) of the quality review framework was built on the data quality dimensions adopted from the SASQAF (2010) (Section 3.7.3) and DAMA UK (2017) (Section 2.3.3.1). The quality measuring framework (Section 3.8.5) includes many measuring standards that were taken from the literature review. The process is represented by the quality process model when these requirements are applied and fully complied with.

The quality process model has to be modified, which proves that the hypothetical component of the previous research sub-questions is real. As a result, change management is a legitimate concern. The development of the quality process model in Section 3.8 of the study demonstrated the multi-tier structure that the scope and strategies of RDM may adopt. A similar process model was also demonstrated to be made up of the various research data lifecycle stages, which were streamlined and grouped into various layers that interact with one another (Section 6.7). The discussion in Sections 2.11 and 6.10 made it clear that senior management must model and promote the change they want to see in an institution in order for it to happen. Research teams must fully comprehend the advantages of implementing new RDM techniques so they can generate high-quality data. Taking the various recommendations that were given in this study, a conclusion can be drawn that extra training is needed at the social sciences organisation. It is thus recommended that a training unit be established at the organisation. The training school will provide on-going training (in all stages of the research data lifecycle) to researchers and curators on all matters related to RDM, quality assurance and data quality.

The responses to the sub-questions revealed that a variety of RDM stakeholders' decisions influence a number of factors that have an effect on data quality. Data quality dimensions and measuring criteria provide the necessary context and structure for the effective management of research data throughout its many lifecycle stages when integrated (i.e. the quality process model). In order to reduce the complexity of the first iteration of the quality process framework (discussed in Chapter 3, Section 3.8.8), layers were employed to manage the tasks associated with each phase of the research data lifecycle.

In summary: The final quality process framework works on a simple and more logical structure:

- A researcher must outline the necessary data management planning strategies before beginning a research project (as discussed in Section 6.7.3).
- Data will be managed using the quality process framework and the suggestions (described in Section 6.7) appropriate for each phase after the Data Management Planning phase is completed.
- Each phase is represented as a layer that needs to be approached and handled as a standalone entity before moving on to the following phase.
- Every associated activity (in all the phases) is assigned someone who is responsible for carrying out that activity. The progress of all activities is monitored and an Interim report is produced.

- A researcher must respond to and address each criterion in accordance with the quality process framework and recommendations offered for each phase in Section 6.7 before moving on to the next phase of the data lifecycle.

The adoption of the quality process framework will not occur spontaneously, so it is advised that researchers and research teams from the social sciences organisation be given the opportunity to attend seminars that outline and explain the quality process model in an effort to contextualise it.

In Chapter 3, Section 3.8.8 (Figure 3.8), the first version of the quality process model was developed while taking all of the characteristics discussed in Chapter 3, Sections 3.2 – 3.6 into account.

Participants in the research provided comments on how to make the model more user-friendly and straightforward because the initial version was used to gather data for the study. It was feasible to create the final version of the quality process model after conducting empirical research, putting the research participants' proposals into action, and analysing the results). The following section discusses the final quality process model created for this research project (complete with descriptions and remedies).

6.7.3 Remedies for the revised quality-focused process model - Planning

This section discusses the necessary documentation (a DMP) that researchers must complete before beginning a research project and revising it as the research project progresses. This is an important step since it will enable researchers to oversee, manage, and carry out activities that have an impact on the quality of the data. The objective of how to practically manage data and get rid of any potential issues that can lead to errors being made is discussed in the next section.

6.7.3.1 Data management planning

As soon as a researcher decides to embark on a research project, comprehensive data management planning must be completed. A researcher must describe how they will handle data to ensure that it is of good quality. This includes outlining their data management strategy, how their data quality evaluation and review will be carried out, and how they will define the data quality indicators that apply to the research data they will use. To make the quality process model understandable in the setting of the social sciences institution, modifying the model was a necessary step. A general modification was made to every stage of the quality process model which will be explained in detail.

A researcher is expected to define the quality assurance processes they will use after outlining how data quality and management planning will be handled. The name and cataloguing rules that will be used for data files, supporting documents and all other research materials must be disclosed by the researcher. An explanation of the metadata schema they intend to utilise is also needed as part of the data management planning process. Data planning's final criteria is for a description of the strategies they will employ to encourage the development of high-quality data.

6.7.3.2 Data quality and management planning

Researchers must properly address a crucial step called data management planning before starting a research project. The techniques that a researcher specifies in a DMP are ultimately what determine how accurate, consistent, and reliable a data file is. As stated in Chapter 2 (Section 2.9), a research data management plan is an important document in laying out the initial course a piece of research will take to address data management issues that may adversely affect data quality. Data management planning is not a straightforward administrative operation that can be completed by copying pre-written text into a template without thinking about how the data management procedures will be implemented. Researcher consideration of the procedures and methods for data handling does not begin until the ethical application form is submitted (during a study project). The study ethics statement should provide a brief description of the procedures to be followed, including how the data will be maintained, formatted, and distributed. It was established (refer to Section 5.8.4) that submitting a DMP is a requirement for getting ethics approval, therefore, it is recommended that there should be a quarterly refresher training programme about how to effectively create a DMP and what a DMP should cover (such areas are discussed in Chapter 4, Section 4.12) before starting a research project. In addition to the basic fields of a conventional DMP (like the DCC DMP), researchers should also include the following topics in their DMPs:

6.7.3.3 Planning how data will be managed

Planning how to store, enter, update, and change data before it is collected will help a researcher save time and will have a direct impact on how a secondary user can use the data. The scope of a DMP should not be limited to topics having an operational focus and the technical aspects of managing data should be included in a DMP as well. Dealing with technical problems enables a researcher to control all potential sources of data inaccuracies. The recommendations that were made in Section 5.6 (i.e. under the Uniqueness, Validity, and Flexibility of the data) should be implemented to assist researchers in having a working and effective plan for managing data across the different stages of the research data lifecycle. Additionally, it was suggested (in Chapter 6, Section 6.14) that greater interactions between data curators and researchers be encouraged in order for curators to advise researchers on trends and best practices in an effort to sustain the integrity of data over time. This suggestion is also made in an effort to give researchers greater RDM help so that they can be guided in finishing a DMP.

6.7.3.4 Data quality assessment and review (as a component of the DMP)

A research project should be managed properly from the beginning to the end. It was found that some data management tasks, such as data cleaning, data storage, data validation, conversion, and passive analysis, are among the quality control techniques that data curators use to manage a data file (as discussed in Chapter, Section 5.7.4). When using data quality assessment and review, research team members must do frequent data assessments during a research project to identify mistakes before

project completion. Unnecessary and premature errors can be prevented by including a section in a DMP for the assessment and review of data quality. The recommendation that was made in Section 5.6 (under the Interpretability quality dimension) of the data curation unit adopting, using and sharing data quality guidelines with researchers in the various research units in the organisation, will assist researchers in selecting and using the best strategies when making data assessments in the course of a research project.

6.7.3.5 Using data quality indicators by ensuring the quality of data at all stages

Researchers will be able to set up a method that will continuously evaluate the degree of accuracy when data is collected and processed because of quality indicators. Another crucial component of research is quality control of the data, which researchers should take into account while developing their data management plans, especially before data collection begins. How to apply data quality indicators and how to ensure the quality of data at all stages for quantitative and qualitative data is covered in the sections that follow and these are the recommendations for researchers:

Quantitative data

The precision and consistency of variable descriptions with value labels, codes and the data collection instrument determine the quality of quantitative data. The following suggestions are offered as quality indicators as part of the quality process model in order to improve the management of variables in a data file:

- To find errors and unusual numbers, statistical analyses should be used, such as frequencies, means, ranges, or clustering.
- Where a researcher is deciding whether replies belong inside the defined categorisation group, quality indicators are believed to assist potential data consumers in avoiding confirming errors that have already been justified and to pick which values are appropriate for specific uses (e.g., working class, age group, salary level, health status).
- In conclusion, it is advised that researchers maintain quality indicators that are kept in separate fields or columns from the actual data. Implementing data cleaning strategies should serve as the foundation for the quality indicators. Using quality data flags is one method for cleaning data; for example, each column of values in a newly formed data file must have the words "Flagging column data" next to it. Data that have been found to be flawed should be properly associated using a naming system, e.g. *Age > flag_Age*. All of the flagged variables have to be corrected before a fieldworker leaves a household or ends an interview with a research participant.
- Another data cleaning technique is the "Identifying missing values" (Chapter 2, Section 2.3.5): knowing which variables represent specific characteristics, such as employment status or economic scale, can get complex. It is advised that researchers include a description field at the

start of each record to account for all identification factors as part of data cleaning. To reflect each case with its identifying element, a unique research number and respondent number should be utilised, e.g. the description field for the variable *EmplyPem* needs to be clearly marked as Employment status = permanent. A data collector will be able to gather accurate replies since they will be aware of what each variable implies.

- Preserving original information: an extensive set of detailed codes in a data file helps data analyses. As a conclusion, researchers should record the coding of original data as part of data cleaning. Recording raw data, such as age and income, is preferable to compressing or bracketing it.

Qualitative data

The precision and consistency of the formatting and grammar used in transcriptions determine the quality of qualitative data. The following conclusions are offered as quality recommendations to improve the quality of transcriptions:

- For the data to be accurate, there needs to be a connection between a questionnaire and the data being collected. A data-collecting program can then be used to incorporate consistency checks into the data-gathering process to address issues that arise during an interview. If the standard archive hanging indent set has not been used, the researcher's indents should have been substituted and correctly and consistently applied. To provide a consistent chronological order between the interviewer and interviewee, speaker identification tags should be added and used consistently.
- Addressing transcription errors as and when data is gathered.
- It is essential to run a spell check (as discussed in Section 6.7.6) on a data file to find glaring grammar and spelling problems (such as *your/you're* and *its/it's*). Correcting spelling and grammar problems, however, might occasionally need more resources. The Readme file that is included with the transcripts (data) should include a disclaimer saying that the transcripts are provided verbatim based on comments from respondents. All annotations, highlighting, time coding, and other markings should be removed from transcripts.

The quality of qualitative data, such as interview transcripts, is significantly enhanced by proper formatting. It is crucial to maintain the integrity of important facts and avoid altering the narrative. Researchers should also anonymise sensitive data, such as names, businesses, localities, positions, and dates. An anonymisation log should be created, recording the original names and applying pseudonyms to them. This technique allows the research team to know the background information of the original interview subjects, enabling accurate extrapolation and analysis of the data.

6.7.3.6 *Create informed file naming and cataloguing conventions*

Little advice is provided to researchers regarding the best naming and cataloguing conventions to adopt; the data curation unit uses its own naming convention. As a recommendation, curators should train researchers on how to choose the appropriate name and cataloguing conventions. Software compatibility, data analysis, software migration, and accessibility for secondary users are all made easier when researchers are taught and follow the proper naming and cataloguing rules.

As a recommendation, researchers must describe in the DMP the naming and cataloguing conventions they plan to utilise for their research project. It is important for researchers to design a descriptive reporting or file system that is user- and future-friendly, consistent, and has straightforward filenames. Choose file types that enable users to look for and find data sets with the intention of long-term access.

6.7.3.7 *Carefully selecting metadata for usage*

To enable secondary usage, researchers must outline the information schema and structure they will use in a DMP. Depending on the type of research, the metadata schema should contain the following fields relevant to data quality (Section 2.9.4):

- Identifier type
- The content of the data
- Structure
- Permissions – so it is discoverable for future use
- Resource Type (with general type description sub-property)
- Format
- Subject classification code
- Format
- Version
- Date

A researcher is in a better position to foresee any implications that might have an impact on the quality of the data by selecting a metadata schema with the aforementioned attributes. As part of the data planning process, a researcher must specifically know the metadata schema to use and whether the metadata schema offers quality control features.

6.7.3.8 *Focus on data quality*

For research data to be correctly handled and to be of high quality, the idea of data quality and how it will be maintained must be specified early on. Researchers must outline the procedures and tactics

they'll employ to promote the production of high-quality data. In conclusion, a researcher should explain the following in the DMP:

- The kind of routine accuracy checks they will employ.
- The type of training in data management that will be given to the study team and data collectors.
- How discrepancies (missing values, spelling errors, and duplicate responses) will be handled during data collection.
- The kinds of codes to be used, or definitions and coding schemes.
- Grammar and language.

This conclusion comes from the recommendation in Chapter 6 (Section 6.7.3).

6.7.3.9 *Quality control measures during data collection*

To guarantee the correctness and dependability of research findings, quality control procedures during data collection are essential. To find and prevent errors, inconsistencies, and biases in the data, these techniques comprise methodical inspections and procedures. Researchers can strengthen the validity and generalisability of their conclusions as well as the integrity of their data by putting strong quality control methods in place.

When it comes to quality assurance procedures during the data collection process, data curators can be a valuable resource for researchers. By offering advice on best practices, detecting potential sources of errors, and tracking data quality over time, they can assist in making sure that data is collected reliably and consistently. Performing routine checks for missing or incomplete data, confirming the accuracy of data entries, and running statistical analyses to find any outliers or abnormalities are a few examples of what this can include. Data curators can also help with documentation and metadata maintenance, which is important for making sure that results are understandable and replicable. Overall, data curators can assist researchers in ensuring that conclusions are supported by trustworthy and high-quality data, which can increase the validity and significance of study findings.

The following data-checking quality control measures should be used by researchers in conjunction with the data curation unit during data collection:

- Verifying the coding of observations, responses, and out-of-range values a second time.
- Checking data completeness.
- Adding variable and value labels where appropriate.
- Verifying random samples of the digital data against the original data.
- Double entry of data.

- Statistical analyses, such as frequencies, means, ranges or clustering to detect errors and anomalous values.
- Correcting errors made during transcription.
- Peer review.

In summary: Researchers are obligated to prioritise data quality when planning a research project in a DMP. A researcher can collect and clean data using the procedures that would be in place for this phase while adhering to best practices. The accuracy and quality of the data will be maintained and restored as a result. As in Chapter 5, Section 5.8.11 and Chapter 6, Section 6.7.5, it is recommended that researchers acquire data electronically in order to emphasise the importance of improving data quality at the time of data collection. The data planning step of the quality process framework was updated to be more structured and quality assurance-based using the aforementioned guidelines.

Modification applied to the Data Management Planning phase: The modified final version is depicted below:

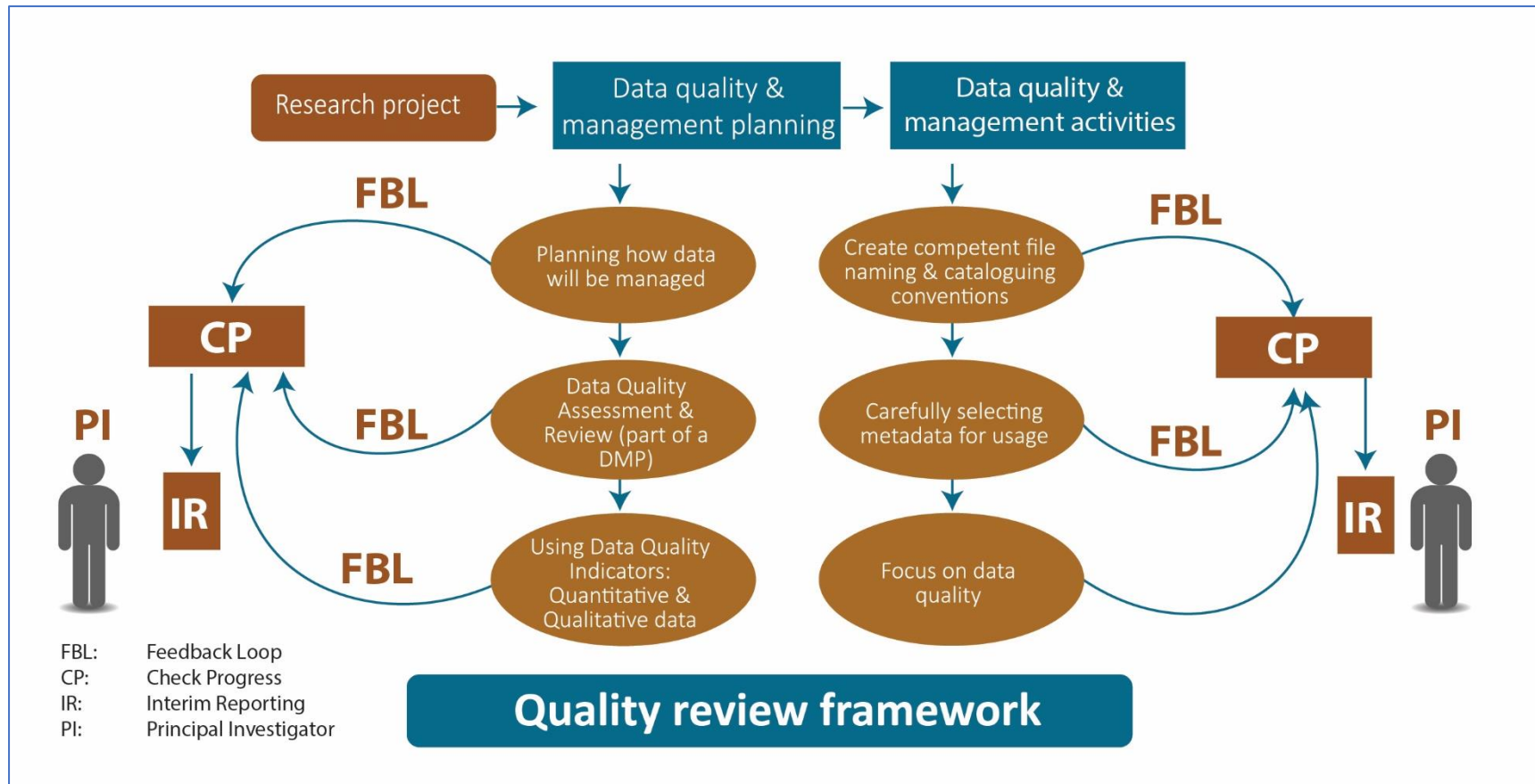


Figure 6.2 - Revised final version of the Data Management Planning phase of the quality process model

The model was modified using the suggestions of research participants after obtaining feedback from them (see Chapter 5, Section 5.8). Every stage will detail the modifications that were made and how they were implemented. The changes include:

In Chapter 5, Section 5.8.3, research participants indicated that a Feedback Loop (FBL), a Check Progress (CP) element, numbers showing where to start and how the model progresses, as well as Responsible Person (RP), and Interim Reporting (IR) components have to be introduced to the quality process model. The components function as described below:

The Check Progress element is linked to the Feedback Loop (curly line), which is used to communicate decisions made and actions taken when handling data files. The Check Progress feature enables the Responsible Person (i.e. the Principal Investigator (PI), Data Manager (DM), or Data Curator (DC) – represented in the model by a stick man and the appropriate abbreviation) to obtain real-time (live) updates on the data management decisions and activities taking place, which in turn results in Interim Reporting. The Interim Reporting function creates an interim report that must be approved by the person in charge of that task for any work done on the data to be taken as legitimate. The quality process model can now go on to the next phase if all tasks are completed correctly. The Feedback Loop generates the Check Progress and Interim Reporting on activities that were carried out in each stage, as explained in this chapter. In each stage, all the descriptions (i.e. data management activities) of that stage apply and should be carried out properly. All phases use this method, and references to this description of the modifications are made. Although the aforementioned adjustments (from study participants) are theory-based, the author must take the practicality of the suggestions into consideration. The author recommends that, in order to incorporate the Feedback Loop, Check Progress, and Interim Reporting components into the quality process model, the IT department of the social sciences organisation should develop a reporting system that can be utilised as each of these components. Allowing researchers and curators to input and update the system with their work would simplify the entire process that the research participants suggested.

6.7.4 Remedies for the revised quality-focused process model – data collection

The primary activity responsible for setting the initial data quality trajectory of research data is the Data Collection phase. Research participants in the study claimed to utilise only basic data-cleaning techniques when collecting data. The recommendation made in Chapter 5 is reiterated as a conclusion for the Creating Data phase – i.e. it is recommended that customised training should be given for the different research units in the organisation. The scope of each research project should be assessed, including the RDM needs of research teams. The training will hence have a focused purpose and train researchers to use best practice data collection methods that are customised and specific to their research projects. The next section explains the concluding quality assurance for both quantitative and qualitative data. The actions listed below are the best practices that are advised as part of the updated Data Collection phase of the quality process framework.

6.7.4.1 Quantitative data

The statistical component of quantitative data necessitates that a researcher manage data using procedures that produce a systematic and regulated management framework. The following tips are intended to help researchers manage their quantitative data:

- As a recommendation, variable names should, ideally, be unambiguous and clearly indicate accurate information. The variable name should be selected by the researcher considering its long-term usability for the widest possible user base and that it should be consistent in length with the program requirements (e.g., SPSS, SAS, STATA). When there are a lot of variables in a dataset, they should be grouped together. A researcher will be able to do this by making a list in a Codebook file that secondary analysts can utilise to rapidly get a grasp of a dataset. Checking for consistency means ensuring that fields that a respondent has completed have accurate values and that fields with missing values are recorded as such.
- To anonymise data, variables should be removed or aggregated, or the precision or precise textual meaning of a variable decreased. A dataset should be free of direct identifiers. Such identifiers are frequently not used in secondary research.
- The ICPSR (2018) codes (discussed in Chapter 2, Section 2.3.5.2) and coding recommendations should be provided by data curators as part of "in-the-field" support services. Researchers would then be able to use these codes for guidance when completing their DMPs. These guidelines will encourage the adoption of uniform, clear, and structured codes, which will make data understandable and simple to use. The ICPSR code and coding rules discussed in Chapter 2 (Section 2.3.5.2 – Table 2.3) are the recommended standards that the social sciences organisation should use.

6.7.4.2 Qualitative data

In comparison to quantitative data, the textual nature of qualitative data makes resolving concerns less difficult. Grammar, spelling, and consistency errors are not acceptable in qualitative transcriptions. The following suggestions are meant to assist researchers handle their qualitative data:

- Proofreading transcripts: after data has been collected, proofreading should always be done to ensure the accuracy of the findings. When data is gathered in the field, the initial round of editing should be done. This will guarantee that participants' responses are checked as they are sent. Reading the textual material in order to fully understand and comprehend the data is necessary to ensure the quality of qualitative transcriptions.
- Creating a data listing and anonymisation log document: once the transcripts are correctly anonymised, they should be thoroughly examined and validated, and a data listing log with the genuine names of research participants should be prepared. A data listing and an anonymisation log document can be created once the data has been understood. These

documents include a description of the transcription content, including the date of the interview, the number of pages in each transcription, the type of interview (individual vs. group), and a list of the anonymisation pseudonyms employed.

- The usage of accurate keyword descriptions: while reading each transcript, a researcher should make notes on the major study themes that were covered in order to help with HASSET Thesaurus keyword indexing. The HASSET Thesaurus is the leading British English thesaurus for the social sciences (University of Essex, 2024). As a recommendation, the data curation unit should train researchers on selecting keywords that are used in the HASSET Thesaurus so that researchers can use the most relevant and included keywords to facilitate discoverability when secondary users search for data.
- Using consistent pseudonyms for anonymisation: all copies of a data file that make up a collection should have the same pseudonym. This will guarantee that data users can comprehend data in the context for which it is used. Logical flow and clarity of the transcriptions are ensured by using consistent pseudonyms.
- Using automation for collecting data: as with quantitative data, it is recommended as a best practice that researchers should automate the collection of qualitative data in order to reduce transcribing errors.
- Testing the data collection instrument before data collection: the data collection tool should be validated before starting data collection to ensure that data will be recorded accurately and that any skip patterns will be handled appropriately.

The quality process framework (discussed in Chapter 3, Section 3.8) was used to interview research participants after the quality review framework served as the initial assessment tool. As a result, the 'Check Accuracy and Reliability of the Data' component was modified to include additional measures for both qualitative and quantitative assessment. The quality process framework requires that all applicable quality aspects (as indicated in Appendix 2) be completely answered for the Creating Data phase. This means that each of the quality dimensions (relevant to the Creating Data phase) listed in Appendix 4's scientific and technical subsections must be handled. Additionally, this covers the specifications for the improved "Check Accuracy and Reliability of the Data" component that was described in the preceding section.

As a conclusion, researchers are urged to use the best practices data-gathering methods covered in this section. Data integrity concerns should be addressed throughout data collection, i.e. make sure that value and variable names, labels, codes, skip patterns and missing data are accurate. In conclusion, researchers should make notes on any field activities they have conducted as part of quality-based data management, and then translate those notes into relevant documentation (such as a user guide) to help make the data easier to comprehend and use. As recommended in Chapter 6 (Section 6.7.3.3), it is also recommended here that more engagement between data curators and researchers needs to take place in

order for curators to recommend trends and best practices to researchers in an effort to maintain data authenticity over time. Before researchers begin a research project, the data curation unit should be integrated into the data planning stage.

In summary: Data curators must advise researchers on how to handle problems that would weaken data quality. It is also recommended that the data curation unit appoint someone or create a new role for existing curators to constantly monitor international RDM trends. A report on new RDM trends should be produced every quarter to allow the data curation unit to make new recommendations to researchers. This will ensure that the social sciences institution stays on par with all the latest RDM trends and best practices. Using a layered approach, the Creating Data phase was adjusted to consider all of the suggestions mentioned.

Modification applied to the Creating Data phase: The modified final version is depicted below:

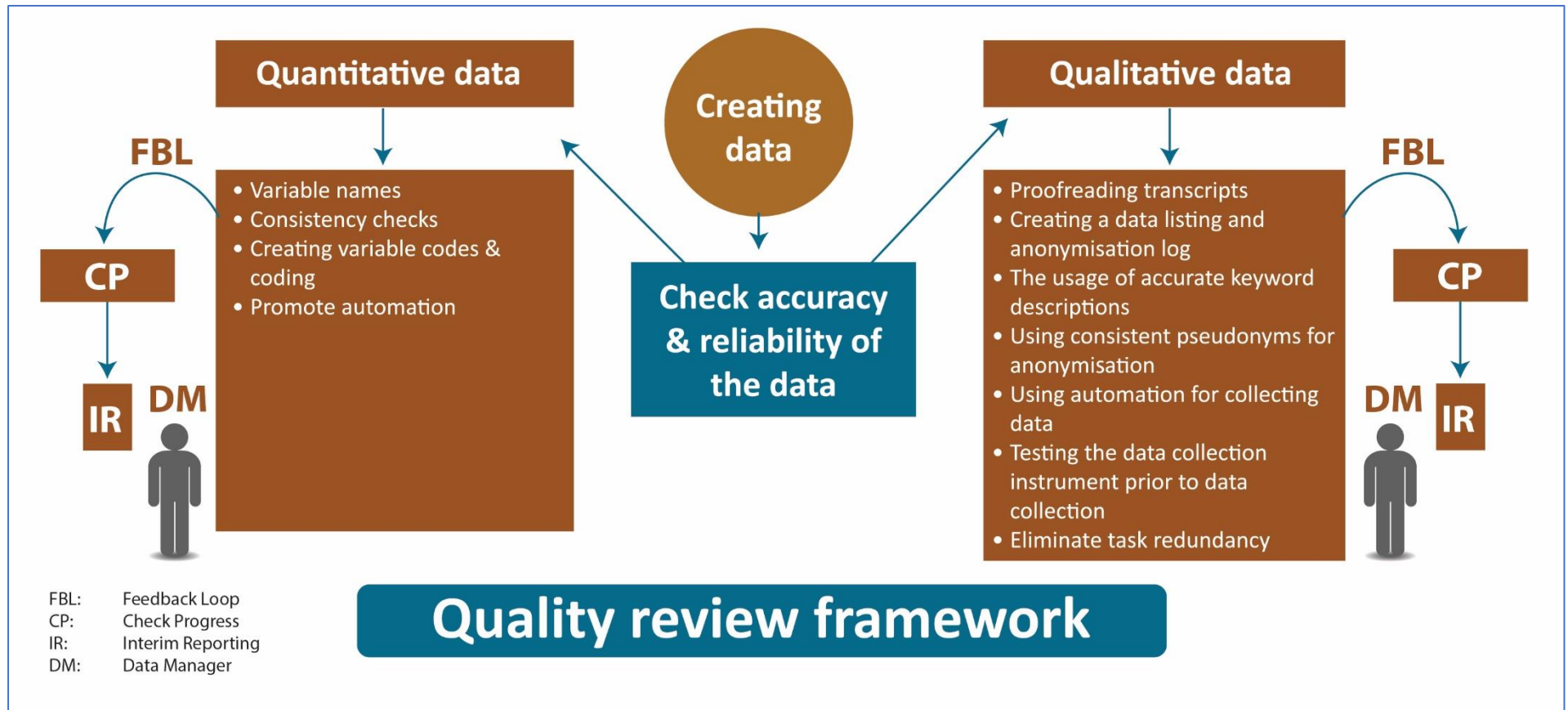


Figure 6.3 - Revised final version of the Creating Data phase of the quality process model

The same modifications and principles (i.e. Feedback Loop, Check Progress, and Interim Reporting) are applied to all stages of the quality process model, as was stated in the Data Management Planning phase. The Data Manager must approve all of the actions that data collectors carry out in the field during the period of creating data in order to monitor data collectors' progress and get interim reports. The tasks for the next stage can be completed if all the activities are completed correctly.

6.7.5 Processing data

A researcher who is processing data should not compromise the originality, relevance, or quality of the data. Processing data is a practical task that requires data management skills and competence. The following steps are suggestions on how data curators might help researchers:

- Review data for confidentiality issues
When using personal data from research subjects in studies, researchers must ensure that the data is correctly and consistently anonymised.
- Convert hardcopy documentation to electronic form.
Any hardcopy documentation that is required to understand a data file should be converted to electronic form. This is because hard-copy files can be lost or physically damaged. Researchers do not incur the risk of losing such documentation when it is in electronic form (and is securely preserved).
- Convert software-specific documentation to PDF/A.
Researchers should convert such documentation to a long-term preservation format that is easily accessible because some secondary users and researchers do not have the luxury of being able to finance pay-as-you-use software.
- Generate multiple data formats for dissemination and preservation
For various uses, researchers should establish alternative data formats; nevertheless, they should ensure that the dimensions of each file are the same and that the results can be replicated by others. The data will become trustworthy, adaptable, and accurate as a result, proving its high-quality value.
- Create a metadata record.
Users can access the characteristics and attributes of a data file using a metadata record. Researchers should describe how particular parts (that contain a quality assurance component) were managed when they developed a metadata record.
- Assign a Digital Object Identifier (DOI).
Researchers should create a DOI for their data files to increase the searchability, discoverability, and use of their research data. All of this is done to make data reusable and to show data findings.
- Verify the anonymisation of data in the Data Collection phase.

Researchers should verify for confidentiality when processing data and keep track of the kinds of information that need to be made anonymous. Giving the remarks to the data curators at the time of data deposit will guarantee that they merely check and do not perform the real anonymisation. This method calls on data curators to verify that anonymisation has been carried out correctly and consistently. For instance, mistakes should be checked when identifiers that should have been anonymised before deposit are found to still be present or when pre-deposit confidentiality editing removes so much information that parts of the data are rendered meaningless.

- Use different processing methods for qualitative and quantitative research

Researchers should employ several techniques for processing qualitative and quantitative data while managing research data. The most popular techniques for qualitative research are content analysis, narrative analysis, and thematic analysis. Regression analysis, descriptive statistics, and inferential statistics are some of the statistical methods used in the processing of quantitative research. It is important to thoroughly document the use of these methods and broadly disseminate the information within the social sciences institution. In their capacities as researchers, the following guidelines are recommended:

- Recode variables to address confidentiality concerns;
- Check for undocumented/out-of-range codes;
- Add question text to variables;
- Create unique variable labels;
- Create consistent value labels;
- Identify and address foreign language characters;
- Adjust format widths;
- Optimise file size;
- Standardise missing values;
- Check for consistency and skip patterns (UK Data Archive, 2017).

In summary: The aforementioned recommendations will help researchers describe and apply precise variables, codes, and values. These three factors contribute to the poor quality of a data file. The above recommendations will aid researchers in better managing a data file. All of the previous suggestions were taken into consideration, and a layered approach was used to modify the Processing Data phase.

Modification applied to the Processing Data phase:

The modified final version is depicted below:

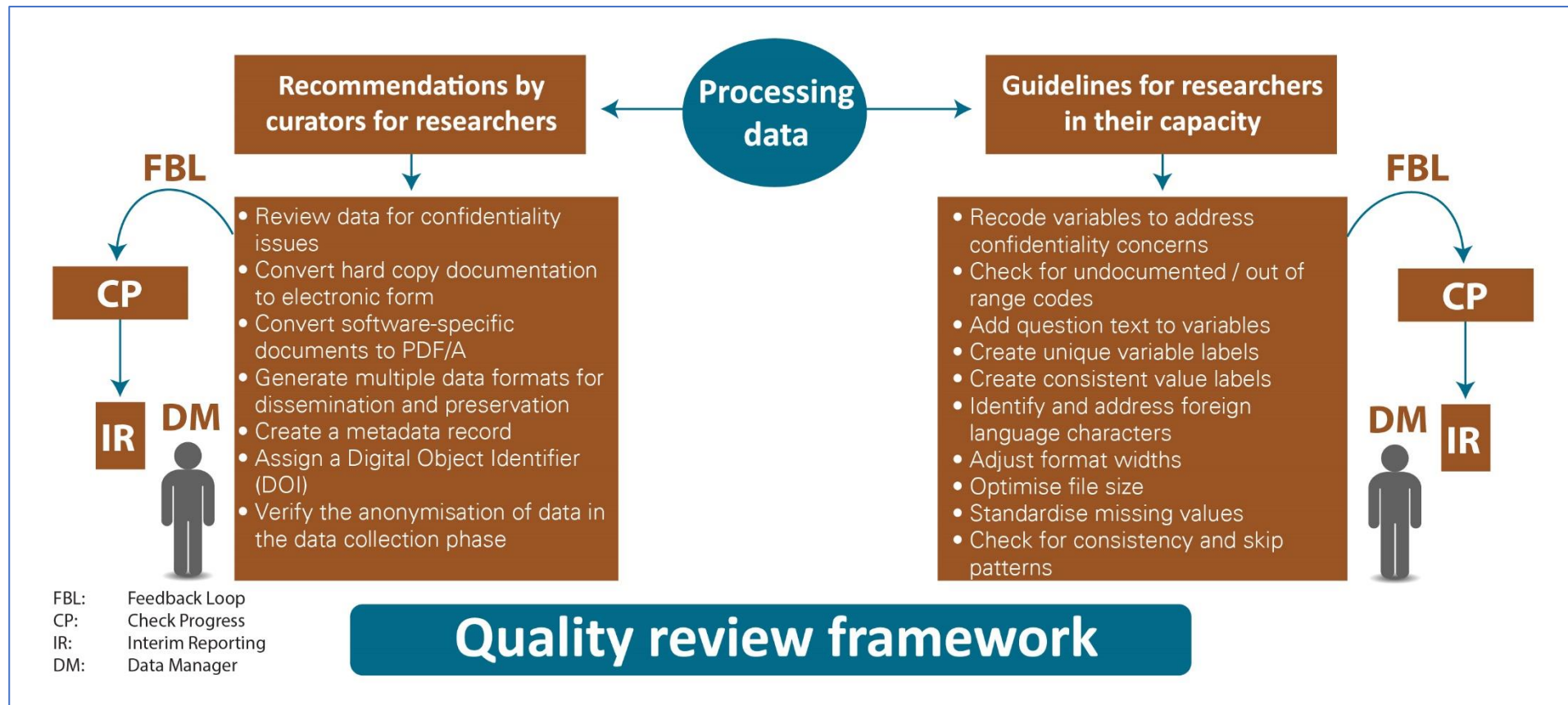


Figure 6.4 - Revised final version of the Processing Data phase of the quality process model

The same improvements and principles (such as a Feedback Loop, Check Progress, and Interim Reporting) are applied to all stages of the quality process model, as was stated in the Data Management Planning phase. The Data Manager must ensure that all duties related to data management and cleaning have been completed effectively during the processing of the Data phase. To do this, the Data Manager must monitor the progress and examine the interim report. The data manager is also responsible for managing the data integration and conversion process to make sure that data can be efficiently merged and analysed. If all the actions are appropriately accomplished, the duties for the following phase can be finished.

6.7.6 Analysing data

Data are evaluated at this stage of the research data lifecycle to create interpretations and discover helpful information. Now more than ever, data quality is crucial since it affects how useful the analyses are. Users can apply statistical and analytical techniques to analyse, transform, and model high-quality data. Here are some suggestions for analytical procedures:

- Users of data are advised to take the following actions to preserve and not alter the quality of the data: Adding title rows: To make it simple to grasp the information in a spreadsheet, add title rows. This suggestion will aid in classifying variables in accordance with the relevant sections of the questionnaire used to gather the data and will facilitate extrapolations.
- Remove duplicate rows or columns: If it is discovered that a data set contains several instances of the same record, it can be helpful to eliminate duplicate rows or columns before doing an analysis. This typically occurs when a variable contains a follow-up question that is incomprehensible on its own (e.g., *where did you sleep in the past 24 hours? Are you the owner of the property?*) Each row represents a separate element, duplicate rows or columns make data analysis easier to interpret.
- Delete any add-ons that happen when data is exported: If the user has exported data, any rows or columns that will not be used should be removed. For instance, many programs include timestamps or an "ID" column in data exports that are not needed for analysis. In some statistical programs, when data is exported to a different format, extra rows and columns are added (e.g. from SPSS to STATA). Users should remove any rows or columns that were added to a data file to preserve the data file's original form after exporting the data. As a result, the data quality is preserved and the data file can be analysed to produce accurate results.
- Standardising data: Data should be standardised to ensure that all numerical values, including numbers, dates, and currencies, are expressed uniformly.
- Record all changes to master files: It is crucial to keep track of every change made to a master file for various reasons. Firstly, this offers a transparent audit trail of all file alterations, which might be helpful for accountability and transparency needs. Secondly, enabling users to track changes and spot any mistakes or inconsistencies can help to maintain data integrity. Finally,

giving team members a central spot to save and retrieve the most recent version of the file can enhance collaboration and communication among them. Overall, maintaining accuracy and effectiveness in data management requires that all changes to a master file be recorded.

- **Maintain old master files in case later ones contain errors:** Keeping an old master file is crucial for several reasons. If later versions include errors or get corrupted, having kept such a file means that there is a backup. Having access to earlier versions is also advantageous for historical or reference purposes. This can be particularly crucial if the data in the file is essential to the company or organisation concerned.
- **Run a spell check to correct all spelling and grammar mistakes:** Spell checking aids in finding and fixing grammatical, typographical, and other mistakes in research data (in qualitative and quantitative data). This results in data that is cleaner and more accurate, lowering the possibility of misunderstanding or misinterpretation. Making sure spelling and grammar problems are fixed makes the data clear and simple to interpret. Reduced linguistic misunderstandings are essential for efficient data analysis and decision-making.
- **Archive copies of master files at regular intervals:** Archiving copies of master files at regular intervals is crucial for ensuring the long-term stability and security of your data. Regularly backing up files and storing them in secure, off-site locations brings protection from data loss due to hardware failures, natural disasters, or other unforeseen events. Additionally, archiving copies of your master files can help an institution maintain compliance with industry regulations and protect the intellectual property of your organisation. Overall, it is important to establish a regular archiving schedule and ensure that all relevant personnel are trained on proper backup and storage procedures.

Following the implementation of the quality process framework to interview research participants, an extra quality measure called "Recommendations for data users" was added to this phase, one that demands adherence to the prior criteria (compliance with the quality dimensions of this phase). The quality review framework demands that all applicable quality aspects (as indicated in Appendix 2) be fully answered during the Analysing Data phase.

In summary: To ensure that a data file is properly analysed and not altered, which would change its meaning and reduce the quality of the data, data consumers should abide by the instructions provided above.

Modification applied to the Analysing Data phase: The modified final version is depicted below:

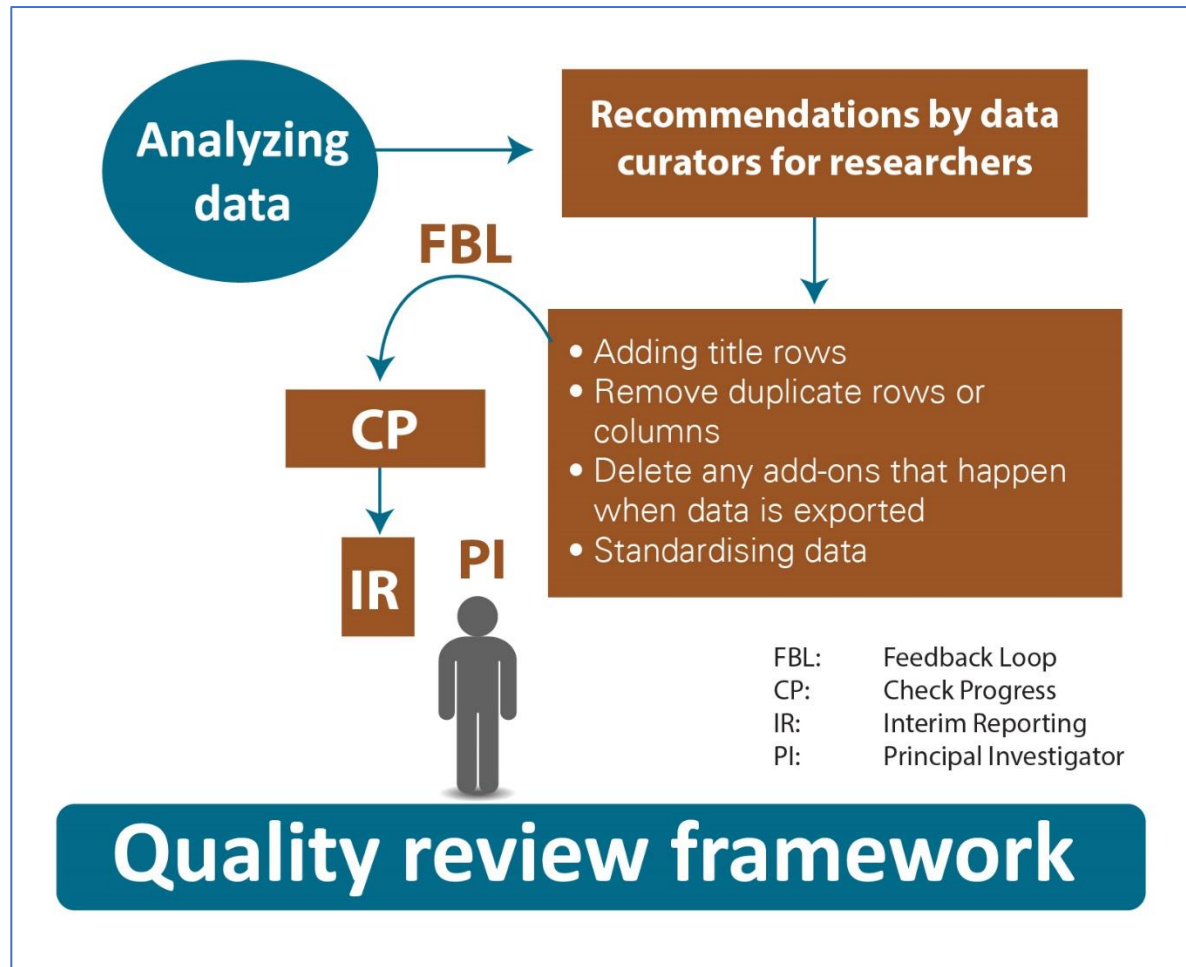


Figure 6.5 - Revised final version of the Analysing Data phase of the quality process model

The same improvements and principles (such as a Feedback Loop, Check Progress, and Interim Reporting) are applied to all stages of the quality process model, as was stated in the Data Management Planning phase. The Principal Investigator must make changes to the data (as part of the analysis) during the Analysing Data phase to make it easy to understand and useful. The tasks for the following phase can be completed if all the activities are completed correctly.

6.7.7 Preserving data

The data that needs to be retained at this point of the research data lifecycle was managed utilising quality control procedures and best practices. A researcher's or organisation's choice of preservation approach will depend on the preservation policies that apply to their organisation or funding requirements. The data curation unit should suggest the following preservation practices to researchers:

- Obtain the level of control necessary to ensure long-term preservation of the information provided.
- Establish the legal right to retain and store data. If data can be shared publicly, that should be determined.
- Check to see if any copyright protections are in place that would prevent the use of the data. Changing the format of data, for instance, to ensure its long-term preservation. These matters should be covered in the deposit agreement or archive contract.
- The intended community should be recognised and best served by the preservation strategy. This entails being aware of the demands and specifications of the community it serves.
- It is not enough for archives simply to store data for users; they also have a duty to ensure that the data they save is understandable to those users and this involves providing descriptive metadata.
- Follow the written rules and processes that guarantee the information is preserved against all foreseeable risks, such as the closure of the Archive, and that it is never erased unless it is part of a plan that has been approved.
- Determine if there are plans in place to ensure, for instance, that the archive's collections are not lost if it is unable to continue operating.

Researchers must be aware of the technological, legal, and best practices for data preservation if they want to ensure that their research data is correctly saved. Following the implementation of the quality process framework to interview research participants, this phase was given an extra quality measure called "Recommendations by Data Curators for Researchers" which calls for adherence to the prior preservation criteria. All applicable quality dimensions (listed in Appendix 13) must be fully addressed for the Preserving Data phase according to the quality review framework. This means all the scientific

and technical quality dimensions (listed in Appendix 2) of each quality dimension (applicable to the Preserving Data phase) must be addressed.

In summary: Controlled actions and management are necessary to ensure that data is secure, useable, and accessible across time. Researchers and the data curation unit must collaborate to guarantee that data is kept without sacrificing data quality, which is a vital aspect of data management.

Modification applied to the Preserving Data phase: The modified final version is depicted below:

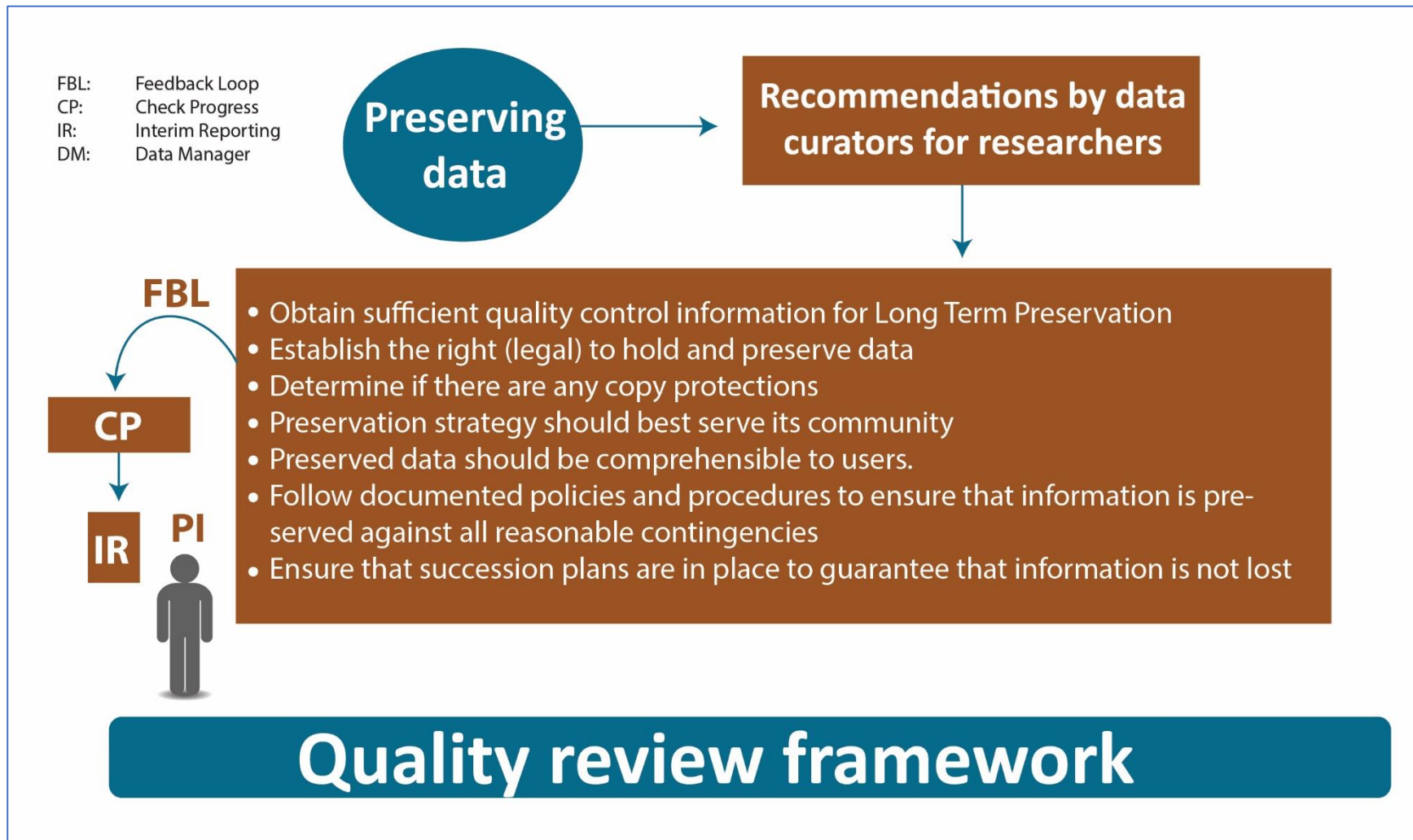


Figure 6.6 - Revised final version of the Preserving Data phase of the quality process model

The same improvements and principles (such as a Feedback Loop, Check Progress, and Interim Reporting) are applied to all stages of the quality process model, as was stated in the Data Management Planning phase. The Principal Investigator must prepare data for adequate preservation during the Preserving Data phase; as the PI completes these tasks, he or she can evaluate the Interim Report to see how far along they are in the process. Before moving on to the next phase, the PI will then be able to correct any incorrectly completed actions.

6.7.8 Giving access to data

Making new discoveries and disseminating knowledge are two of the fundamental goals of research. Giving users access to data necessitates a systems organisational strategy; this method must be backed up by pertinent documentation that secondary users can review to better comprehend the data. Researchers and data curators who promote data access from a technical standpoint should consider the following suggestions. Data depositors should create Data Documentation Initiative (DDI) compliant documentation, but they should also convert any existing metadata that is in another format to DDI as necessary. The following are recommendations for researchers:

- Adopt and employ tools for publishing data (such as Nesstar Publisher) to write XML, a markup language that is comparable to HTML but that lacks predefined elements. As an alternative, create your own tags that are tailored to your needs and help data be found.
- Create security groups, add users to the relevant groups, and give each group the proper access to shared data to establish a security framework.
- Verify that all operations are approved by auditing the data owners' actions. This will stop unauthorised access to data and potential data tampering. The audit will keep track of both modifications to and access to sensitive data.
- Create an access request workflow so people can quickly seek access to the data they require to perform their duties (such as a request portal).
- Make an exhaustive list of users and resources and keep it current.
- Determine the location and owner of sensitive data by working with department managers and other business owners.
- Establish who in the organisation owns the various types of data and who has access to them.

The main focus of the Giving Access to Data phase is on carrying out technological procedures that will make sure that data is accessible, understandable, and usable. The first evaluation tool for the Giving Access to Data phase was the quality review framework. This phase acquired an additional quality measure, called "Recommendations for Researchers and Curators who Facilitate Giving Access to Data from a Technical Perspective"; after the quality process framework was used to interview research participants. Answers and justifications must be provided for all the previously mentioned topics by curators and researchers. The quality review framework mandates that all

applicable quality dimensions (mentioned in Appendix 13) be completely addressed for this phase. This implies that each quality dimension (relevant to the Giving Access to Data phase) of each of the scientific and technical quality dimensions (in Appendix 2) must be addressed.

In summary: Giving someone or something access to data requires granting them the right to see, utilise, or change a certain set of data while also protecting its security and privacy. Access can range from publicly available to highly restricted access, depending on the situation and the type of data.

Modification applied to the Giving Access to Data phase: The modified final version is depicted below:

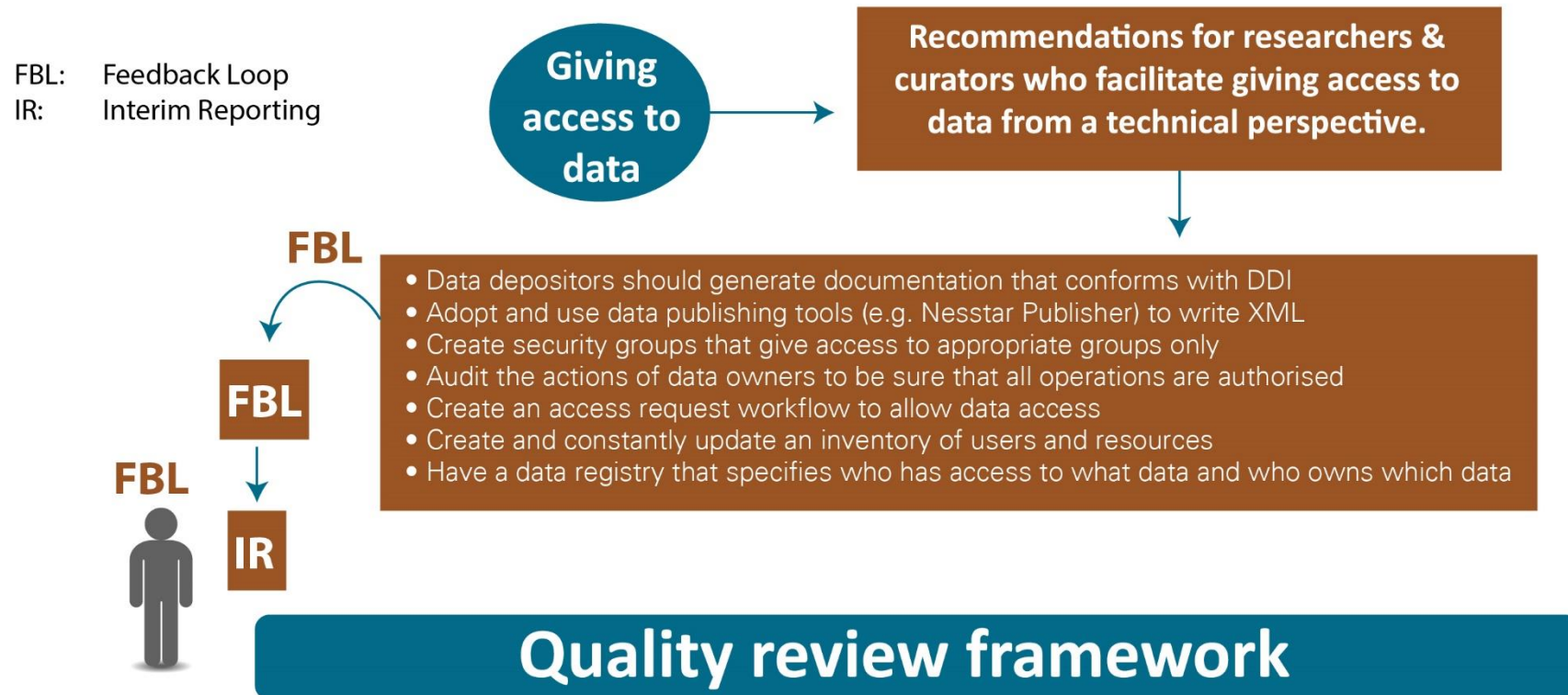


Figure 6.7 - Revised final version of the Giving Access to Data phase of the quality process model

The same improvements and principles (such as a Feedback Loop, Check Progress, and Interim Reporting) are applied to all stages of the quality process model, as was stated in the Data Management Planning phase. The Data Curator (DC) checks to see if the data depositor, the Principal Investigator, did everything necessary to make the data available during the Giving Access to Data phase; this is more of a quality control task. By checking the Interim Report frequently as progress is achieved, the Data Curator can monitor his or her progress. Upon successful completion of this phase, the process model advances to the last stage.

6.7.9 Reusing data

The discussion in Section 3.7.1 demonstrated that, in order to effectively manage data, a sequence of data actions or steps is needed to validate, organise, transform, incorporate, and extract information in an appropriate output form for subsequent use. However, it is crucial to be aware of copyright and intellectual property rights that could prevent data from being used unethically or in a way that degrades the quality of the data. To encourage correct secondary usage of the data, the data curation unit should suggest the following strategies to researchers and secondary users:

- The usage of metadata
In some cases, secondary researchers and users are unable to contact the original data authors with their questions, therefore metadata serves as their de facto means of communication. An improved comprehension of a dataset, data searches that incorporate variables, and a variety of Web presentation options are all made possible by comprehensive and precise metadata. Since it provides information needed to effectively utilise the data's analytical capabilities, metadata is crucial for optimal data utilisation.
- Be mindful of intellectual property rights
You may be violating the intellectual property rights of another party if you reuse something that is not legally yours or that is not publicly accessible to the public. When building a process, it is crucial to ensure that the reusable requirements specified do not include user requirements as these will vary depending on who is using the process and in what context.
- Consistent maintenance of the repository
Once a set of reusable criteria has been categorised, maintenance is necessary. Regardless of the style or context of the requirements, the repository should always be kept current. This reduces the possibility of out-of-date information being used and helps to ensure that all changes can be tracked. To ensure that the requirements remain reusable in the future, update them as new information or data becomes available.
- Do a reusability retrospective review
Frequent retrospective sessions are another excellent practice to keep the requirements reusability process up to date. The procedure should include suggestions and a record of knowledge gained, as well as a way to monitor which suggestions are being implemented to

enhance the procedure. Do this periodically to monitor problems and work with others to find solutions.

After the quality process framework was utilised to conduct research participant interviews, this phase gained a new quality measure component titled "Recommendations by Data Curators to Researchers to ease Secondary Usage of the Data." Every one of the aforementioned topics calls for a response and an explanation from the researcher. For the Reusing Data phase, the quality review framework requires that all applicable quality elements (as listed in Appendix 13) be fully addressed. This means that each quality parameter stated in the scientific and technical subsections of Appendix 2 that is pertinent to the Reusing Data phase must be addressed.

In summary: Effective data reuse can result in cost savings, improved decision-making, and innovative thinking. However, it necessitates careful preparation, data quality assurance, and adherence to ethical and legal guidelines.

Modification applied to the Reusing Data phase: The modified final version is depicted below:

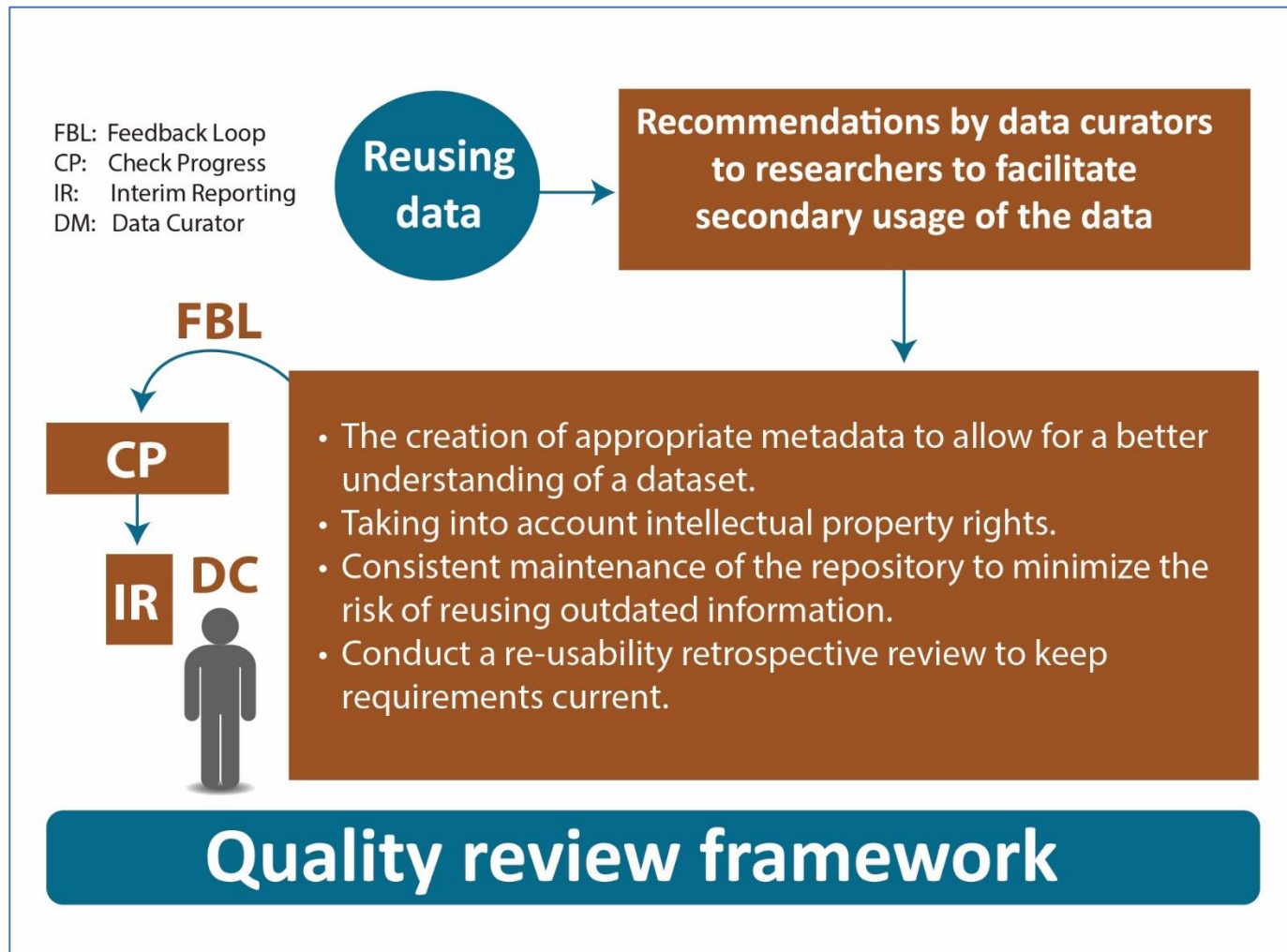


Figure 6.8 - Revised final version of the Reusing Data phase of the quality process model

The same improvements and principles (such as a Feedback Loop, Check Progress, and Interim Reporting) are applied to all stages of the quality process model, as was stated in the Data Management Planning phase. The Data Curator completes crucial tasks throughout the Reusing Data phase that will make data comprehensible and useable, like the generation of relevant metadata. The curator will receive an interim report, in which he or she evaluates whether or not all tasks have been completed appropriately.

This entire section has provided information on how the quality process framework was modified after the initial version was tested with research participants. The initial draft was complicated and ambiguous, according to the study participants who were asked about it. To make the quality process framework easier to understand and apply, it was divided into many stages. The goal of this method is to reduce the number of arrows and loops that the original quality process framework contained by evaluating and applying each step in a more linear manner. The figure below displays the entire redesigned quality process framework, with each phase displayed in a layered fashion as indicated in the preceding section.

Following the evaluation of the first iteration of the quality process model, it became clear that a framework for data quality-focused RDM that addresses data quality issues requires a simplified approach that users can easily relate to. The quality process model is still built on the UK Data Archive (2017) research data lifecycle. The modification exercise enabled the development of the framework for data quality-focused RDM, which has the following features:

- Data quality dimensions are descriptions of the characteristics of the data that can be quantified or compared to current standards for determining the quality of the data.
- Measuring criteria taken from: How key role-players determine data quality. These are concepts and theories that organisations have defined and that are used in their RDM programs and strategies.
- Measuring criteria taken from: RDM initiatives by relevant stakeholders.
- Measuring criteria taken from: Social sciences research funders.
- Measuring criteria taken from: Various evaluation systems.
- Measuring criteria taken from: Books on social science methodology.

These elements were used to enable the adoption of quality-based methods for the assessment of operations throughout the different stages of the lifecycle of research data. It was possible to build a framework that has a strong emphasis on evaluating RDM activities using quality assurance techniques.

6.7.10 The final quality process framework

A simplified version of the quality process framework was developed due to the findings of the empirical research that involved research subjects. The new version takes a layered approach, with various tasks for every phase. The enhanced quality process model is depicted in the following figure:

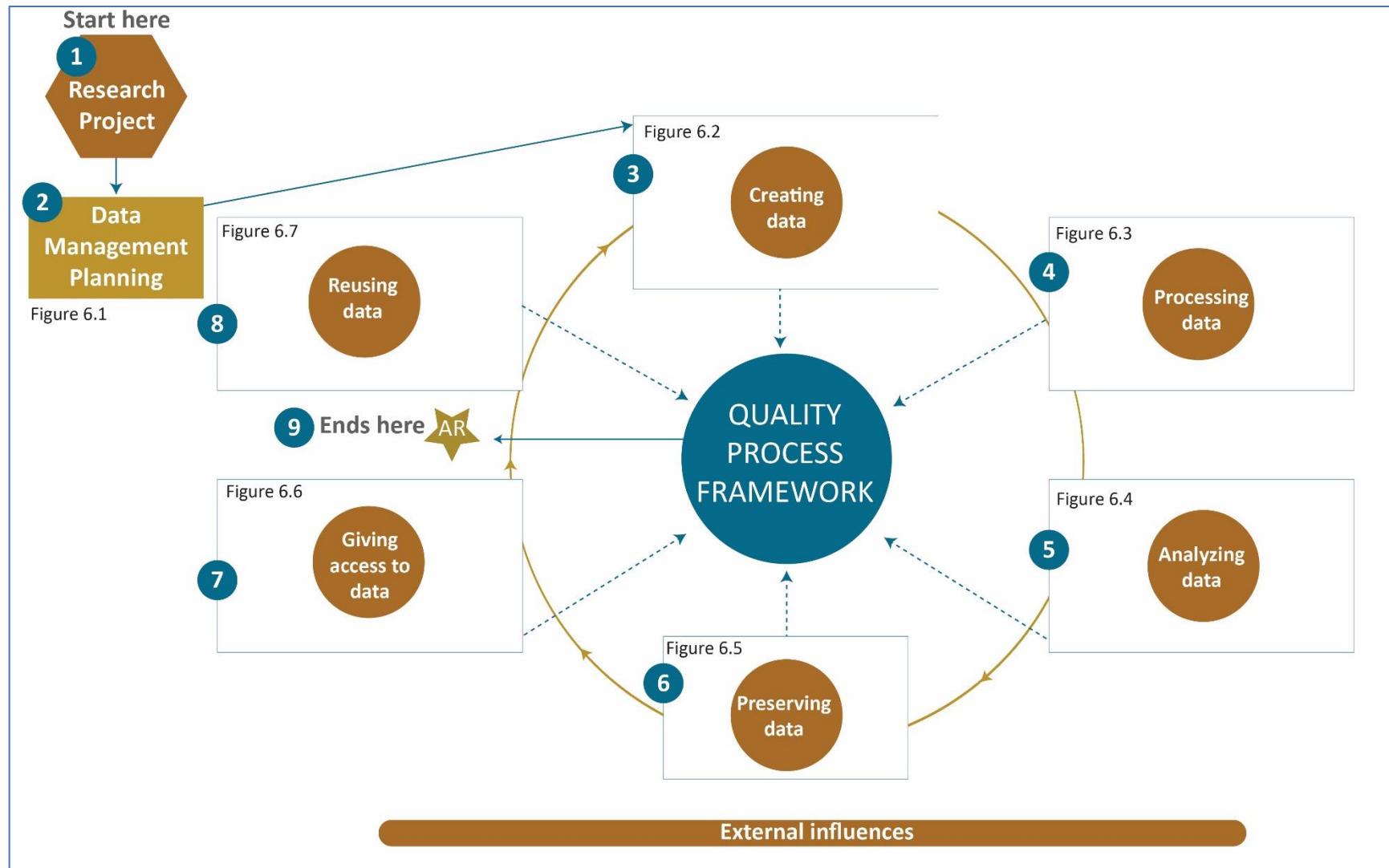


Figure 6.9 - Final quality process framework

6.8 An overview of the functioning of the quality process framework

A researcher is expected to plan for data management as soon as a project has begun; this planning is shown in Figure 6.2 (Section 6.7.3). This section covered all of the necessary steps that needed to be considered in data management planning as well as the guidelines that needed to be adhered to. Following the completion of data management planning, the quality process framework advances to the Creating Data phase. Best practices for data collection should be adhered to; these were covered in Section 6.7.4 (Figure 6.3). After data has been successfully gathered, processing can begin. The quality process framework is at present in the Processing Data phase. Section 6.7.5 (Figure 6.4) discusses best practices and principles that researchers should adhere to when processing data. The next step in the quality process model is Data Analysis; Section 6.7.6 (Figure 6.5) offers a thorough description of the recommended analytical techniques for the analysis of data. The next step is the Preserving Data phase, where researchers should utilise the preservation techniques outlined in Section 6.6.7 (Figure 6.6) to ensure reliable access to and preservation of the data that has been managed thus far. Data must be accessed when preserved for use by subsequent users. This leads to Giving access to data the following step in the quality process framework. Researchers should adhere to the guidelines provided in Section 6.6.8 (Figure 6.7) when disseminating their research data. The Reusing Data phase (Figure 6.8) is the last stage of the quality process framework. Section 6.6.9 covers strategies that researchers and secondary users should be recommended to use by the data curation unit.

An Assessment Review is the end outcome of the quality process framework, which is informed by all of the previously stated stages. The Assessment Review illustrates that throughout the research data lifecycle, data quality-based data management techniques were applied to manage data, resulting in the production of high-quality data.

The implementation of the above quality process framework requires a certain level of change management in the institution. The following section discusses the features of an effective, quality-focused framework to guide the data creation process in a social sciences organisation.

6.9 Features of an effective, quality-focused framework

This section answers the main research question – What are the features of an effective, quality-focused framework to guide the data creation process of a Social Sciences Research Organisation?

The main research question was addressed in the context of the research data lifecycle applicable to the social sciences organisation's data quality assurance, and RDM. The author defined data quality as "fitness for use", as determined by a variety of quality metrics in the form of best practices, recommendations, and standards that are connected to the technique used to assess or enhance data quality. Since the proposed quality process model (as discussed in Section 3.8) does not alter pre-existing models, the initial iteration of the quality process model required revisions. Thus, the

suggestions provided in the preceding sub-question (i.e. change management) represent a justifiable advancement in the application of the quality process model's final version. The study's assessment of the literature demonstrated that the idea of "data quality" has several sides and a variety of characteristics, all of which contribute to the definition of the term. As was also suggested by the literature review in Section 2.3, data quality incorporates principles that convert data into a commodity with clients who give a price and value to it. Different universities, RDM initiatives, evaluation systems, and large research funders have been proven to approach RDM in different ways. RDM was further demonstrated to include quality dimensions that are primarily focused on evaluating and enhancing data quality. The study also demonstrated the various frameworks and a framework development process used in Section 3.2 to create the quality review process and quality review framework (Section 3.8.4 and 3.8.5). The primary measurement basis (i.e. quality dimensions) of the quality review framework was built on the data quality dimensions adopted from the SASQAF (Section 3.7.3) and DAMA UK (Section 2.3.3.1). The quality review framework (Section 3.8.5) includes many measuring standards that were taken from the literature review

The quality process model has to be modified, which proves that the hypothetical component of the previous research sub-questions is real. As a result, change management is a legitimate concern. The development of the quality process model in Section 3.8 of the study demonstrated the multi-tier structure that the scope and strategies of RDM may adopt. A similar process model was also demonstrated to be made up of the various research data lifecycle stages, which were streamlined and grouped into various layers that interact with one another (Section 6.7.3). The discussion in Sections 2.11 and 6.10 made it clear that senior management must model and promote the change they want to see in an institution in order for it to happen. Research teams must fully comprehend the advantages of implementing new RDM techniques so they can generate high-quality data. Taking the various recommendations that were given in this study, a conclusion can be drawn that extra training is needed at the social sciences organisation. It is thus recommended that a training unit be established at the organisation. The training school will provide on-going training (in all stages of the research data lifecycle) to researchers and curators on all matters related to RDM, quality assurance and data quality.

The responses to the sub-questions revealed that a variety of RDM stakeholders' decisions influence a number of factors that have an effect on data quality. Data quality dimensions and measuring criteria provide the necessary context and structure for the effective management of research data throughout its many lifecycle stages when integrated (i.e. the quality process model). In order to reduce the complexity of the first iteration of the quality process framework (discussed in Chapter 3, Section 3.8.8), layers were employed to manage the tasks associated with each phase of the research data lifecycle.

In summary: The final quality process framework works on a simple and more logical structure:

- A researcher must outline the necessary data management planning strategies before beginning a research project (as discussed in Section 6.7.3).
- Data will be managed using the quality process framework and the suggestions (described in Section 6.7) appropriate for each phase after the Data Management Planning phase is completed.
- Each phase is represented as a layer that needs to be approached and handled as a standalone entity before moving on to the following phase.
- Every associated activity (in all the phases) is assigned someone who is responsible for carrying out that activity. The progress of all activities is monitored and an Interim report is produced.
- A researcher must respond to and address each criterion in accordance with the quality process framework and recommendations offered for each phase in Section 6.7 before moving on to the next phase of the data lifecycle.

The adoption of the quality process framework will not occur spontaneously, so it is advised that researchers and research teams from the social sciences organisation be given the opportunity to attend seminars that outline and explain the quality process model in an effort to contextualise it. The following section takes the form of a change management recommendation for the social sciences institution, which serves as the answer to sub-research question 6. The change management recommendations for the social sciences institution are covered in the section that follows.

6.10 Change management recommendation for the social sciences institution

The responses from the social sciences research participants who were interviewed allowed for the formulation of both objective and practical recommendations that will aid in resolving the RDM problems encountered by the institution. In the event that this research indicates that the social sciences organisation needs to implement a new model, the question of "how should change management occur" arises. The following are some suggestions based on best-practice principles and guidelines taken from the literature that was reviewed:

- Address the human side of change systematically (Section 2.11.1): along with data quality, researchers and research teams need to be informed about the advantages of RDM. Thus, in order to make sure that the change is in the research team's best interests, it is vital to tell them of the type of change that is coming.
- Change management calls for a proactive approach: that handles problems one at a time for researchers and research teams hoping to attain good data quality. The management team should be involved in the development of a formal approach to managing change, which

should be created early on and frequently modified as the change spreads throughout the organisation.

- Change starts at the top and begins on day one (see Section 2.11.2): In order to develop a data management culture that supports high data quality, senior management must clearly show the change they desire. For researchers to study what it really means to have systems in place that aim for high data quality, the proper learning environment must be established. In order to change the behaviours, activities, and working principles of employees, or any other type of management division in a company, researchers must accept the principles of data quality. This learning environment, as well as the change that the learning environment is intended to foster, must come from top management from the beginning of change management.
- Real change happens at the bottom (Section 2.11.3): The accountable custodians who carry the baton are data managers and the team on the ground. Any programme that encourages researchers to offer high-quality data will be successful if it includes plans for identifying leaders and distributing design and execution duties throughout the organisation.
- Confront reality, demonstrate faith and craft a vision (Section 2.11.4): Any achievable improvements would be simple for the research teams to comprehend if they were provided with a defined case for change management. The opportunity to develop (or force) management team alignment through the development of a formal case for change and a documented vision statement is valuable.
- Create ownership, not just buy-in (Section 2.11.5): Give researchers and research teams the freedom to pursue the vision of improving RDM and data quality. Enthusiasts who over-perform during the transition and are the catalysts for change in the workforce are the change leaders. Employees are more likely to feel comfortable making changes if they think they can own the change they are being asked to make, and this would help the organisation present a united front.
- Using targeted overcommunication as a strategy (Section 2.11.6): Organisations must establish strong coordination, effective leadership, and transparent communication in order to fully utilise and develop their resources. This strategic leadership approach is crucial for convincing all employees (researchers) to concur on a fair arrangement that will increase the effectiveness of work processes.
- Explicitly address culture and attack the cultural centre (Section 2.11.7): Culture has the potential to become a phenomenon that takes a lot of time and effort to change or eradicate. Determining a baseline through a cultural/organisational diagnosis, defining a clear end-state or desired culture, and creating specific transition plans are therefore crucial for the social sciences organisation.

- Assess the cultural landscape early (Section 2.11.8): Although research data management practices have their own culture, they must be consistent with the company culture. A thorough cultural diagnosis must be carried out early on to accurately assess organisational readiness to change, bring significant issues to the fore, reveal cultural components that support or impede change, and pinpoint leadership and resistance sources.
- Prepare for the unexpected (Section 2.11.9): Unexpected reactions will occur from people, areas of predicted resistance will disappear, and the external environment will change. As a result, it is necessary to manage and oversee the organisational culture of the envisaged RDM in order to anticipate potential unforeseen events.
- Speak to the individual as well as to the institution (Section 2.11.10): Personal adjustments must be performed in order to ensure that high-quality data is generated. The knowledge of how their work will change, what will be required of them throughout and after the change programme, how they will be judged, and what success or failure will entail for them is crucial for researchers and research teams. In the same way that researchers need to be interested and involved as they engage with their work, they also need to be interested in and involved in the process of change.

The suggestions above are regarded as being precise and realistic, and they address significant issues that, if addressed incorrectly, could endanger the development of RDM operations. RDM has the potential to demonstrate the integrity of science and the validation of research findings, as stated in Chapter 2 (Section 2.1). The recommendations made above are suitable for changing RDM processes to incorporate a human interaction-based strategy for data quality. As discussed in Chapter 2 (Section 2.11) and Chapter 6 (Section 6.10), in order to change organisational culture, senior management must demonstrate that the change they require must occur.

The issues of change management that are expected to arise when such a framework is implemented in South African Social Sciences Research Organisations are discussed in Section 2.11. The discussion on change management proved that for change to occur in an organisation, senior managers and staff must push for the desired outcomes and make the necessary modifications. A Social Sciences Research Organisation may face the following change management issues if these changes and those covered in Section 2.11 are not adequately managed.

- The willingness of employees to change

The discussion in Section 2.11 demonstrated that researchers employ their own RDM procedures and are at ease using the data management tools that are available inside their organisations. Persuading researchers and research teams to accept the change and align with senior management in order to anticipate its long-term benefits is one of the most challenging components of adopting the recommended quality process framework. As a conclusion, the social sciences institution should hold

a symposium whereby researchers will be informed about the transformation process and the benefits that come with it by senior management. The symposium should allow participants to have open discussions in order to make sure that the change is in the research teams' best interests. It is vital to tell researchers about the type of change that is coming, according to the discussion handling the human side of change methodically in Chapter 2 (Section 2.11.1). When researchers know what is expected of them and what the change will bring, they will be more open to change. Researchers may believe that implementing the quality process framework and the study's recommendations will cause them to miss their deadlines for both research and publications. As a conclusion, it is crucial for researchers to completely comprehend the advantages of appropriate RDM and how the quality process framework works to encourage better RDM and quality assurance across the entire research data lifecycle.

- Lack of executive support and active sponsorship

It was demonstrated that the idea of senior management is another idea closely tied to change management. The explanation of change management in Section 2.11.2 was given in the context of how top management can either facilitate or impede the adoption of change in an organisation. Many executives are unaware of their position in change management, the resources needed, or the change's objective. Because of this misperception, the organisation's communication, visibility, and transparency are all skewed. Dwindling support after go-live, a general lack of buy-in for the change, and infrequent involvement are all impediments that can be attributed to executive support. Top management must show effective implementation and roll-out of the desired change management activities in order for employees/researchers to fully embrace change management. This is vital because researchers and research teams can comprehend the level of importance required when senior management improves the way they conduct things.

- Change-resistant culture and attitude

According to the discussion of how to deal with culture in an organisation in Sections 2.11.7 and 2.11.8, there are cultural practices that have ingrained themselves into researchers' work procedures. Another problem in change management is changing the mindset and culture of researchers and research teams. Lower performance, trust, and involvement inside the organisation are the outcome of a history of failed changes, the prevalence of silos, and resistant behaviours. Individual objectives, disruptive thinking, and resistance to change are all encouraged by a culture filled with bad attitudes and internal politics. Shifting the organisation's entire culture and atmosphere would require a transformation in employee perspectives. As a conclusion, the social sciences institution should introduce monthly engagement workshops that will highlight the importance of change and what change means in the context of RDM and introducing new processes.

In summary: To improve operations and accommodate changing needs, change management advice for a social sciences institution needs to call for careful preparation and execution of changes. A clear vision for the reforms and precise objectives must be established by the institute for social sciences. To ensure that everyone is on the same page with the expected RDM and data quality results, the institution has to make sure to communicate these goals to all parties involved. To help faculty and staff develop the skills necessary to adapt to new processes or technologies brought about by the changes, continuing training and support should be offered in order to support researchers in becoming aligned with the changes that are occurring.

6.11 Suggestions for further research

It has not been possible in this study to address every potential question concerning a data quality-focused RDM to guide the data management planning process in the Social Sciences Research Organisation concerned. Accordingly, the following topics are offered as suggestions for areas of future research in the field:

- The role data curators play in helping researchers manage their data before, during, and after a study.
- The formation of new curricula for teaching incoming researchers about quality control and data integrity.
- The establishment of a formal structure for data quality and quality assurance at a social sciences organisation and how this could be made applicable to academic institutions.
- Data managers' technical responsibilities in managing research data to provide high-quality data.
- An investigation into the standards that various government departments may employ to enhance their RDM efforts. Such an investigation could focus on data management skill development, infrastructure planning, support for research services, and data management planning.
- The creation of quality assurance process techniques to be used when performing quality control.
- The necessity of developing fully the measuring criteria for an acceptable data quality dimension framework in order to establish a measuring system that will reliably determine levels of data quality compliance.

These topics have the potential to contribute towards allowing institutions and researchers to use RDM best practices and guidelines to improve data management practices. The next section looks at the contribution of this study to the subject field.

6.12 Contribution of this study to the subject field

This study has made several contributions to the subject field, including the following: The development of a quality process model made up of several layers, as indicated in Section 6.7 and validated by the empirical study documented in Chapter 5. This includes all phases of the research data lifecycle (as shown in Figure 1.1). Further, the study has demonstrated, as noted in Sections 2.2 - 2.7, the critical role that data quality dimensions and RDM in general play in managing data throughout the research data lifecycle. Based on the preceding answer to the main research question and the sub-questions, this research study has also made the following contributions to the field of study:

- More involvement from data curators in the data planning stages of research projects being conducted by researchers. When data is collected in the field, it is advised that research teams and research managers incorporate quality assurance techniques.
- The data curation unit should distribute templates for data management planning to all research units within the organisation. Completing the template should become mandatory when seeking ethical clearance.
- As part of data planning, researchers should indicate which best practice recommendations for each phase of the research data lifecycle they will use. This exercise should be done by using the criteria of a quality process model.
- Providing real-time data management support during field operations (data collection).
- The use of measurement criteria during data processing. The ability of researchers to conduct independent data quality assessment.
- Researchers receiving training in performing RDM responsibilities. Provision of data anonymisation techniques, code and coding best practices.
- The social sciences institution should develop an RDM training unit that will occasionally host renowned speakers who will offer insightful commentary and recommendations on RDM.
- Curators and researchers should embrace open access to research data because it encourages collaboration and elevates the bar for science as a whole.
- A point that needs to be stressed is that data quality should be considered as a benefit rather than a burden. In order to solve data management issues and make sure that researchers are providing evidence for their claims, it is also critical to dedicate time, personnel, and resources to data quality evaluations. In this approach, improving data quality will be perceived as a success rather than a corrective action
- Promoting data quality will be simpler if top management is more aware of RDM and quality control.
- When it comes to planning data management and during the Data Collection phase, data managers and curators should take on new supportive and active roles.

6.13 Critical evaluation of the research methodology

The study adopted a qualitative research approach that enabled a thorough examination and comprehension of the research problem by looking at the preferences of research participants. Interviews were used to gather data for the study, and the results offer deep and thorough insights into the participants' viewpoints, experiences, and difficulties with data quality and RDM. The qualitative research methodology employed in the study appears to be well-suited for examining complex phenomena and comprehending the nuances of data quality management. The methodology was developed on the basis of creating a framework for data quality-focused research.

This qualitative research methodology excels at capturing the nuances and contextual influences that affect how curators and researchers manage the quality of their data. Direct interaction with participants allowed the author to learn more about the challenges of managing research data from their perspective. This methodology's adaptability is another advantage. This researcher was able to explore new themes and modify the approach as a result of the iterative data-gathering and analysis methods available with qualitative research. With such adaptability, the study can respond to changing research topics and participant experiences.

It's crucial to recognise that the qualitative research methodology has several drawbacks. The subjectivity inherent in qualitative research is one potential drawback. The personal biases and viewpoints of the researcher affect how the data is interpreted, which could bring some subjectivity to the conclusions. In qualitative research, the sample size is frequently quite small as well. Even though the study may have useful insights on managing data quality-focused research in the given context, it may not be entirely representative of the general population or relevant to other contexts.

Researchers in the future should approach studies on data management and data quality by using a methodology that incorporates qualitative research techniques. This can entail conducting focus groups or interviews with persons who have first-hand knowledge of problems with data management and data quality. To comprehend the research problem more fully, it is critical to collect their viewpoints and perspectives. To make sure the results are representative and generalisable to other contexts and settings, researchers should also think about examining various contexts and settings. The next generation of researchers can make significant contributions to the fields of data management and data quality by using a strong qualitative research methodology.

Overall, the qualitative research approach employed in this study provided useful insights into the intricate field of data management for research with a focus on data quality. It offers a thorough examination of the experiences and viewpoints of participants, enabling a thorough comprehension of the research problem.

This study has shown that for an organisation in a developing country, the idea of data quality and quality assurance is still a relatively new one. It has also demonstrated that although researchers are aware of the significance of RDM, they require institutional support to manage their data management planning more effectively.

The developed quality process model, with its various layers and quality assurance components, originally described in Chapter 3 (Section 3.8) and further updated had to be modified in light of the comments and feedback provided by research participants (documented in Chapter 5). The feedback showed that the initial quality process framework was too complex, making it difficult for responders to understand and follow. The use of various loops was found to make it challenging to understand the process framework. The quality process model was then adjusted in Sections 6.7.10 - 6.9 to make the model easier to follow based on the suggestions made by the research participants. Implementing a framework for data quality-focused research data management can greatly benefit the Social Sciences Research Organisation. This framework prioritises data quality, has established clear guidelines for data management planning, and ensures that researchers and curators are aware of proper data handling procedures. By implementing such a framework, the social sciences institution can ensure that their research data is well-managed, stored securely, and easily accessible when needed.

6.14 Chapter summary

The main research question and the sub-research questions allowed for the completion of this research project. Each chapter made an effort to address the primary research topic, highlight the importance of the study to the discipline, and make suggestions for further study.

Major international RDM role-players are actively addressing data quality issues through a number of approaches and initiatives. In order to maintain the validity and reliability of research findings, it is essential that research methodologies that are becoming more digital are linked with best practices. When researchers and organisations are informed and proficient in adopting RDM best practice procedures, they are better equipped to provide data that is of high quality.

Pilot and pre-testing of research documents and data-gathering techniques should be done by researchers. The structure of all the documents that will support quality assurance should be decided, as well as their various forms. Data curators have to play a more active role in providing RDM and quality control support services to researchers when they are planning data management, and when they are collecting data. The establishment of a training unit will provide an ongoing training platform for researchers. This will have a great impact on the RDM skills development of researchers and improve their knowledge of RDM best practices, standards and guidelines.

To ensure the validity and legitimacy of research findings, policy choices, and public services, South African academic institutions are actively addressing data quality challenges. However, the variance in

RDM maturity levels amongst institutions has resulted in an unbalanced RDM environment across the nation. Government departments should broaden the scope of their data management efforts and work more closely with groups that have already started to advance their RDM activities. Governmental agencies will be better able to manage data quality, accessibility, and consumption through a shift in perspective.

Researchers need guidance from data curators on how to address issues that could compromise the quality of the data. It is also advised that the data curation unit designate someone or give current curators a new task to regularly monitor global RDM trends. To enable the data curation unit to provide fresh recommendations to researchers, a report on new RDM trends should be generated every three months. This will guarantee that the institution for social sciences remains current with all emerging RDM trends and best practices.

Change management recommendations for a social sciences institution call for thorough planning and execution of changes in order to enhance operations and adapt to changing needs. The Institute for Social Sciences must have a clear vision for the reforms and specific goals. The institution must make sure to convey these objectives to all stakeholders involved in order to guarantee that everyone is on the same page with regard to the anticipated RDM and data quality results. In order to support researchers in getting aligned with the changes that are occurring, ongoing training and support should be provided to enable faculty and staff to gain the skills necessary to adapt to new processes or technology brought about by proposed changes.

By offering a framework to evaluate and enhance the overall quality of the data, the data quality dimensions that were selected for this research study can help with data quality. Using the data quality dimensions and associated measuring criteria (Appendix 2), researchers can increase data quality by identifying and fixing mistakes in data collection and processing. The data quality dimensions are an essential criterion for ensuring that researchers can produce relevant findings that are unambiguous, can be applied in many scientific contexts, and can be used to address open research topics.

References

- Abawi, K. 2013. *Data collection instruments (Questionnaire and Interview)*. Available from <https://www.gfmer.ch/SRH-Course-2012/Geneva-Workshop/pdf/Data-collection-instruments-Abawi-2013.pdf> [accessed 27 September 2018].
- Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M. & Tang, N. 2018. Detecting data errors: where are we and what needs to be done? *VLDB Endowment*, Vol. 9, no. 12.
- Abiteboul, S., Arenas, M, Barcelo, P., Bienvenu, M., Calvanese, D., David, C., Hull, R. & Hullermeier, M. 2016. Research directions for principles of data management (Abridged). *SIGMOD Record*, Vol. 45, no. 4.
- Abraham, R., Schneider, J. & vom Brocke, J. 2019. Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, Vol. 49, pp. 424-438.
- Abrams, S. 2020. *Merritt and Dryad certified by CoreTrustSeal as trustworthy repositories*. Available from <https://cdlib.org/cdlinfo/2018/08/30/coretrustseal/> [accessed 9 April 2020].
- Adom, D., Yeboah, A. & Ankrah, A.K. 2016. Constructivism philosophical paradigm: implication for research, teaching and learning. *Global Journal of Arts Humanities and Social Sciences*, Vol. 4, no.10.
- Adom, D. & Hussein, E.K. 2018. Theoretical and conceptual framework: mandatory ingredients of a quality research. *International Journal of Scientific Research*, Vol. 7, no. 1.
- Adu, P. 2015. *Conducting qualitative research: decisions, actions, and implications*. Available from <https://www.youtube.com/watch?v=z4YIrykEL8M> [accessed 17 March 2020].
- Agnemyr, H. & Bhargava, A. 2023. *Planning for the unexpected*. Available from https://opticos.se/uploads/2023/04/Planning-for-the-unexpected_14-April-2023-FINAL.pdf [accessed 08 July 2024].
- Algonquin College. 2023. *Research data management strategy*. Available from <https://www.algonquincollege.com/applied-research/files/2023/09/F23-AC-Institutional-RDM-Strategy.pdf> [accessed 18 November 2023].
- Alshenqeeti, H. 2014. Interviewing as a data collection method: A critical review. *English Linguistics Research*, Vol. 3, no. 1.

- Alvi, M. 2016. *A manual for selecting sampling techniques in research*. Available from https://mpra.ub.uni-muenchen.de/70218/1/MPRA_paper_70218.pdf [accessed 7 October 2020].
- American Productivity & Quality Center. 2022. *What are the different types of process models?* Available from <https://www.apqc.org/blog/what-are-different-types-process-models> [accessed 21 February 2020].
- Anderson, R. 2014. Research on structure in teaching. *Journal of Research in Science Teaching*, Vol. 11, no. 3.
- Andrews, L., Higgins, A., Andrews, M.W. & Lalor, J.G. 2012. Classic grounded theory to analyse secondary data: reality and reflections. *The Grounded Theory Review*, Vol. 11, no. 1.
- Anfara, V.A. 2008. *Theoretical frameworks*. Thousand Oaks: Sage.
- Armenakis, A.A. & Harris, S.G. 2009. Reflections: our journey in organizational change research and practice. *Journal of Change Management*, Vol. 9, no. 2, pp. 127-142.
- Asher, A., Deards, K., Esteva, M., Halbert, M., Jahnke, L., Jordan, C., Keralis, S., Kulasekaran, S., Moen, W., Stark, S., Urban, T. & Walling, D. 2013. *Council on library and information resources research data management principles, practices, and prospects*. Available from <https://www.clir.org/wp-content/uploads/sites/6/pub160.pdf> [accessed 2 February 2023].
- Asiamah, N., Mensah, H.K. & Oteng-Abayie, E. 2017. General, target, and accessible population: demystifying the concepts for effective sampling. *The Qualitative Report*, Vol. 22, no. 6, pp. 1607-1621.
- Atieno, O. 2009. *An analysis of the strengths and limitation of qualitative and quantitative research paradigms*. Available from http://www.scientiasocialis.it/pec/node/files/pdf/Atieno_Vol.13.pdf [accessed 6 June 2020].
- Austin, Z. and Sutton, J. 2014. Qualitative research: getting started. *Canadian Journal of Hospital Pharmacy*, Vol. 67, no. 6.
- Australian Data Archive (ADA). 2018. *Australian data archive: learn more*. Available from <https://ada.edu.au> [accessed 23 April 2020].
- Australian National Data Service. 2015. *Research Data Management Framework for Institutions*. Available from <https://ardc.edu.au/resource/research-data-management-framework-for-institutions/#> [accessed 10 July 2024].
- Australian Research Council. 2016. *Future fellowships: Australian Research Council*. Available from <http://www.arc.gov.au/future-fellowships> [accessed 7 May 2018].

Australian Research Council. 2018. *Australian Code for the Responsible Conduct of Research*. Available from

<https://www.nhmrc.gov.au/sites/default/files/documents/attachments/grant%20documents/The-australian-code-for-the-responsible-conduct-of-research-2018.pdf> [accessed 10 October 2023].

Australian Research Council. 2022. *About the ARC*. Available from <https://www.arc.gov.au/about-arc> [accessed 10 October 2023].

Awre, C., Baxter, J., Clifford, B., Colclough, J., Cox, A., Dods, N., Drummond, P., Fox, Y., Gill, N., Gregory, K., Gurney, A., Harland, J., Khokhar, M., Lowe, D., O'Beirne, R., Proudfoot, R., Schwamm, H., Smith, A., Verbaan, E., Waller, L., Williamson, L., Wolf, M., & Zawadzki, M. 2015. Research Data Management as a “wicked problem”. *Library Review*, Vol. 64, no. 4/5, pp. 356-371. DOI: 10.1108/LR-04-2015-0043

Azimaee, M. 2015. *Good decisions require good data: process for data quality assurance*. Available from <https://studylib.net/doc/18797109/process-for-data-quality-assurance> [accessed 7 September 2016].

Ball, A. 2012. *Preservation and curation in institutional repositories, digital curation centre technology watch report*. Available from www.dpconline.org/advice/technology-watch-reports [last accessed 7 February 2016, since deactivated].

Balogun, T. 2023. Data management of digitized indigenous knowledge system in repositories. *Information Development*, Vol. 39, no. 3.

Ban, N.C. & Cox, M. 2017. Advancing social-ecological research through teaching: summary, observations, and challenges. *Ecology and Society*, Vol. 22, no. 1.

Barreiro, P. & Albandoz, J. 2001. *Population and sample: sampling techniques*. Available from <http://www.evisdom.com/pdf/000504273.pdf> [last accessed 25 August 2020, since deactivated].

Barsky, E. 2015. *Research data management dataguide*. Available from <https://open.library.ubc.ca/media/stream/pdf/494/1.0340263/5> [accessed 09 May 2018].

Bauman, J. 2019. *Data quality management: What you need to know*. Available from https://www.sas.com/en_zh/insights/articles/data-management/data-quality-management-what-you-need-to-know.html# [accessed 12 February 2020].

Becker, T. 2019. *Discussing data: evaluating data quality*. Available from <https://www.kansascityfed.org/documents/4986/tb-Discussing%20Data:%20Evaluating%20Data%20Quality.pdf> [accessed 31 October 2023].

- Beckhard, R. & Pritchard, W. 2012. *Changing the essence: The art of creating and leading fundamental changes in organizations*. Jossey-Bass: San Francisco.
- Beedham, H., Missen, J., Palmer, M. & Ruusaleep, R. 2015. *Assessment of UKDA and TNA compliance with OAIS and Mets standards*. Available from <https://www.yumpu.com/en/document/read/31308545/oais-mets-report-pdf-uk-data-archive> [accessed 22 November 2021].
- Bellamy, N. 2015. *Criterion validity*. Available from <https://www.sciencedirect.com/topics/nursing-and-health-professions/criterion-related-validity> [accessed 01 July 2020].
- Berg, B.L. 1998. *Qualitative research methods for the social sciences*. Boston: Allyn and Bacon.
- Bertino, E. 2014. Data security: challenges and research opportunities. In: Jonker, W. & Petkovic, M. eds. *SDM 2023*. Cham: Springer International Publishing 9–13.
- Bishop, B.W., Neish, P., Kim, J.H., Bats, R., Million, A.J., Carlson, J., Moulaison-Sandy, H. & Pham, M.T. 2023. Data management plan implementation, assessments, and evaluations: implications and recommendations. *Data Science Journal*, Vol. 22, no.27, pp. 1–8.
- Bishop, L. 2012. *Archiving your data: planning and managing the process*. Available from https://eprints.ncrm.ac.uk/id/eprint/3071/1/archiving_your_data29may12.pdf [accessed 24 October 2021].
- Black, A. & van Nederpelt, P. 2020. *Dimensions of data quality (DDQ)*. Available from <https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf> [accessed 31 October 2023].
- Bloom, T., Ganley, E. & Winker, M. 2014. Data access for the open access literature: PLOS's data policy. *PLoS Biology*, Vol. 12, pp. 2.
- Bolognese, A.F. 2002. *Employee resistance to organisational change*. Available from <https://www.newfoundations.com/OrgTheory/Bolognese721.html> [accessed 22 April 2020].
- Borghi J., Abrams, S., Lowenberg, D., Simms, S. & Chodacki, J. 2018. *Support your data: A research data management guide for researchers*. Available from <https://riojournal.com/article/26439/download/pdf/> [accessed 13 February 2020].
- Borgman, C.L. 2015. *Big data, little data, no data: scholarship in the networked world*. Cambridge: MIT Press.
- Bowen, G.A. 2009. Document analysis as a qualitative research method. *Qualitative Research Journal*, Vol. 9, no. 2.

- Brase, J. 2010. *Datacite: A global registration agency for research data*. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 149.
- Bryant, R., Brian L. & Rhinehart, A.K. 2023. Building research data management capacity: case studies. In *Strategic Library Collaboration*. Dublin: OCLC Research.
- Bryant, R., Lavoie, B. & Malpas, C. 2017. *The realities of research data management*. Dublin: OCLC Research.
- Buckey, D.J. 2016. *Data accuracy and quality*. Available from http://planet.botany.uwc.ac.za/nisl/gis/gis_primer/page_08.htm [last accessed 26 October 2018, since deactivated].
- Bui, Y. & Park, J-R. 2013. *An assessment of metadata quality: A case study of the National Science Digital Library metadata repository*. Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI. (Oct. 2013). DOI: 10.29173/cais166.
- Burns, N. & Grove, S.K. 1999. *Understanding nursing research*. Philadelphia: W.B. Saunders Company.
- Burton, P. 2023. Agile and organizational change management to increase project success. *PM World Journal*, Vol. 12, no. 8.
- Business Research Methodology. 2018. *Positivism research philosophy*. Available from <https://research-methodology.net/research-philosophy/positivism/> [accessed 6 March 2021].
- Cai, L. & Zhu, Y. 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, Vol. 14, no. 2, pp. 1-10.
- Cambridge University Press. 2020 *Preservation*. Available from [https://dictionary.cambridge.org/dictionary/english/preservation?q=preservation+ Oxford](https://dictionary.cambridge.org/dictionary/english/preservation?q=preservation+Oxford) [accessed 6 February 2024].
- Cambridge University Press & Assessment. 2024. *Meaning of researcher in English*. Available from: <https://dictionary.cambridge.org/dictionary/english/researcher> [accessed 10 July 2024].
- Canadian Institute for Health Information (CIHI). 2005. *Framework for data quality*. Available from https://secure.cihi.ca/free_products/dq-data_quality_framework_2009_en.pdf [accessed 3 April 2020].
- Carlson, J., Rolando, L., Hswe, P., Parham, S., Westra, B. & Whitmire, L. 2015. Data management plans as a research tool. *Association for Information Science and Technology*, Vol. 41, no. 5.

Carter, C.R. & Rogers, D. S. 2008. A framework of sustainable supply chain management: moving toward new theory. *International Journal of Physical Distribution and Logistics Management*, Vol. 38, no. 5.

Caulfield, J. 2023. *How to do thematic analysis: Step-by-step guide & examples*. Available from <https://www.scribbr.com/methodology/thematic-analysis/#:~:text=Thematic%20analysis%20is%20a%20method,meaning%20that%20come%20up%20repeatedly> [accessed 10 May 2022].

Center for Community Health and Development, University of Kansas. (n.d.). *Conducting Concerns Surveys*. Available from <http://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/conduct-concerns-surveys/main> [accessed 07 April 2021].

CEOS. 2012. Data life cycle models and concepts CEOS Version 1.1. Available from https://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/WGISS_Data-Lifecycle-Models-And-Concepts.pdf [accessed 23 June 2018].

Chand, M. 2017. *New paradigms in qualitative research*. Available from <https://www.sfu.ca/educ867/htm/paradigm.htm> [accessed 22 April 2020].

Chapman, A.D. 2005. *Principles of data quality*. Copenhagen: Global Biodiversity Information Facility.

Charles, V., Stiller, J., Kiraly, P., Bailer, W. & Freire, N. 2017. *Data quality assessment in Europeana: metrics for multilinguality*. Available from <http://ceur-ws.org/Vol-2038/paper6.pdf> [accessed 23 April 2020].

Chigwada, J., Chiparausha, B. & Kasiroori, J. 2017. Research Data Management in Research Institutions in Zimbabwe. *Data Science Journal*, Vol. 16, no. 31.

Chiware, E. & Mathe, Z. 2015. Academic libraries' role in research data management services: A South African perspective. *South African Journal of Libraries and Information Science*, Vol. 81, no. 2.

Clarke, R.A. 2014. *Data integrity means and practices*. Santa Clara: SNIA Data Management Forum. Available from http://www.digitalpreservation.gov/meetings/documents/othermeetings/3-4_Clarke-SUN-Data_Integrity.pdf [accessed 11 March 2018].

Cohn, M. 2022. *What is quality?* Available from <https://www.mountangoatsoftware.com/blog/what-is-quality> [accessed 07 January 2022].

College of the Rockies. 2023. *Researcher data management planning guidance*. Available from <https://cotr.bc.ca/wp-content/uploads/COTR-DMP-Plan-Guidance-13-January-2023 -w-Formatting-23-Jan-23.pdf> [accessed 1 December 2023].

Consortium of European Social Science Data Archives (CESSDA)/ represents the European Research Infrastructure Consortium (ERIC). 2017. *Delivering Research Data Management Services*. Available from <https://www.cessda.eu/Research-Infrastructure/Training/Research-Data-Management> [last accessed 19 June 2021, since deactivated].

Consultative Committee for Space Data Systems. 2011. *Audit and certification of trustworthy digital repositories*. Available from public.ccsds.org/publications/archive/652x0m1.pdf [accessed 14 April 2016].

Cooper, J.S. & Kahn, E. 2012. Commentary on issues in data quality analysis in life cycle assessment. *Springer-Verlag*, Vol. 17.

Corti, L., van den Eynden, V., Bishop, L. & Woollard, M. 2014. *Managing and sharing research data: a guide to good practice*. New York: Sage. 22.

Coulehan, M. & Wells, J. 2012. *Guidelines for responsible data management in scientific research*. Available from <https://ori.hhs.gov/images/ddblock/data.pdf> [accessed 23 April 2018].

Council for Scientific and Industrial Research (CSIR). 2019. *DIRISA launches a data management planning tool*. Available from <https://www.csir.co.za/dirisa-launches-data-management-planning-tool> [accessed 23 April 2020].

Crawford, L.M. 2020. Conceptual and theoretical frameworks in research. Available from https://uk.sagepub.com/sites/default/files/upm-assets/105274_book_item_105274.pdf [accessed 01 July 2021]

Creswell, J.W. 2009. *Research design: qualitative, quantitative, and mixed methods approaches*, 3rd edn. Thousand Oaks: Sage.

Creswell, J.W. 2014. *Research design: qualitative, quantitative and mixed methods approaches*. Thousand Oaks: Sage.

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A. & Sheikh, A. 2011 The case study approach. *BioMed Central Medical Research Methodology*, Vol. 11, no.100.

Curdt, C. & Hoffmeister, D. 2015. *Research data management services for a multidisciplinary, collaborative research project: design and implementation of the TR32DB project database*. Available from https://www.researchgate.net/profile/Dirk-Hoffmeister-2/publication/282657848_Research_data_management_services_for_a_multidisciplinary_collaborativ

[e research project Design and implementation of the TR32DB project database/links/565d675c08ae4988a7bbe0a9/Research-data-management-services-for-a-multidisciplinary-collaborative-research-project-Design-and-implementation-of-the-TR32DB-project-database.pdf? tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19](https://www.research-project.com/Design-and-implementation-of-the-TR32DB-project-database/links/565d675c08ae4988a7bbe0a9/Research-data-management-services-for-a-multidisciplinary-collaborative-research-project-Design-and-implementation-of-the-TR32DB-project-database.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) [accessed 13 December 2023].

Curtis, J. 2021. *How to measure data quality*. Available from <https://www.itpro.co.uk/business-intelligence-bi/29773/how-to-measure-data-quality> [accessed 07 January 2022].

DAMA UK Ltd. 2017. *About*. Available from <https://www.dama-uk.org/about-DAMA-uk> [accessed 13 August 2018].

Daniel, H. 2023. Enhancing data integrity through data governance in clinical trials: A computational approach. *International Journal of Computer Science and Technology (IJCT)*, Vol. 7, no. 3.

Data Archiving and Networked Services (DANS). 2010. *DANS Data Guide 8*. Leiden: Pallas Publications.

DataFirst. 2020. *Welcome to DataFirst*. Available from www.datafirst.uct.ac.za [accessed 26 April 2020].

DataOne. 2017. *Best practice: Discover*. Available from https://dataoneorg.github.io/Education/bp_step/discover/ [accessed 09 July 2024].

Data Seal of Approval (DSA) – World Data System (WDS) (DSA-WDS). 2016. *Core Trustworthy Data Repositories Requirements*. Available from https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf [accessed 21 July 2018].

David, R., Richard, A., Connellan, C., Lauer, K. & Chiusano, M., L. 2023. Umbrella Data Management Plans to Integrate FAIR Data: Lessons from the ISIDORE and BY-COVID Consortia for Pandemic Preparedness., *CODATA Data Science Journal*.

Decman, M. & Vintar, M. 2013. A possible solution for digital preservation of e-government: A centralised repository within a cloud computing framework. *Emerald Group Publishing Limited*, Vol. 65, no. 4.

De Lemos, R. & Potena, P. 2017. Identifying and handling uncertainties in the feedback control loop. In: Misrik, I., Ali, N., Kazamn, R., Grundy, J. & Schmerl, B. (eds.). *Managing trade-offs in adaptable software architectures*. Amsterdam: Elsevier. 353-367.

Denzin, N.K. & Lincoln, Y.S. 2000. *Handbook of qualitative research*. London: Sage.

- Digital Curation Centre (DCC). 2008. *The DCC curation lifecycle model*. Available from <https://www.era.lib.in.c.uk/bitstream/handle/1842/3350/Higgins%20DCCLifecycle.pdf?sequence=1> [accessed: 22 January 2018].
- Digital Curation Centre. 2023. *About DMPonline*. Available from https://dmponline.dcc.ac.uk/about_us [accessed 13 December 2023].
- Digital Repository Ireland. 2023. *About Us*. Available from <https://dri.ie/about-us-2/> [accessed 10 October 2023].
- DIRISA. 2017. *About us*. Available from <https://www.dirisa.ac.za/aboutus/> [accessed 19 December 2017].
- Divecha, V. 2016. *Data integrity: view of the manufacturer*. Available from https://www.unicef.org/supply/files/2_Data_Integrity_Manufacturer.pdf [last accessed 29 August 2018, since deactivated].
- Domingus, M. 2015. *Checklist for quality assurance and quality control: An essential part of Research Data Management (RDM) prior to, during and after a research project*. Research Support Office EUR. Available from https://www.eur.nl/sites/corporate/files/2017-11/Checklist_QA_QC_RDM_0_2_MD_20150129.pdf [accessed 11 February 2024]
- Donalek, J. G. 2004. Demystifying nursing research: phenomenology as a qualitative research method. *Urologic Nursing*, Vol. 24, pp. 516–517.
- Donnelly, M., Jones, S. & Pattenden-Fail, J.W. 2010. DMP online: The Digital Curation Centre's web-based tool for creating, maintaining and exporting data management plans, *International Journal of Digital Curation*, Vol. 5, no. 1.
- Donner, E.K. 2022. Research data management systems and the organization of universities and research institutes: A systematic literature review. *Journal of Librarianship and Information Science*, Vol. 1, no. 21.
- Doorn, P. & Dillo, I. 2017. *Assessing the FAIRness of datasets in trustworthy digital repositories: A 5-star scale*. Available from https://indico.cern.ch/event/588219/contributions/2384979/attachments/1426152/2188462/Dillo_Door_n_-_Assessing_FAIRness_CERN_Geneva_13-03-2017_-3.pdf [accessed 13 October 2019].
- Dora, M. & Kumar, H.A. 2015. *Managing research data in academic institutions: role of libraries*. Available from <http://eprints.rclis.org/24911/1/50.pdf> [accessed 18 July 2020].
- Dörnyei, Z. 2007. *Research methods in applied linguistics: quantitative qualitative, and mixed methodologies*. Oxford: Oxford University Press.

- Downs, R., Ramapriyan, H.K., Peng, G. & Wei, Y. 2021. Perspectives on citizen science data quality. *Frontiers in Climate*, Vol. 3, no. 6.
- Dunphy, D.C. 2006. Organizational change in corporate settings. *Human Relations*, Vol. 49, no. 5, pp. 541-542.
- Duranti, L., Suderman, L. & Todd, M. 2008. *A framework of principles for the development of policies, strategies and standards for the long-term preservation of digital records*. Available from [http://www.interpares.org/public_documents/ip2\(pub\)policy_framework_document.pdf](http://www.interpares.org/public_documents/ip2(pub)policy_framework_document.pdf) [accessed 12 August 2020].
- Edge Hill University. 2013. *Research data management guidelines*. Available from <https://www.edgehill.ac.uk/wp-content/uploads/documents/RDM-Guidance-RO-GOV-14-2.docx> [accessed 24 October 2018].
- Edwards, J. 2020. *Creating an organizational structure*. Available from <https://opentextbc.ca/strategicmanagement/chapter/creating-an-organizational-structure/> [accessed 23 November 2021].
- Eisenhardt, K.M. 2009. Building theories from case study research. *Academy of Management Review*, Vol. 14, no. 4.
- Elite Institute. 2019. *Systematic sampling, convenience sampling, and purposive sampling*. Available from <http://www.elitemv.com/2019/09/systematic-sampling-convenience.html#:~:text=Purposive%20sampling%20is%20a%20non,in%20saving%20time%20and%20money%E2%80%9D>. [last accessed 02 March 2022, since deactivated].
- Emerson, M.S. 2022. *7 reasons why change management strategies fail and how to avoid them*. Available from <https://professional.dce.harvard.edu/blog/7-reasons-why-change-management-strategies-fail-and-how-to-avoid-them/> [accessed 18 December 2023].
- Etheridge, A. 2016. *Bringing data to life: lessons from the UK Data Service*. Available from <https://www.questia.com/magazine/1P3-4070835901/bringing-data-to-life-lessons-from-the-uk-data-service> [last accessed 26 July 2018, since deactivated].
- European Commission. 2023. *Archiving by design*. Available from https://commission.europa.eu/system/files/2023-06/Whitepaper%20AbD_en.pdf [accessed 18 November 2023].
- European Research Council (ERC). 2018. *Open research data and data management plans*. Available from https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf [accessed 3 March 2019].

Ewim, D.R., Orikpete, O.F., Scott, T.O., Onyebuchi, C.N., Onukogu, A.O., Uzougbo, C.G. & Onunka, C. 2023. Survey of wastewater issues due to oil spills and pollution in the Niger Delta area of Nigeria: a secondary data analysis. *Bulletin of the National Research Centre*, Vol. 47, no. 116.

Explorable.com. 2009. *Non-probability sampling*. Available from <https://explorable.com/non-probability-sampling> [accessed 01 March 2022].

Farha, R., Sekeris, E. & Hermansson, P.D. 2017. *The hidden cost of reputation risk*. Available from <https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2017/jul/Reputational%20Risk.pdf> [accessed 3 May 2019].

Farid, G., Warraich, N.F. & Iftikhar, S. 2023. Digital information security management policy in academic libraries: A systematic review (2010 – 2022). *Journal of Information Science*. DOI: 10.1177/01655515231160026.

Farooq, U. 2013. *What is interview schedule, definitions and types*. Available from <http://www.studylecturenates.com/social-research-methodology/what-is-interview-schedule-definition-types> [accessed 13 January 2021].

Farren, A.T. 2015. Leininger's ethnonursing research methodology and studies of cancer survivors: A review. *Journal of Transcultural Nursing*, Vol. 26, no. 4, pp. 418-427. DOI: 10.1177/1043659614524254

Federation of American Scientists. 2020. *Process models: Introduction to process models*. Available from <https://irp.fas.org/doddir/army/strap/strpsec3.htm#:~:text=Process%20Models%20are%20graphical%20representations,information%20needs%20within%20the%20enterprise> [accessed 23 November 2021].

Fellows, R. F., & Liu, A., M., M. 2015. *Research Methods for Construction*. 4th edn. New Jersey: Wiley Blackwell.

Finlay, L. 2006. "Rigour", "ethical integrity" or "artistry"? Reflexively reviewing criteria for evaluating qualitative research. *British Journal of Occupational Therapy*, Vol. 69, no. 7.

Fischer, J. Egli, L., Groth, J., Barrasso, C., Ehrmann, S., Figgemeier, H., Henzen, C., Meyer, C., Muller-Pfefferkorn, R., Rummler, A., Wagner, M., Bernard, L. & Seppelt, R. 2023. Approaches and tools for user-driven provenance and data quality information in spatial data infrastructures. *International Journal of Digital Earth*, Vol. 16, no. 1, pp. 1510-1529.

Flick, U. 2013. *The SAGE handbook of qualitative data analysis Uwe Flick*. Available from http://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/qualitative_sozial-

[bildungsforschung/Medien/58869_Flick_The_SAGE_HB_of_Qualitative_Data_Analysis_Chapter1_mapping-the-field.pdf](#) [accessed 18 September 2018].

Flores, J.R., Brodeur, J.J., Daniels, N. & Turnator, E. 2015. Libraries and the research data management landscape. In: Maclachlan, J.C., Waraksa, E.A & Williford, C. eds. *The process of discovery: The CLIR postdoctoral fellowship program and the future of the academy*. Alexandria: Council on Library and Information Resources. 82-102.

Ford, D. & Saren, M. 2006. *Technology strategy for business*. Cambridge: International Thompson Business Press.

Fowler, F.J. 1995. *Improving survey questions: design and evaluation*. Thousand Oaks: Sage.

Frankfort-Nachmias, C. & Nachmias, D. (1996). *Research methods in the social sciences*. New York: St. Martin's Press.

Friesike, S. 2017. Creative commons licences. Available from https://www.researchgate.net/profile/Sascha-Friesike/publication/312685639_Creative_Commons_Licences/links/59945113458515c0ce6528bd/Creative-Commons-Licences.pdf [accessed 24 January 2022].

Gabriel, A. 2008 The meaning of theory. *Sociological Theory*, Vol. 26.

Galli, B.J. 2021. Effective strategies for communicating and managing communication in a project team: my perspective. *International Journal of Applied Industrial Engineering*, Vol. 8, no. 1.

Gangrade, K.D. 2012. Methods of data collection: questionnaire and schedule. *Journal of the Indian Law Institute*, Vol. 24, no. 4.

Gassman, J., Walter, W., Kuntz, T. & Amoroso, M. 2015. *Data quality assurance, monitoring, and reporting*. Amsterdam: Elsevier.

Glasow, P.A. 2005. *Fundamentals of survey research methodology*. Available from https://www.mitre.org/sites/default/files/pdf/05_0638.pdf [accessed 24 September 2021].

Goertz, G. & Mahoney, J. 2012. *A tale of two cultures: qualitative and quantitative research in the social sciences*. Princeton: Princeton University Press.

Grant, C. & Osanloo, A. 2014. *Understanding, selecting, and integrating a theoretical framework in dissertation research: creating the blueprint for your "house"*. Available from <https://files.eric.ed.gov/fulltext/EJ1058505.pdf> [accessed 18 March 2020].

Green, H.E. 2014. *Use of theoretical and conceptual frameworks in qualitative research*. Available from <https://www.researchgate.net/profile/Helen-Green->

[8/publication/264245293_Use_of_theoretical_and_conceptual_frameworks_in_qualitative_research/links/5694cbf408ae3ad8e33c41db/Use-of-theoretical-and-conceptual-frameworks-in-qualitative-research.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19](https://www.researchgate.net/publication/264245293_Use_of_theoretical_and_conceptual_frameworks_in_qualitative_research/links/5694cbf408ae3ad8e33c41db/Use-of-theoretical-and-conceptual-frameworks-in-qualitative-research.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) [accessed 24 August 2020].

Gutema, D.M. 2023. *Data governance practices in higher education institutions: opportunities and challenges*. Available from https://www.doria.fi/bitstream/handle/10024/187361/gutema_dandi.pdf?sequence=3 and [isAllowed=y](#) [accessed 19 November 2023].

Guy, M. 2013. *RDM Training for Librarians*. Available from <http://www.dcc.ac.uk/sites/default/files/documents/publications/case-studies/Training-for-librarians-case-study.pdf> [accessed: 6 May 2018].

Hagedorn, G., Mietchen, D., Morris, R.A., Agosti, D., Penev, L., Berendsohn, W.G. & Hobern, D. 2011. Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *Biodiversity Science*, Vol. 150, pp. 127-149.

Hancock, D.R. & Algozzine, B. 2006. *Doing case study research: A practical guide for beginning researchers*. New York: Teachers College Press.

Handy, C. 1978. The gods of management. *Executive Book Summaries*, Vol. 18, no. 2, pp. 1-8.

Haradhan, M. 2017. Two criteria for good measurements in research: validity and reliability. Available from https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3152355_code1353801.pdf?abstractid=3152355&mirid=1 [accessed 10 February 2024].

Harding, J. 2013. *Qualitative data analysis: From start to finish*. New Delhi: Sage.

Harling, K. 2012. *An Overview of Case Study*. Available from https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2141476_code1284569.pdf?abstractid=2141476&mirid=1 [accessed 19 June 2018].

Harrell, M.C. & Bradley, M.A. 2009. Data collection methods, semi-structured interviews and focus groups: Available from https://www.rand.org/pubs/technical_reports/TR718.html [accessed 10 February 2024].

Hassenstein, M.J. & Vanella, P. 2022. Data quality: concepts and problems. *Encyclopedia*, Vol.2, pp. 498-510.

Hermelin, S. (1993). Scientific quality in the eyes of the scientist. *Scientometrics*, Vol. 27, no. 1, pp. 3-18.

- Higgins, S. 2008. The DCC curation lifecycle model. *The International Journal of Digital Curation*, Vol. 3, no. 1.
- Hodson, S. & Molloy, L. 2014. *Current best practice for research data management policies*. Available from <https://zenodo.org/records/27872s> [accessed 14 November 2018].
- Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. 2018. *The Dataset nutrition label: A framework to drive higher data quality standards*. Available from <https://arxiv.org/ftp/arxiv/papers/1805/1805.03677.pdf> [accessed 17 May 2020].
- Honebein, P.C. 1996. *Seven goals for the design of constructivist learning environments*. New Jersey: Educational Technology Publications.
- Hubbart, J.A. 2023. Organizational change: The challenge of change aversion. *Administrative Sciences*, Vol. 13, no. 162.
- Hugo, W. 2012. *Towards 2011: An IT strategy for SAEON*. Available from <http://www.saeon.ac.za/enewsletter/archives/2008/november-2008/saeon-it-strat.pdf> [accessed 16 April 2018].
- Hultman, K.E. 1995. Scaling the wall of resistance. *Training and Development*, Vol. 49, no. 10.
- Intergovernmental Panel on Climate Change (IPCC). (1997). *IPCC guidelines for national greenhouse gas inventories*. Available from https://www.ipcc-nggip.iges.or.jp/public/gp/english/8_QA-QC.pdf [accessed 07 January 2022].
- Inter-university Consortium for Political and Social Research (ICPSR). 2018. *Guide to Social Science Data Preparation and Archiving*, 6th edn. Available from <https://www.icpsr.umich.edu/files/deposit/dataprep.pdf> [accessed 21 April 2020].
- International Organization for Standardization (ISO). 2018. *Technical specification ISO/TS 8000-150 Data Quality: Part 150: Master data: Quality management framework*. Available from <https://www.iso.org/obp/ui/#iso:std:iso:ts:8000:-150:ed-1:v1:en> [accessed 14 April 2020].
- International Organization for Standardization (ISO). 2019. *Data quality: Part 63: Data quality management: Process measurement*. Available from <https://www.iso.org/obp/ui/en/#iso:std:iso:8000:-63:ed-1:v1:en> [accessed 17 November 2023].
- Inter-university Consortium for Political and Social Research (ICPSR). 2018. *Guide to social science data preparation and archiving: best practice throughout the data life cycle*. 5th edn. Ann Arbor: University of Michigan.
- Jabareen, Y. 2009. Building a conceptual framework: philosophy, definitions, and procedure. *International Journal of Qualitative Methods*, Vol. 8, no. 4.

- Jayawardene, S., Sadiq, S. & Indulska, M. 2016. *An Analysis of data quality dimensions*. Available from https://espace.library.uq.edu.au/data/UQ_312314/UQ312314_UPDATED_2015_02.pdf?dsi_version=ef3ceaacb5d3741215cbb5d5fed2af2&Expires=1707854371&Key-Pair-Id=APKAJKNB4MJBNC6NLQ&Signature=hFouK~o~VEzO9wfQi4~MSKQmyWakzKgZUamhIdfwrNDHZwrNVTzq7V-N8CheU6jW8ZyqI8ujEZF2WbhaDMUkeV96hU5l59KkMUg9Ztp8YtFm86gEqmrP3-hbvwMGE5fN9OrD15MWVf7FOXn97sSne6PN7iXNSMdlI5ydpzl5Kwg3RAyH75bbqvsDpz55nlsgRM4sz3tAwV~JOQBKvl0xVbWZIFS-TMPiwV93j0SXout9qIleWk3l9b2jDvlOteaoe-eFtfVq2a4JJGyO3Ihtmob9NitA~eH750b3ysMZlypN~J4j0~gYd89QOAmcHyaaCRnB-KSQcWldH5~D-Y5Lsg [accessed 10 February 2024].
- Jeyadevi, C. 2023. *The significance of data analysis in decision making: An insight*. Available from https://www.researchgate.net/profile/Christina-Jeyadevi-J/publication/373825278_The_significance_of_Data_Analysis_in_Decision_Making_-_An_Insight/links/64feef91849bbb203b8fcd91/The-significance-of-Data-Analysis-in-Decison-Making-An-Insight.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 30 November 2023].
- JISC. 2017. *Security of research data*. Available from <https://www.jisc.ac.uk/guides/data-protection-and-research-data/security-of-research-data> [accessed 20 January 2018].
- Johnsson, M. & Ahlfeldt, J. 2015. *Research libraries and research data management within the humanities and social sciences: project report*. Available from <http://lup.lub.lu.se/record/5050462> [accessed 2 January 2017].
- Johnston, M.P. 2014. Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries (QQML)*, Vol.3, no. 619, pp. 619-626.
- Kabanda, S.M., Cengiz, N., Rajaratnam, K., Watson, B.W., Brown, Q., Esterhuizen, T.M. & Moodley, K. 2023. Data sharing and data governance in sub-Saharan Africa: Perspectives from researchers and scientists engaged in data-intensive research. *South African Journal of Science*, Vol. 119, no. 5/6.
- Kaboub, F. 2008. *Positivist paradigm*. Available from <http://personal.denison.edu/~kaboubf/Pub/2008-Positivist-Paradigm> [accessed 6 April 2021].
- Kahn, M., Higgs, R., Davidson, J. & Jones, S. 2014. Research data management in South Africa: How we shape up. *Australian Academic and Research Libraries*, Vol. 45, no. 4, pp. 296-308.

- Kalender, M. 2007. Applying the subject cell through constructivist approach during science lessons and the teacher's view. *Journal of Environmental and Science Education*, Vol. 2, no. 1.
- Kaplan, D. 2004. *The SAGE handbook of quantitative methodology for the social sciences*. New York: SAGE College.
- Kaplan University. 2017. *Protecting human subjects in research*. Available from <https://slideplayer.com/slide/9713004/> [accessed 18 April 2020].
- Karimova, Y. 2023. *Research data description in multiple domains: supporting researchers with data management plans*. Available from <https://repositorio-aberto.up.pt/bitstream/10216/152809/2/642174.pdf> [accessed 09 July 2024].
- Kent, F. 2023. *Understanding the basics: What are digital file formats?* Available from <https://www.folderit.com/blog/digital-file-formats-2023-updated-guide-to-conversion-projects/> [accessed 11 December 2023].
- Kim, Y. 2010. *The pilot study in qualitative inquiry*. Available from <https://www.coursehero.com/file/112638800/qualitative-Pilot-studypdf/> [accessed 24 October (2021)].
- Kivunja, C. & Kuyini, A.B. 2017. Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, Vol. 6, no. 5.
- Kivunja, C. 2018. Distinguishing between theory, theoretical framework, and conceptual framework: A systematic review of lessons from the field. *International Journal of Higher Education*, Vol. 7, no. 6.
- Knight, G. 2015. Building a research data management service for the London School of Hygiene and Tropical Medicine. *Emerald*, Vol. 49, no. 4, pp. 424–439.
- Kotter, J.P. 1997. Leading change: A conversation with John P. Kotter. *Strategy and Leadership*, Vol. 25, no. 1, pp. 18-23.
- Kumarasinghe, H.P.N.I. & Dilan, H.K.T. 2022. Organizational change and change management. In: *Contemporary Innovations in Management*. Tamilnadu: Global Learners Academy of Development: 49-69.
- Kumar, R. 2011. *Research methodology: A step-by-step guide for beginners*. 3rd edn. London: Sage.
- Kvale, S. & Brinkmann, S. 2009. *InterViews: Learning the craft of qualitative research*. 3rd edn. Los Angeles: Sage.

- Laliberté, L., Grünewald, W. & Probst, L. 2004. *Data quality: a comparison of IMF's data quality assessment framework (dqaf) and eurostat's quality definition*. Available from <https://www.imf.org/external/pubs/ft/bop/2003/dataq.pdf> [accessed 31 October 2023].
- Latham, N. 2014. *A Practical guide to evaluating systems change in a human services system context*. Available from http://www.pointk.org/resources/files/Latham_Human_Services_Systems.pdf [accessed 3 February 2018].
- Lawton, G. & Pratt, M.K. 2023. *What is change management?* Available from <https://www.techtarget.com/searchcio/definition/change-management>. [accessed 13 December 2023].
- Leedy, P.D. & Ormrod, J.E. 2016. *Practical research: Planning and design*. 11th edn. Boston: Pearson.
- Lewis, M.J. 2010. *Libraries and the management of research data*. In: McKnight, S. ed. *Envisioning future academic library services: Initiatives, ideas and challenges*. London: Facet Publishing.
- LIASA. 2017. *About*. Available from <http://www.liasa.org.za/about/> [accessed 3 February 2018]
- Lin, W. & Yao, G. 2020. *Concurrent validity*. Available from https://link.springer.com/referenceworkentry/10.1007%2F978-94-007-0753-5_516 [accessed 01 July 2020].
- LinkedIn Corporation. 2023. *Trends in data quality management for 2023*. Available from <https://www.linkedin.com/pulse/trends-data-quality-management-2023-master-data-management/> [accessed 12 December 2023].
- Lötter, L. 2014. *Reflections on the RDM position in South Africa*. Paper presented at the LIASA Research Data Management Workshop, Cape Town, South Africa, March 27. Available <https://repository.hsrc.ac.za/bitstream/handle/20.500.11910/2405/8757.pdf?sequence=1&isAllowed=y> [accessed 22 December 2017].
- Lyke-Ho-Gland, H. 2015. A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, Vol. 47, no. 2.
- Mabin, V.J., Forgeson, S. & Green, L. 2001. Harnessing resistance: Using the theory of constraints to assist change management. *Journal of European Industrial Training*, Vol. 25, no.2, pp.168-191.
- Maggetti, M., Gilardi, F. & Radaelli, C.C. 2013. *Designing research in the social sciences*. New York: Sage.
- Mandinach, E.B., Honey, M. & Light, D. 2006. *A theoretical framework for data-driven decision making*. Available from

https://cct.edc.org/sites/cct.edc.org/files/publications/DataFrame_AERA06.pdf [accessed 23 July 2020].

MANTRA. 2017. *About*. Available from <https://mantra.ed.ac.uk/about.html> [accessed 15 March 2018].

Margoni, T. & Schirru, L. 2023. *The role of licensing in data FAIRization*. Available from <https://lirias.kuleuven.be/retrieve/694798> [accessed 11 February 2024].

Marton, A. 2013. *Purposive selection and evaluation criteria for qualitative research*. Available from https://www.researchgate.net/profile/Attila-Martons/publication/255717521_Purposive_Selection_and_Evaluation_Criteria_for_Qualitative_Research/links/00b4952063398c87c9000000/Purposive-Selection-and-Evaluation-Criteria-for-Qualitative-Research.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 22 November 2023].

Masinde, J., Chen, J., Wambiri, D. & Mumo, A. 2021. Research librarians' experiences of research data management activities at an academic library in a developing country. *Data and Information Management*. Vol. 5, no. 4, pp. 412–424. doi.org/10.2478/dim-2021-0002.

Mathers, N., Fox, N. & Hunn, A. 2002. *Trent focus for research and development in primary health care: Using interviews in a research project*. 2nd edn. Nottingham: Trent Focus.

Mathiyazhagan, T. & Nandan, D. 2010. Survey research method. *Media Mimansa*, Vol. 4, pp. 34-45.

Maydanchik, A. 2007. *Data quality assessment*. Bradley Beach: Technics.

Mayernik, M.S. 2015. Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, Vol. 67, no. 4.

Mayson, J. 2002. *Qualitative researching*. 2nd edn. London: Sage.

Mazikana, A.T. 2023. *Promoting the use of change management strategy to achieve sustainable competitive advantage of organisation: A case of banking organizations in Harare, Zimbabwe*. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4377799 [last accessed 11 February 2024, general access since restricted].

McCord, S.E., Welty, J.L., Courtwright, J., Dillon, C., Traynor, A., Burnett, S.H., Courtwright, E.M., Fults, G., Karl, J.W., van Zee, J.W., Webb, N.P. & Tweedie, C. 2022. Ten practical questions to improve. *Rangelands*, Vol. 44, no. 1, pp. 17–28.

McDonald, K. 2018. *Process models*. Available from <https://www.kbp.media/process-model/> [accessed 24 November 2019].

- McKnight, W. 2010. *Improving data quality through data modeling*. Available from https://www.odbms.org/wp-content/uploads/2013/11/Data_Quality_McKnight_ERwin.pdf [accessed 15 March 2018].
- McMaster University. 2023. *McMaster University Research Data Management (RDM) institutional strategy 2023 -2025*. Available from <https://macsphere.mcmaster.ca/handle/11375/28392> [accessed 19 November 2023].
- Menna, I., Yehia, M. & Doaa, S.E. 2021. Data quality dimensions, metrics, and improvement techniques. *Future Computing and Informatics Journal*, Vol. 6, no. 3.
- Merriam, S.B. 2001. *Qualitative research and case study applications in education*. San Francisco: Jossey-Bass.
- Merriam-Webster. 2020. *Storing*. Available from <https://www.merriam-webster.com/dictionary/storing> [accessed: 15 June 2020].
- Micheli, M., Hupont, I., Delipetrev, B. & Soler-Garrido, J. 2023. The landscape of data and AI documentation approaches in the European policy context. *Ethics and Information Technology*, Vol. 25, no. 56.
- Middleton, F. 2023. *The 4 types of validity in research: Definitions and examples*. Available from <https://www.scribbr.com/methodology/types-of-validity/> [accessed 6 December 2023].
- Mills, A.J., Durepos, G. & Wiebe, E. 2010. *Encyclopedia of case study research*. Thousand Oaks: Sage.
- Mohajan, H. 2017. *Two Criteria for Good Measurements in Research: validity and reliability*. Available from https://mpira.ub.uni-muenchen.de/83458/1/MPRA_paper_83458.pdf [accessed 11 April 2019].
- Mohebi, F. 2023. Managing unanticipated changes by adopting agile project management requirements: A case study of a large-scale project in banking. *Journal of Applied Intelligent Systems and Information Sciences*, Vol. 4, no. 2, pp. 30-37.
- Monette, D.R., Sullivan, T.J. & DeJong, C.R. 2002. *Applied social research: tool for the human services*. New York: Harcourt Brace College Publishers.
- Moreno, H. 2017. The importance of data quality: Good, bad or ugly. *Forbes Newsletter*. Available from <https://www.forbes.com/sites/forbesinsights/2017/06/05/the-importance-of-data-quality-good-bad-or-ugly/?sh=3f858c5710c44> [accessed 21 March 2019].
- Morehouse, R. 2012. *Beginning interpretative inquiry: A step-by-step approach to research and evaluation*. New York: Routledge.

- MRC Regulatory Support Centre. 2022. *Retention framework for research data and records*. Available from <https://www.ukri.org/wp-content/uploads/2023/03/MRC-100323-RegulatorySupportCentre-RetentionFrameworkResearchDataRecords.pdf> [accessed 18 November 2023].
- Muijs, D. 2004. *Doing quantitative research in education*. London: Sage.
- Nadler, D.A. & Tushman, M.L. 2015. *Types of organizational change: From incremental improvement to discontinuous transformation*. New York: Jossey-Bass.
- Nasr, M.M, Shaaban, E. & Ibrahim, M. 2020. *Data quality dimensions*. Available from https://www.researchgate.net/profile/Menna-Ibrahim-7/publication/340422622_Data_Quality_Dimensions/links/637e702e37878b3e87d5e85f/Data-Quality-Dimensions.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 31 October 2023].
- National Health and Medical Research Council. 2018. *Australian Code for Responsible Conduct of Research*. Available from: <https://www.nhmrc.gov.au/file/14384/download?token=gje4DNtT> [accessed 10 October 2023]
- National Library of Medicine. 2023. *Data management plan*. Available from <https://www.nlm.gov/guides/data-glossary/data-management-plan> [accessed 12 December 2023].
- National Research Foundation (NRF). 2017. *National Research Foundation Strategy 2020*. Available from <https://www.nrf.ac.za/wp-content/uploads/2021/05/NRF-Strategic-Plan-2020-2025.pdf> [accessed 26 March 2018].
- National Science Foundation (NSF). 2024. *Funding at NSF*. Available from <https://new.nsf.gov/funding/getting-started> [accessed 12 July 2024].
- Neubauer, B.E., Witkop, C.T. & Varpio, L. 2019. How phenomenology can help us learn from the experiences of others. *Perspectives on Medical Education*, Vol.8, pp. 90-97.
- Ng'eno, E. & Mutula, S. 2018. Research Data Management (RDM) in agricultural research institutes: A literature review. *Journal of Humanities and Social Sciences*, Vol. 10, no. 1.
- Ngulube, P. & Mathipa, E.R. 2015. *Theoretical and conceptual frameworks in the social and management sciences*. Available from https://www.researchgate.net/profile/Patrick_Ngulube/publication/278961764_Theoretical_and_Conceptual_Frameworks_in_the_Social_and_Management_Sciences/links/5587e07b08aeb0cdade0e78f/Th

[eoretical-and-Conceptual-Frameworks-in-the-Social-and-Management-Sciences.pdf](#) [accessed 22 May 2020].

Nguyen, D. 2023. Information use and data quality assurance practices in Rwanda. Available from <https://www.duo.uio.no/bitstream/handle/10852/103790/5/master-danielng.pdf> [accessed 12 December 2023].

Nhendodzashe, N. & Pasipamire, N. 2017. *Research data management services: Are academic libraries in Zimbabwe ready? The case of University of Zimbabwe library*. Available from <http://library.ifla.org/id/eprint/1728/1/S06-nhendodzashe-en.pdf> [accessed 17 November 2021].

Norwegian National Research Ethics Committees. 2022. *Guidelines for Research Ethics in the Social Sciences and the Humanities*. Available from <https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/guidelines-for-research-ethics-in-the-social-sciences-humanities-law-and-theology/> [accessed 14 July 2020].

Nwabugwu, M.J. & Godwin, L.S. 2020. Research Data Management (RDM) services in libraries: Lessons for academic libraries in Nigeria. *Library Philosophy and Practice*, Vol. 4238. Available from <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=7915&context=libphilprac> [accessed 10 February 2024].

Nyaboga, A.B. & Mwaura, M.F. 2009. Strategies for gaining competitive advantage in a dynamic environment thru data quality. *International Journal of Management and Information Systems*, Vol. 13, no. 1.

Ogier, A., Monena, H., Bailey, A. & Stovall, C. 2014. Data management inside the library: Assessing electronic resources: data using the Data Asset Framework methodology. *Journal of Electronic Resources Librarianship*, Vol. 26, no. 2, pp. 101-113.

Onwuegbuzie, A., and Collins, K., M., T. 2007. A Typology of Mixed Methods Sampling Designs in Social Science Research. *The Qualitative Report*, Vol. 12, no. 2, pp. 281-316. DOI:10.46743/2160-3715/2007.1638

Open University. 2023. *Data quality*. Available from <https://www5.open.ac.uk/library-research-support/research-data-management/data-quality> [accessed 11 December 2023].

Open Risk Academy. 2016. *Data completeness*. Available from https://www.openriskmanual.org/wiki/Data_Completeness [accessed 14 June 2020].

Oracle. 2009. *Managing attribute uniqueness constraint entries*. Available from https://docs.oracle.com/cd/E16764_01/oid.1111/e10029/atruniq.htm [accessed 10 February 2024].

Organisation for Economic Co-operation and Development (OECD). 2017. *Data quality*. In: *Handbook for Internationally Comparative Education Statistics: Concepts, Standards, Definitions and Classifications*. Paris: OECD.

Oxford Pocket Dictionary. 2002. 'Archiving'; 'methodology'; 'preservation'; 'storing'. 3rd edn. Oxford: Oxford University Press.

Palinkas, L.A., Horwitz, S.M., Green, C.A., Wisdom, J.P., Duan, N. & Hoagwood, K. 2015. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health and Mental Health Services Research*, Vol. 42, pp. 533-544.

Palmer, C.B. An approach for managing data quality [dissertation]. Bruce: University of Canberra. 2011.

Park, Y.S., Konge, L. & Artino, A. 2018. The positivism paradigm of research. *Academic Medicine*, Vol. 95, no. 5.

Partelow, S. 2023. What is a framework? Understanding their purpose, value, development and use. *Journal of Environmental Studies and Sciences*, Vol. 13, no.4. DOI:10.1007/s13412-023-00833-w.
Patel, D. 2016. Research data management: A conceptual framework. *Library Review*, Vol. 65, no. 4, pp. 226-241.

Perera, S., Jim, X., Maurushat, A. & Opoku, D.J. 2022. Factors affecting reputational damage to organisations due to cyberattacks. *Informatics*, Vol. 9, no. 28.

Peters, D. 2016. *Research data management: Advocacy and outreach programme*. Available from http://www.digitalservices.lib.uct.ac.za/sites/default/files/image_tool/images/346/RDM_FRB_2017.pdf [last accessed 28 February 2018, since deactivated].

Phrasisombath, K. 2009. *Sample size and sampling methods*. Available from https://www.gfmer.ch/Activites_internationales_Fr/Laos/PDF/Sample_size_methods_Phrasisombath_Laos_2009.pdf [accessed 24 June 2020].

Pienaar, H. 2014. *Developing an institutional research data management plan*. Available from <https://uir.unisa.ac.za/handle/10500/13909> [accessed 8 February 2018].

Pinnick, J. 2017. Exploring digital preservation requirements: A case study from the National Geoscience Data Centre (NGDC). *Records Management Journal*, Vol. 27, no. 2, pp. 175-191.

Pinsonneault, A. & Kraemer, K.L. 1993. Survey research methodology in management information systems: an assessment. *Journal of Management Information Systems*, Vol. 10, no. 2.

Ponelis, S.R. 2015. Using interpretive qualitative case studies for exploratory research in doctoral studies: A case of information systems research in small and medium enterprises. *International Journal of Doctoral Studies*, Vol. 10, pp. 535-550.

Porras, J.I. & Robertson, J.P. 1992. Organizational development: Theory, practice, and research. In M.D. Dunnette & L.M. Hough eds. *Handbook of Industrial and Organizational Psychology*, 3.

Price, P.C., Jhangiani, R. & Chiang, I. 2013. *Correlational research*. Available from <https://opentextbc.ca/researchmethods/chapter/correlational-research/#:~:text=Correlational%20research%20is%20a%20type,effort%20to%20control%20extraneous%20variables> [accessed 4 March 2020].

Price Waterhouse Coopers (PWC). 2004. *Ten guiding principles of change management*. Available from <https://www.strategyand.pwc.com/gx/en/insights/2002-2013/ten-guiding-principles/strategyand-ten-guiding-principles-of-change-management.pdf> [accessed 23 November 2018].

Pryor, G. 2012. *Managing research data*. Available from Pryor, G. (2012) *Managing Research Data*. Facet Publishing, Cambridge. DOI: 10.29085/9781856048910 [accessed 16 March 2018].

QuestionPro Survey Software. 2021. *Questionnaire: The ultimate guide, advantages & examples*. Available from <https://www.questionpro.com/blog/what-is-a-questionnaire/> [accessed 2 February 2021].

Quinn, P. & Cochran, M. 2002. *A Guide to using qualitative research methodology*. Available from <https://evaluation.msf.org/sites/default/files/2021-12/An%20MSF%20guide%20to%20Using%20Qualitative%20Research%20Methodology.pdf> [accessed 11 September 2020].

Rajhans, K. 2018. *Effective communication management: A key to stakeholder relationship management in project-based organisations*. Available from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.projektassistenz-blog.de/wp-content/uploads/2020/02/Effective_Communication_with_Stakeholders.pdf [accessed 09 July 2022].

Rass, L., Treur, J., Kucharska, W. & Wiewiora, A. 2022. Adaptive dynamical systems modelling of transformational organizational change with focus on organizational culture and organizational learning. *Cognitive Systems Research*, Vol. 79.

Ravitch, S. & Riggan, M. 2012. *Reason and rigor: How conceptual frameworks guide research*. London: Sage.

Ray, J.M. 2014. *Introduction to research data management*. West Lafayette: Purdue University Press.

Reeves, S., Peller, J., Goldman, J. & Kitto, S. 2013. Ethnography in qualitative educational research: AMEE guide no. 80. *Medical Teacher*, Vol. 35, no. 8.

Reisch, J. & Webster, J. 2004. *Quality control measures*. Available from <https://www.alz.washington.edu/NONMEMBER/FALL04/texas.pdf> [accessed 13 April 2020].

Reitz, J.M. 2002. *Online Dictionary of Library and Information Science (ODLIS)*. Available from <http://vlado.fmf.uni-lj.si/pub/networks/data/dic/odlis/odlis.pdf> [accessed 10 February 2024].

Research Data Alliance (RDA). 2016. *The Research Data Alliance Magazine*. Available from https://rd-alliance.org/sites/default/files/attachment/RDA_magazine_Final_web.pdf [accessed 12 August 2019].

Rice, R., Ekmekcioglu, C., Haywood, J., Jones, S., Lewis, S., Macdonald, S. & Weir, T. 2011. *Opportunities and Challenges in Research Data Management: Developing a Roadmap and Services at the University of Edinburgh*. Available from <https://ijdc.net/index.php/ijdc/article/download/8.2.194/327/1172> [accessed 20 December 2017].

Rinne, A. 2023. *Change management requires a change mindset*. Available from <https://hbr.org/2023/05/change-management-requires-a-change-mindset> [accessed 13 December 2023].

Robbins, S.P. 1990. *Organisational theory: Structure, design, and application*. New Jersey: Prentice Hall.

Roberts, D. & Khattri, N. 2012. *Designing a results framework for achieving results: A how-to guide*. Washington, DC: Independent Evaluation Group.

Rocco, T. & Plakhotnik, M.S. 2009. Literature reviews, conceptual frameworks, and theoretical frameworks: Terms, functions, and distinctions. *Human Resource Development Review*, Vol.8, no. 1, pp. 120-130.

Rod, A.B., Zhou, B. & Rousseau, M. 2023. There's no "I" in research data management: Reshaping RDM Services toward a collaborative multi-stakeholder model. *Journal of eScience Librarianship*, Vol. 12, no. 1.

Rogers, L., De Brun, A. & McAuliffe, E. 2020. Development of an integrative coding framework for evaluating context within implementation science. *BMC Medical Research Methodology*, Vol. 20. doi.org/10.1186/s12874-020-01044-5.

Rohman, F., Noermijatia, N., Mugionoa, M. & Soelton, M. 2023. The role of quality assurance in improving the distribution of organizational performance. *Uncertain Supply Chain Management*, Vol. 11.

- Rolando, L., Carlson, J., Hswe, P., Parham, S., Westra, B. & Whitmire, L. 2015. Data management plans as a research tool. *Association for Information Science and Technology*, Vol. 41, no. 5.
- Rolland, B. 2016. *Data sharing and reuse: Expanding our concept of collaboration: Team science toolkit*. Available from <https://www.teamsciencetoolkit.cancer.gov/public/ExpertBlog.aspx?tid=4> [accessed 12 January 2018].
- Rose, S., Spinks, N. & Canhoto, A.I. 2015. *Management research: Applying the principles*. Available from <https://studylib.net/doc/8746573/case-study-research-design> [accessed 23 June 2021].
- Rouchon, O., Prat, P. & Cloirec, M. 2013. Best practices on standards, policies and quality assurance in digital repositories for long term preservation. *Partnership for Advanced Computing in Europe*. Available from https://prace-ri.eu/wp-content/uploads/Best_Practices_on_Standards_Policies_and_Quality_Assurance_in_Digital_Repositories_for_Long_Term_Preservation-2.pdf [accessed 10 February 2024].
- Rowley, J. 2002. Using case studies in research. *Management Research News*, Vol. 25, no. 1.
- Rudolph, J.E., Zhong, Y., Duggal, P., Mehta, S.H. & Lau, B. 2023. Defining representativeness of study samples in medical and population health research. *BMJ Medicine*, Vol. 2, no.1. DOI: 10.1136/bmjmed-2022-000399.
- Russell, R.J. 2011. *Quality assurance*. Available from https://www.researchgate.net/profile/Ronald-Russell/publication/237287165_Chapter_4_QUALITY_ASSURANCE/links/54b3998c0cf28ebe92e2edee/Chapter-4-QUALITY-ASSURANCE.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 31 October 2023].
- Salian, D. 2023. *Usability of open data*. Available from https://www.researchgate.net/publication/375788279_Usability_of_Open_Data/fulltext/655ca9483fa26f66f41bcd1f/Usability-of-Open-Data.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 18 October 2023].
- Samuels, E. 2015. *Terrorism: A paradigm shift in corporate risk management*. New York: Page Publishing.
- Sananikone, J., Arnaud, E., Norvez, O., Pamerlon, S., Archambeau, A. & Le Bras, Y. 2022. From raw data to data standards through quality assessment and semantic annotation. *Biodiversity Information Science and Standards*, Vol. 6.

- Sanjeeva, M. 2018. *Research data management: A new role for academic/research librarians*. Conference: Reshaping the Academic Libraries, Trends and Issues UGC Sponsored Seminar at JM Patel College on 12th Sept 2014 At: JM Patel College, Mumbai.
- Sarpatwari, A. 2014. Ensuring patient privacy in data sharing for postapproval research. *The New England Journal of Medicine*, Vol. 371, no. 17.
- Scannapieco, M., Missier, P. & Batini. 2005. Data quality at a glance. *Datenbank-Spektrum*, Vol. 14, no. 4, pp. 6-14.
- Scarborough, E. & Tanenbaum, E. 1998. *Research strategies in the social sciences*. Oxford: Oxford University Press.
- Schmerse, D., Dominke, H., Mohr, J., & Steffensky, M. (2024). Children's understanding of scientific inquiry: The role of instructional support and comparison making. *Journal of Educational Psychology*, Vol. 116, no. 2, pp. 233–255. doi.org/10.1037/edu0000836.
- Schneider, J., Seidel, S., Basalla, M. & vom Brocke, J. 2023. Reuse, reduce, support: Design principles for green data mining. *Business and Information Systems Engineering*, Vol. 65, no. 1.
- ScienceDirect. 2012. *Explanatory model*. Available from <https://www.sciencedirect.com/topics/psychology/explanatory-model> [accessed 9 March 2021].
- Sengupta, D., Gurtoo, A. & Manjunath, B.S. 2023. *Identifying data quality dimensions*. Available from https://csp.iisc.ac.in/wp/wp-content/uploads/2023/01/ReportDataQuality_DebarunSengupta13012023.pdf [accessed 13 December 2023].
- Shanmugan, D.B., Dhilipan, J., Prabhu, T., Sivasankari, A. & Vignesh, A. 2023. The management of data quality assessment in big data presents a complex challenge, accompanied by various issues related to data quality. *Research Highlights in Mathematics and Computer Science*, Vol. 8, pp. 78-91. DOI: 10.9734/bpi/rhmcs/v8/18858D.
- Sidi, F., Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H. & Mustapha, A. 2012. *Data quality: A survey of data quality dimensions*. Available from https://www.researchgate.net/profile/Payam-Hassany-Shariat-Panahy/publication/241631252_Data_quality_A_survey_of_data_quality_dimensions/links/00463521d14b7923fd000000/Data-quality-A-survey-of-data-quality-dimensions.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 1 November 2023].

Sigmoid. 2023. *Data quality: A strategic imperative for modern enterprises*. Available from https://www.sigmoid.com/wp-content/uploads/2023/11/PoV_Data-Quality_Strategic-imperative-for-modern-enterprises.pdf [accessed 30 November 2023].

Silverman, D. 2017. *Doing qualitative research*. London: Sage.

Simon, M.K. & Goes, J. 2011. Correlation research. Available from <https://pdf4pro.com/cdn/correlational-research-dissertationrecipes-com-17d6e5.pdf> [accessed 25 June 2021].

Simplilearn. 2023. *What is data collection: Methods, types, tools*. Available from <https://www.simplilearn.com/what-is-data-collection-article> [accessed 19 November 2023].

Simplilearn. 2023. *What is Quality Control (QC)*. Available from <https://www.simplilearn.com/what-is-quality-control-article> [accessed 23 November 2023].

Singleton, A., Straits, B.C. & Straits, M.M. 1993. *Approaches to social research*. New York: Oxford University Press.

Socioeconomic Data and Applications Center (SEDAC). 2020. *Data submission*. Available from <https://sedac.ciesin.columbia.edu/data-submission> [accessed 1 May 2020].

Soelton, M., Noermijati, N., Rohman, F. & Mugiono, M. 2023. To improve the quality management of children welfare institutions and provide better services. *Quality Access to Success*, Vol. 24, no.195.

South African Data Archive (SADA). 2020. *Repository Details*. Available from <https://www.re3data.org/repository/r3d100010554> [accessed 22 March 2022].

South African Medical Research Council (MRC). 2022. *Who we are*. Available from <https://www.samrc.ac.za/about-us/who-we-are> [accessed 1 September 2022].

Stake, R.E. 2005. Qualitative case studies. In Denzin, N.K. and Lincoln, Y.S. eds. *The Sage handbook of qualitative research*. 3rd edn. Thousand Oaks: Sage.

Stanford Libraries. 2020. *Data management plans*. Available from <https://laneguides.stanford.edu/DataManagement/Planning> [accessed 14 January 2021].

Statistics South Africa. 2010. *South African Statistical Quality Assessment Framework (SASQAF)*. Available from http://www.statssa.gov.za/standardisation/SASQAF_OpsGuidelines_Edition_1.pdf [accessed 08 April 2020].

- Statistics South Africa. 2018. *Standardisation*. Available from http://www.statssa.gov.za/?page_id=371 [accessed 24 April 2020].
- Statistics South Africa. 2018. *Vision and Mission*. Available from http://www.statssa.gov.za/?page_id=560 [accessed 24 April 2020].
- Stedman, C. & Vaughan, J. 2023. *Data quality*. Available from <https://www.techtarget.com/searchdatamanagement/definition/data-quality> [accessed 23 November 2023].
- Stiles, P.G. & Boothroyd, R.A. 2015. Ethical use of administrative data for research purposes. In *Actionable Intelligence*. New York: Palgrave Macmillan 125-155.
- Strasser, C., Cook, R., Michener, W. & Budden, A. 2012. *Primer on data management: What you always wanted to know: A DataONE publication*. Available from <https://escholarship.org/uc/item/7tf5q7n3>
- Stvilia, B., Gibradze, L. & Lee, D.J. 2022. *What is quality to research data repositories?* Available from https://www.researchgate.net/profile/Besiki-Stvilia/publication/365713429_What_is_Data_Quality_to_Research_Data_Repositories/links/637fa29c7b0e356feb7ca52a/What-is-Data-Quality-to-Research-Data-Repositories.pdf [accessed 4 December 2023].
- Subaveerapandian, A. 2023. Research data management practices and challenges in academic libraries: A comprehensive review. *Library Philosophy and Practice*. Available from <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=15119&context=libphilprac> [accessed 10 February 2024].
- Suber, P. 2009. Ten challenges for open-access journals. *SPARC Open Access Newsletter*, Vol. 138. Available from https://dash.harvard.edu/bitstream/handle/1/4316131/suber_10challenges.html?sequence=2 and isAllowed=y [accessed 23 August 2019].
- Sunday, C. 2016. *Qualitative Data Analysis (QDS)*. Available from <https://www.uwc.ac.za/Students/Postgraduate/.../Qualitative%20data%20analysis.pdf> [last accessed 18 September 2018, since deactivated].
- Surbhi, S. 2017. *Validity vs reliability*. Available from <https://keydifferences.com/difference-between-validity-and-reliability.html> [accessed 11 July 2024].
- Surkis, A. & Read, K. 2015. Research data management. *Journal of the Medical Library Association*, Vol. 103, no. 3.

Swedberg, D. & Douglas, J. 2005. Transforming by design: An innovative approach to implementation of e-government. *Eight-Japan Engineering Consultants Inc (EJEC)*, Vol. 6, no. 1, pp. 51-56.

Syafwan, F. & Fathia, S. 2023. *The role of communication in organizational change*. Available from <https://kth.diva-portal.org/smash/get/diva2:1766525/FULLTEXT01.pdf> [accessed 18 December 2023].

Taherdoost, H. 2021. Data collection methods and tools for research: A step-by-step guide to choosing data collection technique for academic and business research projects. *International Journal of Academic Research in Management (IJARM)*, Vol. 10, no. 1, pp. 10-38

Tashakkori, A. & Teddlie, C. eds. 2003. *Handbook of mixed methods in social and behavioral research*. Thousand Oaks: Sage.

Teegavarapu, S., Summers, J.D. & Mocko, G.M. 2008. Case study method for design research: A justification. In Volume 4: 20th International Conference on Design Theory and Methodology, *Second International Conference on Micro- and Nanosystems*.

Terrell, S R. 2012. Mixed-methods research methodologies. *The Qualitative Report*, Vol. 17, no. 1, pp. 254-280.

Terrell, S.R. & Edmonds, A. 2017. *Mixed-methods research methodologies*. Available from https://education.nova.edu/summer/2017/sessionmaterials/Elective_Mixed-Methods_Edmonds_Terrell.pdf [last accessed 3 February 2019, since deactivated].

Texas State Library and Archives Commission. 2023. *Best practices for long-term preservation of electronic records*. Available from <https://tsl.access.preservica.com/tslac-digital-preservation-framework/best-practices-for-long-term-preservation/> [accessed 18 November 2023].

Thanh, N.C. & Thanh, T.T.L. 2015. The interconnection between interpretivist paradigm and qualitative methods in education. *American Journal of Educational Science*, Vol. 1, no. 2.

The Society for Human Resource Management. 2023. *Managing organizational change*. Available from <https://www.shrm.org/topics-tools/tools/toolkits/managing-organizational-change#:~:text=An%20employer%20that%20is%20serious,plan%20for%20dealing%20with%20resistance> [accessed 11 July 2024].

Toni, L. 2014. *Managing reputation risk*. Available from <https://www2.deloitte.com/content/dam/Deloitte/ro/Documents/about-deloitte/ro-managing-reputation-risk-laura-toni.pdf> [accessed 11 April 2020].

- Tripathi, M., Shukla, A. & Sonker, K. 2017. Research data management practices in university libraries: A study. *Journal of Library and Information Technology*, Vol. 37, no. 6, pp. 417-424.
- Tsang, D.C. 2014. Research data management: Challenges and opportunities. Presentation at Singapore Management University. Li Ka Shing Library, January 7.
- Tushman, M., Kahn, A. & Porray, M. 2017. Introduction: Why does change fail, and what can we do about it? *Journal of Change Management*, Vol. 11, no. 4, pp. 445–450.
- UK-DAMA Working Group. 2013. *The six primary dimensions for data quality assessment: Defining data quality dimensions*. Available from <https://silo.tips/download/the-six-primary-dimensions-for-data-quality-assessment> [accessed 13 August 2021].
- UK Data Archive. 2011. *Activity-based data management costing tool for researchers*. Available from http://www.data-archive.ac.uk/media/257647/ukda_jiscdmcosting.pdf [last accessed 8 November 2017, since deactivated].
- UK Data Archive. 2017. *Research data lifecycle*. Available from <https://www.data-archive.ac.uk/create-manage/life-cycle> [last accessed: 19 December 2017, since updated].
- UK Data Service. 2014. *Documenting and contextualising your data*. Available from <https://www.ukdataservice.ac.uk/media/440277/documentingdata.pdf> [accessed 8 February 2018].
- UK Data Service. 2019. *The importance of managing and sharing data*. Available from <https://www.ukdataservice.ac.uk/manage-data/plan/planning.aspx> [accessed 29 August 2019].
- UK Data Service. 2020. *QAMyData*. Available from <https://ukdataservice.ac.uk/about/research-and-development/past-projects/qamydata/#:~:text=The%20UK%20Data%20Service%20developed,duplication%2C%20outliers%20and%20direct%20identifiers> [accessed 26 April 2020].
- UK Data Service. 2023. *Quality*. Available from <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/quality> [accessed 8 May 2023].
- UK Essays. 2017. *Purpose and advantages of primary data*. Available from <https://www.ukessays.com/essays/marketing/purpose-and-advantages-of-primary-data-marketing-essay.php> [accessed 8 August 2018].
- Umanailo, M.C.B. 2019. *Overview phenomenological research*. Available from https://www.researchgate.net/profile/Muhamad-Chairul-Basrun-Umanailo/publication/335230717_Overview_Phenomenological_Research/links/5d8af7c5a6fdcc255496e0f7/Overview-Phenomenological-

[Research.pdf? tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19](#) [accessed 11 November 2022].

Unal, Y., Chowdhury, G., Kurbanoglu, S., Boustany, J. & Walton, G. 2019. Research data management and data sharing behaviour of university researchers. *Information Research*, Vol. 24, no. 1.

United Nations Children's Fund (UNICEF). 2021. *Data quality framework*. Available from <https://data.unicef.org/resources/data-quality-framework/#:~:text=The%20Data%20Quality%20Framework%20is,confidence%20in%20our%20data%20products> [accessed 31 October 2023].

University of California. 2023. *What gets measured gets done: Key performance indicators*. Available from <https://cdlib.org/cdinfo/2010/09/15/what-gets-measured-gets-done-key-performance-indicators/> [accessed 26 November 2023].

University of Cape Town. 2018. *UCT RDM Policy*. Available from https://uct.ac.za/sites/default/files/content_migration/uct_ac_za/39/files/TGO_Policy_Research_Data_Management_2018.pdf [accessed: 28 March 2018].

University of Essex. 2024. *HASSET Thesaurus*. Available from <https://hasset.ukdataservice.ac.uk/hasset/en/> (accessed 07 July 2024).

University of Hertfordshire. 2023. *Corporate records management standards: Standard naming conventions for electronic files, folders and records*. Available from https://www.herts.ac.uk/_data/assets/pdf_file/0009/361656/IM11-Apx5-Standard-naming-conventions.pdf [accessed 29 November 2023].

University of Johannesburg. 2023. *Policy: Research Data Management*. Available from <https://www.uj.ac.za/wp-content/uploads/2023/02/research-data-management.pdf> [accessed 17 November 2023].

University of Kansas. 2023. *Collecting and analyzing data*. Available from <https://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/collect-analyze-data/main> [accessed 13 November 2023].

University of Leicester. 2017. *Data management planning*. Available from <https://www2.le.ac.uk/services/research-data/create-data/DMPlan> [accessed: 13 May 2018].

University of Michigan. 2018. *Data management and curation*. Available from <https://www.icpsr.umich.edu/web/pages/datamanagement/index.html> [accessed 16 August 2018].

University of Michigan. 2022. *Additional ICPSR services*. Available from <https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/services.html> [accessed 07 July 2022].

University of Michigan. 2023a. *Deposit data*. Available from <https://www.icpsr.umich.edu/web/pages/DSDR/deposit.html#:~:text=Why%20Deposit%20Data%20With%20DSDR,discovery%20and%20protect%20respondent%20privacy> [accessed 05 December 2023].

University of Michigan. 2023b. *Inter-university Consortium for Political and Social Research (ICPSR)*. Available from <https://deepblue.lib.umich.edu/handle/2027.42/57738> [accessed 29 November 2023].

University of Pretoria. 2023. *Application for approval of a research project*. Available from https://www.up.ac.za/media/shared/7/ZP_Files/EBIT-EthicsFiles/applicationform.zp95756.docx [accessed 23 September 2021].

University of Pretoria. 2023. *Research compliance regulation*. Available from https://www.up.ac.za/media/shared/1/ZP_Files/s5103_23-research-compliance-regulation.zp233199.pdf [accessed 12 May 2022].

University of Oxford. 2023. *DMPonline*. Available from <https://researchdata.ox.ac.uk/dmp-online> [accessed 17 January 2023].

University of Pittsburgh. 2024. *What is research data management?* Available from <https://pitt.libguides.com/researchhelpuls/managing> [accessed 08 July 2024].

University of Sheffield. 2019. *Why manage your data?* Available from <https://www.sheffield.ac.uk/library/rdm/benefits> [last accessed 26 November 2019, since deactivated].

University of Southern California. 2016. *Social behavioral research*. Available from <https://hrpp.usc.edu/research/sbirb-social-behavioral-research/> [accessed 14 July 2018].

University of Southern California (USC) Libraries. 2020. *Organizing your social sciences research paper: Theoretical framework*. Available from <https://libguides.usc.edu/writingguide/theoreticalframework> [accessed 12 May 2020].

University of Wisconsin. 2020. *Quality assurance in research*. Available from <https://researchdata.wisc.edu/uncategorized/quality-assurance-in-research/> [accessed 29 November 2023].

University of Wollongong Australia. 2017. *Research data management guidelines*. Available from <https://policies.uow.edu.au/document/view-current.php?id=141> [accessed 29 March 2020].

USAID. 2010. *Performance monitoring and evaluation tips: Conducting data quality assessments*. Available from https://pdf.usaid.gov/pdf_docs/pnadw118.pdf [accessed 27 March 2020].

USAID. 2019a. *Conducting data quality assessments*. Available from <https://www.usaid.gov/project-starter/program-cycle/implementing-monitor-evaluation-commitments/monitoring-performance/conducting-data-quality-assessments> [accessed 18 April 2019].

USAID. 2019b. *Data quality and quality improvement training*. Available from https://pdf.usaid.gov/pdf_docs/PA00ZT94.pdf [accessed 31 October 2023].

USAID. 2022. *Logical framework assumptions*. Available from <https://www.usaid.gov/logical-framework-assumptions#:~:text=Assumptions%20reflect%20our%20recognition%20that,identify%20and%20assess%20critical%20assumptions>. [accessed 8 October 2023].

USC Libraries. 2021. *Research guides: Organizing your social sciences research paper: Theoretical framework*. Available from <https://libguides.usc.edu/writingguide/theoreticalframework> [accessed 11 February 2022].

Vale, S. 2010. *Statistical data quality in the UNECE*. Available from <https://unstats.un.org/unsd/dnss/docs-nqaf/UNECE-Quality%20Improvement%20Programme%202010.pdf> [accessed 31 October 2023].

Van den Eynden, V. 2011. *Managing and sharing data: best practice for researchers*. 3rd edn. Essex: UK Data Archive.

Van der Graaf, M. 2023. *Research data management support services*. Available from https://libereurope.eu/wp-content/uploads/Liber_ABDU1920X1080_FINAL.pdf [accessed 15 January 2024].

Van Deventer, M. & Pienaar, H. 2015. Research data management in a developing country: a personal journey. Available from https://www.researchgate.net/profile/Martie-Van-Deventer/publication/281223665_Research_Data_Management_in_a_Developing_Country_A_Personal_Journey/links/5745f08208ae9f741b431113/Research-Data-Management-in-a-Developing-Country-A-Personal-Journey?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 12 December 2023].

Van Teijlingen, E. & Hundley, V. 2001. The importance of pilot studies. *Sociology at Surrey*, Vol. 35.

- Vaughan, R. 2008. *Conceptual framework*. Available from https://assets.bournemouth.ac.uk/news-archive/services-management/news_events_conferences/PPTs/r_vaughan.ppt [accessed 6 October 2020].
- Vetro, A., Canova, L., Minotas, C.O., Lemma, R. & Morando, F. 2016. Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, Vol. 33, no. 2.
- Victoria State Government. 2018. *Data quality guideline: Information management framework*. Available from <https://content.vic.gov.au/sites/default/files/2019-07/IM-GUIDE-09-Data-Quality-Guideline.pdf> [accessed 28 May 2020].
- Vollstedt, M. & Rezar, S. 2019. *An introduction to grounded theory with a special focus on Axial coding and the coding paradigm*. Available from https://link.springer.com/content/pdf/10.1007%2F978-3-030-15636-7_4.pdf [accessed 24 September 2020].
- Vosloo, J.J. 2014. *Data analysis and interpretation*. Available from https://www.researchgate.net/profile/Mohamed-Hammad-4/post/What_instrument_is_used_to_analyze_qualitative_descriptive_data/attachment/59d659ca79197b80779af218/AS%3A543750520664064%401506651634578/download/CHAPTER+6-+DATA+ANALYSIS+AND+INTERPRETATION.pdf [accessed 08 July 2024].
- Waddell, D. & Sohal, A.S. 1998. Resistance: A constructive tool for change management. *Management Decision*, Vol. 36, no. 8, pp. 543-549.
- Walden University. 2024. *Theories and frameworks: Introduction*. Available from <https://academicguides.waldenu.edu/library/theory#:~:text=A%20theoretical%20framework%20is%20a,problem%20is%20understood%20and%20investigated> [accessed 08 July 2024].
- Walwyn, D. & Chan, A. 2023. *Research guide for masters and doctoral students in the graduate school of technology management*. Available from <https://www.up.ac.za/media/shared/26/2023/research-guide-for-post-grad-students-issue-27-january-2023.zp232086.pdf> [accessed 05 December 2023].
- Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S. & Aggarwal, S. 2023. Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. *Journal of the Knowledge Economy*, Vol. 15, no. 1. DOI:[10.1007/s13132-022-01096-6](https://doi.org/10.1007/s13132-022-01096-6)
- Wang, R.Y., Storey, V.D. & Firth, C.P. 1995. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, no. 4.

- Whatfix. 2022. *6 Types of organizational change, explained*. Available from <https://whatfix.com/blog/types-of-organizational-change/> [accessed 18 December 2023].
- Whyte, A. & Tedds, J. 2011. Making the case for research data management. Available from https://www.researchgate.net/profile/Jonathan_Tedds/publication/252931138_Making_the_Case_for_Research_Data_Management/links/00b7d51f53d2e7b867000000/Making-the-Case-for-Research-Data-Management.pdf [accessed 17 April 2020].
- Wichita State University. 2023. *Data governance by-laws, policies, standards and coordinating bodies for Wichita State University data systems and reporting standards*. Available from https://www.wichita.edu/services/DataGov/WSU_DataGovernance.pdf [accessed 12 December 2023].
- Wiggins, A. Bonney, R., Graham, E. & Henderson, S. 2013. *Data management guide for public participation in scientific research: DataONE Public Participation in Scientific Research Working Group*. Available from <https://education.ucdavis.edu/sites/main/files/file-attachments/dataone-ppsr-datamanagementguide.pdf> [accessed 15 September 2018].
- Wikipedia. 2020. *Government of Victoria*. Available from https://en.wikipedia.org/wiki/Government_of_Victoria [accessed 23 June 2020].
- Wilkinson, M., Dumontier, M. & Aalbersberg, I. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, Vol. 3.
- Willis, J. W. 2007. *Foundations of qualitative research: interpretive and critical approaches*. London: Sage.
- Wilms, K., L., Stieglitz, S., Buchholz, A., Vogl, R. & Rudolph, D. 2018. *Do researchers dream of research data management?* Available from https://www.researchgate.net/profile/Konstantin-Wilms/publication/322676062_Do_Researchers_Dream_of_Research_Data_Management/links/5a6849acaca2720266b6e972/Do-Researchers-Dream-of-Research-Data-Management.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 [accessed 13 December 2023].
- Wisdom, J. & Creswell, J.W. 2013. *Mixed methods: Integrating quantitative and qualitative data collection and analysis while studying patient-centred medical home models*. Available from <https://www.ahrq.gov/sites/default/files/wysiwyg/ncepcr/tools/PCMH/mixed-methods.pdf> [accessed 5 February 2021].
- Wissik, T. & Durco, M. 2015. *Research data workflows: From research data lifecycle models to institutional solutions*. Available from <https://ep.liu.se/ecp/123/008/ecp15123008.pdf> [accessed 26 June 2020].

- Witte, K. 2015. *Theories of behaviour change*. Available from <http://documents1.worldbank.org/curated/en/456261468164982535/pdf/526140BRI0Beha10Box345574B01PUBLIC1.pdf> [accessed 16 July 2018].
- Woods, A.D., Gerasimova, D., van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P.E., Halvorson, M., King, K.M., Logan, J.A.R., Xu, M., Vasilev, M.R., Clay, J.M., Moreau, D., Joyal-Desmarais, K., Cruz, R.A., Brown, D.M.Y., Schmidt, K. & Elsherif, M.M. 2023. Best practices for addressing missing data through multiple imputation. *Infant and Child Development*.
- World Bank Group. 2015. *Grant guidelines*. Available from http://www.worldbank.org/content/dam/Worldbank/document/Gender/SVRI%20and%20WBDM%20Grant%20Guidelines_v9.pdf [accessed 26 January 2019].
- World Data System (WDS). 2015a. *WDS annual report*. Available from <https://zenodo.org/record/165821> [accessed 26 November 2016].
- World Data System (WDS). 2015b. *WDS data sharing principles*. Available from <https://www.icsu-wds.org/services/data-sharing-principles> [accessed 17 April 2020].
- Yin, R.K. 2003. *Case study research: design and methods*. 3rd edn. Thousand Oaks: Sage.
- York University. 2018. *Open access and open data steering committee*. Available from <https://www.library.yorku.ca/web/open/overview/rdmbenefits/> [last accessed 01 September 2020, since updated].
- Youngson-Baines, S. & Lee, A. 2023. *Research data management policy*. Available from <https://www.ucem.ac.uk/wp-content/uploads/2023/10/Data-Management-Policy.pdf> [accessed 18 November 2023].
- Zainal, Z. 2007. Case study as a research method. *Educational Technology and Society*, Vol. 9.
- Zhou, X., Xu, Z. & Kogut, A. 2023. Research data management needs assessment for social sciences graduate students: A mixed methods study. *PLoS ONE*, Vol. 18, no. 2.

Appendix 1: Extended background information

This document serves as a discussion document about the RDM stakeholders reviewed in this research study. It provides background information on the organisation's purpose, importance activities, data quality issues, and graphical representation of the stakeholder data quality framework. It serves as a support document for Chapter 2 and aims to scrutinise the data quality aspects in the reviewed documentation and web content. The document aims to provide a comprehensive understanding of the stakeholders involved, as mentioned in Chapter 2.

1. Statistics South Africa (SASQAF)

Purpose/Mandate of the organisation

Statistics South Africa (Stats SA) is the national South African statistics agency set up under the Statistics Act (Act No.6 of 1999) with the goal of generating official statistics in a timely, reliable and accessible manner (Statistics South Africa, 2010). In relation to the numerical and arithmetical elements, the characteristics of social science data and statistical data vary slightly; however, researchers must handle data (irrespective of the type) in the best way possible in order to generate quality data. Stats SA's mission is to guide and cooperate on evidence-based decisions in statistical systems (Statistics South Africa, 2010). To realise this vision, Stats SA aims to produce statistical products that comply with relevant quality standards. In formulating and using quality standards to handle its statistical data, Stats SA has made great strides. Stats SA has provided guiding principles for relevant practices, such as the standards infrastructure in Stats SA, non-Stats SA standards (including those in the NSS) and the development of standards expertise in Stats SA (Statistics South Africa, 2018). The numerous standards used in Stats SA allow the agency to generate quality statistical data that meets the statistical community's user needs.

Brief description of stakeholder activities According to Statistics South Africa (2010:10) “*the main purpose of SASQAF is to provide a flexible structure for the assessment of statistical products*”. SASQAF covers various quality aspects relevant to the entire statistical value chain (i.e. need, design, construct, collect, process, evaluate and disseminate) and certifies national statistics (Statistics South Africa, 2010).

Stats SA has developed the South African Statistical Quality Assessment Framework (SASQAF) “*which is one of the statistical tools at the centre of the statistical system, enabling the implementation process of statistical coordination and thus of the South African National Statistics System (NSS)*” (Statistics South Africa, 2010:5). SASQAF is made up of eight dimensions of quality, which include *relevance, accuracy, timeliness, accessibility, interpretability, coherence, methodological soundness* and *integrity* (Statistics South Africa, 2010). These dimensions are all intended to make data as useful as possible by accurately describing the phenomena they were intended to measure. In addition, the

SASQAF 8 quality dimensions overlap with the current data quality dimension in literature (as discussed in Section 2.3 – i.e. The DAMA UK Quality Dimensions).

How does this stakeholder address data quality issues?

SASQAF has four levels at which statistical certification is granted, based on certain criteria being met. The levels include:

Level Four: Quality Statistics - *“These are statistics that meet all the quality requirements as set out in SASQAF. They are designated as quality statistics to the extent that deductions can be made from them, and are ‘fit for use’ for the purpose for which they were designed”* (Statistics South Africa, 2010:5).

Level Three: Acceptable Statistics - *“These are statistics that meet most, but not all the quality requirements as stipulated in SASQAF. They are designated as acceptable to the extent that, despite their limitations, deductions can be made from them, and are ‘fit for use’ for the purpose for which they were designed”* (Statistics South Africa, 2010:5).

Level Two: Questionable Statistics - *“These are statistics that meet few of the quality requirements as stipulated in SASQAF. They are designated as questionable to the extent that very limited deductions can be made, and they are therefore not ‘fit for use’ for the purpose for which they were designed”* (Statistics South Africa, 2010:5).

Level One: Poor Statistics: *“These are statistics that meet almost none of the quality requirements as stipulated in SASQAF. They are designated as poor statistics to the extent that no deductions can be made from them, are not ‘fit for use’ for the purpose for which they were designed”* (Statistics South Africa, 2010:5).

Consistency is a key component of each of the aforementioned four levels; basic prerequisites must be satisfied in order for statistical data to be accepted and assigned a grade at one of the SASQAF's four levels.

Why is this stakeholder important?

“SASQAF covers the various quality aspects of the entire statistical value chain (i.e. need, design, build, collection, processing, analysis and dissemination)” (Statistics South Africa, 2010:13). SASQAF brings quality and quality assurance at the forefront of data management, specifically in the context of this study of social science research data. Another component of SASQAF is the Prerequisites of quality – which refer to *“the institutional conditions that have an impact on data quality. It defines the minimum set of necessary conditions that have to be met in order to produce good quality statistics. It therefore serves as the foundation on which all other dimensions of data quality should be premised on”* (Statistics South Africa, 2010:14). The Prerequisites of quality are highly valuable at the institutional level in order

to establish the correct level of competence and quality that can be controlled by statistics and data. The eight prerequisites include the following quality indicators:

- *“The responsibility for producing statistics is clearly specified”*
- *“Standards and policies are in place to promote consistency of methods and results”*
- *“Data sharing and coordination among data-producing agencies are clearly specified”*
- *“Measures are in place to ensure that individual data are kept confidential and used for statistical purposes only”*
- *“Measures to oblige response are ensured through law”*
- *“Resources are commensurate with the needs of statistical programmes”*
- *“Measures to ensure efficient use of resources [are] implemented”*
- *“Processes are in place to focus on, monitor and check quality”* (Statistics South Africa, 2010:6).

The indicators and associated standards that are employed as quality dimensions in the suggested quality framework are listed in the table below. Since these metrics are based on statistical data, an interpretation that shows how the standards resemble the setting of this study (social science research data) is given; however, where the standard is regarded “as is”, no interpretation is given.

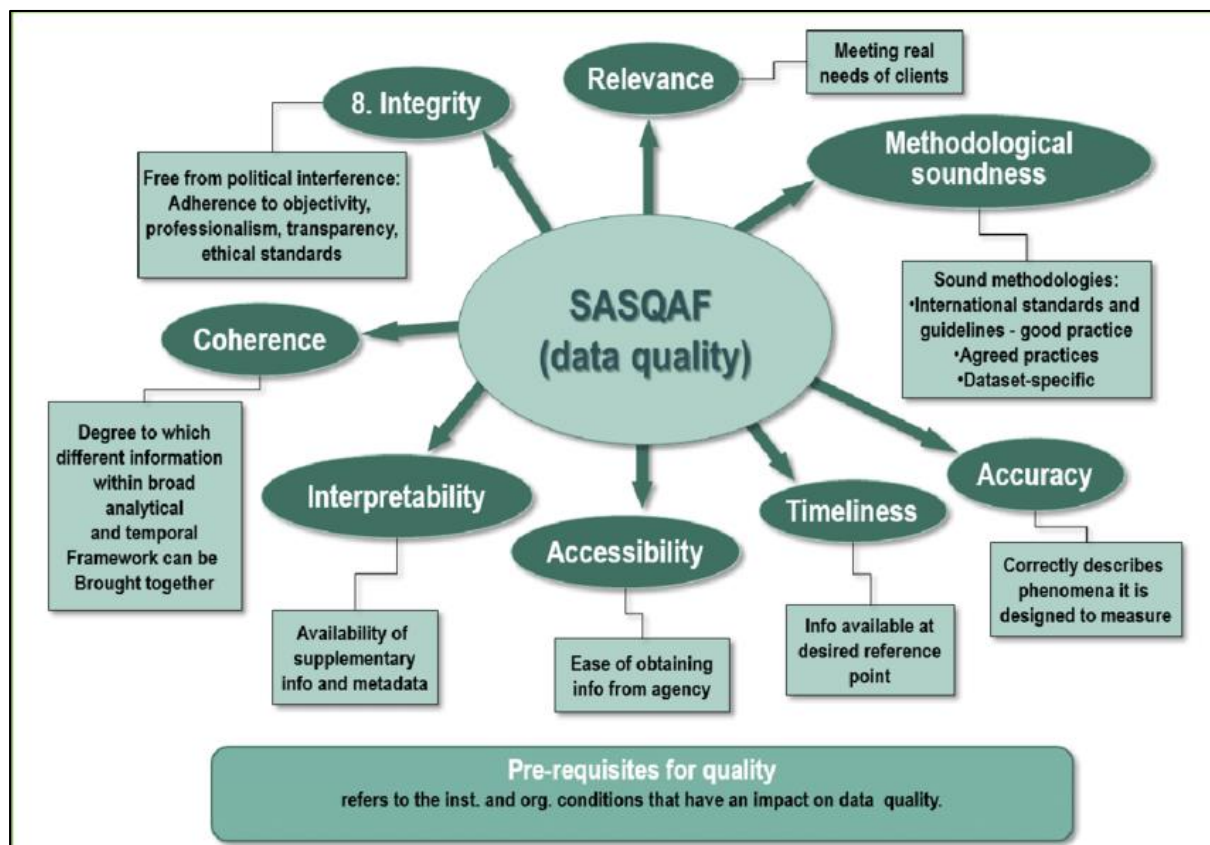
Table 1: Adopted SASQAF Prerequisites indicators and standards

Indicator	Standard
1.2 <u>Standards and Policies</u> : are in place to promote consistency of methods and results.	1.2.1 A set of policies must exist which cover all aspects of the statistical value chain. <u>Interpretation</u>: An RDM policy must exist, and must cover all aspects of the research data cycle; this standard is used as a measuring criterion for the Standards and policies quality dimension. 1.2.2 A set of standards related to appropriate policies must exist. <u>Interpretation</u>: An RDM policy has to be informed by best practice standards and guidelines; this standard is used as a measuring criterion for the Standards and policies quality dimension.
1.8 <u>Quality Check Processes</u> : Processes are in place to focus on, monitor and check quality	1.8.1 The agency must have a quality management system in place; staff members in production areas must have a data quality management requirement as part of their performance agreements or job descriptions. <u>Interpretation</u>: There has to be a clear distinction between the roles of data managers and curators responsible for data quality; this standard is used as a measuring criterion for the Standards and policies quality dimension. 1.8.2 The data-producing agency must have an independent data quality audit process. <u>Interpretation</u>: The guidelines and operating procedures that researchers use have to be reviewed in terms of quality; this standard is used as a measuring criterion for the Standards and policies quality dimension. 1.8.3 Staff members in production areas must have a data quality management requirement as part of their performance agreements or job descriptions.

	<u>Interpretation:</u> At a minimum, it is imperative for research teams and project leaders to have data management background to assist with proper RDM; this standard is used as a measuring criterion for the Standards and policies quality dimension.

The metrics and their related criteria referred to above discuss institutional areas that are not covered by other dimensions of quality. In order to ensure a high level of quality compliance from researchers while using these dimensions, it is necessary that an organisation has the infrastructure in place to motivate researchers to produce high-quality data.

What graphical representation does this stakeholder use for its data activities?



2. The Australian Research Council (ARC)

Purpose/Mandate of the organisation

The Australian Research Council (ARC) is a Commonwealth organisation established under the Australian Research Council Act, 2001, tasked with fostering knowledge and innovation for the Australian community by funding high-quality research, assessing its quality, engagement, and impact, and providing advice on research matters, with the Minister for Education reporting to the ARC (Australian Research Council, 2022).

Why is this stakeholder important?

The Australian Code for the Responsible Conduct of Research aims to promote ethical, responsible, and honest research practices in Australia and globally. This code establishes a framework for ethical research conduct, ensuring high-quality research, credibility, and community trust. It outlines standards for conducting research in Australia or under Australian institution sponsorship, ensuring a responsible and honest research culture (Australian Research Council, 2018).

Brief description of stakeholder activities

The best fundamental and applied research as well as research training are supported by the ARC. The organisation supports national competitions in all fields except clinical and related medical research. The ARC manages the National Competitive Grants Program (NCGP), which awards around \$800 million annually to Australia's most promising researchers, and we counsel the government on research-related issues. The Excellence in Research for Australia (ERA), which evaluates the calibre of research conducted at Australia's postsecondary educational institutions, is provided by the ARC. By research discipline, ERA offers a national inventory that is compared to international standards (Australian Research Council, 2022).

How does this stakeholder address data quality issues?

The Australian Code for Responsible Conduct of Research outlines best practices for managing research data, ensuring high-quality research and fostering an honest, ethical, and diligent research culture. It establishes a framework for ethical research conduct, ensuring credibility and community trust in the research endeavour. The Code describes the standards for conducting research, ensuring a high-quality and trustworthy research experience. The applicable guidelines of the Australian Code for Responsible Conduct of Research include:

P1 Honesty in the development, undertaking and reporting of research.

P2 Rigour in the development, undertaking and reporting of research

P3 Transparency in declaring interests and reporting research methodology, data and findings

P7 Accountability for the development, undertaking and reporting of research: comply with relevant legislation, policies and guidelines.

P8 Promotion of responsible research practices promote and foster a research culture and environment that supports the responsible conduct of research
(Australian Research Council, 2018).

What else is worth knowing about this stakeholder?

The Australian Research Integrity Committee uses the Investigation Guide as a benchmark for reviewing how an institution funded by the National Health and Medical Research Council (NHMRC) or Australian Research Council (ARC) has managed a potential breach of the Code (Australian Research Council, 2018).

3. Inter-university Consortium for Political and Social Research (ICPSR)

Data Lifecycle

Purpose/Mandate of the organisation

For a varied and growing community of social science researchers, the Inter-university Consortium for Political and Social Research (ICPSR) offers guidance and instruction in data access, curation, and analytical techniques. The depth and scope of the ICPSR data archive are unmatched; a variety of disciplines are covered by its data holdings. ICPSR data enable academics worldwide to perform secondary analysis, which advances scientific understanding (ICPSR, 2018).

Why is this stakeholder important?

The Inter-university Consortium for Political and Social Research (ICPSR) supports a growing and diverse community of social science researchers by offering leadership and training in data access, curation, and analysis techniques. ICPSR data contribute to scientific knowledge by enabling secondary analyses by researchers worldwide. In order to facilitate the replication, validation, and extension of original research findings and to ensure that these can continue throughout time, ICPSR makes data and methodologies available to the research community (ICPSR, 2018).

Brief description of stakeholder activities

ICPSR facilitates the social sciences by making original research findings replicable, verifiable, and extendable over time, and by making data and methodologies available to the research community. Long-term preservation of data is ensured by archiving with ICPSR, shielding it from irreversible damage, loss, and obsolescence. Knowledgeable staff is available to offer user support, which is an additional benefit of archiving data with ICPSR. Furthermore, more and more federal institutions that

support research, such as the National Institutes of Health (NIH), the National Science Foundation (NSF), and others, are advising or mandating that procedures for data management and sharing be included in all projects that get funding. These needs can be satisfied by archiving with ICPSR, which consistently complies with accepted standards and practices for long-term preservation (ICPSR, 2018).

How does this stakeholder address data quality issues?

Since data quality issues are covered by the phases of the ICPSR Data Lifecycle, as stated in Chapter 3 (Section 3.7.2), all other phases of the ICPSR Data Lifecycle (starting with Phase 2) will be covered in this section.

- Phase 2: Project Start-Up

Importance of Good Data Management

A researcher should plan for the final data collection, including metadata, after funding and starting the research project. A comprehensive management and preservation plan is crucial for the project's success, as early planning can significantly reduce costs. It is essential to include data and documentation topics related to the core data management plan in the project plan. The topics that have to be included in the DMP include:

- **Data and file structure.** *How will the data file be arranged and what format will it take? What is the unit of analysis? Will there be multiple shorter data records or just one large one? (ICPSR, 2018).*
- **Naming conventions.** *How will files and variables be named? What naming conventions will be used to achieve consistency? (ICPSR, 2018).*
- **Data integrity.** *How will data be input or captured? Will the variable formats be numeric or character? What checks will be used to find invalid values, inconsistent responses, incomplete records, etc.? What checks will be used to manage the data versions? For example, archives increasingly use checksums and other techniques for ensuring integrity (ICPSR, 2018).*
- **Preparing dataset documentation.** *What will the dataset documentation or metadata look like and how will it be produced? How much is necessary for future retrieval and archival processing? What documentation standard will be used? (ICPSR, 2018?)*
- **Variable construction.** *What variables will be constructed following the collection of the original data? How will these be documented? (ICPSR, 2018).*
- **Project documentation.** *What steps will be taken to document decisions that are made as the project unfolds? How will information be recorded on field procedures, coding decisions, variable construction, and the like. (ICPSR, 2018).*

- Phase 3: Data Collection and File Creation

The majority of errors are created during the data collection stage. The guidelines provided by ICPSR for accurately managing the data collection phase are explained in the next section.

Quantitative Data

Dataset creation and integrity: Many forms of errors, such as typographical errors, incorrect codes, and inconsistent records, can arise while transcribing information from a questionnaire or interview schedule to an actual data record. Because of this, it is advised to use a data collecting technique that directly collects data throughout the interview process. Then, using CATI/CAPI software, consistency checks can be incorporated into the data collection procedure to address issues that may arise during an interview (ICPSR, 2018). Nonetheless, even when data transcription is being done (from survey forms or publicly available tables), there are a few proactive measures that can be implemented to reduce the likelihood of errors:

- *Separate the coding and data-entry tasks as much as possible. Coding should be performed in such a way that distractions to coding tasks are minimized (ICPSR, 2018).*
- *Arrange to have particularly complex tasks, such as occupation coding, carried out by people specially trained for the task (ICPSR, 2018).*
- *Use a data-entry program that is designed to catch typing errors, i.e., one that is pre-programmed to detect out-of-range values (ICPSR, 2018).*
- *Perform double-entry of the data, in which each record is keyed in and then re-keyed against the original – several standard packages offer this feature and, in the re-entry process, the program catches discrepancies immediately (ICPSR, 2018).*
- *Carefully check the first 5 to 10 percent of the data records created, and then choose random records for quality-control checks throughout the process (ICPSR, 2018).*

As explained in Chapter 2 (Section 2.3.5) the recommended code and coding guidelines for researchers include:

Table 2: Recommended code and coding guidelines for researchers

Code and coding guideline	Description
Identification variables	Give each record the necessary fields at the outset to accommodate all identification factors. A distinct research number and respondent number to reflect each case are frequent identification factors.
Code categories	Code categories should be extensive, clearly defined, and mutually exclusive. Each interviewee should offer only one category of response. Ambiguity will make coding challenging and interfere with data interpretation.
Preserving original information	Code as much information as you can. It is more beneficial to record raw data rather than to compress or bracket it – such as in the cases of age and income. Secondary analysts are not limited to the brackets selected by others but can instead choose additional meaningful brackets with original or thorough data.

(ICPSR – University of Michigan, 2022)

All statistical software packages will be able to manage the data if consistent codes and coding are created, especially when data is translated from one format to another (e.g. SPSS to STATA). Codes and coding make it easier to analyse data files since survey responses are represented by numerical codes with defined meanings.

Encourage automation: During data collection, data capturers (field workers) must switch from manual to automated data capture techniques. Typographical errors, fatigue, misinterpretation, misunderstanding, and improper use of the data collection tool are just a few of the causes of data entry errors. The most effective way to reduce errors when collecting data (from field activities) is to automate the data flow.

Utilise mobile data collection apps where appropriate: Since skilled data capturers should have a backup automation system that verifies the accuracy of the data recorded, using automation techniques means

that data capturers will have a limited involvement in the physical data collection process. The reliability of the data is directly related to how few manual interactions are necessary. When it comes to the time and resources needed to gather field data, technology will also allow researchers to save a sizable sum of money.

Qualitative Data

With proper and complete documentation, archived qualitative data can provide a rich source of research material to be reanalysed, reworked, and compared to other data.

Confidentiality in qualitative data: Data depositors should ideally take care to delete any information that may allow any of their research subjects to be identified before uploading qualitative data to an archive. By developing an anonymisation plan before data collection and anonymising the data when the qualitative files are being prepared for analysis, this procedure can be made less laborious (ICPSR, 2018).

- Phase 4: Data Analysis

Master Datasets and Work Files

A master dataset should preferably be updated by one or two employees only and kept in read-only storage. This dataset should serve as the foundation for all analysis, and copying ought to be avoided. It is up to the researchers to choose whether to generate and save new variables in a work file or add them to the master dataset. If the latter is chosen, thorough documentation needs to be created and the new variables need to go through the required checks for anomalies and discrepancies. The final dataset representing published analyses is the version to archive (ICPSR, 2018).

- Phase 5: Prepare data for sharing

Respondent Confidentiality

The majority of the ICPSR guide is devoted to data preparation techniques that can be utilised to support secondary data analysts' as well as the principal investigators' research needs. However, one area of divergence has to be addressed that is required by the need to maintain respondent confidentiality in this section. This is a topic that requires special attention from researchers. After data are made available to the public, it is hard to keep an eye on how they are used to make sure that other researchers respect respondent privacy. Therefore, it is standard procedure when creating datasets for public use to make changes to the files so that any material that would jeopardise study subjects' confidentiality is either removed or hidden before the dataset is released. However, caution must be taken to ensure that the

changes do not unreasonably limit the researcher's capacity to replicate or expand the results of the initial research (ICPSR, 2018).

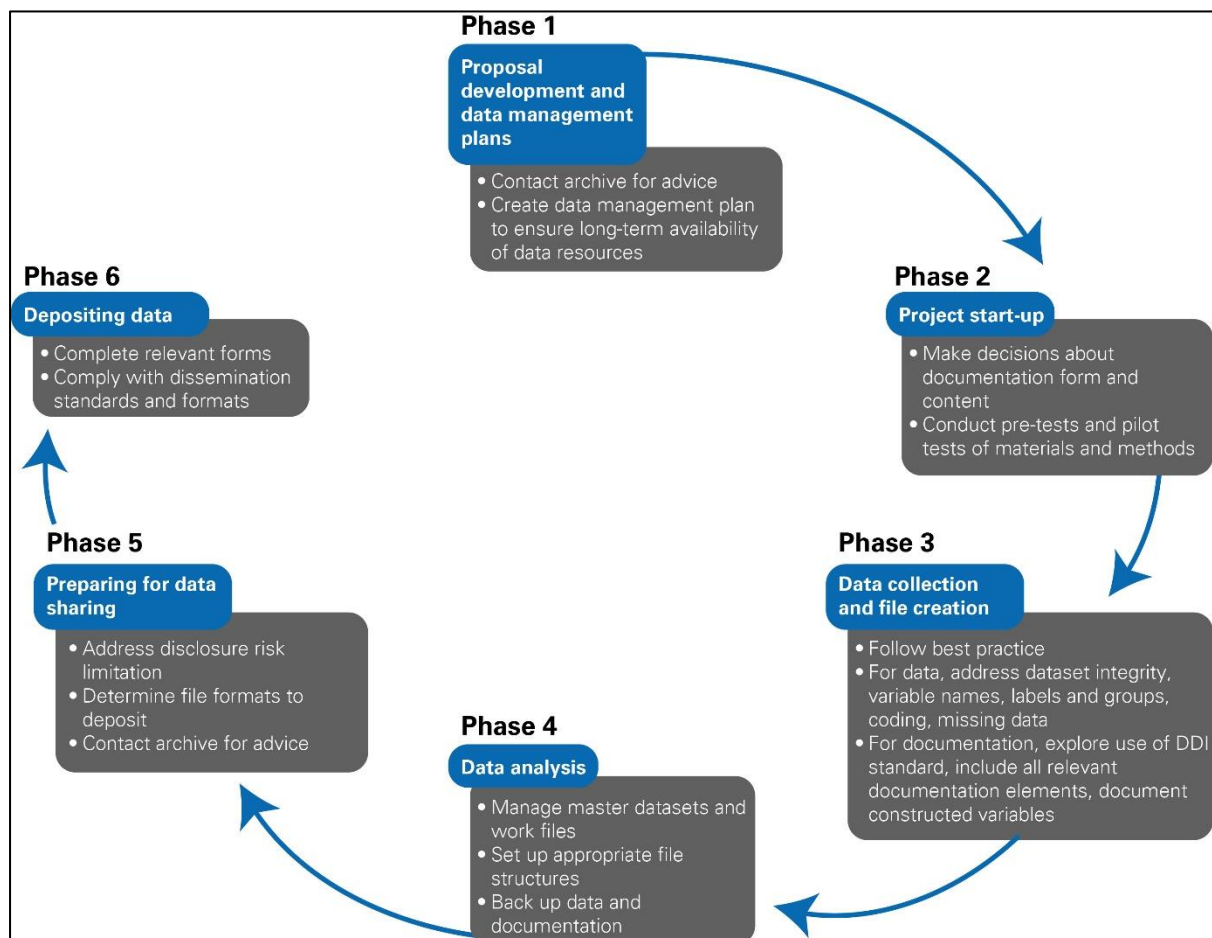
- Phase 6: Depositing data

The following elements should be taken into consideration while preparing data for final deposit in a data archive. Data creators should keep an eye out for changes regarding the digital preservation of social science data, as the list below is not exhaustive.

File Formats

It is necessary to arrange a dataset such that others may read it if it is to be archived. Ideally, A standard statistical format, such SAS, SPSS, or Stata, should be used. Provide the data in one of three common formats for data file preparation: (1) in raw ASCII format with setup files to read the data into standard statistical programs; (2) as a system file inside a particular analysis program; or (3) as a portable file generated by a statistical program.

What graphical representation does this stakeholder use for its data activities?



4. World Data System (WDS)/Core Trustworthy Data Repositories Requirements (Core Trust Seal)

Purpose/Mandate of the organisation

The World Data System's goal is to improve members' data services and repositories' capacities, influence, and sustainability through:

- *Creating trusted communities of scientific data repositories* (WDS, 2022).
- *Strengthening the scientific enterprise throughout the entire lifecycle of data and all related components creating first-class data that feeds first-class research output* (WDS, 2022).
- *Advocating for accessible data and transparent and reproducible science* (WDS, 2022).

Why is this stakeholder important?

Funders, both national and international, are more likely to require open data and data management practices that require data to be stored and made accessible for extended periods. Data must be kept in a reliable data repository if it is to be shared. Researchers should organise, curate, and archive their data in a way that protects the initial investment made in gathering it. Researchers need to be sure that information stored in archives will continue to be relevant and helpful in the future. Funding agencies have made this a crucial component of data management plans since they are requiring more and more ongoing access to the data generated by the projects they support. These days, several funders require that the data they support be stored in a trustworthy repository.

Repository sustainability brings up a lot of difficult questions in a variety of domains, including organisational, technical, financial, and legal. In order to guarantee the longevity and dependability of data repositories and, consequently, the possibility of data sharing over an extended length of time, certification might play a significant role. Repositories can prove to their users and funders that they are trustworthy by obtaining certification, which is a mark of approval from an impartial body (WDS, 2015).

Brief description of stakeholder activities

The Core Trust certification is a relatively simple process in which data repositories provide proof of their sustainability and reliability. A repository evaluates itself internally first, and then peers in the community examine it. These evaluations support the data communities (producers, repositories, and consumers) in raising knowledge of and adhering to set standards, as well as improving the quality and transparency of their processes. An inclusive environment with close interaction between the reviewers and the candidate repository is ensured by this community method (WDS, 2015).

How does this stakeholder address data quality issues?

The WDS quality management focus is based on the Core Trust Seal requirements (R7, R8, R11), as stated in Chapter 2 (Section 2.7.2). Each criterion is covered in the section that follows:

R7: The repository guarantees the integrity and authenticity of the data.

The repository must offer proof that it runs a metadata and data management system fit for guaranteeing authenticity and integrity during data access, ingest, and archival storage. Integrity guarantees that alterations to data and metadata are recorded and can be linked back to the person who made the change and their reasoning (WDS, 2015).

The degree of trustworthiness of the original data deposited and its provenance, as well as the connection between the original and distributed data and the preservation of any pre-existing linkages between datasets and/or metadata, are all included in the concept of authenticity.

Responses pertaining to data integrity for this requirement must provide proof of the following:

- *Description of checks to verify that a digital object has not been altered or corrupted (i.e., fixity checks) (WDS, 2015).*
- *Documentation of the completeness of the data and metadata (WDS, 2015).*
- *Details of how all changes to the data and metadata are logged (WDS, 2015).*
- *Description of version control strategy (WDS, 2015).*
- *Usage of appropriate international standards and conventions (which should be specified) (WDS, 2015).*

R8: The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

When it comes to figuring out whether data satisfy all requirements to be included in the collection and setting up the right management for their preservation, the evaluation function is essential. It is important to make sure that the data are clear and pertinent to the designated community that the repository serves.

Evidence pertaining to the following questions should be included in responses to this requirement:

- *Does the repository use a collection development policy to guide the selection of data for archiving? (WDS, 2015).*
- *Does the repository have quality control checks to ensure the completeness and understandability of data deposited? If so, please provide references to quality control standards and reporting mechanisms accepted by the relevant community of practice, and include details of how any issues are resolved (e.g., are the data returned to the data provider for rectification, fixed by the*

repository, noted by quality flags in the data file, and/or included in the accompanying metadata?) (WDS, 2015).

- *Does the repository have procedures in place to determine that the metadata required to interpret and use the data are provided? (WDS, 2015).*
- *What is the repository's approach if the metadata provided are insufficient for long-term preservation? (WDS, 2015).*
- *Does the repository publish a list of preferred formats? (WDS, 2015).*
- *Are quality control checks in place to ensure that data producers adhere to the preferred formats? (WDS, 2015).*
- *What is the approach towards data that are deposited in non-preferred formats? (WDS, 2015).*

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

To make sure that there is adequate information about the data available so that the Designated Community can evaluate the substantive quality of the data, repositories and depositors must collaborate. When the Designated Community is multidisciplinary and researchers may lack the necessary personal knowledge to assess quality solely based on data, such quality assessment becomes even more pertinent. The technical quality of data deposits must also be assessed by repositories in terms of the precision and standard of the materials supplied as well as the calibre of the metadata.

If a user can make an informed judgment about the credibility of the data through the available documentation, then their use in science is not prohibited even if the data or accompanying metadata have quality issues related to their research value. Regarding this prerequisite, scholars ought to describe:

- *The approach to data and metadata quality taken by the repository (WDS, 2015).*
- *Any automated assessment of metadata adherence to relevant schema (WDS, 2015).*
- *The ability of the Designated Community to comment on, and/or rate data and metadata (WDS, 2015).*
- *Whether citations to related works or links to citation indices are provided (WDS, 2015).*

What graphical representation does this stakeholder use for its data activities?

N/A

5. Digital Repository of Ireland

Purpose/Mandate of the organisation

The Digital Repository of Ireland (DRI) is a national Trusted Digital Repository (TDR) for Ireland's social science, humanities, and cultural heritage data, as well as a research-performing organisation (Digital Repository Ireland, 2023).

Why is this stakeholder important?

DRI offers access to social and cultural digital data as well as reliable, long-term, continuous digital preservation. In accordance with the FAIR data principles of findability, accessibility, interoperability, and reusability, DRI makes data publicly available. The goal is to actively manage digital content throughout time in order to preserve Ireland's social, cultural, and historical record and make it available to scholars, enthusiasts of cultural heritage, and the general public for years to come (Digital Repository Ireland, 2023).

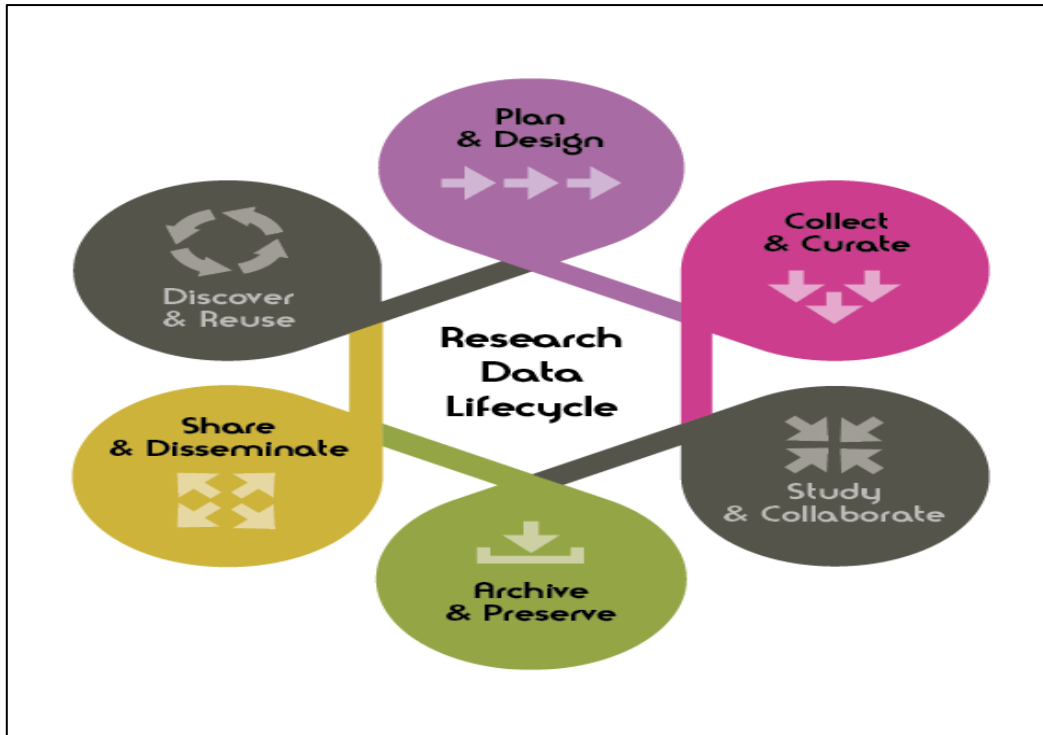
Brief description of stakeholder activities

Best practices in digital archiving, digital preservation, open access, open research, and FAIR data sharing are supported by the Digital Repository of Ireland. DRI uses a membership-based business model. They offer digital data stewardship services to a variety of member organisations, including government agencies, councils, institutes of higher learning, and cultural heritage organisations (the GLAM sector, which consists of galleries, libraries, archives, and museums). DRI encourages the diversity of datasets and provides non-funded organisations who are a part of the DRI Community with free DRI Memberships and associated advantages (Digital Repository Ireland, 2023).

How does this stakeholder address data quality issues?

A collection of guiding principles designed to make data and other research outputs Findable, Accessible, Interoperable, and Reusable, the FAIR principles are the focus of the DRI's lobbying and policy work. It has also been in the forefront of its application in Ireland as a reliable and accredited repository for data related to the social sciences and humanities, both nationally and internationally. The repository's work with FAIRification has been on optimising rules, ensuring that persistent identifiers, such as DOIs and ORCIDs, are applied consistently, producing rich, standardised, searchable information, and encouraging the reuse of vocabularies and licensing (Digital Repository Ireland, 2023).

What graphical representation does this stakeholder use for its data activities?



Appendix 2: Quality Process Framework

Quality Dimension	Potential scientific measures for the dimension	Potential technical measures for the dimension
1. Completeness	- Are there review processes in place that discover gaps or coverage issues in the data? (Victoria State Government, 2018) - Checking 5-10% of created records for consistency. (ICPSR, 2018) - Are known gaps in the data well documented for secondary users to be aware of? (Victoria State Government, 2018)	- Was a spell-check conducted to pick up obvious spelling and grammatical mistakes (i.e. your/you're; its/it's)? (UK Data Archive, 2011) - Have time coding, marginalia, annotations, highlighting or any other markings left by the depositor/researcher as part of their initial analysis been removed? (UK Data Archive, 2011) - Were responsibilities for data management and curation in the research team specified? (DCC, 2021)
2. Uniqueness	- What type of metadata accompanies the data? (DMPOnline, 2020) - Source data based on common identifiers documented. (Statistics South Africa, 2010) - How are variable names and groups handled for consistency? (ICPSR, 2018) - A data file does not have two attributes or two variables with the same ID. (DAMA UK Working Group, 2013)	- Is the confidential respondent data removed before the dataset is made public?? (Oracle, 2009) - Have the variable and column characteristics been documented for interpretation purposes? (University of California, Santa Barbara, 2016)
3. Timeliness	- File formats of data files should be defined in the DMP. (Rolland, 2016) - Is data processed and delivered rapidly without delays? (Jayawardene, 2016)	- Does the data represent the reality of the required point in time? (DAMA UK Working Group, 2013) - Will the data be made available as soon as possible both internally and externally? (York University, 2018)

Quality Dimension	Potential scientific measures for the dimension	Potential technical measures for the dimension
	- How is data quality monitored when working on several sites, and in a timely manner? (DCC, 2021)	
4. Validity	- What is the degree to which attributes are not missing from a database schema? (Open Risk Academy, 2016) - Which methods were used to identify missing data? (Cooper & Kahn, 2012) - Was coding done separately from data entry to minimise distractions and making errors? (ICPSR, 2012) - Were missing values identified and missing value codes defined? (DataOne - Wiggins et al., 2013)	- Were syntax definitions and identifiers in related documents (e.g. user guide) accurately documented? (DAMA UK Working Group, 2013) - Was the data checked, validated and cleaned? (UK Data Archive, 2017)
5. Accuracy	- Are the value labels used a true reflection of what was investigated? (Buckey, 2016) - Were records entered twice to check for duplication? (ICPSR, 2018) - Was data management planning (formats, storage) done? (UCSB, 2016) - Does the data meet a certain set of accepted quality specifications? (Wang et al., 1995)	- Are definitions of data collection and processing available widely in the organisation to provide a common understanding? (Victoria State Government, 2018) - Was a review done to make sure the data and documentation correspond? (MANTRA, 2017) - What types of checks were used for invalid and incomplete records in the data file? (ICPSR, 2018)
6. Consistency	- Documented workflow process during data processing. (CoreTrustSeal -Abrams, 2020) - Have common data definitions and standards been used across multiple data sets? (Victoria State Government, 2018)	- Are data files organised using active versions and file names? (MANTRA, 2017)

Quality Dimension	Potential scientific measures for the dimension	Potential technical measures for the dimension
	- Accurate description of each variable. (Holland et al., 2018) - Can the original dataset remain consistent and be traced over time even when it has been changed? (Victoria State Government, 2018) - Describing the nature and scale of the data. (ICPSR, 2012)	- Have data quality control processes been documented and communicated to research teams? (DataOne - Wiggins et al., 2013) - Are correct definitions used in metadata to help interpret the data? (UK-DAMA, 2013)
7. Interpretability	- Were consistency check programmes used for quality control during data collection? (ICPSR, 2018) - Are consistent codes used to represent missing values? (ICPSR, 2018) - Are definitions used in the methodology of data collection and processing documented? (Statistics South Africa, 2010) - To what degree was the data anonymised to avoid multiple data collection among vulnerable groups (RDA, 2018)	- Is the data adequately described in a related document (e.g. user guides)? (UK Data Archive, 2017) - Establish quality control research methods for data management procedures. (York University, 2018) - Were quality control methods used throughout the data lifecycle? (Reisch & Webster, 2004)
8. Integrity	- Are there automated ways to conduct research and produce quality data? (RDA, 2018) - Is the researcher familiar with different data formats in order to use the best format? (ICPSR, 2018). - Are your data handling skills good in various software environments? (MANTRA, 2017) - Use of a data-entry program to pick up typing errors. (ICPSR, 2018)	- Values and related practices that maintain users' confidence in the agency producing the product/data file. (Statistics South Africa, 2010) - Identifying, designing, and implementing data quality rules. (Maydanchik, 2007) - Naming conventions: Which naming convention standards were used to achieve consistency? (DANS, 2010)

Quality Dimension	Potential scientific measures for the dimension	Potential technical measures for the dimension
	- Naming conventions: How were variables named? (DANS, 2010) - Was the provisional hypothesis tested using a different case to verify the results? (DataOne Wiggins et al., 2013)	
9. Relevance	- Is the data relevant to user needs and does it reflect the real-world situation? (Victoria State Government, 2018) - Good research practice/RDM to mitigate risks. (Knight, 2015) - Were data entry errors verified through a data quality audit process? (Statistics South Africa, 2010)	- Face validity: involves assessing whether a logical relationship exists between the variable and the proposed measure. (Monette et al., 2002) - Data quality: were correct data management practices used from initial development of the project? (Australian Data Archive, 2018)
10. Comparability and Coherence	- Was basic quality control ensured during data collection and data processing? (UK Data Archive, 2011) - Does the data collection method provide validity regardless of the sample type and instrument used? (Creswell, 2014) - Data can be understood and viable for secondary usage. (Domingus, 2015) - Are you aware of the strengths and limitations of the data that affect the truthfulness of the data? (USAID, 2019)	- Is the data available in a variety of other file formats? (Statistics South Africa, 2010) - Was a quality approach used to manage and preserve data? (Rouchon et al., 2013) - Were accepted research processes and protocols followed? (Johnston, 2014) - Executing a recurrent data quality assessment. (Maydanchik, 2007)
11. Accessibility	- Does the selected data repository ensure authenticity of digital objects and metadata? (DSA-WDS, 2016)	- Can data be accessed easily using the appropriate medium? (Statistics South Africa, 2010) - Is research data available in a usable format? (Tsang, 2014)

Quality Dimension	Potential scientific measures for the dimension	Potential technical measures for the dimension
	- What type of repository will the data be deposited in (i.e. long-term preservation or access repository)? (University of Leicester, 2017) - Are data licensing issues (restrictions on use, redistribution of data, permissions, terms and conditions for derivative works) addressed? (Patel, 2016) - Terms-of-use: is there a statement clarifying how data can be used by others? (The ERC Scientific Council, 2018)	
12. Methodological Soundness	- Authenticity: is research data the same genuine object it was originally and not changed over time? (Clarke, 2014)	- Copyright of the data must be clearly explained. (Patel, 2016) - Have international/national or peer-agreed standards been used for managing the data? (Statistics South Africa, 2010) - Is data copyright and data ownership understood and stipulated in the research contract? (Coulehan & Wells, 2012) - Were organisational methods and procedures implemented when managing data? (Reisch & Webster, 2004)
13. Quality Check Processes	- Are data dictionaries available that explain validations, specifications, and glossary terms? (Victoria State Government, 2018) - Organising rule and error catalogues to help secondary users on how to deal with possible errors. (Maydanchik, 2007)	- Does your institution provide ethical guidelines for research? (Coulehan & Wells 2012) - Do members of the research team have relevant experience and expertise to successfully manage the project (data)? (World Bank Group, 2015) - How were field operations, coding decisions, variable construction documented? (DANS, 2010)

Quality Dimension	Potential scientific measures for the dimension	Potential technical measures for the dimension
14. Standards and Policies	- Follow documented institutional policies and procedures when conducting a research project. (UK Data Archive - Lavoie, 2014) - Using an independent data quality audit process. (Statistics South Africa, 2010)	- Were policies informing how a repository and institution at large should manage its research data followed? (CCSDS, 2011) - Is there a policy covering all aspects of the data lifecycle? (Statistics South Africa, 2010)
15. Flexibility of the data	- Outlining strategies for handling research data. (Australian Research Council, 2016) - Researcher uses a data file format recommended by a data repository. (DSA-WDS, 2016) - Did the data collection instrument measure the intended concept or the concept is open to interpretation? (Singleton et al., 1993)	- How was field procedure information recorded? (DANS, 2010)
16. Usability of the data	- Adherence to funder requirements on dissemination and sharing of research findings. (Bryant et al., 2017) - Does the data comply with FAIR (Findable, Accessible, Interoperable and Reusable) principles? (ERC, 2018) - Has a Codebook containing codes and their definition been created? (Creswell, 2014)	- Data sharing clause in the consent form. (Ball, 2012) - The management of the Master data file. (ISO, 2018)

Appendix 3: Data lifecycles with an emphasis on quality

#	Lifecycle	Source
1	The UK Data Archive research data lifecycle	Ball, A. 2012. Preservation and Curation in Institutional Repositories, Digital curation centre technology watch report. Available from: www.dpconline.org/advice/technology-watch-reports [accessed 07 February 2016].
2	Ellyn Montgomery, USGS, Data Lifecycle Diagram	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
3	FGDC Stages of the Geospatial Data Lifecycle	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
4	University of Oxford Research Data Management Chart	Wissik, T. & Durco, M. 2015. Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions. Available from: http://www.ep.liu.se/ecp/123/008/ecp15123008.pdf [accessed 16 July 2018].

#	Lifecycle	Source
5	NOAA Environmental Data Life Cycle Functions	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
6	Open Archival Information System (OAIS) Framework	Lavoie, B. 2014. The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). Available from: https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file [accessed 14 July 2018].
7	USGS Scientific Information Management Workshop Vocabulary	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
8	Peter Fox Lifecycle Diagrams	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
9	National Science Foundation	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].

#	Lifecycle	Source
10	NDIIPP Preserving Our Digital Heritage	The National Digital Information Infrastructure. 2011. Preserving Our Digital Heritage. Available from: http://www.digitalpreservation.gov/documents/NDIIPP2010Report_Post.pdf [accessed 22 February 2020].
11	What Researchers Want	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
13	IWGDD's Digital Data Life Cycle Model	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
14	Data Science Lifecycle	Microsoft Azure. 2020. Data acquisition and understanding stage of the Team Data Science Process. Available from: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle-data [accessed 21 February 2020].
15	Linear Data Life Cycle	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].

#	Lifecycle	Source
16	Scientific Data Lifecycle	Mario Valle Web. 2019. Scientific Data Management. Available from: http://mariovalle.name/sdm/scientific-data-management.html [accessed 14 July 2018]
17	Cassandra Ladino Hybrid Data Lifecycle Model	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
18	Ray Obuch Data Management – A Lifecycle Approach	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
19	USGS Data Management Plan Framework (DMPf) – Smith, Tessler, and McHale	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
20	BLM Data Management Handbook	Ustinovičius, L., Rasiulis, R., Nazarko, L., Vilutiene, T. & Reizgevicus, M. 2015. Innovative research projects in the field of Building Lifecycle Management. Available from: https://www.researchgate.net/publication/283962838

#	Lifecycle	Source
		Innovative Research Projects in the Field of Building Lifecycle Management [accessed 14 July 2018].
21	ARL Joint Task Force on Library Support for E-Science	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
22	U.S. Department of Health and Human Services Key Components	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
23	ICPSR Guide to Social Science Data Preparation and Archiving	The Regents of the University of Michigan. 2020. Guide to Social Science Data Preparation and Archiving. Available from: https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/ [accessed 17 July 2018].
25	IBM Aspects of Lifecycle Management - Research	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].

#	Lifecycle	Source
26	University of California San Diego Digital Curation Lifecycle	UC San Diego. 2020. Data Management. Available from: https://library.ucsd.edu/lpw-staging/research-and-collections/data-curation/data-management/index.html [accessed 14 July 2018].
27	University of Miami Scientific Data Lifecycle	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
28	Managing Research Data Lifecycles through Context	Oregon Health & Science University. 2020. Plan and Organize. Available from: https://www.ohsu.edu/library/plan-and-organize [accessed 14 July 2018].
29	Purdue University Research Repository – Curation Service Model	Witt, M. 2012. Curation Service Models: Purdue University Research Repository. Available from: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?referer=https://www.google.co.za/&httpsredir=1&article=1002&context=lib_fspres [accessed 21 July 2018].
30	Data Lifecycle Management at Dell	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].

#	Lifecycle	Source
31	John Faundeen & Ellyn Montgomery “Spins”	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
32	DataTrain Model 1	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
33	DataTrain Model 2	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
34	Steve Tessler Data Management Cycles	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
35	Peter Fox Full Lifecycle of Data	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].

#	Lifecycle	Source
36	Steve Tessler Data and System Lifecycle Models	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
37	University of Virginia Library lifecycle	University of Virginia. 2020. Steps in the Data Life Cycle. Available from: https://data.library.virginia.edu/data-management/lifecycle/ [accessed 14 July 2018].
38	Mario Valle	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
39	Michigan State University Records Life Cycle Model	ResearchGate. 2019. The Evolution, Approval and Implementation of the U.S. Geological Survey Science Data Lifecycle Model. Available from: https://www.researchgate.net/figure/Michigan-State-University-Records-Life-Cycle-http-archivesmsuedu-records_fig1_320360683 [accessed 14 July 2018].
40	GeoMAPP Geoarchiving Process Lifecycle	http://www.geomapp.net/docs/GeoMAPP_FGDC_UHDWG_20110728.pdf

#	Lifecycle	Source
41	Jeff de La Beaujardière's proposed Data Management Framework	https://slideplayer.com/slide/6227956/
42	Sarah Demb Continuum Model	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
44	ARMA International Life Cycle	https://www.arma.org/page/Perspective
45	JISC Simon Hodson Research Data Lifecycle	https://www.researchgate.net/figure/JISC-Webinar-Simon-Hodson-Programme-Manager-Managing-Research-Data-Digital_fig4_320360683
46	United Kingdom Digital Archive Research Data Lifecycle	https://www2.le.ac.uk/services/research-data/images/new_datalifecycle.PNG/view

#	Lifecycle	Source
47	Lisa Dunn Data Flow Model	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].
48	JISC Information Environment	https://researchportal.bath.ac.uk/files/456723/TSreview-jisc-final-Sept.doc
49	Make it Digital New Zealand's Digital Content Life Cycle	http://apennings.com/global-e-commerce/digital-content-flow-and-life-cycle-the-value-chain/
50	CASPAR Preservation Work Flow	https://www.researchgate.net/figure/CASPAR-Information-Flow-Architecture_fig3_250150721
51	DataOne Lifecycle	https://www.dataone.org/data-life-cycle
53	IWGDD Scientific Data Life Cycle Model	CEOS. 2012. Data Life Cycle Models and Concepts CEOS Version 1.1 Available from: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v11.docx?api=v2 [accessed 14 July 2018].

#	Lifecycle	Source
54	DDI version 3.0 Combined Life Cycle Model	https://www.ddialliance.org/sites/default/files/Concept-Model-WD.pdf
55	Data-Information- Knowledge Ecosystem	http://www.thinkinc.com/blog/introducing-the-flowscore-a-digital-ecosystem-analytics-framework/
56	USGS lifecycle	https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle
57	PII lifecycle	https://blog.lansa.com/rapid-application-development/protect-customers-pii
58	Enterprise data lifecycle	https://blog.lansa.com/rapid-application-development/protect-customers-pii
59	CRUD lifecycle	https://searchdatamanagement.techtarget.com/definition/CRUD-cycle
60	DIKW Information pyramid	https://www.i-scoop.eu/big-data-action-value-context/dikw-model/

Appendix 4: Measurements for each quality dimension

4.1 Scientific measures (researchers)

Quality Dimension	Measure	Level 1	Level 2	Level 3	Level 4
1. Completeness	Internal review process completed and documented	No tolerance for defect; Internal review completed and documented	No tolerance for defect; Internal review completed and documented	No tolerance for defect; Internal review completed and documented	Internal review completed and documented
	Individual records checked for completeness	5-10% checked	At least 25% checked	At least 50% checked	All records checked
	Gaps in the data documented	No tolerance for defect; All gaps documented	No tolerance for defect; All gaps documented	No tolerance for defect; All gaps documented	All gaps documented in a related document
2. Uniqueness	Metadata standard followed	Own unique variables indicated	Dublin core was used and all fields match	DDI was used, some fields not completed	DDI was used, all fields completed
	Source data use common identifiers and identifiers are documented	Identifiers are coded	Identifiers are documented	Common identifiers were used but there is no documented proof	Source data use common identifiers and identifiers are documented
	Handling of Variables and group names for consistency	Variables and group names are 10% reliable	Variables and group names are 25% reliable	Variables and group names are 50% reliable	All Variables and group names are reliable

	Variables do not have the same ID	50 % of variables share the same ID	25% of variables share the same ID	10% of variables share the same ID	No variables have the same ID
3. Timeliness	Data file format defined in DMP	Data collection format used in the field defined	Data format used when saving data defined	Data format used when processing data defined	All data formats for managing data defined in DMP
	Data processed and delivered without delays	Delivery of data suspended due to handling challenges	Data processing delayed due to technical problems	Data processed and delivered with retrospective changes	Data processed without experiencing delays
	Monitoring data quality while on several sites in a timely way	Data quality not monitored while on several sites	Data quality only monitored in sites where there's network coverage	Data quality monitored in sites without data managers	Data quality monitored in all sites in a timely manner
4. Validity	Degree to which attributes are not missing	Degree of missing characteristics not known	Degree of missing characteristics to be determined by the user	Degree of missing characteristics known but not documented	Degree of missing characteristics fully accounted for
	Methods used to identify missing data	No tolerance for defect; Identifying missing data methods used	No tolerance for defect; Identifying missing data methods used	No tolerance for defect; Identifying missing data methods used	Missing data methods are in place and are used
	Coding done separately from data-entry to minimise errors	No coding done	Coding done during data-entry	Coding done during data entry and data cleaning	Coding done separately from data entry
	Missing values identified and defined	Missing values not accounted for	Only numeric missing values defined	Only string missing values defined	All missing values identified and defined

5. Accuracy	True reflections of value labels	5-10% of value labels are a true reflection	At least 25% of value labels are a true reflection	50% of value labels are a true reflection	All value labels are a true reflection
	Checking for duplication by entering a record twice	Duplication not checked	25% of records checked using double-entry	50% of records checked using double-entry	All records entered twice to check for duplication
	Data management planning done	No tolerance for defect; Data management planning done	No tolerance for defect; Data management planning done	No tolerance for defect; Data management planning done	Data management planning done
	Data complies with quality specifications	Data does not comply with any quality specification	Compliance with quality specification not acknowledged	Compliance with general quality specifications	Data fully complies with quality specifications
6. Consistency	Documenting data processing workflow	Data processing workflow not documented	Data processing workflow documented with unresolved corrections	Data processing workflow documented excluding QA	Data processing workflow documented including QA
	Common data definitions and standards used across multiple data sets	Common data definitions and standards not used	Different data definitions and standards used for multiple data sets	Similar (50%) data definitions and standards used for multiple data sets	The same data definitions and standards used across multiple data sets
	Accurate description of variables	Variables not accurately described	Variable description accuracy describes with spelling errors	Variables description open for interpretation	Variable descriptions accurately described

	Consistency of original dataset when changes have been made	Original dataset inconsistent after changes have been made	Changes creating misunderstanding of original dataset	Consistency of dataset differs from original dataset	Consistency of original dataset maintained when changes have been made
	Description of the nature and scale of a data file	Nature and scale of data not described	25% of the nature and scale of data described	50% of the nature and scale of data described	Full description of the nature and scale of a data file
7. Interpretability	Quality control consistency checks programmes used during data collection	No tolerance for defect; Quality control checks used during data collection	No tolerance for defect; Quality control checks used during data collection	No tolerance for defect; Quality control checks used during data collection	Consistent quality control checks used during data collection
	Consistent codes used for missing values	No tolerance for defect; Consistent codes used for missing data	No tolerance for defect; Consistent codes used for missing data	No tolerance for defect; Consistent codes used for missing data	Consistent codes used for missing values
	Definitions used in the methodology of data collection and processing documents	No tolerance for defect; Correct definitions used in methodology of data collection and processing documents	No tolerance for defect; Correct definitions used in methodology of data collection and processing documents	No tolerance for defect; Correct definitions used in methodology of data collection and processing documents	Correct definitions used in the methodology of data collection and processing documents
	Degree of anonymisation to prevent multiple data collection among vulnerable groups	Anonymisation not done	Anonymisation done excluding identification information of participants	Anonymisation done excluding location and organisation names	Full anonymisation done to prevent multiple data collection among vulnerable groups

8. Integrity	Automated methods used to conduct research and produce quality data	Automated methods not used	The usage of automated methods still being investigated	The usage of automated methods used on a trial basis	Full usage of automated methods to conduct research and produce quality data
	Researcher familiarity with different data formats to facilitate selection of the best data format	No tolerance for defect; Familiarity with different data formats	No tolerance for defect; Familiarity with different data formats	No tolerance for defect; Familiarity with different data formats	Researcher familiar with different data formats
	Data handling skills in different data software programs	No data handling skills in different data software programs	Data handling skills in one data software program	Data handling skills in proprietary data software programs	Data handling skills in different data software programs
	Using data-entry programs to identify typing errors	Data-entry programs not used to identify capturing errors	Use of data-entry programs is being investigated	Data-entry programs are used only for confidential research projects	Data-entry programs are used to identify capturing errors
	Naming convention for variable names	Naming convention not used for variable names	Naming convention used only on biographic variables	In-house naming convention used for variable names	Best practice naming convention used for variable names
	Verification of research results using hypothesis testing (through a different case)	Verification of research results not done	Verification of research results done only for quantitative research	Verification of research results done only for qualitative research	Verification of research results done using a different case

9. Relevance	Relevance of data to user needs in real-world situation	Data not relevant to user needs	Data relevant to user needs based on interpretation	Data answers certain areas of users-needs	Data fully relevant to user needs in real-world situation
	Good RDM to mitigate risks	RDM not mitigating risks	RDM mitigates only operational risks	Good RDM used but doesn't mitigate risks	Risks mitigated by good RDM
	Data quality audit process to verify data entry errors	No tolerance for defect; Data entry audit process to verify data entry errors	No tolerance for defect; Data entry audit process to verify data entry errors	No tolerance for defect; Data entry audit process to verify data entry errors	Data entry audit process conducted to verify data entry errors
10. Comparability and Coherence	Basic quality control during data collection and processing	No tolerance for defect; Basic quality control used for data collection and processing	No tolerance for defect; Basic quality control used for data collection and processing	No tolerance for defect; Basic quality control used for data collection and processing	No tolerance for defect; Basic quality control used for data collection and processing
	Validity of data collection method irrespective of sample and instrument used	Data collection method not valid	Validity of the data collection method verified using certain questions in the data collection method	Validity of the data collection method verified using only the sample	Validity of the data collection method verified using only the instrument used
	Data understandable and practical for secondary usage	No tolerance for defect; Data is logical and practical for secondary use	No tolerance for defect; Data is logical and practical for secondary use	No tolerance for defect; Data is logical and practical for secondary use	Data is logical and viable for secondary usage

	Aware of the strengths and limitations of the data that impacts data truthfulness	No awareness of strengths and limitations of data	25% awareness of data strengths and limitations	50% awareness of data strengths and limitations	Full awareness of data strengths and limitations
11. Accessibility	Data repository ensuring authenticity of digital objects and metadata	Data repository doesn't ensure authenticity of data and metadata	Data repository only provides metadata fields to be captured	Data repository allows for version control and preserves research data	Data repository ensuring validity of digital objects and metadata
	Type of repository the data will be deposited in	Type of repository the data will be deposited in is irrelevant	Type of repository data will be deposited in promotes data discovery	Type of repository data will be deposited in promotes data interoperability	Type of repository the data will be deposited in is relevant and promotes secondary usage
	Addressing data licensing issues	Data licensing issues not addressed at all	Data licensing issues not addressed formally	Data licensing issues informed by institutional policies	Data licensing issues informed by funder requirements and institutional policies
	Statement explaining data terms-of-use by others	No tolerance for defect; There is a statement that explains how data can be used by others	No tolerance for defect; There is a statement that explains how data can be used by others	No tolerance for defect; There is a statement that explains how data can be used by others	Statement explaining how data can be used by others available on the repository and related documentation of the data

12. Methodological Soundness	Authenticity of research data over time	The truthfulness of the originality of research data has changed over time	The truthfulness of the originality of research data has been affected by technological changes	The truthfulness of the originality of research data has not changed due to secure preservation	The truthfulness of the originality of research data remains intact
13. Quality Check Processes	Availability of data dictionary that explains validations, specifications, and glossary terms	No tolerance for defect; Validation, specification and glossary terms of data are available	No tolerance for defect; Validation, specification and glossary terms of data are available	No tolerance for defect; Validation, specification and glossary terms of data are available	Data dictionaries form part of related documents that accompany research data.
	Rule and error catalogue to assist secondary users in dealing with possible errors	No tolerance for defect; Rule and error catalogue to assist secondary users in dealing with possible errors accompanies the data	No tolerance for defect; Rule and error catalogue to assist secondary users in dealing with possible errors accompanies the data	No tolerance for defect; Rule and error catalogue to assist secondary users in dealing with possible errors accompanies the data	Rule and error catalogues to assist secondary users in dealing with possible errors are part of related documents
14. Standards and Policies	Following documented policies and procedures when conducting a research project	Documented policies and procedures do not exist	Documented policies and procedures not followed	Documented policies and procedures not fully followed	Documented policies and procedures fully adhered to
	Independent data quality audit process implemented	Independent data quality audit process not acknowledged in the research process	Independent data quality audit process not implemented	Independent data quality audit process implemented when	Independent data quality audit process accepted and implemented

				errors are identified after data is made available	
15. Flexibility of the data	Outlining strategies for handling research data	No tolerance for defect; Strategies for handling research data are outlined	No tolerance for defect; Strategies for handling research data are outlined	No tolerance for defect; Strategies for handling research data are outlined	Strategies for handling research data are clearly outlined in related documentation
	Using a data file format recommended by a data repository	Using a data file format not recommended by a data repository	A combination of using a data file format recommended by a data repository and a data file format preferred by a researcher	Using a data file format recommended by the data repository manager/administrator	Only using a data file format recommended by a data repository
	Data collection instrument measuring intended concepts	Data collection instrument not measuring intended concepts	Data collection instrument measuring 25% of intended concepts	Data collection instrument measuring 50% of intended concepts	Data collection instrument measuring all of the intended concepts
16. Usability of the data	Adherence to funder requirements on dissemination and sharing of research findings	Non-adherence to funder requirements on dissemination and sharing of research findings	Adherence to 25% of funder requirements on dissemination and sharing of research findings	Adherence to 50% of funder requirements on dissemination and sharing of research findings	Adherence to all the funder requirements on dissemination and sharing of research findings
	Compliance with FAIR principles	Non-compliance with FAIR principles	25% compliance with FAIR principles	50% compliance with FAIR principles	Full compliance with FAIR principles

	Creation of a Codebook containing codes and definitions	No tolerance for defect Codebook with definition codes created	No tolerance for defect Codebook with definition codes created	No tolerance for defect Codebook with definition codes created	Codebook with accurate codes and their definitions is created

4.2 Technical measures (curators)

Dimension	Measure	Level 1	Level 2	Level 3	Level 4
1. Completeness	Spell check done to pick up mistakes	No tolerance for defect; Spell check conducted	No tolerance for defect; Spell check conducted	No tolerance for defect; Spell check conducted	Spelling and grammatical mistakes corrected through a spell-check
	Remove markings in the data left by a researcher as part of their initial analysis	Time coding markings removed	Annotations and highlighting removed	Other kinds of markings removed	All data markings removed
	Specification of data management and curation responsibilities communicated to research teams	Responsibilities not communicated	Data management responsibilities communicated	Curation responsibilities communicated	Data management and curation responsibilities communicated
2. Uniqueness	Data are anonymised (respondent confidentiality)	Zero tolerance for deviation. All data anonymised	Zero tolerance for deviation. All data anonymised	Zero tolerance for deviation. All data anonymised	Zero tolerance for deviation. All data anonymised

	Variable and column characteristics documented for interpretation purposes	No tolerance for defect; Variable and column characteristics documented	No tolerance for defect; Variable and column characteristics documented	No tolerance for defect; Variable and column characteristics documented	Variable and column characteristics documented in a related document
3. Timeliness	Data representing required point of time reality	No tolerance for defect; Data represents required point of reality	No tolerance for defect; Data represents required point of reality	No tolerance for defect; Data represents required point of reality	Data represents the required point of reality
	Data being made available as soon as possible	Availability of data delayed indefinitely	Data is only available after an embargoed period	Data is available upon request to registered users	Data is available as soon as possible
4. Validity	Syntax definitions and identifiers accurately documented	Syntax definitions and identifiers not documented	Syntax definitions and identifiers partially documented	Syntax definitions and identifiers documented with changes to be made	Syntax definitions and identifiers correctly documented in a User guide
	Checking, validating and cleaning of data	Checking, validating and cleaning of data not done	Data checking, validating and cleaning done superficially	Data checking, validating and cleaning done up to 50%	100% of Data checking, validating and cleaning done
5. Accuracy	Data collection and processing definitions available in the organisation	Data collection and processing definitions not available	Data collection and processing definitions are in a draft form	Data collection and processing definitions available on an ad-hoc basis	Data collection and processing definitions readily available

	Review to check if data and documentation correspond	Data and documentation review not done	Data and documentation review done with only data collection documents	Data and documentation review done with only project team documents	Data and documentation review done with all documentation
	Type of checks for invalid and incomplete records	No tolerance for defect; Checks for invalid and incomplete records done	No tolerance for defect; Checks for invalid and incomplete records done	No tolerance for defect; Checks for invalid and incomplete records done	Checks for invalid and incomplete records done
6. Consistency	Active versions and file names used to organise data files	No version control and active file names used to organise data files	Inconsistent versions and file names used to organise data files	Correct version and file names used without following a best practice naming convention	Best practice naming convention versions and file names used to organise data files
	Documenting and communicating data quality control process documents with research teams	No tolerance for defect; Sharing data quality control processes with research teams	No tolerance for defect; Sharing data quality control processes with research teams	No tolerance for defect; Sharing data quality control processes with research teams	Data quality process documents shared with research teams
	Correct definitions used in metadata for interpretation purposes	No tolerance for defect; Correct definitions used in metadata to interpret data	No tolerance for defect; Correct definitions used in metadata to interpret data	No tolerance for defect; Correct definitions used in metadata to interpret data	Correct definitions are used in metadata to interpret data
7. Interpretability	Sufficient description of data in related documents	No descriptions of data in related documents	25% accurate descriptions of data in related documents	50% accurate descriptions of data in related documents	Satisfactory descriptions of data in related documents

	Establishing quality control methods for data management procedures	No tolerance for defect; Quality control methods established for data management procedures	No tolerance for defect; Quality control methods established for data management procedures	No tolerance for defect; Quality control methods established for data management procedures	Quality control methods established and used for data management procedures
	Quality control methods applied throughout the data lifecycle	Quality control methods not applied throughout the data lifecycle	Quality control methods applied only in the Creating Data phase	Quality control methods applied only in the Creating and Processing Data phases	Quality control methods applied in all stages of the data lifecycle
8. Integrity	Values and related practices to maintain secondary user confidence in the institution	Institution's values and related practices are questionable	Secondary users have minor trust in the values and related practices of the institution	Values and related practices are considered reliable by secondary users	Secondary users are confident in the values and related practices of the institution
	Identifying, designing, and implementing data quality rules	Data quality rules do not exist	Research teams using their own data quality rules	Data quality rules identified and designed but not implemented	Identifying, designing, and implementing data quality rules
	Naming convention standard used	Naming convention standard not used	Naming convention standard used only for related documents	Naming convention standard used only for data files	Best practice naming convention standard used for all files
9. Relevance	Logical relationship between variables and the proposed measures	No tolerance for defect; Variable and proposed measures have a logical relationship	No tolerance for defect; Variable and proposed measures have a logical relationship	No tolerance for defect; Variable and proposed measures have a logical relationship	Logical relationship exists between variables and proposed measures

	Usage of correct data management practice from project onset	Data management practice not used from project onset	Data management practice used in certain stages of the research project	Correct data practice only used as part of funder requirements	Usage of correct data management project throughout the research project.
10. Comparability and Coherence	Data available in a variety of file formats	No tolerance for defect; Data is available in a variety of file formats	No tolerance for defect; Data is available in a variety of file formats	No tolerance for defect; Data is available in a variety of file formats	Data available in proprietary and non-proprietary formats
	Quality approach used to manage and preserve data	Quality approach not used to manage and preserve data	Quality approaches defined in the DMP but not used to manage and preserve data	General quality approach used to manage and preserve data	Recognised quality approach used to manage and preserve data
	Accepted research processes followed	No tolerance for defect; Accepted research processes used	No tolerance for defect; Accepted research processes used	No tolerance for defect; Accepted research processes used	Accepted research processes and protocols used
	Executing a recurrent data quality assessment	Recurrent data quality assessment not executed	Researchers leaving the data quality assessment task to curators	Researcher executing data quality assessment on an ad-hoc basis	Researchers and curators executing data quality assessment
11. Accessibility	Accessibility of data using appropriate medium	Data cannot be accessed through appropriate medium	Data can be accessed by registered users	Confidential data can be accessed by registered users in a data enclave	Data can be accessed through an appropriate medium
	Data available in a usable format	No tolerance for defect Data is available in a usable format	No tolerance for defect Data is available in a usable format	No tolerance for defect Data is available in a usable format	Data is readily available in a usable format

12. Methodological Soundness	Explanation of data Copyright	Data copyright not explained	Data copyright not explicitly explained	Data copyright informed by the funder	Data copyright stated in research contract
	Usage of international/national or peer-agreed standards for managing data	International /national or peer-agreed standards for managing data not used.	International/national or peer-agreed standards for managing data only acknowledged in grant application	International/national or peer-agreed standards for managing data consulted when facing a data management challenge	International/national or peer-agreed standards for managing data used throughout a research project
	Data copyright and ownership stipulated in the research contract	Data copyright and ownership is not stipulated in research contract	Data copyright and ownership stipulated in accepted research contract	Data copyright and ownership stipulated in research contract that is not signed	Data copyright and ownership stipulated in a signed research contract
	Usage of organisational methods and procedures to manage data	organisational methods and procedures to manage data not used	Preference of using funder methods and procedures as compared to organisational methods and procedures	Usage of organisational methods and procedures to manage data	Organisational methods and procedures to manage data consulted and used
13. Quality Check Processes	Provision of institutional ethical guidelines for research	Institutional ethical guidelines for research not provided	Institutional ethical guidelines for research provided only to data managers	Institutional ethical guidelines for research provided upon request	Institutional ethical guidelines for research are readily available and provided to everyone
	Experience and expertise of research team	Experience and expertise of research team	Experience and expertise of research team	Experience and expertise of research team	Experience and expertise of research team

	members to manage research data effectively	members to manage research data effectively not assessed	members to manage research data effectively only assessed when a research project has stated	members to manage research data effectively are assessed during research training	members to manage research data effectively are determined during the data planning stage
	Documenting field operations, coding decisions and variable construction	No tolerance for defect; Field operations, coding decisions and variable construction all documented	No tolerance for defect; Field operations, coding decisions and variable construction all documented	No tolerance for defect; Field operations, coding decisions and variable construction all documented	All field operations, coding decisions and variable construction all well documented
14. Standards and Policies	Adherence to policies that inform how a repository and institution should manage data	Non-adherence to policies that inform how a repository and institution should manage data	Adherence only to funder policies that inform how a repository and institution should manage data	Adherence only to academic institution policies that inform how a repository and institution should manage data	Full adherence to policies that inform how a repository and institution should manage data
	Policy covering all aspects of the data lifecycle	Policy not covering aspects of the data lifecycle	Policy covering certain phases of the data lifecycle	Policy only covering phases relating to the creation and dissemination of data	Policy covering all aspects of the data lifecycle with no exceptions
15. Flexibility of the data	Recording of field procedure information	Field procedure information not recorded	Field procedure information recorded but	Field procedure information	All field procedure information is correctly documented

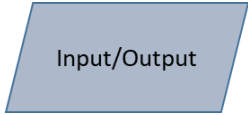
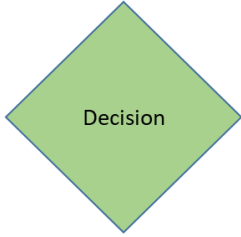
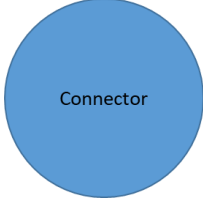
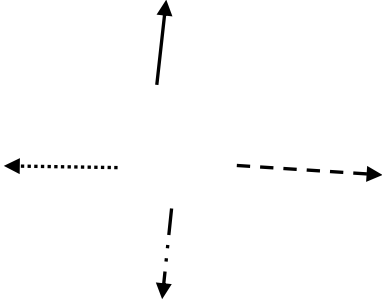
			not transferred to an official related document	communicated to the research manager	
16. Usability of the data	Data sharing clause in the consent form	No data sharing clause in the consent form	Data sharing clause in the consent form does not explicitly detail how the data will be shared	Data sharing clause in the consent form is explicit but does not cover confidentiality	Data sharing clause and confidentiality in the consent form are fully explained
	Methods of managing the Master data file	Methods of managing the Master data file not known	Methods of managing the Master data file not properly differentiated from the dissemination file	Methods of managing the Master data file available only to curators	Methods of managing the Master data file widely available for the institution

Appendix 5: Criteria used to identify relevant models

There is already a wide variety of frameworks that deal with RDM and quality assurance is addressed differently by each of them. For this research, a selection checklist was used to identify data quality and quality assurance. The table below shows the values that were used to select the relevant data lifecycle models (refer to Appendix 2) that were considered for the proposed framework for quality-focused RDM. The checklist also includes three exclusion criteria that were used to identify the models to be excluded from this research.

Selection Criteria	Compliance Y/N
A. The model includes the completion of a data management plan. (An explanation of how data will be managed when data is collected and processed is required.)	
B. Data collection guidelines, that address the risk of making errors when data are created, are provided.	
C. Data processing guidelines, that address the risk of making errors when data are processed, are provided.	
D. The model puts forward recommendations to promote the preservation of data.	
Exclusion Criteria	
E. The model focuses only on the theoretical aspects of what a lifecycle is.	
F. The model is generic with no reference to quality assurance and/or data quality.	
G. There is no correlation between the model and how to promote data that are of good quality.	

Appendix 6: Flow diagram symbol description – initial Quality Process Framework

Flowchart diagram symbol	Name and description
	Oval: represents a beginning or an end in a flowchart diagram.
	Square: represents a step in a flowchart diagram
	Parallelogram: represents either the input or output of information in a flowchart diagram
	Diamond: indicates a decision point between two or more paths in a flowchart diagram
	Circle: used to represent different parts in a flowchart diagram.
	Arrows: These are arrow mark symbols used to connect two boxes and to indicate the direction of data or processing flow. There are many different types of arrow tips to choose from

Appendix 7: Data Management Plan

Developing a framework for data quality-focussed RDM to guide the data management planning process in a Social Sciences Research Organisation

PI: Glenn Tshweu

Data Collection

What data will you collect or create?

The type of data that I am going to collect are semi-structured qualitative interviews that will be transcribed (in .word format).

How will the data be collected or created?

The data will be collected using an interview schedule as the data collection tool. The interviews will be voice recorded.

Documentation and Metadata

What documentation and metadata will accompany the data?

Documents that will accompany the data will include: the Interview Schedule, Consent Form, Readme file and a Study Background Document.

Ethics and Legal Compliance

How will you manage any ethical issues?

In order to meet ethical issues, all identifying information (i.e. peoples names, institutions, positions of research participants) in the transcripts will be anonymised. The data will be stored in a secure organisational IT location (server). Participants will be required to give their consent in order to take part in the study.

De-identified data will be archived and used for research now and in future in ways that will not reveal who participants are within the limits imposed by the confidentiality regulations pertaining to the University of Pretoria and its stakeholders.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The copyright will be owned by the University of Pretoria (as the main affiliated institution). The usage of the data will require users to register in order to receive access to the data.

Storage and Backup

How will the data be stored and backed up during the research?

The data will be stored on an external hard drive and backup copies will be saved on an institutional server that has security measures in place.

How will you manage access and security?

Data access will be managed through the University's institutional repository access management processes. Users who want the data will have to be registered to access the data.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The qualitative transcripts will have long-term value and should be shared and preserved.

What is the long-term preservation plan for the dataset?

The dataset will be preserved in the University's repository.

Data Sharing

How will you share the data?

The data will be shared through the university's institutional repository.

Are any restrictions on data sharing required?

N/A

Responsibilities and Resources

Who will be responsible for data management?

The author will be responsible for data management.

What resources will you require to deliver your plan?

University ethical clearance; a storage medium (such as an external hard drive); a Laptop; related documentation; voice recorder; stationery.

Appendix 8: Interview schedule - Researchers

Developing a framework for data quality-focussed RDM to guide the data management planning process in a Social Sciences Research Organisation

Introduction

I have developed a quality process model as well as a measuring framework. It is anticipated that adhering to the proposed quality process model and data quality dimensions would improve quality assurance across the research data lifecycle and ultimately improve the quality of data produced by the social sciences institution when it is made accessible for reuse.

The intention with this interview is to gain insight into your requirements for data quality. The feedback you provide will be used to modify the initial quality process model as well as the quality measurement framework.

The questions of the interview schedule are divided into three sections. Section A collects demographic information as well as educational background information. Section B contains questions that relate to the data quality management dimensions while Section C will require that you critique the quality process model. The quality process model and data quality dimensions are explained in the handout (given to you before this interview). Your opinion about these will be asked. The aim is to understand which quality dimensions you accept as being relevant to [the social sciences organisation's] data. Once we have determined which quality dimensions are relevant we will also determine how you foresee us measuring compliance. These measures will be put into a measuring framework that will be offered to the [the social sciences organisation] as a measuring framework. Your feedback will provide me with an understanding of your data quality assurance needs and practices.

It is anticipated that the interview will last between 30 and 60 minutes.

Now that you understand what the research is about and you know you have the right to refuse to participate, please sign the informed consent form and return it to me (gtshweu@gmail.com) and my supervisors, using the following contact details:

Prof T Bothma

Theo.Bothma@up.ac.za

Dr Martie Van Deventer

Vandeventer.martha@up.ac.za

Section A: Demographic and educational information

1. **Education:** What is the highest degree or level of school that you have completed?

Bachelor's degree	
Master's degree	
Doctorate degree	
Other (please specify)	

2. **Number of years as a researcher**

For how long have you been conducting research?

0-5 years	
6-15 years	
More than 15 years	

The quality process model and a list of data quality dimensions were sent to you prior to the interview. Is there anything you would like me to explain before we start?

Section B: Measuring data quality

This section of the interview will address the dimensions that are used to determine data quality. It is accepted that the re-use quality of data is expressed both as scientific quality and its technical quality. To better understand what each dimension means, please have a look at the list of quality dimensions in the handout.

3. **What quality dimensions should have been listed but were not?**

I would like to create a list of those dimensions that you think determine the scientific quality of the data. You may use any of the dimensions listed and you could also add further dimensions.

A definition of scientific quality is: *“judgements that are made when quality indicators are examined, when the researcher's competence is in focus, or when extra scientific effects are under scrutiny. Consequently, judgements of scientific quality are made in a context of influencing factors in which the criteria for evaluation occasionally change according to the factor emphasized”* (Hermlin, 1993:4).

4. Which of the dimensions mentioned (your own included) would determine the scientific quality of the data?

Factors that determine scientific quality of data

A definition of technical quality is: Technical quality is achieved by direct RDM interventions and the steps a researcher takes before embarking on the study journey – these steps include “*writing DMPs and support for various data management components such as creating metadata or choosing appropriate data repositories*” (Rolando et al., 2015). RDM activities undertaken by either a researcher or a curator that ensure data comply with FAIR principles – i.e. data should be findable, accessible, interoperable and re-usable.

5. Which of the dimensions mentioned (your own included) would determine the technical quality of the data?

Factors that determine technical quality of data

Measuring Data Quality Dimensions

6. Considering the data quality dimensions you identified above, please indicate how we could measure the dimension

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
1. Completeness		- How are gaps in the data file documented? - Where were the results of the internal review processes documented?
2. Uniqueness		- Are the variables and group names consistent? - Where is the list of variables with unique IDs?
3. Timeliness		- Have both the data capture and data processing dates been captured? - Where was the quality control date captured?
4. Validity		- What methods were used to identify missing data? Where were the methods and the values specified? - Was coding done separately from data-entry to minimise errors
5. Accuracy		- How was de-duplicating done/Where is the process documented?

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
		- Indicate how many capturers captured the data. (i.e. do value labels provide a true reflection of what was investigated?).
6. Consistency		- How is the accuracy of variables descriptions checked? - Are common data definitions and standards used across multiple data sets?
7. Interpretability		- How do you verify that consistent codes have been used for missing values? - Where are definitions used in the methodology of data collection & processing documented?
8. Integrity		- How do you determine which is the best format to use for your data? - What method do you use to identify typing errors?
9. Relevance		- How do you determine the relevance of your data to meet user-needs? - What type of audit process do you do to verify data entry errors?

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
10. Comparability & Coherence		- What are the typical quality control steps you follow during data collection and processing? - How do you validate your data collection method irrespective of the sample and instrument used?
11. Accessibility		- How do you determine the type of repository the data will be deposited in? - What informs the statement explaining data terms-of-use by others?
12. Methodological Soundness		-How is the authenticity of research data maintained over time?
13. Quality Check Processes		- What do data dictionaries that accompany data files contain? - How are rule and error catalogues (to assist secondary users in dealing with possible errors) shared?
14. Standards & Policies		- Which policies & procedures do you follow when conducting a research project? - At what stage (in the data lifecycle) do you implement an independent data quality audit process?

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
15. Flexibility of the data		- How are strategies for handling research data outlined? - How to you verify that your selected data format is recommended by a data repository?
16. Usability of the data		- How do you confirm that funder requirements (on dissemination and sharing of research findings) have been adhered to? - What does a Codebook accompanying a data file contain?

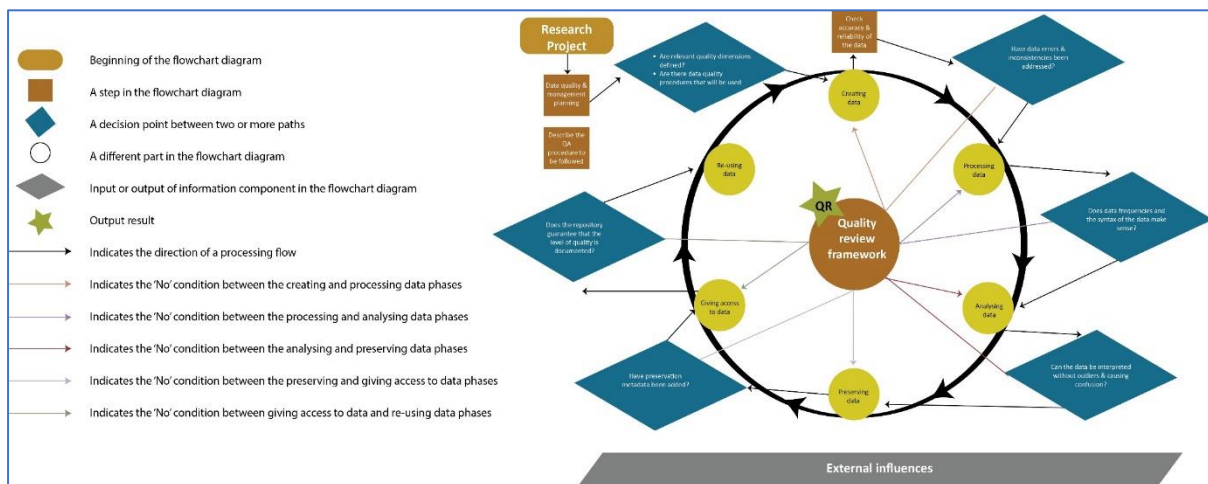
Personal Data quality and management planning:

7. What quality dimension/measuring criteria do you usually address when working with your own research data?

8. What quality procedures do you make use of/ are you aware of?

Section C: The Quality Process Model

Please have a look at the quality process model provided in your handout. Do you have any questions before we start?



Proposed quality process model

The researcher will share a full-size model handout during the interview.

9. What is your opinion – will it be acceptable to plan data quality at the same time as when data management is planned? If so, please elaborate

10. When do you propose we measure the scientific quality of the data?

11. When should the technical quality of the data be measured?

12. Is the model too complex/too simple for practical use? Please explain your viewpoint.

13. How should the model be changed to comply with your requirements?

14. Taking your research work into consideration, which areas do you think the model will assist with?

15. Which data management areas do you think the model covers well?

16. Taking your research work into consideration, which areas of concern do you think the model will NOT address?

Closure

- 17. Is there anything else, regarding the quality of research data, that you would like to add before we close our session?**

Thank you for sharing your knowledge and experience. The time that you spent thinking and talking about data quality is also appreciated.

----- END -----

Appendix 9: Interview schedule – Data curators

Developing a framework for data quality-focussed RDM to guide the data management planning process in a Social Sciences Research Organisation

Introduction

I have developed a quality process model as well as a measuring framework. It is anticipated that adhering to the proposed quality process model and data quality dimensions would improve quality assurance across the research data lifecycle and ultimately improve the quality of data produced by the social sciences institution when it is made accessible for reuse.

The intention with this interview is to gain insight into your requirements for data quality. The feedback you provide will be used to modify the initial quality process model as well as the quality measurement framework.

The questions of the interview schedule are divided into three sections. Section A collects demographic information as well as educational background information. Section B contains questions that relate to the data quality management dimensions while Section C will require that you critique the quality process model. The quality process model and data quality dimensions are explained in the handout (given to you before this interview). Your opinion about these will be asked. The aim is to understand which quality dimensions you accept as being relevant to [the social sciences organisation's] data. Once we have determined which quality dimensions are relevant to data curators, the author will also determine how you foresee us measuring compliance. These measures will be put into a measuring framework that will be offered to the [the social sciences organisation] as a measuring framework for all [its] data shared openly via [the] data repository. Your feedback will provide me with an understanding of what you think are the quality contributions data curators make when ensuring that the quality of our data enhances the reputation of [the social sciences organisation] as a research organisation.

It is anticipated that the interview will last between 30 and 60 minutes.

Now that you understand what the research is about and that you know you have the right to refuse to participate, please sign the informed consent form and return it to me (gtshweu@gmail.com) and my supervisors, using the following contact details:

Prof T Bothma

Theo.Bothma@up.ac.za

Dr Martie Van Deventer

Vandeventer.martha@up.ac.za

Section A: Demographic and educational information

1. Education: What is the highest degree or level of school you have completed?

Bachelor's degree	
Master's degree	
Doctorate degree	
Other (please specify)	

2. Number of years as a curator

0-5 years	
6-15 years	
More than 15 years	

The quality process model and a list of data quality dimensions were sent to you prior to the interview. Is there anything you would like me to explain before we start?

Section B: Measuring data quality

It is difficult to quantify data quality. As a result, several dimensions of quality have been identified and defined. These dimensions are measurable and serve as indicators of the data quality. It is generally accepted that the re-use quality of data is expressed both as scientific quality (the responsibility of the researcher) and its technical quality (mostly the responsibility of the curator). To better understand what each dimension means, please have a look at the list of quality dimensions in the handout.

3. What quality dimensions should have been listed but were not?

I would like to create a list of those dimensions that you think determine the scientific quality of the data. You may use any of the dimensions listed and you could also add further dimensions.

A definition of scientific quality is: *“judgements that are made when quality indicators are examined, when the researcher's competence is in focus, or when extra scientific effects are under scrutiny. Consequently, judgements of scientific quality are made in a context of influencing factors in which the criteria for evaluation occasionally change according to the factor emphasized”* (Hermlin, 1993:4)

4. Which of the dimensions mentioned (your own included) would determine the scientific quality of the data?

Factors that determine scientific quality of data

A definition of technical quality is: is achieved by direct RDM interventions and the steps a researcher takes before embarking on the study journey – these steps include “*writing DMPs and support for various data management components such as creating metadata or choosing appropriate data repositories*” (Rolando et al., 2015:43). RDM activities undertaken by either a researcher or a curator that ensure data comply with FAIR principles – i.e. data should be findable, accessible, interoperable and re-usable.

5. Which of the dimensions mentioned (your own included) would determine the technical quality of the data?

Factors that determine technical quality of data

Measuring Data Quality Dimensions

6. Considering the technical data quality dimensions, you identified above, please indicate how we could measure the dimension

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
1. Completeness	.	- When do you run a spell check on a data file? - How do you manage markings in a data file left by a researcher as part of their initial analysis?
2. Uniqueness		- What strategy do you use to anonymise the identities of respondents? - How do you document variable and column characteristics of a data file?
3. Timeliness		- How do you ensure that your data represents the relevant point of time in reality? - How do you ensure that data is made available as soon as possible?
4. Validity		- How do you guarantee that syntax definitions and identifiers are accurately documented? - How is your data checked, validated and cleaned?
5. Accuracy		- How do you check if data and documentation correspond?

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
		- What approach do you use to check for invalid and incomplete records?
6. Consistency		- What are the data quality control processes and documents you share with research teams? - How do you verify that the correct definitions have been used in metadata for interpretation purposes?
7. Interpretability		- How do you determine that sufficient descriptions of data are used in related documents? - How often do you review and change quality control methods for data management procedures?
8. Integrity		- How do you identify, design and implement data quality rules? - Which naming convention standard do you use?
9. Relevance		- How do you determine the logical relationship between variables and the proposed measure? - How do you oversee that researchers use correct data management practices from the project onset?

Dimension	How should the dimension be measured?	Prompting options if the participant is not able to identify any
10. Comparability & Coherence		- Which file formats are data made available in? - Which approach do you use to manage & preserve data (in terms of quality assurance)?
11. Accessibility		- Accessibility of data using appropriate medium. - Data available in a usable format
12. Methodological Soundness		- Which international/national or peer-agreed standards do you use when curating and managing research data? - How are organisational data management procedures used when curating data?
13. Quality Check Processes		- Which experience and expertise is required to effectively curate and manage research data? - Do you have a method in place that requires researchers to document field operations, coding decisions & variable construction?

14. Standards & Policies		- Is there a policy that informs how a repository and institution should manage research data? - Is there an organisational policy covering all aspects of the data lifecycle?
15. Flexibility of the data		- How do you authenticate that field procedure information is well recorded by researchers?
16. Usability of the data		- Who is the responsible party to ensure that a consent form has a data sharing clause? - How is the Master data file managed as compared to the dissemination data file?

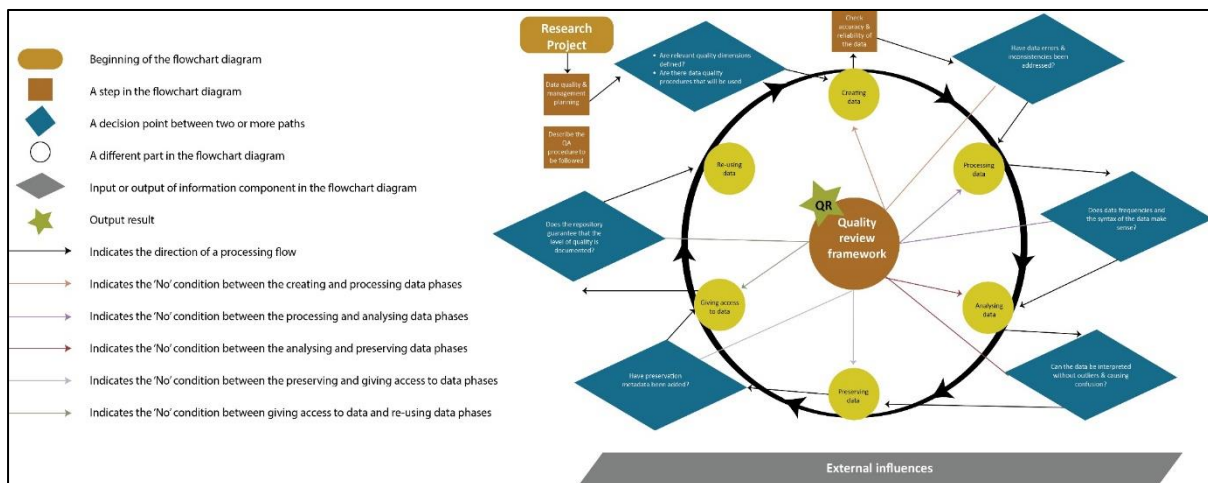
Personal experience in data quality assurance

7. What quality dimension/measuring criteria do you usually use when curating research data?

8. What quality control processes do you use to ensure that deposited data files are accurately processed to make it distinguishable from other data files/versions?

Section C: The Quality Process Model

Please have a look at the quality process model provided in your handout. Do you have any questions before we start?



Proposed quality process model

The researcher will share a full-size model handout during the interview.

9. Is there a procedure in place that confirms that data quality preparation was completed by researchers? Please elaborate on your answer.

10. When do you propose we measure the scientific quality of the data?

11. When should the technical quality of the data be measured?

12. Is the model too complex/too simple for practical use? Please explain your viewpoint.

13. How should the model be changed to comply with your requirements?

14. Taking your curation work into consideration, which areas do you think the model will assist with?

15. Which data management areas do you think the model covers well?

16. Taking your research work into consideration, which areas of concern do you think the model will NOT address?

Closure

17. Is there anything else, regarding the quality of research data, that you would like to add before we close our session?

Thank you for sharing your knowledge and experience. The time that you spent thinking and talking about data quality is also appreciated.

----- END -----

Appendix 10: Research participant handout

Developing a framework for data quality-focussed RDM: guiding the data management planning process at a Social Sciences Research Institution

My name is Glenn Tshweu. I am a PhD student at the University of Pretoria. I am developing a process model that will focus on quality and that could be used to guide Research Data Management planning and practices at a Social Sciences Research Organisation.

Introduction

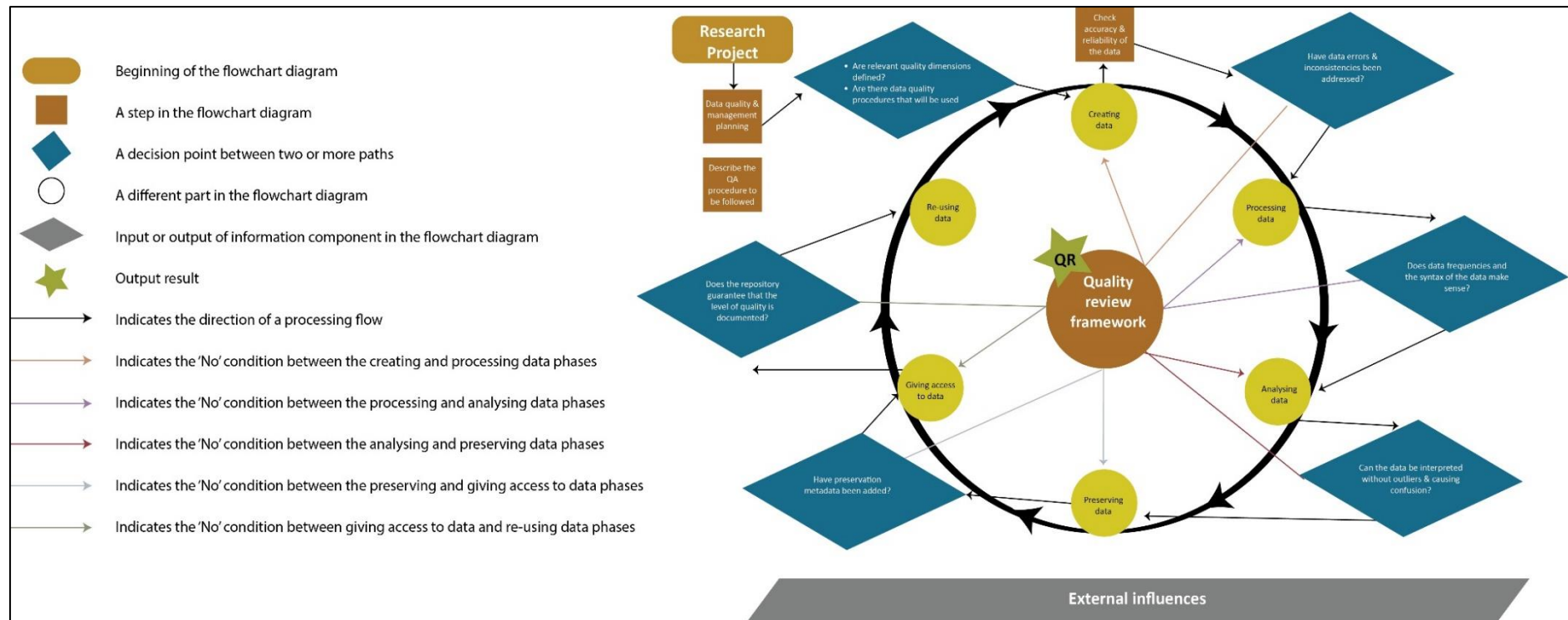
Explaining the proposed quality process model: The model assumes that a variety of data quality dimensions have been accepted across the [the social sciences organisation] and all its research projects. The model also assumes that each quality dimension is informed by relevant measuring criteria. A last assumption is that the quality of research data could be measured both in terms of its scientific and technical quality.

A generic list of quality dimensions was created from the literature consulted. This document provides descriptions for each of the identified quality dimensions. The first thing to do is to check whether the dimension of quality that is important to you does appear in the list.

The questions that you will be asked during the interview intends to identify those dimensions that you would like to see associated with the quality control for all our research data. We will then discuss how we could go about measuring compliance. You will also be asked if there could be tolerance for deviation on the measurement or whether compliance should be 100%.

Lastly, it is possible to measure quality as data moves through the data lifecycle or the measurement can be done at specific points. These points are indicated in the model. The model assumes that the data will be accompanied by a data quality measurement framework that will keep track of the quality process. This framework would be incorporated into the data management plan (DMP) so that it is not an additional document to be completed. You will be asked to comment on this practice.

Adhering to the proposed quality process model would improve quality assurance across the research data lifecycle and ultimately improve the quality of data made accessible for reuse. The figure below is the proposed quality process model:



The proposed quality process model

Description of the quality process model

The foundation of the proposed quality process model lies in the research data lifecycle adopted from the UK Data Archive. Managing research data should be done using methods and criteria that are based on ensuring quality. To achieve this, proper data management planning is required before a researcher begins with a research project. It is suggested that this is also the point of introducing a data quality framework with quality dimensions and measures indicated. Having quality assurance steps and using RDM best practices ensures that the research data is free of errors which, in turn, facilitates successful secondary usage.

The completion of a ‘data quality preparation’ section in the DMP allows a researcher to identify and use only applicable data quality dimensions. In the DMP, a researcher has to account for the relevant quality dimensions and data quality procedures that will be used in the research project. If a researcher has not defined the relevant quality dimensions and data quality procedures (as part of Data quality & management planning) the researcher has to describe the quality assurance procedure to be followed before proceeding to the Creating Data phase – this course is represented by the torque arrow in the quality process model. Once a researcher has a plan in place, the researcher has to make informed quality assurance decisions to use, these will ultimately impact the quality of research data. After data management planning and quality assurance decisions are taken, the research project can begin by proceeding to the ‘Creating Data’ phase.

The ‘quality review framework’ acts as a best practice and quality assurance reference point for data management in all the stages of the data lifecycle. The ‘quality review framework’ constitutes the data quality dimensions and all the applicable measuring criteria. The intention is for researchers and data curators to consult (by assessing their current RDM practices and making adjustments where necessary) RDM discipline guidelines, standards and best practices by using the quality dimensions as a reference point.

In the course of a research project, certain disruptions may occur which may derail or negatively affect data management. These disruptions are represented as ‘External influences’ in the proposed quality process model and are represented at the bottom of the model. External influences are ongoing and apply to all the stages of the research data lifecycle because they may impact any RDM activity as data is managed in the data lifecycle. Overcoming external influences requires a researcher to have management and quality assurance action plans in place (that are informed by best practices).

Once data has been collected, a researcher has to verify that it is accurate and reliable. This involves running frequencies and spell checks to identify and address errors and inconsistencies. As data is being managed across the different stages of the research data lifecycle each phase of the lifecycle requires

unique quality dimensions to be addressed. The quality model is comprised of suggested dimensions that may adapted.

When data has been successfully managed throughout the research data lifecycle the quality process model produces a ‘Quality Review’ (represented as QR). The Quality Review is closely associated with the quality review framework because the review itself is informed by data quality dimensions which ensure a quality-based assessment. The quality dimensions and their definitions are listed below:

Generic list of quality dimensions

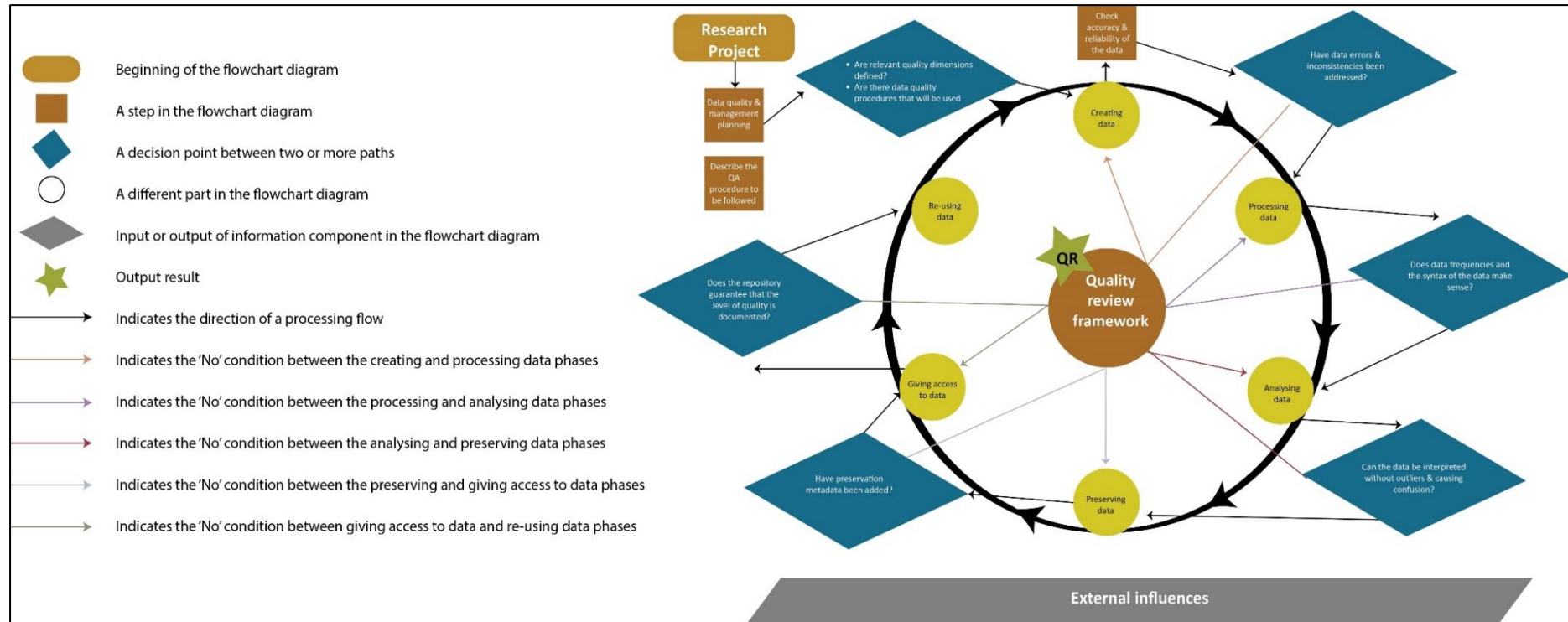
Determining the quality of research data is difficult. As a result, evaluators have started to use quality dimensions as a proxy to indicate the level of quality associated with a collection of data. A generic list was compiled and a definition for each was developed.

Dimensions of Quality	Definitions
1. Completeness	Completeness refers to “ <i>the proportion of stored data against the potential of ‘100%’ complete</i> ” (UK-DAMA Working Group, 2013:8). Accounting for missing, null, invalid and incorrect values determines if a dataset is complete and can be used by the next person.
2. Uniqueness	Every variable in a data file should be distinct in its subject. Uniqueness means that “ <i>nothing will be recorded more than once based upon how that thing is identified</i> ” (DAMA UK Working Group, 2013:9). If variables are duplicated, their uniqueness is compromised and data quality is negatively impacted.
3. Timeliness	Timeliness has a direct impact on when data was collected as compared to when data is made available. Timeliness “ <i>is the information processed and delivered rapidly without delays?</i> ” (Jayawardene, et al., 2016:12). Timeliness represents “ <i>the degree to which data represents reality from the required point in time</i> ” (DAMA UK Working Group, 2013:10).
4. Validity	Data validity means that “ <i>data is valid if it conforms to the syntax (format, type, range) of its definition</i> ” (DAMA)
5. Accuracy	This aspect refers to incorrect data entries and how they were corrected – the process followed when mistakes were identified. This aspect also refers to fieldwork errors and how these are managed. Accuracy also refers to incorrect data created by input errors (i.e.

	contradictory information, typographical errors, obsolete data attributes) <i>Accuracy determines whether the information you hold is correct or not, and isn't to be confused with validity, a measure of whether the data is actually the type you wanted (DAMA UK Working Group, 2013).</i>
6. Consistency	This aspect refers to the reliability of the data and how reliability was tested. This aspect also refers to the similarity between the question asked by interviewer and the questions in the data collection instrument. <i>For anyone trying to analyse data, consistency is a fundamental consideration. Basically, you need to ensure you can compare data across data sets and media (whether it's on paper, on a computer file, or in a database) - is it all recorded in the same way, allowing you to compare the data and treat it as a whole? (DAMA UK Working Group, 2013).</i>
7. Interpretability	This dimension has a close link with related documentation that explains the data. Interpretability refers to “<i>the ease with which users can understand data/statistics through provision of metadata. This information normally includes the underlying concepts, definitions and classifications used in the methodology of data collection and processing</i>” (Statistics South Africa, 2010:4).
8. Integrity	Integrity is one of the distinct dimensions of data quality. It refers to “<i>values and related practices that maintain users’ confidence in the agency producing statistics/data and ultimately in the statistical product</i>” (Statistics South Africa, 2010:4). This also resembles social science research data because data also has to have integrity as a final research output.
9. Relevance	Research data is potentially beneficial because it can help open up the doors of new knowledge discovery and better decision making. Data has to be relevant if it stands a chance of achieving this. Relevance is “<i>the degree to which it meets the real needs of clients. It is concerned with whether the available data sheds light on the issues of most importance to users</i>” (Statistics South Africa, 2010:4).

10. Comparability & Coherence	The management of research data doesn't happen in isolation, the comparability and coherence of data refers to <i>“the ease with which data can be successfully brought together with other statistical/research information within a broad analytical framework and over time”</i> (Statistics South Africa, 2010:4).
11. Accessibility	Data accessibility restrictions prevent secondary data usage. External users have to be able to access all data aspects to conveniently use data. Accessibility refers to <i>“the ease with which it can be obtained from the agency. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed”</i> (Statistics South Africa, 2010:4).
12. Methodological Soundness	The research field is informed by recognised standards and best practices. Methodological soundness refers to <i>“the application of international, national or peer-agreed standards, guidelines and practices to produce outputs”</i> (Statistics South Africa, 2010:4).
13. Quality Check Processes	In the course of managing research data, a research institution must have a quality management system in place (Statistics South Africa, 2010:13). These systems act as quality check processes.
14. Standards & Policies	This quality dimension requires that a set of standards and policies be in place which cover all aspects of the statistical value chain/research data lifecycle and are adhered to (Statistics South Africa, 2010:7).
15. Flexibility of the data	A research project can have multiple versions of data files that differ on a few modules and scope. The entire data collection has to be flexible, meaning that it has <i>“to be comparable and compatible with other data including useful groupings, classifications, repurposing and easy to manipulate”</i> (DAMA UK Working Group, 2013:14).
16. Usability of the data	Usability of the data should not be hampered by syntax or any technicalities; <i>“is data understandable, simple, relevant, accessible, maintainable and at the right level of precision?”</i> (DAMA UK Working Group, 2013:14). Usable data should facilitate easy secondary usage.

Appendix 11: The initial quality process model



The initial quality process framework

Appendix 12: Results of the criteria used to identify relevant role players/models

Legend: Evaluation criteria:

- A: The model includes the completion of a data management plan
- B: Data collection guidelines that address the risk of making errors when data are created are provided.
- C: Data processing guidelines that address the risk of making errors when data are processed are provided.
- D: The model puts forward recommendations to promote the preservation of data.
- E: The model focuses only on the theoretical aspects of what a lifecycle is.
- F: The model is generic with no reference to quality assurance and/or data quality.
- G: There is no correlation between the model and how to promote data that are of good quality.

Legend: Symbols

√: Complies/Yes

N: Doesn't comply/No

⌘: Neutral/Makes reference to the criteria, but doesn't provide a full description of quality assurance

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
The UK Data Archive research data lifecycle	√	√	√	√	N	N	√	√
Open Archival Information System (OAIS) Framework	⊗	√	⊗	√	N	⊗	⊗	⊗
ICPSR Guide to Social Science Data Preparation and Archiving	√	√	N	√	⊗	N	√	√
DataOne Lifecycle	√	⊗	√	⊗	N	N	√	√
DDI version 3.0 Combined Life Cycle Model	N	√	√	N	N	N	√	√

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
USGS Data Lifecycle Model	√	√	√	√	N	N	√	√
The Federal Geographic Data Committee (FGDC) Geospatial Data Lifecycle	√	√	√	√	N	N	√	√
Ellyn Montgomery, USGS, Data Lifecycle Diagram	N	✱	✱	✱	N	√	N	N
University of Oxford Research Data Management Chart	√	N	✱	N	√	√	N	N
NOAA Environmental Data Life Cycle Functions	✱	√	N	✱	N	✱	✱	✱

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
USGS Scientific Information Management Workshop Vocabulary	√	⊗	√	⊗	N	N	√	√
Peter Fox Lifecycle Diagrams	⊗	√	⊗	N	N	N	⊗	⊗
National Science Foundation	⊗	√	⊗	N	N	√	⊗	N
NDIIPP Preserving Our Digital Heritage	N	⊗	N	√	√	√	N	N
What Researchers Want	N	N	N	N	N	N	N	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
The Digital Curation Centre Model	✳	✳	N	√	N	✳	✳	✳
IWGDD's Digital Data Life Cycle Model	✳		N	N	✳	√	N	N
Scientific Data Management Plan Guidance		✳	N		✳	√	N	N
Linear Data Life Cycle	✳	√	✳	N	√	√	✳	N
Generic Scientific Data Lifecycle	✳	✳		✳	√	√	N	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
Cassandra Ladino Hybrid Data Lifecycle Model	√	N	N	N	N	⊗	N	N
Ray Obuch Data Management – A Lifecycle Approach	N	⊗	⊗	√	⊗	√	⊗	N
USGS Data Management Plan Framework (DMPf) – Smith, Tessler, and McHale	√	√	N	√	N	⊗	⊗	⊗
BLM Data Management Handbook	√	√	⊗	⊗	N	N	⊗	⊗
ARL Joint Task Force on Library Support for E-Science		⊗	⊗	⊗	⊗	√	√	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
US Department of Health and Human Services Key Components	N	√	⊗	⊗	⊗	⊗	N	N
IBM Aspects of Lifecycle Management - Research	N	⊗	N	N	⊗	√	⊗	N
University of California San Diego Digital Curation Program	N	⊗	⊗	√	N	⊗	⊗	N
University of Miami Scientific Data Lifecycle	N	⊗	⊗	⊗	√	√	N	N
Managing Research Data Lifecycles through Context	N	⊗	N	N	⊗	√	√	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
Purdue University Research Repository Collaborative Model	N	✪	✪	✪	√	✪	√	N
Data Lifecycle Management at Dell	N	N	✪	√	√	√	✪	✪
John Faundeen & Ellyn Montgomery “Spins”	√	✪	✪	✪	N	✪	✪	✪
DataTrain Model 1	✪	✪	N	N	✪	√	N	N
DataTrain Model 2	N	✪	N	✪	✪	√	N	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
Steve Tessler Data Management Cycles	N	✪	✪	N	✪	✪	✪	N
Peter Fox Full Lifecycle of Data	N	✪	N	N	✪	√	N	N
Steve Tessler Data and System Lifecycle Models	✪	✪	N	✪	✪	√	✪	N
University of Virginia Library lifecycle	✪	√	✪	N	✪	✪	✪	✪
Mario Valle	N	✪	N	N	N	√	√	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
Michigan State University Records Life Cycle Model	N	✳	N	✳	√	√	N	N
GeoMAPP Geoarchiving Process Lifecycle	✳	✳	✳	✳	✳	√	✳	N
Jeff de La Beaujardière's proposed Data Management Framework	√	√	√	√	N	✳	✳	✳
Sarah Demb Continuum Model	N	✳	N	N	N	√	✳	N
ARMA International Life Cycle	N	✳	✳	✳	√	√	N	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
JISC Simon Hodson Research Data Lifecycle	√	✖	✖	N	N	✖	✖	N
United Kingdom Digital Archive Research Data Lifecycle	√	√	√	√	N	N	√	√
Lisa Dunn Data Flow Model	N	✖	N	✖	N	√	√	N
JISC Information Environment	N	✖	N	✖	✖	√	N	N
Make it Digital New Zealand's Digital Content Life Cycle	✖	✖	✖	√	✖	✖	√	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
CASPAR Preservation Work Flow	N	✱	✱	✱	N	N	✱	✱
IWGDD's Digital Data Life Cycle Model	✱	✱	N	✱	N	N	✱	N
Data-Information-Knowledge Ecosystem	N	✱	✱	N	N	√	√	N
PII lifecycle	N	✱	✱	N	√	√	N	N
Enterprise data lifecycle	N	✱	√	✱	N	√	N	N

Models Identified	Evaluation criteria							
	A	B	C	D	E	F	G	Quality explicit
CRUD lifecycle	N	✪	✪	N	N	√	N	N
DIKW Information pyramid	N	✪	N	N	N	√	N	N

Appendix 13: Grouping of quality dimensions with stages of the research data lifecycle that form the quality process framework

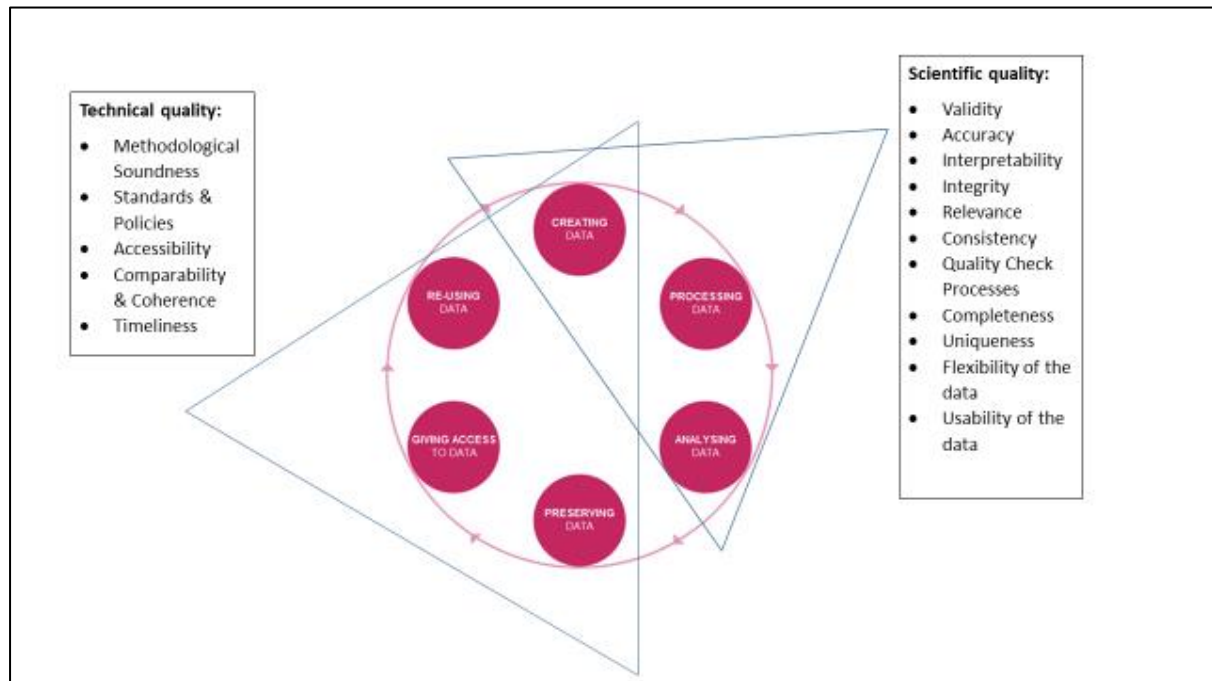


Figure: Quality dimensions grouped with stages of the data lifecycle

The following table shows the quality dimensions assigned to each stage of the research data lifecycle, the assignment is based on the nature and level of a quality dimension in relation to assuring that all the activities of that particular stage are performed optimally to produce good data quality and quality assurance:

Table: Quality dimensions assigned to each stage of the research data lifecycle

Research data lifecycle phase	Applicable quality dimension
Creating data	Validity, Accuracy, Interpretability, Integrity, Relevance, Consistency, Quality check processes, Completeness, Uniqueness, Flexibility of the data, Usability of the data
Processing data	Validity, Accuracy, Interpretability, Integrity, Relevance, Consistency, Quality check processes, Completeness, Uniqueness, Flexibility of the data, Usability of the data

Analysing data	Validity, Accuracy, Interpretability, Integrity, Relevance, Consistency, Quality check processes, Completeness, Uniqueness, Flexibility of the data, Usability of the data
Preserving data	Methodological Soundness, Standards & Policies, Accessibility, Comparability & Coherence, Timeliness
Giving access to data	Methodological Soundness, Standards & Policies, Accessibility, Comparability & Coherence, Timeliness
Re-using data	Methodological Soundness, Standards & Policies, Accessibility, Comparability & Coherence, Timeliness

Appendix 14: UK Data Quality Dimensions

Determining the quality of research data is difficult. As a result, evaluators have started to use quality dimensions as a proxy to indicate the level of quality associated with a collection of data. A generic list was compiled and a definition for each was developed.

Dimensions of quality	Definition
1. Completeness	Completeness refers to <i>“the proportion of stored data against the potential of ‘100%’ complete”</i> (UK-DAMA Working Group, 2013:8). Accounting for missing, null, invalid and incorrect values determines if a dataset is complete and can be used by the next person.
2. Uniqueness	Every variable in a data file should be distinct in its subject. Uniqueness means that <i>“nothing will be recorded more than once based upon how that thing is identified”</i> (DAMA UK Working Group, 2013:9). If variables are duplicated, their uniqueness is compromised and data quality is negatively impacted.
3. Timeliness	Timeliness has a direct impact on when data was collected as compared to when data is made available. Timeliness <i>“is the information processed and delivered rapidly without delays?”</i> (Jayawardene, et al., 2016:12). Timeliness represents <i>“the degree to which data represents reality from the required point in time”</i> (DAMA UK Working Group, 2013:10).
4. Validity	Data validity means that <i>“data is valid if it conforms to the syntax (format, type, range) of its definition”</i> (DAMA)
5. Accuracy	This aspect refers to incorrect data entries and how they were corrected – the process followed when mistakes were identified. This aspect also refers to fieldwork errors and how these are managed. Accuracy also refers to incorrect data created by input errors (i.e. contradictory information, typographical errors, obsolete data attributes) <i>Accuracy determines whether the information you hold is correct or not, and isn't to be confused with validity, a measure of whether the data is actually the type you wanted</i> (DAMA UK Working Group, 2013).

6. Consistency	This aspect refers to the reliability of the data and how reliability was tested. This aspect also refers to the similarity between the question asked by interviewer and the questions in the data collection instrument. <i>For anyone trying to analyse data, consistency is a fundamental consideration. Basically, you need to ensure you can compare data across data sets and media (whether it's on paper, on a computer file, or in a database) - is it all recorded in the same way, allowing you to compare the data and treat it as a whole?</i> (DAMA UK Working Group, 2013).
7. Interpretability	This dimension has a close link with related documentation that explains the data. Interpretability refers to <i>“the ease with which users can understand data/statistics through provision of metadata. This information normally includes the underlying concepts, definitions and classifications used in the methodology of data collection and processing”</i> (Statistics South Africa, 2010:4).
8. Integrity	Integrity is one of a distinct dimension for data quality. It refers to <i>“values and related practices that maintain users’ confidence in the agency producing statistics/data and ultimately in the statistical product”</i> (Statistics South Africa, 2010:4). This also resembles social science research data because data also has to have integrity as a final research output.
9. Relevance	Research data is potentially beneficial because it can help open up the doors of new knowledge discovery and better decision making. Data has to be relevant if it stands a chance of achieving this. Relevance is <i>“the degree to which it meets the real needs of clients. It is concerned with whether the available data sheds light on the issues of most importance to users”</i> (Statistics South Africa, 2010:4).
10. Comparability & Coherence	The management of research data doesn’t happen in isolation, the comparability and coherence of data refers to <i>“the ease with which data can be successfully brought together with other statistical/research information within a broad analytical framework and over time”</i> (Statistics South Africa, 2010:4).

11. Accessibility	Data accessibility restrictions prevent secondary data usage. External users have to be able to access all data aspects to conveniently use data. Accessibility refers to “ <i>the ease with which it can be obtained from the agency. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed</i> ” (Statistics South Africa, 2010:4).
12. Methodological Soundness	The research field is informed by recognised standards and best practices. Methodological soundness refers to “ <i>the application of international, national or peer-agreed standards, guidelines and practices to produce outputs</i> ” (Statistics South Africa, 2010:4).
13. Quality Check Processes	In the course of managing research data, a research institution must have a quality management system in place (Statistics South Africa, 2010:13). These systems act as quality check processes.
14. Standards & Policies	This quality dimension requires that a set of standards and policies be in place which cover all aspects of the statistical value chain/research data lifecycle and are adhered to (Statistics South Africa, 2010:7).
15. Flexibility of the data	A research project can have multiple versions of data files that differ on a few modules and scope. The entire data collection has to be flexible, meaning that it has “ <i>to be comparable and compatible with other data including useful groupings, classifications, repurposing and easy to manipulate</i> ” (DAMA UK Working Group, 2013:14).
16. Usability of the data	Usability of the data should not be hampered by syntax or any technicalities; “ <i>is data understandable, simple, relevant, accessible, maintainable and at the right level of precision?</i> ” (DAMA UK Working Group, 2013:14). Usable data should facilitate easy secondary usage.

Appendix 15: Data quality evaluations by different stakeholders

Appendix 15 part A: Victoria State Government data quality dimensions

<i>Dimension & description</i>	<i>Example properties</i>	<i>Dimension usage description</i>
Accuracy: Is the data accurate and valid, and to what level? Accuracy refers to the degree to which the data correctly portrays the real-world situation that it was originally designed to measure.	Data collection: Errors may occur at the source of collection, usually by providers of data or during data entry. They can occur at multiple stages of the collection process and may be caused by multiple factors including poor form design, providers having an unclear understanding of concepts and/or the form, and human error.	The ‘data collection’ property is used as a measuring criterion for the Accuracy quality dimension.
Completeness: How complete is the data? Are there known gaps? Completeness relates to the extent the data is complete, i.e. is it 100% complete (if	Process: Feedback and review processes should be implemented to ensure that any gaps or coverage issues are discovered and resolved Gaps: This property relates to the ability to determine and document gaps in the data. Gaps may be identified due to a breakdown in system or processes, lack of available resourcing, or when data is collected for a defined time period. In some cases, 100% completeness may not be required to be fit for purpose, and this	The ‘process’ and ‘gaps’ properties are used as measuring criteria for the Completeness quality dimension.

<i>Dimension & description</i>	<i>Example properties</i>	<i>Dimension usage description</i>
100% is required)? This dimension also reflects the ability to determine what data is missing, and whether omissions are acceptable (for example, optional data).	should also be noted. Known gaps should be documented so that users of the data can be aware and compensate if needed.	
Representative: Is the dataset representative of the conditions or scenario to which it refers? Representativeness relates to the relevance of data and the extent to which it meets the defined purpose which initiated its collection or creation. For data to be representative it must be indicative of the environment in which it	Coverage: Coverage relates to the proportion of a population or sample that have been incorrectly included or excluded. This can be discovered when comparing the actual results against expected results, response rates and totals. Relevance: The degree to which the data reflects the real-world situation meeting the needs of users. Data with high relevance efficiently tells you what you want to know without having to sort through irrelevant data and information.	The ‘coverage’ property is used as a measuring criterion for the Validity quality dimension. The ‘relevance’ property is used as a measuring criterion for the Relevance quality dimension.

<i>Dimension & description</i>	<i>Example properties</i>	<i>Dimension usage description</i>
was collected or created as well as reflective of the situation it is attempting to describe.		
Timeliness/currency: Are the timeliness and currency of the data appropriate? Timeliness refers to how quickly data can be made available when required, and the delay between the reference period (period to which data refers, such as a financial year) and the release of information. Factors that may impact this include collection method and processing.	Availability: Data must be made available as soon as possible both internally to the department as well as externally to public entities (where applicable). Delays commonly occur between the release date of the data and the reference period it spans. Currency: Based on the needs of the user, if data is no longer current, it may no longer be fit for its intended use. For example, data for activities occurring within a set period of time, or where data expires or becomes irrelevant after a set date or time. Where it references specific time periods, data should be current and released before it is superseded.	The ‘availability’ property is used as a measuring criterion for the Timeliness quality dimension. The ‘currency’ dimension is used as a measuring criterion for the Timeliness quality dimension.

<i>Dimension & description</i>	<i>Example properties</i>	<i>Dimension usage description</i>
Collection: What was the collection method and was it consistent? Data collection methods must be appropriate depending on the type of data being collected.	Data is commonly understood: Definitions of the data itself, and of the data collection and processing methods should be made available organisation-wide to provide a common understanding. When this is well documented, collection over time is consistent independent of the person or system used to collect the data. Appropriateness: Data must be collected using the most appropriate method possible for the data being recorded. Some possible methods of data collection include surveys, forms, data feeds, and reporting from service providers. Collection methods must be chosen depending on the level of data quality required for analysis, and the level of risk.	The organisation-wide availability of definition of data collection and processing methods is a measuring criteria for the Accuracy quality dimension. The ‘Appropriateness’ property is used as a measuring criterion for the Methodological Soundness quality dimension.
Consistency: Is the data consistent with related datasets, agreed standards and formats?	Comparison with similar data assets: There must be the capacity to make meaningful comparisons across multiple data assets. This is achieved through common data definitions and standards. Common data definitions should be agreed and shared across the department, and any inconsistencies should be managed.	The ‘<i>Comparison with similar data assets</i>’ property is used as a measuring criterion for the Consistency quality dimension.

<i>Dimension & description</i>	<i>Example properties</i>	<i>Dimension usage description</i>
Consistency of data means that the data is collected, grouped, structured and stored in a consistent and standardised way. This requires standard concepts, definitions and classifications to be implemented across departments, and agreed upon as to their meanings and interpretation.	Consistency over time: This refers to the ability to track a data asset over time. Changes may be made regarding scope, definition and collection; however, the data must remain consistent so it can be compared with previous data assets. Documentation: Changes to data and any related processes should be well documented and traceable. Data dictionaries specifying business rules, validations, specifications, and glossary terms should be available, regularly maintained and updated with any changes made to data.	The ‘<i>consistency over time</i>’ property is used as a measuring criterion for the Consistency quality dimension. The ‘<i>documentation over time</i>’ property is used as a measuring criterion for the Quality Check Processes quality dimension.

Appendix 15 part B: International role-players and their RDM initiatives

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
1. The Data Management Association United Kingdom (DAMA UK)	The DAMA UK data quality dimensions are an appropriate reference point for this study, since they provide a constructive and specific context for assessing any proposed efforts and practices in data quality. Apart from the dimensions of data quality, DAMA UK also aims to influence future directions in various areas of information management, including education, standards, technology, regulation and business practices (DAMA UK Ltd, 2017). Therefore, it can be concluded confidently that the organisation will continue to control data quality activities.	They have documented six dimensions of data quality: Completeness, Uniqueness, Timeliness, Validity, Accuracy, and Consistency. In this study, these dimensions are the epitome of measuring data quality and quality assurance and they are addressed in more detail in the following section.
2. The International Science Council (ICS)	The ICS uses an overarching and collaborative approach to seeking solutions to data management problems; this technique is applied by 'WDS Working Groups'. The WDS Working Groups discuss data quality concerns directly related to Trustworthy Repositories certification processes. The WDS has reported the concepts of data sharing. Principle 3 deals with factors of data quality (ICSU – WDS, 2016)	The ICS has adopted the requirements of the Core Trust Seal which relate to confidence in repositories and their content. Many Core Trust Seal criteria are to ensure integrity and authenticity; hence this criterion is a good indicator for evaluating consistency.

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
3. The Research Data Alliance (RDA)	RDA Working and Advocacy Groups concentrate on particular concerns and interests related to RDM. Each group is comprised of both practitioners and academics (RDA, 2016). In essence, the Alliance is driving the international agenda when it comes to RDM.	Various RDA Working Groups and Community Groups discuss data quality and quality assurance problems, respectively. The working groups include: – The RDA Dynamic Citation Working Group: “[which] developed a comprehensive specification for citing everything from a subset of a data set to data generated dynamically, on-the-fly” (RDA, 2016) Knowing how to cite and record each data file specification means that all potential errors are detected and addressable. This, in essence, protects data quality. This focus (specification for citing data) of the WG is used as a measuring criterion for the Uniqueness of the Data quality dimension. – RDA-WDS Publishing Data Workflows Working Group: “provide an analysis of a reasonably representative range of existing and emerging workflows and standards for data publishing, including deposit and citation, and to provide components of reference models and implementations for application in new workflows” (RDA, 2016) The information behind data access/publication workflows needs to be aligned with best practices and standards; this WG enables researchers to have access to guidance that will increase the quality of data publishing. This focus (using workflows and standards for data

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
		publishing) of the WG is used as a measuring criterion for the Accessibility quality dimension. WDS/RDA Assessment of Data Fitness for Use Working Group: Data accessibility and fitness for use are focused on issues related to data objects, access systems, and data management processes such as data description, curation, peer review, and machine readability. (RDA, 2015). This WG focuses on determining the quality requirements to be included and the weighting of each of those requirements. Preferably, the method should lead to a corresponding metric (RDA, 2016). Research Data Collections Working Group This WG focuses on defining object collections by making them actionable through automated processes in order to be able to cope with the ever increasing quantities and volumes of data (RDA, 2016). <u>Data Versioning Working Group:</u> Reproducibility of research findings is urgently required. This WG focuses on the capacity of researchers to quote the exact extract of the data set used to substantiate their study publication (RDA, 2016). The WG is confident that data sets and data products need some sort of systematised way to reference the exact version of the data used to support the research findings and/or to produce higher-level products (RDA, 2016). The concept of versioning is a measure of quality assurance since versioning concepts developed for research data must

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
		be consistent with versioning concepts used in persistent identification systems (RDA, 2016). Versioning of data used to underpin research findings is used as a measuring criterion for the Interpretability quality dimension. Data for Development Interest Group: Many researchers work on subjects that require information from areas and population groups where there is only limited data (RDA, 2016). The goal of this WG is to show that data can be anonymised to the degree that it safeguards the right of the person to anonymity and still gives access to the data in order to avoid multiple data collection among these vulnerable groups (RDA, 2016). This is a quality assurance operation, since it preserves the privacy of participants in research and is compatible with ethical research practice. This aspect (anonymization and not creating multiple data collection among vulnerable groups) of the IG is used as a measuring criterion for the Integrity quality dimension. Education and Training on handling of research data Interest Group The vast quantity of research data generated in the scientific community has created a demand for a new and evolving range of skills and competencies for researchers who generate and use the data and for the professionals who support them (RDA, 2016). This IG aims to share information on current technologies and initiatives and to

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
		encourage training/education in order to handle data across the data lifecycle. The level at which data can be handled across the data lifecycle is used as a measuring criterion for the Methodological Soundness quality dimension. Metadata Interest Group: Metadata is a fundamental feature of RDM, The Metadata IG deals with all aspects of analysis data metadata (RDA, 2016), which leads to data quality assurance. This IG focuses on metadata to create a coherent metadata approach covering metadata definition, restriction, navigation, provenance, preservation and use for metadata exploration, contextualisation, validation, analytical analysis, simulation, visualisation and interoperation purposes (RDA, 2016). This component (creating a coherent metadata approach) of the IG is used as a measuring criterion for the Usability of Data quality dimension. Social Science Research Data Interest Group: The wide reach of social science research needs more narrow viewpoints on how to interpret data from research in the social sciences. Social sciences and humanities research data spans several fields, exists in several forms of data, deals with various artefacts and levels, and is widely distributed – from different sources (RDA, 2016). This IG prioritises three particular areas based on data populations for study in the social sciences and humanities:

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
		– <u>Quality of data</u>: Find automated ways to conduct research and provide quality information; this area of the IG is used as a measuring criterion for the Integrity of Data quality dimension. – <u>Data policy</u>: Align and simplify data policies and their implications for making data available and for data use wherever possible; this area of the IG is used as a measuring criterion for the Timeliness quality dimension.
4. Data Observation Network for Earth (DataONE)	DataONE has developed a multi-institutional, institutional and interdisciplinary partnership to develop a cyber infrastructure that supports the lifecycle of the information.	These guidelines simplify the process of data entry by recommending simple practical measures, thereby reducing the likelihood of a researcher making mistakes.
5. UK Data Service	The UK Data Service provides one unified access point to a wide variety of secondary social science data (Bishop, 2015).	In an attempt to improve data quality The UK Data Service has a tool called QAMyData which provides numeric data health checks. The tool uses automated methods to identify and report some of the most common issues in survey or numerical data, such as missing values, repetition, outliers and direct identifiers (UK Data Service, 2020). QAMyData tackles key technological data concerns that arise when handling quantitative and survey data. The usage of automated methods to identify issues in data is used as a measuring criterion for the Quality Check Processes quality dimension.

Organisation/Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
	The UK Data Service has developed a 'Formatting data' module that specifically looks at ensuring quality of data at all stages (UK Data Service, 2023)	The UK Data Service has developed suitable procedures that are aimed to guide researchers along the data collection. The procedures are discussed in more detail in Appendix 1.
6. The European Research Council (ERC)	Together with the FAIR data principles the ERC Open Access movement offers an effective structure for managing data quality and quality assurance efforts.	The ERC provides key data quality guidelines which were consulted. These are discussed in more detail in Appendix 1. These guidelines concentrate mainly on data interoperability and how that process can be promoted. Combining various datasets requires accuracy and consistency.
7. ICPSR	The ICPSR uses a data lifecycle approach to reflect and explain key archiving considerations at any stage of the data creation process (ICPSR, 2018). Their comprehensive review of what data archiving means provides a framework for quality assurance that effectively protects the data quality.	Comprehensive detailed information supports the ICPSR goal of properly archiving records. Quality assurance to monitor data quality means data at ICPSR useful knowledge is improved to make it detailed, self-explanatory and accessible for prospective researchers (University of Michigan, 2018). The ICPSR provides specialised data services that facilitate quality assurance while archiving research data. These are discussed in more detail in Section 2.3.4.2

Appendix 15 part C: Stakeholders and their RDM initiatives

Organisation	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
The Australian Research Council (ARC) now known as ARDC	The ARDC grant application allows applicants to explain how the data will be handled, the following questions relate to data quality and quality assurance: – Researchers must outline strategies for handling the data generated as a result of the proposed study, including but not limited to arrangements for storage, access and re-use. Outlining strategies for handling data is used as a measuring criterion for the Flexibility of the data quality dimension. – The ARDC finds data management preparation to be an essential part of responsible research behaviour and strongly promotes the depositing of data resulting from a project into a suitable subject and/or institutional repository that is publicly available.	As part of getting processes in place to ensure the quality of research data management, the ARDC has criteria that are consistent with the obligations outlined in the Australian Code for Research Responsibility. These include the proper management by researchers of research data and primary materials, as well as institutional policies on data ownership, storage, retention and adequate access by the research community (Australian Research Council, 2016). The Australian Code for Responsible Conduct of Research has specific guidelines on what constitutes good research data management; hence these guidelines have a direct effect on enhancing the quality of the data. These guidance documents are discussed in Appendix 1.
DataFirst	Knowing how DataFirst operates in data management is important because DataFirst is currently the only trusted data repository in South Africa. It complies with the specifications of the Core Trust Seal. Understanding the management processes provides recommendations for quality assurance from the perspective of a trusted repository.	DataFirst has a mandate to assist organisations with data quality assessments as well as with South African data series harmonisation and sub-national data harmonisation in South Africa. Their contributions are further discussed in Section 2.10.

Organisation	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
SEDAC	The Socioeconomic Data and Applications Center (SEDAC), which is operated by Columbia University's Center for International Earth Science Information Network (CIESIN), collects and disseminates scientific data for application and study uses (SEDAC, 2020). To control data SEDAC uses clear data policies, data collection requirements and data quality specifications.	As a repository for NASA and a member of the International Science Council (ISC) – World Data System (WDS), SEDAC is notified by the highest quality standards and procedures. The data management strategy employed by SEDAC is constructive, useful and ensures good data quality (SEDAC, 2020).
Data Archiving and Networked Services (DANS)	Since 2005, DANS has helped researchers, data practitioners, other data libraries, research institutions and research financiers with data management, certification and subjects such as FAIR, open access and software sustainability (DANS, 2010). This support is critical to helping researchers to successfully carry out their research projects. DANS provides realistic guidelines which could be applied at the [the social sciences organisation].	The DANS guidelines use a basic approach to cater for data quality. DANS offers a comprehensive overview of the steps to be taken when handling data quality. These have been discussed in Section 2.15.
Australian Data Archive (ADA)	The ADA places critical importance on data quality and description. This is useful because practitioners will perform a data quality evaluation on their data that will tell you how effective data is in meeting the requirements of your stakeholders and help you	The ADA views data quality as a continuous process of management, involving data management practice from its initial development to its possible destruction (Australian Data Archive, 2018); this aspect is used as a measuring criterion for the Relevance quality dimension. The ADA

Organisation	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
	prioritise remediation on high-value datasets (Australian Data Archive, 2018).	provides useful advice on how best to handle data quality. These guidelines are dealt with in Appendix 1
California Digital Library (CDL)	The CDL Merrit repository and Dash publishing platform have a CoreTrustSeal certification.	The CDL uses a ‘Digital Object Management’ process. The process deals with data integrity and authenticity, assessment, storage, conservation planning, quality assurance, documented workflows, identification and discovery (Abrams, 2020). Documenting workflow processes when data is processed improves decision-making on how to best represent data for it to be understood. Documenting workflow processes is used as a measuring criterion for the Consistency quality dimension. The identification and discovery of research data should be done using methodological approved methods, this process is used as a measuring criterion for the Methodological Soundness quality dimension.
Digital Repository of Ireland (DRI)	Some countries have national repositories which hold nationally significant resources. The Digital Repository of Ireland (DRI) has specific processes in place that help its researchers manage and preserve research outputs better.	DRI's acknowledgement and adoption of the FAIR principles shows the wider importance of quality assurance in RDM processes. For more information, please refer to Appendix 1.
InterPARES research projects	The InterPARES research projects have examined the creation, maintenance and preservation of digital records. A major finding of the research is that, to preserve trustworthy digital records (i.e., records that	The findings of the InterPARES 1 research were confirmed by the research conducted by the InterPARES 2 Policy team, which further concluded that it is possible to develop such a framework of principles

Organisation	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
	can be demonstrated to be reliable, accurate and authentic), records creators must create them in such a way that it is possible to maintain and preserve them. This entails that a relationship between a records creator² and its designated preserver³ must begin at the time the records are created.	to support record creation, maintenance and preservation, regardless of jurisdiction. The Policy team developed two complementary sets of principles, one for records creators and one for records preservers, which are intended to support the establishment of the relationship between creators and preservers by demonstrating the nature of that relationship. The principles for records creators are directed to the persons responsible for developing policies and strategies for the creation, maintenance and use of digital records within any kind of organisation, and to national and international standards bodies. The principles for records preservers are directed to the persons responsible for developing policies and strategies for the long-term preservation of digital records within administrative units or institutions that have as their core mandate the preservation of the bodies of records created by persons, administrative units or organisations external to them, selected for permanent preservation under their jurisdiction for reasons of legal, administrative or historical accountability.

Appendix 15 part D: Data quality and social sciences research funders

Research funder	Reason for reviewing the research funder for clues on data quality	Documented proof of how data quality is being addressed
World Bank	The World Bank is using research results developed by researchers in the social sciences to introduce policies and programs to achieve its strategic objectives. As part of the application process for research funding, applicants are expected to supply a project plan detailing who will have data management responsibilities and at what level. Specifying certain obligations is a viable way of understanding who is responsible for ensuring good quality of the data that is generated.	Project Plan: A project plan must accompany the application (World Bank Group, 2015). The project plan contains a 'capacity statement' which gives the World Bank Group details of the extent to which applicants have the relevant expertise/experience to successfully manage/complete the project (World Bank Group, 2015). The 'capacity statement' component is a measuring criterion for the Quality Check Processes quality dimension.
United States Agency for International Development (USAID)	USAID has a 'Conducting Data Quality Assessment Framework' which offers "<i>practical guidance to the assessment of data quality and helps managers to understand how confident they should be in the data used to manage a program and report on its success</i>" (USAID, 2010:1). The main purpose of the USAID Data Quality Assessment Framework is to make data managers	USAID has a 'Conducting Data Quality Assessment Framework' that informs the performance management and measurement assessment of quality assurance. The USAID 'Conducting Data Quality Assessment Framework' provides useful guidance to help data managers (researchers) resolve quality problems when gathering their data. The USAID recommendation for this study states that the Assessor of Data Quality (DQ) should ensure that they understand the precise definition of the

	aware of “<i>the strengths and weaknesses of the data, as determined by applying the five data quality standards and are aware of the extent to which the data integrity can be trusted to influence management decisions</i>” (USAID, 2010:1). Reviewing the USAID Data quality Assessment was important to recognise the data quality measures undertaken by research funders. USAID Data Quality Assessment Framework enables data managers to be aware of the strengths and limitations of the data, allowing them to make strategic decisions that will have a positive effect on the quality of the data they plan to produce. The five data quality standards that USAID uses are <i>validity, reliability, precision, integrity and timeliness</i>.	indicator by reviewing the Performance Indicator Reference Sheet and addressing any ambiguity issues before the DQA is conducted (USAID, 2019). It is wise that researchers have a complete understanding of all the indicators and measures in a data file, this helps them to eliminate any contradictions that may arise when collecting the data. The USAID Data Quality Assessment Framework is a measuring criteria for the Comparability and Coherence quality dimension. The USAID grant application process emphasises quality assurance to ensure that grant holders resolve major issues that may hinder the research process due to data management errors that affect the quality of the data.
The National Science Foundation (NSF)	The NSF has a merit process that allows candidates to draft their proposals with due diligence to determine how quality assurance during the study process can be ensured. Potential applicants must meet the criteria set out in the NSF Proposal & Award Policies & Procedures Guide (PAPPG). The main NSF PAPPG requirement that is important for this study is the Data Management Plan.	The NSF Data Management Plan is an important quality assurance document because it demands that an applicant outline all the measures that will be taken while carrying out their study. The DMP should clarify how the proposal would adhere to the NSF policy on dissemination and sharing of research findings as described on the NSF website. Proposals which do not include a DMP cannot be submitted (The National Science Foundation, 2018). In this study, the NSF DMP was considered to consolidate the essential components a DMP should have which affect

		the quality of the data. The scope of a DMP is a measuring criterion for the Usability of Data quality dimension, the following NSF DMP sections were considered as measuring criteria for this study: – <i>The data forms, samples, physical collections, software, curriculum materials and other materials that will be generated during the project</i> (The National Science Foundation, 2018). The type of research materials that will be generated during a research project is used as a measuring criterion for the Usability of Data quality dimension – <i>The standards to be used for data and metadata format and content in the absence or considered insufficient of established standards should be recorded along with any suggested solutions or remedies</i> (The National Science Foundation, 2018); This section is used as a measuring criterion for the Comparability and Coherence quality dimension.
European Research Council (ERC)	The ERC uses a general collection of criteria to be met by DMPs to illustrate efforts to promote the highest quality research. The ERC Open Data and Data Management Strategy demands that data be in compliance with FAIR principles (Findable, Accessible, Interoperable and Reusable). This requirement is used as a measuring criterion for the Usability of Data quality dimension.	The ERC Open Research Data and Data Management Plan applicants can comply with the FAIR principles by completing the ERC DMP parts. The FAIR principles were acknowledged and are measuring criteria for the Integrity, and Usability of Data quality dimensions.

Appendix 15 part E: Academic and government institutions in South Africa that contribute to data quality management

Initiative	Reason for reviewing the stakeholder for clues on data quality	Documented proof of how data quality is being addressed
The Network of Data and Information Curation Communities (NeDICC)	An established RDM initiative comprising of research specialists (in their various roles) from different academic institutions in South Africa formed a community of practice focusing on managing research data	Members of NeDICC are dedicated in improving the skills of patrons at member institutions. This is done by arranging “<i>seminars, workshops and conferences to promote awareness around RDM and digital (including data) curation aimed at practitioners and managers involved with digital object management and encourages the growth of knowledge in this area</i>” (Kahn et al., 2014:298). NeDICC is dedicated towards improving organisational RDM knowledge through learning from each member institution RDM challenges and successes. Quality assurance is an aspect that the NeDICC group takes seriously and intends for all member organisations to uphold high quality RDM practices.
Data-Intensive Research Infrastructure for South Africa (DIRISA)	DIRISA aims to implement a Certified Tier one (national) Trusted Repository for research data and to operationally deploy and maintain data services and virtual research environments that enable researchers to leverage this ICT platform for data-intensive research (DIRISA, 2017).	The national strategic framework of DIRISA allows it to be able to coordinate the creation and implementation of sound data management and operational policies and practices underpinning efficacious data-intensive research (DIRISA, 2017)

The Association for Library Professionals of South Africa (LIASA)	Represents all institutions and people working in libraries and information services in South Africa (LIASA, 2017)	LIASA is an association that has the mandate of advancing library services to a level that will assist organisations to align library services with best RDM practices
The South African Data Archive (SADA)	SADA provides a central archival facility in which institutions such as market research companies, statistical agencies and academic institutions can preserve their data for future use (SADA, 2020).	SADA provides quality guidelines (i.e. formulation of policies, preservation, usage and dissemination channels of data and related documentation) to organisations that are interested in changing RDM and related policies
DataFirst	Provides Open Research Data infrastructure for discovering and accessing data and developing skills among prospective users, particularly in South Africa (DataFirst, 2020).	Provides training to advance the data management skills of practitioners through intensive curation workshops that address different data quality aspects. The training focuses on policies and methods for managing individual-level data through all the stages of the data lifecycle (DataFirst, 2020)
CPUT	SA University with an established RDM function	The university has implemented an e-strategy that is an integral part of the research data services that enable research publications to be effortlessly linked to datasets (Chiwane & Mathe, 2015).
The University of Cape Town (UCT)	The university has established an eResearch Centre to work and partner with researchers to find appropriate IT solutions for their research activities (Chiwane & Mathe, 2015).	Fundamental parts of the UCT RDM policy include funder responsibilities; researcher responsibilities; university responsibilities; data service provider responsibilities; data management planning; tracking and reporting of validation requirements

University of Pretoria (UP)	The university has developed an RDM policy	Provide RDM recommendations that can be adopted by other universities (such as, Investigate necessary IT infrastructure) and that can accommodate small and big data sets, as well as high-performance computing.
The Human Sciences Research Council (HSRC)	Science Council with an established RDM function	The institution has policies and procedures that facilitate data deposit, preparing data and related documentation and data sharing (Lotter, 2014)
The South African Medical Research Council (MRC)	Data resulting from MRC-funded research must be correctly curated and published with the required high-quality metadata during its lifecycle.	The council has defined research data management principles that are noteworthy; including making data available with minimum restrictions, promoting secondary usage, research data to conform to the highest validity and ethical considerations.
The Council for Scientific and Industrial Research (CSIR)	The Council has embarked on numerous RDM initiatives to understand the data management dynamics of researchers (CSIR, 2019).	The CSIR has set up an RDM policy that has created a new atmosphere in the way researchers do their job and RDM based on best practices and standards.
SAEON	A government research platform that monitors environmental monitoring through data management and usage.	SAEON has implemented data management systems that promote current and future use of data through data networks.

The South African National Biodiversity Institute (SANBI)	Conducts key research on the ecosystem and biodiversity using an information management strategy.	The SANBI research strategy acknowledges tacit and codified management of knowledge through proper RDM.
Department of Science and Innovation (DSI)	The department manages the South African Data Archive (SADA), which encourages and facilitates the exchange of research data and related documentation of large-scale, regional national and international research projects, primarily in the humanities and social sciences (SADA, 2020).	The department has outlined various data management plans to promote the open access movement of data and how to better manage publicly funded research data.
The National Research Foundation (NRF)	The NRF is tasked with facilitating knowledge creation, innovation and advancement in all fields of science and technology (NRF, 2017); this is done using RDM as a driving force.	The NRF has thoroughly discussed the main and majority areas that are covered in the lifecycle of research data and is investigating ways to improve data management.
Department of Health	The department has built and outlined a strategy aimed at spearheading efforts and delivering South Africans with an improved healthcare system	The department has outlined a data management vision to strengthen digital health governance structures (DOH, 2018); whereby quality assurance is a critical component to make such a plan effective.
Department of Environmental Affairs (DEA) – RSA	The DEA is an active department that has developed a data management guideline for good environmental data management (DEA, 2019).	Includes useful data guidelines relying on good quality assurance to guide operations within an organisation.

Appendix 16: Augmented UK DAMA data quality dimensions

Note: that the suggested framework for data quality (Augmented UK Data Archive Lifecycle) did not include all of the data quality characteristics listed in the table below. Overlapping dimensions were aggregated, and only those relevant to social sciences research were used.

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
1. Completeness	Completeness refers to “<i>the proportion of stored data against the potential of ‘100%’ complete</i>” (UK-DAMA Working Group, 2013:8). Stored data does not necessarily mean that comprehensiveness is maintained throughout the dataset. Completeness refers to the rate of missing values for all data elements (Azimae, 2015). Accounting for missing, null, invalid and incorrect values determines if a dataset is complete and can be used by the next person.	<u>Open Risk Academy</u> Completeness denotes the degree to which all relevant data in the dataset are available. A calculation of the completeness of data will be the percentage of the data missing (Open Risk Academy, 2016). In the context of a database/data file completeness is defined as the following: – Schema completeness: “which is the degree to which entities and attributes are not missing from a database schema” (Open Risk Academy, 2016); this completeness is used as a measuring criterion for the validity quality dimension.	Taking a fixed set of fields in a data file “<i>completeness is thus straightforward to measure, and can be expressed as the absolute number or fraction of the fields present and not empty</i>” (Charles et al., 2017). This relates to counting the number of null values in a data set. In order for data to be 100% complete all missing, null and invalid values in a dataset have to be accounted for. Statistics South Africa (2010:42) substantiates that “<i>all data (including source data, related frame data, and related survey data) within the same series must use the same concept and definitions</i>”. This principle has to be

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
		– Column completeness: “<i>is a function of the missing values in a column of a table</i>” (Open Risk Academy, 2016); this completeness strategy is used as a measuring criterion for the validity quality dimension. – Population completeness: “amounts to evaluating missing values with respect to a reference population” (Open Risk Academy, 2016); this completeness strategy is used as a measuring criterion for the completeness quality dimension.	consistent across different data files including the master file. This dimension is concerned with how an institution or a researcher handles missing values in a data set. This aspect is used as a measuring criterion for the Completeness quality dimension.
2. Uniqueness	Every variable in a data file should be distinct in its subject. Uniqueness means that “ <i>nothing will be recorded more than once based upon how that thing is identified</i> ” (DAMA UK Working Group, 2013:9). Not recording a variable more than once is used as a measuring criterion	<u>Oracle, 2009</u> The uniqueness attribute function prevents the repetition of attribute values, both when added and changed (Oracle, 2009). In a data file uniqueness can be defined using two methods:	Identifiers in a dataset are important in distinguishing the unique dimensions and characteristics found in a data file. According to Statistics South Africa (2010:46), “<i>the common identifier must be unique in every dataset. Rules and practices must be agreed upon to ensure</i>

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	for the Uniqueness quality dimension. If variables are duplicated, their uniqueness is compromised and negatively impacts data quality. Therefore <i>“the attribute of uniqueness feature prevents duplication of attribute values, both when adding and modifying them”</i> (Oracle, 2009). Uniqueness is a dimension that confirms the exclusivity of values in a data file.	– Value duplication: Produces multiple points in a data file that result in interpretation and analysis. Duplication compromises data consistency because it repeats information that alters context (Oracle, 2009). This method is used as a measuring criterion for the Uniqueness quality dimension. – Across one object class: If the uniqueness attribute is allowed, then uniqueness prevents the user from adding two attributes or two variables with the same ID (Oracle, 2009). This method is used as a measuring criterion for the Uniqueness quality dimension.	<i>uniqueness”</i>. Such rules enable accountability when tracing how uniqueness was managed. Duplication undermines the accuracy of the data as it repeats details that change significance. Value duplication in a data file creates multiple points which result in interpretation and analysis. This aspect is used as a measuring criterion for the Uniqueness quality dimension
3. Timeliness	Timeliness has a direct impact on when data is collected as compared to when data is made available; <i>“is the information processed and delivered rapidly without</i>	<u>Jayawardene, et al., (2016).</u> <i>“Is the information processed and delivered rapidly without delays?”</i>	After the completion of a research project, it may take a long time before research findings are made available. This dimension measures the time it takes for

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	<i>delays?” (Jayawardene, et al., 2016:12). Timeliness represents “the degree to which data represents reality from the required point in time” (DAMA UK Working Group, 2013:10). This guideline is used as a measuring criteria for the Timeliness quality dimension.</i>	(Jayawardene, et al., 2016:12). The time it takes to access data between the estimated date, and when it is readily available for use, represents timeliness. This feature of timeliness is used as a measuring criterion for the timeliness quality dimension. <u>Sidi et al. (2012)</u> Timeliness is “<i>the extent to which the age of the data is appropriated for the task at hand</i>” (Sidi et al., 2012)	data to be processed and disseminated. How this dimension can be measured is that “<i>data must follow an expected trend established over time any inconsistencies in the key variables must be reconciled</i>” (Statistics South Africa, 2010:43). Questions that assist in measuring this dimension include; <i>Are there likely to be updates or revisions to the data after its release?; What is the reference period for the data?</i> (Bauman, 2019). Answers to these questions indicate the continuity aspect of the data. The time between data collection, processing, and distribution should reflect the real events and facts that occurred throughout the collection period. This dimension is used as a measuring criterion for the Consistency quality dimension

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
4. Validity	Data validity means that “ <i>data are valid if it conforms to the syntax (format, type, range) of its definition</i> ” (DAMA UK Working Group, 2013:11). Data definitions used in related documents must correlate with the actual syntax used in a data file. This dimension is used as a measuring criterion for the validity quality dimension	<u>Sidi et al. (2013).</u> Validity is “ <i>a measure of the existence, structure, content and other basic characteristics of the data</i> ” (Sidi et al., 2013:303). Measuring basic characteristics is used as measuring criteria for the Validity quality dimension. Measurement of validity includes evaluating relevant data specifications (data standards, data model, business rules, metadata, and reference data) In both source and destination repositories (Sidi et al., 2013). This feature of validity is used as a measuring criterion for the validity quality dimension	Validity can be measured through the provision for “ <i>comparable series or source data to be based on common identifiers that are documented in metadata</i> ” (Statistics South Africa, 2010:44). This provision is used as measuring criteria for the Uniqueness quality dimension. Comparison between the data and the metadata or related documentation (i.e. Codebook) for the data item reveals valid data representation. This dimension is used to determine if all elements of a data file are correct and do not contain any inaccurate information that may reduce the precision of variables. This aspect is used as a measuring criterion for the Validity quality dimension
5. Accuracy	The association between the data being collected and a questionnaire have to relate,	<u>Buckey (2016)</u>	Identifiers (such as variable headings and value labels) are a direct representation of

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	otherwise the data is not an accurate representation. Accuracy is “<i>the degree to which data correctly describes the "real world" object or event being described</i>” (DAMA UK Working Group, 2013:12). Accuracy represents the true image of what is being studied. Accuracy is the closeness of the results of observations to the true values or values accepted as being true (Buckey, 2016). Generally, accuracy is the degree to which data correctly reflects the real-world object (Smart Bridge, 2020).	Accuracy is the closeness of observational findings to the true values or values recognised as true (Buckey, 2016). The key type of accuracy that distinguishes the essence of true values is the accuracy of the attributes, representing estimates of the facts. Attribute accuracy is used as a measuring criterion for the Accuracy quality dimension. <u>Sidi et al. (2012)</u> The extent to which data is correct, reliable and certified	the data collection instrument (questionnaire). To measure accuracy “<i>the common identifier must be unique in every dataset</i>” (Statistics South Africa, 2010:46). Questions that assist in measuring this dimension include what are the standard errors for the key data items? What steps have been taken to minimise processing errors? (Bauman, 2019). <i>consistency</i>. through the lifecycle of the research data. Accuracy is a dimension which emphasises the correctness of what has been observed. Accuracy is expected to reduce incoherence in a data file as opposed to what was found during data collection. This aspect is used as a measuring criterion for the Accuracy quality dimension.
6. Consistency	Consistency is “ <i>the absence of difference when comparing two or more</i>	<u>ICPSR</u>	Any element that may influence the reliability and consistency of the data

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	<i>representations of a thing against a definition</i>” (DAMA UK Working Group, 2013:13). Consistency checks can then be integrated into the data collection process through the use of CAPI software in order to correct problems during an interview (ICPSR, 2018). CAPI software is a survey data collection by an in-person interviewer (i.e. face-to-face interviewing) that uses a computer to administer the questionnaire to the respondent and captures the answers onto the computer (SAGE, 2008). The use of consistency checks is used as a measuring criterion for the Interpretability quality dimension. The standards set by a researcher should be the basis to verify consistency because consistency “<i>describes the formal</i>	Consistency tests can be implemented in the data collection process by using software to address issues during an interview (ICPSR, 2018). Consistency checks programmes can be used concurrently for a variety of functions including direct data entry, quality checks, and skips and fills (ICPSR, 2018). The usage of consistency checks programmes is a measuring criterion for the Interpretability quality dimension.	should be documented. A practical way to measure such changes requires “<i>minimum metadata for interpreting the product must be accessible</i>” (Statistics South Africa, 2010:39). Inconsistencies can be observed between a researcher’s syntax and the manner in which a fieldworker interprets the syntax, resulting in inaccurate capturing. Using the same data collecting techniques (i.e. instrument, method), and standards, similar findings should be achieved. This element is used as a measuring criterion for the Consistency quality dimension.

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	<i>conformity to set standards and field rules as well as the logical coherence of the metadata</i> ” (Charles et al., 2017:6).		
7. Usability of the data	The usability of the data should not be hampered by syntax or any technicalities; <i>“is data understandable, simple, relevant, accessible, maintainable and at the right level of precision?”</i> (DAMA UK Working Group, 2013:14). Usable data should facilitate easy secondary usage.	<u>Menna et al. (2021)</u> The extent to which data is clear and easily used (Menna et al.,2021:40). <u>Hassenstein & Vanella (2022)</u> For data to be usable <i>“both the data and the metadata must be depicted in detail, including the statement of usage licenses along with their publication and the exact explanations of the data origin, and should generally be shared following standards used in the particular domains”</i> (Hassenstein & Vanella, 2022:501)	A measuring indicator for this dimension is producing frequencies that reveal the coherence and logical structure used in a data file (Clark, 2014). Data has to be useful to the extent that it can answer questions that are under its scope, which measures its usability.
8. Flexibility of the data	A research project can have multiple versions of data files that differ on a few modules and scope. The entire data	<u>Becker (2019)</u> <i>“Data can be combined in specific instances with a few other data sets; Data</i>	The flexibility of all the components of data is measured if it allows for new developments (e.g. changes in definitions

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	collection has to be flexible, meaning that it has “ <i>to be comparable and compatible with other data including useful groupings, classifications, repurposing and easy to manipulate</i> ” (DAMA UK Working Group, 2013:14).	<i>can be combined in a high-level manner with a few other data sets</i>” (Becker, 2019:8). <u>Nyaboga & Mwaura (2009).</u> Flexibility is the capacity of a view to change to accommodate new users as the application demands this (Nyaboga & Mwaura, 2009:16).	and classifications) (Statistics South Africa, 2010). Adding new developments demonstrates the longevity aspect that a data file has.
9. Relevance	Research data is potentially beneficial because it can help open up the doors of new knowledge discovery and better decision-making. Data has to be relevant if it stands a chance of achieving this. Relevance is “ <i>the degree to which it meets the real needs of clients. It is concerned with whether the available data sheds light on the issues of most importance to users</i> ” (Statistics South Africa, 2010:4).	<u>Nasr et al. (2020)</u> Relevance means “<i>the extent to which data is applicable and relevant for the task at hand</i>” (Nasr et al., 2020:5). <u>OECD (2017)</u> Measuring relevance requires “<i>the identification of user groups and their needs</i>” (OECD, 2017:78).	Data that is error-free can be easily interpreted and used. Relevance can be measured through “ <i>the data producing agency having an independent data quality audit process</i> ” (Statistics South Africa, 2010:13). The audit has to verify if data entry errors average an acceptable accuracy rate. This dimension is used as a measuring criterion for the Relevance quality dimension.

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
			Questions that assist in measuring this dimension include if rates and percentages have been calculated, are the numerators and denominators for the same data source(s)?; Have standard classifications been used? (Bauman, 2019). Such questions contribute to guaranteeing that data is relevant and can be processed/used.
10. Accessibility	Data accessibility restrictions prevent secondary data usage. External users have to be able to access all data aspects to conveniently use data. Accessibility refers to <i>“the ease with which it can be obtained from the agency. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed”</i> (Statistics South Africa, 2010:4). This dimension is	<u>Sidi et al. (2013)</u> <i>“Extent to which information is available, or easily and quickly retrievable”</i> (Sidi et al. 2013:302). <u>Sengupta et al. (2023)</u> . <i>“Data accessibility is the metric of difficulty in accessing the existing data. From a user perspective, accessibility is the primary dimension allowing</i>	A highly practical measure for this dimension is how regular contact between the data-producing agencies and secondary users/agencies to resolve accessibility issues is maintained (Statistics South Africa, 2010). The inability of users to access data hinders secondary usage. Questions that assist in measuring this dimension include can data that hasn't been published be requested?; In which formats is the data available for people to use?

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	used as a measuring criterion for the Accessibility quality dimension.	<i>individuals to use data for their respective analyses</i> ” (Sengupta et al., 2023).	Where and how do you access them? (Bauman, 2019).
11. Interpretability	This dimension has a close link with related documentation that explains the data. Interpretability refers to “<i>the ease with which users can understand data/statistics through provision of metadata. This information normally includes the underlying concepts, definitions and classifications used in the methodology of data collection and processing</i>” (Statistics South Africa, 2010:4). This dimension is used as a measuring criterion for the Interpretability quality dimension. This dimension also addresses the question of whether there is a particular context that this data needs to be considered within; What other information is available to help	<u>OECD (2017)</u> Interpretability “<i>reflects the ease with which users may understand and properly use and analyse the data</i>” (OECD, 2017:78). <u>Jayawardene et al., 2016</u> Interpretability of data refers “<i>to the ease at which the user can understand the data. Is there any ambiguity in understanding the data and is there information available to help the user understand the terminology</i>” (Jayawardene et al., 2016).	Measuring interpretability means that a quality management system needs to be in place that will check the data for “<i>consistency with comparable datasets and any inconsistencies must be reconciled</i>” (Statistics South Africa, 2010). Reconciling inconsistencies, (ambiguous or technical terms that may need further explanation) should be done by providing users with additional information to help users better understand the data source (Asher et al., 2013).

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	users better understand this data source? (Bauman, 2019).		
12. Comparability	The management of research data doesn't happen in isolation, the comparability and coherence of data refers to " <i>the ease with which data can be successfully brought together with other statistical information within a broad analytical framework and over time</i> " (Statistics South Africa, 2010:4). This dimension captures the endurance characteristics of data and if there is technical differences in the data over time, what are the differences and what is their impact (Bauman, 2019).	<u>UNICEF (2021)</u> Comparability "<i>over time means the adoption of international standards or coding where these exist</i>" (2021:16). <u>Vale (2010)</u> Comparability is "<i>the extent to which differences between statistics are attributed to differences between the true values of the statistical characteristic, or to methodological differences</i>" (Vale, 2010:4)	Measuring comparability can be done through the ability to combine data in different contexts (other data) and the data is available in a variety of file formats (Statistics South Africa, 2010). The aspect of data being available in a variety of other formats is used as a measuring criterion for the Comparability and Coherence quality dimension.
13. Methodological soundness	The research field is overseen by recognised standards; methodological soundness refers to " <i>the application of international, national or peer-agreed standards, guidelines and practices to</i>	<u>Laliberté et al. (2004).</u> Methodological soundness refers to "<i>how do the current practices relate to the</i>	Standards, guidelines and best practices offer a corrective pathway for managing data. A measure of this dimension is based on an institution's ability to have in place a set of policies which covers all aspects of

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	<i>produce outputs</i> ” (Statistics South Africa, 2010:4). The usage of recognised standards is used as a measuring criterion for the Methodological soundness quality dimension.	<i>internationally agreed methodological practices for specific datasets?”</i> (2004:5).	the research lifecycle value chain Statistics South Africa, 2010).
14. Integrity	Integrity is one of the distinct dimensions of data quality. It refers to “ <i>values and related practices that maintain users’ confidence in the agency producing statistics and ultimately in the statistical product</i> ” (Statistics South Africa, 2010:4). This also resembles social science research data because data also has to have integrity as a final research output. This dimension is used as a measuring criterium for the Integrity quality dimension	<u>Torana Inc (2023)</u> Integrity is “ <i>the degree to which a defined relational constraint is implemented between two data sets</i> ” (Torana Inc, 2023). <u>USAID, 2019</u> Data “ <i>have integrity when the system used to generate them is protected from any bias or deliberate manipulation for political or personal reasons</i> ” (USAID, 2019:5).	This dimension measures if data is useful and satisfies user needs. Integrity can be measured through studying and reporting on the effects of the data collection mode (UK Data Service, 2019).
15. Quality Check Processes	In the course of managing research data, a research institution must have a quality	<u>Russell (2011)</u>	This dimension measures whether an institution has processes in place to focus

Dimension	Definition	Supporting institution/author description from literature	Possible measures/indicators and interpretation
	management system in place (Statistics South Africa, 2010:13).	Quality Check Processes “is a management method that is defined as “ <i>all those planned and systematic actions needed to provide adequate confidence that a product, service or result will satisfy given requirements for quality and be fit for use</i> ” (Russell, 2011).	on, monitor and check quality (Statistics South Africa, 2010:13).
16. Standards and Policies	This quality dimension requires that a set of standards and policies exist which cover all aspects of the statistical value chain /research data lifecycle and are adhered to (Statistics South Africa, 2010:7)	<u>Black & van Nederpelt (2020)</u> Standards and Policies measure “ <i>the degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use</i> ” (Black & van Nederpelt, 2020:56)	When standards and policies are in place they promote consistency of methods and results (Statistics South Africa, 2010:7)

Source: Compilation based on the following: Asher *et al.* (2013); Azimae (2015); Bauman (2019); Black & van Nederpelt (2020); Buckey (2016); Charles *et al.* (2017); Clark (2014); Jayawardene, *et al.* (2016); Oracle (2009); Russell (2011); SAGE (2008); Smart Bridge (2020) Statistics South Africa (2010); Strasser *et al.* (2012); UK-DAMA Working Group (2013); UK Data Service (2019)