Research data management practices of emerging researchers at a South African research council

By

Louise (L.H.) Patterton

Submitted in fulfilment of the requirements of the degree

MIS (Research)

Department of Information Science

Faculty of Engineering, Built Environment and Information Technology

University of Pretoria

Supervisor: TJD Bothma

Co-supervisor: MJ van Deventer

Date of submission: September 2016

Declaration

I declare that the Master's dissertation, which I hereby submit for the degree MIS (Research) at the University of Pretoria, is my own work and has not been previously submitted by me for a degree at another university.

31 August 2016

_____ _____

Louise (LH) Patterton Date

Abstract

Management of research data is globally being seen as part of good research practice. As a result of this, funders are increasingly insisting on proof of good research data management (RDM) practices when funding proposals are submitted. This study aimed at establishing the data management practices of emerging researchers at the Council for Scientific and Industrial Research (CSIR), South Africa. With no official RDM procedures currently being implemented at the CSIR, it was hoped that by gaining information about the RDM practices of emerging CSIR researchers, as well as insight into the RDM challenges experienced by them, this researcher would be able to put forward recommendations enabling the establishing of an RDM regime at the CSIR.

The study aimed at answering several research questions. The main research question was:

How can an organisation like the CSIR ensure that future researchers apply best practices when managing the CSIR's research data?

Five research sub-questions were identified:

1. What are the international RDM requirements, standards, best practices and expectations that are being developed?
2. What data practices need more formalised support: at CSIR, nationally, internationally?
3. What data are collected and held by emerging researchers in the CSIR?
4. What are the current RDM practices and themes among emerging researchers in the CSIR?
5. What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?

A total of 48 emerging researchers from the Council for Scientific and Industrial Research (CSIR), South Africa completed an online survey investigating their RDM practices. RDM practices investigated included the use of data management plans, data storage and backup locations, creation of metadata, metadata standard adherence, and data sharing practices. Challenges faced when managing research data, as well as RDM needs and requirements, also formed part of the survey. Results of the online questionnaire revealed that the RDM practices of the group studied do not show to differ significantly from experienced CSIR researchers, or from researchers studied elsewhere on the globe. Findings enabled this researcher to put forward several recommendations which would assist in the implementing of a formalised RDM structure at the CSIR. Recommendations addressed, but were not limited to: formalization of RDM procedures, RDM marketing, and RDM training.

Acknowledgement

I wish to express my appreciation to the following organisations and persons who made this dissertation possible:

1. The dissertation is based on a research project at the CSIR. Permission to conduct a study with emerging researchers, and to make use of data and findings, is gratefully acknowledged.
2. The CSIR for financial support.
3. Professor Theo Bothma, my supervisor, and doctor Martie van Deventer, my co-supervisor, for their guidance and support.
4. CSIR emerging researchers for their time, effort and willingness when asked to take part in the study and complete the online survey.

Contents

List of figures and tables

List of figures

List of tables

List of acronyms and abbreviations

CSIR: Council for Scientific and Industrial Research. For more information, see http://www.csir.co.za/

DIRISA: Data Intensive Research Initiative of South Africa. For more information, see http://www.dirisa.ac.za/

DMP: Data Management Plan

DOI: Digital Object Identifier

FAQs: Frequently Asked Questions

FTP: File Transfer Protocol

NeDICC: Network of Data & Information Curation Communities. For more information, see https://nedicc.com/

RDM: Research Data Management

SET: Science, Engineering and Technology

TOdB: Technical Outputs Database. It is the in-house repository of the CSIR

1. Chapter 1: Study background and research questions

1.1 Introduction

This chapter introduces the study, and aims to provide insight into the objectives to be achieved, the research problems to be answered, and the contribution that the study makes to the field of information science. As part of this introductory chapter, this researcher also explains how the remainder of the dissertation will unfold.

1.2 Research data management and its components

Research data management (RDM) is defined as the active management and appraisal of data over the lifecycle of scholarly and scientific interest (Donnelly, 2013). It is a general phrase used to include the following activities: the planning, creating, storing, organising, accessing, sharing, describing, publishing and curating of data used or generated during the research project (Curtin Library, 2015). It would also include any accompanying documentation or information that provides context to the data. RDM is regarded as a standard of good research practice, offering a variety of benefits for researchers and the broader scientific community. Benefits are profuse and include the enabling of verification, sharing and citation of results (Ondracek, 2013), the meeting of funding body grant requirements (University of Oxford, 2016), avoiding unnecessary duplication, preventing data loss, and improving data integrity (Emory Libraries, 2016). In short, the benefits of research data management will ensure the responsible conduct of research in several key areas; these areas being security, compliance, efficiency, access and quality (University of Western Australia, 2015).

An essential part of the RDM process is the formulation of a data management plan: a formal document that outlines what a researcher will do with his/her research data during and after the completion of research (NCSU Libraries, 2014). Following on from this: a RDM-related trend gaining moment world-wide, is the insistence of research funders on the provision of an RDM plan when a research proposal is submitted. Such a plan normally contains information on the following components: description of data, standards for formats and metadata, plans for short-term storage, legal and ethical issues, access policies, and long-term archiving (MIT libraries, n.d.).

Evidence of this amplifying trend abroad is quite prevalent when investigating the funding terms and conditions related to RDM of major funders in the United Kingdom (UK), the United States of America (USA), and Australia. The Digital Curation Centre (DCC), based in the UK, states that funding bodies increasingly have a requirement entailing the

development and implementation of data management and sharing plans (Digital Curation Centre, 2016). When scrutinizing the DCC's analysis of UK funding requirements, it is seen that all nine major UK research funders stipulate requirements with regards to data management plans, access/sharing, long-term curation, and monitoring.

In the USA, a 2013 memorandum by the Office of Science and Technology, and addressed to the heads of executive departments and agencies, announced several objectives regarding the public access to scientific data (Executive Office of the President, 2013). One of the objectives stated that all researchers in receipt of federal grants should develop data management plans. These plans needed to describe how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explain why long-term preservation and access cannot be justified. A recent crowd-sourced table, attempting to collect and consolidate guidelines from US federal funding agencies (Whitmire *et al.,* 2015), reveals that in addition to data management plan requirements, many funders in the United States now also stipulate requirements pertaining to data access, data sharing, and long-term curation.

In a similar vein, recent updates to Australian Research Council funding rules resulted in all funding applications now requiring a brief data management plan (Australian National Data Service, 2014). Moving across to Europe, it is found that participants in the Open Research Data Pilot of Horizon 2020[1] are required to deliver a data management plan within the first six months of the project (European Commission, 2016). In South Africa, a statement was released by the National Research Foundation (NRF), an independent government agency forming an entity of the Department of Science and Technology, declaring that from 1 March 2015, data generated during NRF-funded research should be deposited in an accredited Open Access repository, with the provision of a Digital Object Identifier for future citation and referencing (National Research Foundation, 2015). This requirement entails applicants having to supply details of their intended data management practices, and complete a section titled 'Details of Research: Data Storage and Dissemination', when submitting NRF grant proposals (Pillay, 2016).

when applying for funding (Pillay, 2016). It is anticipated that in the near future, more South African research funders will require similar RDM plans and practices from grant recipients.

While literature pertaining to the adherence of good data management practices in the United Kingdom, the USA, and Australia is plentiful and detailed in its analysis, a similar

[1] Horison 2020: The biggest European Union Research and Innovation Programme spanning from 2014 to 2020 and including funding of 80 billion Euros.
https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020 .

scenario does not exist in South Africa. South African RDM-related study are scarce and often subject to confidentiality issues. Examples of the sensitivity of this RDM information include the non-sharing of data from studies done at the University of Pretoria (Pienaar, 2013) as well as the University of South Africa (Shai, 2014). This researcher was also not able to trace find any RDM study results conducted elsewhere in Africa.

Summarising the above, it can be said that RDM, including the formulation of a data management plan and in most instances, supplying this plan to funders, is becoming an integral and beneficial part of the research process. It would not make sense for any research organization to ignore this trend.

1.3 The CSIR and research data management

The Council for Scientific and Industrial Research (CSIR) is one of the leading research organisations in Africa. The CSIR has a mandate to promote scientific development and stimulate industrial growth for the betterment of South African society and economy. It employs more than 2000 employees who use their expertise to improve, through multidisciplinary research and technological innovation, the quality of life of the people of South Africa (CSIR, 2015). Its research and development output for the 2014/2015 year resulted in 311 published journal articles, 294 conference papers, 45 new technology demonstrators, and 18 new international patents (CSIR, 2015). Considerable amounts of data are produced during research, yet there is currently a very limited picture of how researchers create, use, store, share or otherwise manage research data at the CSIR. CSIR-wide data management policy and procedures do not exist as separate documents. These have been incorporated as part of the Research and Development policy, as well as the Records Management Policy. The current Records Management Policy (CSIR, 2011) addresses RDM by stating that managing, storing and retaining data for specified periods of time forms a CSIR responsibility. This policy views research data as a research record, and while mentioning the need for management thereof, it does not view data as separate from research record, nor does it address the unique or specific characteristics and requirements of research data. It is, however, not clear to what extent the implementation of these policies are being monitored.

CSIR research funding stems from two main sources: public funding and contract funding. Publicly funded research at the CSIR originates from the Department of Science and Technology, meaning that the NRF statement discussed in section 1.2, is highly relevant to CSIR research. Contract research funders are dynamic, numerous, diverse in kind, location (i.e. local as well as international) and requirements, and it is expected that this diversity will

be revealed in RDM study results. It should be mentioned that it is not known/stated/clear whether the two policies mentioned here have been based on RDM requirements stemming from funding sources.

In an exploratory study investigating the RDM practices of experienced researchers at the CSIR (Patterton, 2014a), it was found that the most recent research guide, used by CSIR researchers and assisting them with the entire research process, was published in 2003. The guide does not have a separate section on data management and only contains a few sentences giving basic instructions on backup, metadata and file-naming of 'records' (thus, not data-specific) under a single-paged section titled 'Keeping records' (Scholes, 2003). The absence of any mention of a data management plan, while a normal omission for that era, renders this 2003 guide as mostly unsuitable for current CSIR researchers. Looking at international trends discussed in a previous section, it can be safely said that it is only a matter of time before more CSIR funders and clients will require or request data management plans from CSIR scientists. Bearing this in mind, the current lack of institutionalised RDM at the CSIR is cause for concern.

The factors mentioned above (uncertainty about the extent of CSIR RDM practices, no CSIR-wide RDM policy, no recent CSIR RDM guide, no CSIR-wide practice of supplying detailed RDM plans to clients/funders) and foreseen government requirements have resulted in the perceived necessity to introduce RDM at the CSIR. The appointment of a research data librarian can be seen as one of the first steps in establishing the practice of RDM at this research council. This researcher, as the appointed data librarian, has made the choice of first establishing what data are held by scientists, what the current CSIR RDM practices and trends are, and investigating this information gathered before embarking on the formulation and implementation of RDM-related policy, guidelines and other related procedures and recommendations. The knowledge gained during that process will serve as background to the current study. Part of the reasoning is explained below.

Senior researchers tend to be set in their RDM ways, received their tertiary education when RDM was not an established practice, and have generally not been exposed to formalized RDM practices for the major part of their careers. After conducting an exploratory investigation of the RDM practices of experienced CSIR scientists from September 2013 to December 2013 (Patterton, 2014a), this researcher decided that it would be more beneficial to influence younger researchers, thereby embedding good RDM habits right from the start of their research careers, than to change RDM behaviours of established researchers.

In short: although RDM is practiced in some instances, formal RDM is yet to be an established practice at the CSIR. It is anticipated that by determining the RDM habits and

methods of emerging researchers, and comparing these to the practices of the established researchers, the outcome can be utilized as a first step in developing an effective and formal RDM policy and the necessary procedures to be implemented across the CSIR.

1.4 Research questions

This chapter's previous sections have revealed and discussed the following:

- Providing proof of research data management planning to funders is a requirement found in many parts of the world: the United Kingdom, The United States of America, Australia and the European Union (see section 1.2).
- Providing proof of research data management is a requirement not yet firmly established in South Africa, but is seen to be gaining momentum (see section 1.2).
- Currently, there exists no distinct CSIR-wide RDM policy or procedures, no recent CSIR RDM guide, and no CSIR-wide practice of supplying detailed DMPs to funders (see section 1.3).
- Knowledge about the RDM practices of CSIR researchers is limited to information about RDM practices of CSIR research group leaders, obtained via an earlier study (Patterton, 2014a). Current knowledge of the data used by young CSIR researchers, their RDM practices, the RDM challenges faced by them as well as services required, seem to be non-existent or speculative at best.
- Information and data on the RDM practices of scientists and researchers elsewhere in South Africa are scarce and often confidential. Information on RDM practices of scientists in other parts of the world, in particular the UK, USA and Australia, is plentiful, mostly not restricted, and often quite detailed in its analysis.

As a result of these aforementioned aspects and findings, it was evident that the CSIR could no longer choose to ignore the looming approach of funders' RDM requirements, and that RDM should be introduced as part of the good research process/cycle at the CSIR.

A main research question and several sub-questions were formulated after considering and investigating the following:

- the current absence of RDM-related information pertaining to emerging[2] researchers at the CSIR,

[2] For the purposes of this study, an emerging researcher is a permanent CSIR employee, 35 years or younger, and either busy with, or in possession of a PhD. This definition is in line with the definition used by the research institute; as such, the institute's biennial Emerging Researcher Symposium showcases the research conducted by researchers age 35 years and younger.

- newly-available information on RDM and established researchers (as opposed to 'emerging researchers') at the CSIR, and
- the more readily-available data on RDM practices of researchers and scientists abroad.

1.4.1 Research problem/main question

This study attempted to find an answer to the following question:

How can an organisation like the CSIR ensure that future researchers apply best practices when managing the CSIR's research data?

It was this researcher's intention to answer this research question by putting forward a detailed set of RDM-related guidelines/recommendations (see Chapter 5); these guidelines were based on the information gathered during this study in general, and via the answering of the study's sub-questions, in particular.

1.4.2 Research sub-questions

To be able to answer the main research question, this researcher needed to also address the following five sub-questions:

1. What are the international RDM requirements, standards, best practices and expectations that are being developed?
2. What data practices need more formalised support: at CSIR, nationally, internationally?
3. What data are collected and held by emerging researchers in the CSIR?
4. What are the current RDM practices and themes among emerging researchers in the CSIR?
5. What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?

A short discussion on each of these questions, detailing the reason for it being included in the study, as well as the manner in which each question's answer/s were obtained, follows below.

1.4.2.1 Sub-question 1: What are the international RDM requirements, standards, best practices and expectations that are being developed?

This sub-question was answered via the gathering and analysis of existing information on RDM practices (see Chapter 2: Literature Analysis). In order to accomplish this objective,

this researcher needed to embark on a detailed study into available information into the RDM practices elsewhere in the world. This researcher was intent on gathering and analysing as many existing RDM studies as are available, and determined the major RDM themes and developments as found via these surveys and studies. RDM themes and developments investigated included a range of RDM practices; data sharing, data annotation, data storage, data backups as well as the use of metadata will be included when establishing global themes. In addition, this researcher also aimed at finding information on RDM training undergone, as well as RDM training requirements stated by researchers, elsewhere. Challenges experienced by researchers, as well as any other RDM requirements stated, were also noted when reporting on global RDM themes.

The importance of achieving this objective is the fact that this information was required when determining the state of the 'RDM art', globally. Such information made intergroup comparisons possible, and provided this researcher with RDM behavioural targets to strive towards. It is expected that due to funder requirements being already well-established in other parts of the world (see section 1.2); information gained via literature analysis will indicate the degree to which RDM practices are more established abroad. Additionally, similarities in RDM challenges and needs served as warning to this researcher that similar themes might surface at the CSIR in due course.

1.4.2.2 Sub-question 2: What data practices need more formalised support: at CSIR, nationally, internationally?

In order to answer this research question, this researcher needed to compare her research results with the results of studies conducted elsewhere on the globe. RDM studies were analysed by this researcher; this activity forms the larger part of Chapter 2: Literature analysis. Comparisons between emerging researchers' RDM practices, and RDM practices of researchers elsewhere, were done in Chapter 4: Results and discussion.

The importance of achieving this objective is the fact that intergroup comparisons provided this researcher with RDM behavioural targets to strive towards. In addition, knowledge of challenges and needs stated elsewhere provided this researcher with insight into possible similar issues at the CSIR, once RDM is being implemented.

Information obtained by the answering of this research question enabled this researcher to get a detailed view of the global state of the 'RDM art'. This view provided a glimpse into the RDM practices of not only researchers at the CSIR, but also researchers elsewhere in South Africa, as well as across the globe. This information enabled RDM comparisons, and showed where the CSIR was lacking, was on par, or was ahead of national or global

practices. Such information served as a stepping stone in providing justification for CSIR-related RDM guidelines made towards the end of this dissertation.

Additional outcomes achieved via the answering of this research question, included being able to make comparisons between emerging researchers' RDM practices, and RDM practices of the CSIR's experienced researchers. The importance of achieving this objective is the fact that comparisons indicated how experience, age and post level might influence RDM behaviours. Conversely, it was foreseen that the more recent university attendance and education, might have resulted in more advanced skills, for example. In short: by achieving this objective, a better overall picture of RDM across all post levels, age groups and research experience, could be established.

In order to answer this sub-question, information was obtained via a survey method to be described in Chapter 3: Methodology, a previous RDM study into the RDM practices of experienced CSIR researchers (Patterton, 2014a), as well as the gathering and analysis of existing information on RDM practices (Chapter 2: Literature Analysis). These three modes of data gathering provided insight into the RDM practices of the CSIR's emerging researchers, the RDM practices of the CSIR's experienced researchers, and the RDM practices of researchers elsewhere, respectively.

1.4.2.3 Sub-question 3: What data are collected and held by emerging researchers in the CSIR?

Through answering this research question, this researcher gained an idea of the types of data held by emerging researchers in the CSIR. In particular, aspects such as data format and type, data volumes, as well as software tools used to collect or view data, were clarified. This information not only created a picture of the data held by emerging researchers, but also indicated the storage requirements of emerging researchers, whether formats used are suited to long-term preservation, whether formats used are suited to data sharing, whether formats used are suited to depositing the data in an institutional repository, and whether special RDM arrangements should be made to deal with software tools used or required.

Details of the population being investigated are provided in section 3.7.1: Target population and sampling. More details about the method and questions used to gather information about data held by researchers, is described in section 3.6: Questionnaire planning and design. As mentioned in section 1.4.2.1, this research sub-question was answered via a survey; survey details are described in detail in Chapter 3: Methodology.

1.4.2.4 Sub-question 4: What are the current RDM practices and themes among emerging researchers in the CSIR?

Through answering this question, this researcher gained an idea of the behaviours and habits of the CSIR's emerging researchers when they are dealing with, and managing their data. This information illustrated whether or not emerging researchers are displaying accepted good practice when dealing with their data, and whether their data habits are lacking when compared with researchers elsewhere. In addition, it was also suspected that many similarities with RDM habits elsewhere would be found, but that these habits, here and elsewhere, were not adhering to best practices. The importance of achieving this study objective was the fact that such information was required when comparing RDM practices of emerging researchers with other groups. In addition, information obtained via achieving this objective provided a stepping stone, as well as rationale, for RDM recommendations put forward by this researcher.

It was this researcher's intention that by asking and answering this research question (i.e. what are the current RDM practices among emerging CSIR researchers?), a general understanding of an emerging researcher's RDM habits was obtained. Through asking this question, this researcher wished to obtain information about emerging researchers' every dealing with regards to research data: the data volumes dealt with, the storage locations used, the frequency of data backups, and adding data documentation to datasets. Currently, knowledge of RDM practices of the CSIR's emerging researchers is speculative; relying on the results of a previous CSIR RDM study and extrapolating these findings was not a viable option when the RDM habits across the CSIR were determined.

RDM is a complex activity and consists of many sub-sections; this researcher therefor expended considerable effort into deciding on the survey tool to be used, RDM aspects to be investigated, RDM aspects to be excluded, and formatting of questions.

As mentioned in section 1.4.2.1, this research sub-question was answered via a survey; survey details are described thoroughly in Chapter 3: Methodology. Problems and challenges faced by this researcher when conducting an earlier RDM study were duly taken into account when this survey's data collection method was decided on (Patterton, 2014a). Similar efforts were made when this survey's data collecting method, tool, as well as information-gathering specifics, were determined and created.

1.4.2.5 Sub-question 5: What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?

Through answering this question, this researcher gained insight into the problems faced by emerging researchers and was able to make recommendations aimed at minimising challenges, either by way of training, or establishing and implementing new services and infrastructures. It was this researcher's intention that by asking and answering this research question (i.e. what are the RDM challenges?), an understanding of issues and challenges standing in the way of emerging researchers when managing their data, would be obtained. This researcher was of the opinion that RDM at the CSIR cannot be successfully established, and RDM procedures not followed, should concerns mentioned not be taken note of and/or addressed.

A simple fictitious example would be the following: making use of metadata standards could be stipulated in the planned CSIR RDM procedure. At the same time, the results of this study might have indicated that a metadata standard is not adhered to, and lack of knowledge about metadata and metadata standards, is stated by emerging researchers to be a challenge. Not addressing this challenge either via training or guidelines, but expecting compliance to metadata-related procedures, is setting RDM adherence at the CSIR up for failure.

As mentioned in section 1.4.2.1, this research sub-question was answered via a survey; survey details are described in detail in Chapter 3: Methodology. It was suspected that responses to stated challenges would be diverse and at times incriminating: information to be gathered when answering this question, was best supplied by respondents in a free text version, and anonymity needed to be guaranteed.

1.5 Field of research

This study on RDM practices and behaviour aimed at making an essential contribution to the field of Information Science in South Africa. It was especially aimed at increasing the prevailing knowledge, information and scientific studies available in the subfield: Research Data Management. Looking at the research focus areas at the University of Pretoria's Department of Information Science, this study can be said to fall within the areas: Information and Knowledge Management.

1.6 Value of research/contribution

As discussed earlier (see section 1.2), published information on RDM practices of South African researchers is scarce, mostly confidential, and quite peripheral in nature. Furthermore, it seemed as if the majority of studies involved higher education institutes, and not science councils.

When looking at available studies in South Africa as well as abroad, there did not seem to be a wealth of information available on comparisons made between experienced researchers and emerging researchers.

As a result of this, the study aimed to make the following contributions:

1. Provide the information science community with research findings portraying the research data management practices of a selected group of researchers at a specific South African research institute;
2. Identify areas in the current (at the time) CSIR RDM policy and procedures that would benefit from modification or revision;
3. Provide a basis for change management , to enable best practices with regards to RDM among emerging researchers in the CSIR;
4. Add to the sparse current knowledge and awareness of RDM in South Africa;
5. Promote the area of RDM as an independent focus area in the field of Information Science;
6. Inspire the executing of similar RDM-related studies at universities and science councils in South Africa;
7. Promote RDM in the field of Information Science itself by making non-confidential research data collected during this study available to interested parties, store data on a publicly-accessible platform, and promote and encourage re-use of research data collected during this study;
8. The use of emerging researchers as target population will enable the possibility of a longitudinal study at a later stage.

1.7 Study construction and research steps

This chapter aimed at identifying the research question, as well as research sub-questions to be answered during this study. With the main research question being:

How can an organisation like the CSIR ensure that future researchers apply best practices when managing the CSIR's research data?,

the remainder of this dissertation entailed describing the steps used to answer this research problem. The dissertation consists of five chapters in total.

Broadly stated, the remaining chapters portray the following steps:

- determining the RDM behaviours of researchers elsewhere as reported on, in earlier studies (Chapter 2: Literature Analysis)
- describing the method to be used to gather information about the RDM practices of emerging CSIR researchers: the data gathering tool, study sample to be used, RDM areas to be investigated, question wording and format (Chapter 3: Methodology)
- describing the research results, and comparing the findings pertaining to RDM habits of the CSIR's emerging researchers with those of the CSIR's experienced researchers, as well as researchers elsewhere on the globe. (Chapter 4: Results and discussion). Information on the RDM behaviour of the CSIR's experienced researchers was obtained via an earlier CSIR RDM study (Patterton, 2014a), while information on the RDM practices of researchers elsewhere were reported on in Chapter 2: Literature Analysis
- guidelines given by this researcher with the objective of improving RDM practices at the CSIR (Chapter 5: Recommendations)

1.8 Summary

This chapter briefly defined RDM, and gave a short overview of the global RDM 'state of the art' as it pertains to funder requirements both locally and abroad. With South African research funders starting to implement RDM requirements, and a CSIR RDM policy and procedure not yet formalised CSIR-wide, it was stated that it would no longer be feasible for the CSIR to ignore imminent funder requirements with regards to RDM. By conducting a survey investigating the RDM practices of emerging CSIR researchers, it was possible to gain insight into the current RDM situation at the CSIR, challenges being faced by researchers when managing data, and RDM training requirements. The results of the survey provided this researcher with a platform enabling the submission of guidelines pertaining to the implementation of RDM at the CSIR. In doing so, this researcher was able to address this study's research question: 'How can an organisation like the CSIR ensure that future researchers apply best practices when managing the CSIR's research data?'

The next chapter contains the study's literature analysis, where this researcher will be discussing and analysing the characteristics and findings of RDM studies conducted elsewhere.

2. Chapter 2: Literature analysis

2.1 Introduction

A literature review can be defined as **'…***a search and evaluation of the available literature in a given subject area'* (RMIT University, 2014), and involves surveying available literature, synthesising the gathered information into a summary, analysing the information, and presenting the literature in an organised style. With this definition in mind, this chapter has as its purpose to place the current study in the context of previous research and to establish a firm theoretical foundation on which the intended study is based. This was done by doing a thorough investigation and review of relevant studies by other scholars and researchers, and by establishing a relationship between the literature review and the current study.

The literature review is descriptive, as well as analytical in nature and attempts to identify general themes and positions in the area of researcher RDM. Furthermore, papers, reports and other scholarly documents cited are compared and contrasted. This researcher also attempted to identify gaps necessitating the need for further research, identify possible areas of controversy, and formulate questions for further research.

Planning the literature critique called for careful thought: it needed to be a comprehensive overview, but as RDM is an area which has been researched and implemented for close to a decade now, it was not possible to discuss all material ever written on the topic. Only pertinent material, assisting in conveying to the reader what knowledge and ideas have been established on the topic, as well as showing strengths and weakness of previous relevant research, were reviewed. In addition to this, literature had to be evaluated on applicability and relevance; a study portraying the RDM habits of visual arts researchers in a university setup, while ground-breaking, is not relevant to the context of this research.

The University of Leicester's online study guide states that in addition to the literature review presenting a research context, aspects such as the theoretical context, the methodological context, the practice context and the political context could also be addressed (University of Leicester, n.d.). This researcher realised that the methodological and practice context were issues in need of discussion within this chapter.

In essence, this chapter is a review of earlier work on the research field and pays due attention to contributions, as well as to any methodological problems and limitations involved. The chapter, consisting of a literature survey, is not simply an annotated list of

papers which have been read by the researcher. It is a literature critique covering a wide range of material pertinent to the research topic of the current study.

2.2 Structure of chapter

Sources consulted by this researcher and providing guidance on writing a literature review chapter, are mostly in agreement with regards to the contents of such a chapter. Pickard (2013:35) as well as RMIT (n.d.) propose the use of an introduction, a middle body consisting of methodology and discussion, followed by a conclusion. In addition, online university guides often give more detailed information on what to include in the chapter, and elaborate on the importance of dividing the works, as well as the importance of looking for differences and similarities (Concordia University, 2014). Tips on writing style, emphasizing the avoidance of plagiarism, and providing a literature review checklist (University of Melbourne, 2014), can also be found.

This researcher, after studying mentioned guides, implemented these basic as well as more elaborate principles when writing the literature review chapter. As a result of this, all headings and categories (mentioned below) contain an analysis of the works, an investigation into similarities and differences if applicable, as well as a summary/conclusion.

As far as the basic outline of the literature review goes, it was this researcher's opinion that the previous headings in this chapter constitute the introduction, while this current heading contains an extension of the introduction and the start of the methodology. The remainder of the chapter constitutes the middle body, made up of methodology, discussion and conclusion.

A more detailed outline of the headings and categories to be used in this chapter, is deemed necessary by this researcher. Chronologically, the remainder and body of this chapter will contain the following headings and contents:

The first section will be study-related, with this researcher focusing on issues NOT dealing with RDM behaviour, yet of relevance to this chapter and to a researcher thinking of embarking on an RDM study. As such, it will investigate available literature related to the study and the study tool itself. Numbered as section 2.3, it will contain the following topics:

- overview of RDM literature,
- RDM studies as global phenomenon,
- RDM studies: date range,
- RDM studies: aim/goal of study,
- RDM studies: publication types and sources,

- RDM studies: institutes/disciplines/groups surveyed,
- RDM studies: respondents,
- RDM studies: sample size,
- RDM studies: survey tool used, and
- RDM studies: study frameworks used.

The second section will be RDM-practice related, solely focusing on published RDM behaviours of researchers. In contrast with the section describe above, this section will look at literature findings related to RDM behaviour, such as specific RDM practices, needs and challenges. Numbered as section 2.4, it will contain the following topics:

- introduction,
- overview of RDM practices,
- state of RDM,
- data formats and data size,
- data storage and data backups,
- data preservation,
- data sharing,
- use of metadata,
- data management plans,
- RDM training undergone, and
- RDM services and recommendations.

The rationale behind sectioning the literature, as well as dividing literature findings into a multitude of topics, is that RDM, as well as the measuring of RDM, is a complex subject. Literature analysed in this chapter portray a myriad of survey tools, study frameworks, disciplines and sample sizes, to name but a few aspects revealing great variety among studies investigated. This heterogeneity in study attributes necessitated analysing the various components of literature studies, in an effort to ascertain the differences and similarities in said studies. A thorough investigation into study attributes enabled this researcher to establish cause and effect scenarios, supplied reasons for outliers or unexpected results, and placed this researcher in a more informed position when deciding on study tool, and its myriad of accompanying elements and traits, to be used.

These two main sections of the chapter will be followed by a section dealing with limitations of the literature study, and shortcomings of sources used, as well as a recapitulating conclusion.

2.3 RDM study-related literature

This section of the chapter focuses on, describes and analyses non-RDM-behavioural study components; in other words, aspects forming part of the study itself. RDM practices and behaviour are not included within this section; instead, this researcher categorised study traits in an effort to show differences and similarities within studies. RDM themes are also being identified. A further outcome of this is a better grasp of study limitations, concerns and problems, and recognising how these aspects could assist this researcher in her own studies, when deciding on aspects such as framework, tool, and sample.

2.3.1 Overview of literature

When collecting all available literature on RDM practices prevalent worldwide, this researcher could not fail but be overwhelmed by the voluminous variety of RDM study-related data available. Tracing literature has been an arduous yet prolific task, eased by open publishing and a willingness of authors to make research results readily available.

This researcher has traced and consulted close on 100 English language publications, each one reporting on the RDM behaviour of a certain group of researchers, when collecting RDM-study literature for purposes of this study. The majority of these publications have been published during the last decade, in other words, between 2005 and 2015. These studies, although possessing many common traits, appear in different publication formats, have been conducted in many countries and in several continents, and have investigated various subject disciplines and research institutes. Sample sizes of studies range from five to more than a thousand. Different study tools and study frameworks have been used. In short: when analysing available literature on RDM practices, this researcher was faced with a myriad of publications portraying distinct differences as well as unmistakable similarities. It was this researcher's intention to examine these aspects and ascertain the role played in study findings.

2.3.2 RDM studies: a global phenomenon

Although this researcher only gathered English-language RDM-related literature, it was clear that the use of an RDM study as research/investigative tool is a global phenomenon and it seemed as if it was done by many institutes starting out on a RDM programme.

When using the literature to examine the RDM studies done, it was established that the term 'survey' is the most popular word to describe the practice of investigating RDM behaviour in an institution. Although the practice remains the same, it was found that authors

investigating RDM also made use of the following terms and phrases: '…data audit' (Ekmekcioglu & Rice, 2009:2), '…landscape study of research data management' (Mossink & Bijsterbosch, 2013), '…scoping' (Tam, Fry & Probets, 2014:721), '…study of data curation practices' (Jahnke & Asher, 2012:3), '…needs assessment' (Doty *et al.,* 2013) and '…state of the art of the digital curation of research data' (Ball, 2010) when describing the process of gathering information pertaining to RDM practices and behaviours.

The largest part of English-language literature available on RDM practices and RDM studies originates from the USA and the United Kingdom. When looking at the total number of literature sources used by this researcher, studies investigating RDM practices in the USA account for approximately a third of the total, while UK-studies make up another third. Examples of recent studies in these two countries include Gu & Averkamp (2012), Doty *et al.* (2013), Fearon *et al*. (2013), and Beile (2014) reporting on researchers in the USA, with the Open Exeter Project Team (2012), Parsons, Grimshaw and Williamson (2013), and Tam, Fry & Probets (2014) investigating UK RDM behaviour.

Apart from the two above-mentioned countries, which, when combined make up more than 60% of RDM studies when studies are divided geographically, several countries have produced single-figure RDM investigations. Studies belonging to this group originate from Australia (Henty *et al.*, 2008; Bradbury and Borchert, 2010 as well as Gibson & Gross, 2013), Canada (Mowers, Humphrey & Perry, 2013), Germany (Osswald & Strathmann, 2012), Malaysia (Johare, 2014), the Netherlands (Dillo & Doorn, 2011), and Portugal (Ribeiro & Fernandes, 2011).

Besides studies investigating RDM practices within a single country, several investigations have also been done incorporating the RDM behaviours of researchers in a handful of countries, a whole continent, or in many countries worldwide. Carvalho *et al.* (2010) combined the findings of researchers in a few different countries, Kuipers & van der Hoeven (2009:3) investigated researchers across the European continent, while Enke *et al*. (2012) examined RDM practices in many countries, spread all over the world.

When narrowing RDM research conducted to the South African environment, it was discovered that South African RDM studies have been conducted since 2011, when the RDM practices of researchers at a South African university were reported on (Pienaar, 2011). Although it appears that more South African studies were completed, they were never published, or made available to either the South African or international library and information science communities. At the time of writing, this researcher was not been able to obtain access to either the datasets or full report of RDM studies done at any South African university or research centre, bar her own report with regards to RDM at the CSIR

(Patterton, 2014a), and an internal communication containing the findings of RDM at the University of Pretoria (Pienaar, n.d.). Sensitivity of results seems to be the common reason for the non-availability of findings; the deduction is made that these findings, as is the case with the survey done by this researcher into CSIR RDM behaviour, were subject to confidentiality issues and as such shared and distributed within the organisation only. In addition, this researcher was also not able to trace any publications reporting on RDM studies elsewhere in Africa.

Although satisfied with the number and geographic spread of RDM-practice literature collected, and being able to use it as proof of RDM-studies being conducted worldwide, this researcher was nevertheless curious about similar studies being conducted in non-English speaking countries. With studies in Malaysia and Portugal being published in English, it is hoped that other studies might follow suit. Conducting internet searches while being assisted by foreign-language scholars could also be considered, as could requesting relevant publications from foreign research institutes.

This section has attempted to illustrate the fact that RDM studies are currently a global phenomenon. This researcher has been successful in tracing publications pertaining to the findings of RDM studies conducted in nine separate countries, a study investigating several countries within the continent of Europe, and studies examining the RDM practices of many countries worldwide. The conclusion can be made that auditing the RDM practices of an organization is a global practice.

2.3.3 RDM studies: date range

Investigating RDM practices appears to be a relatively new concept, with the majority of studies collected by this researcher conducted in the last decade, meaning 2005 up to 2015. In more precise and absolute terms: RDM study dates range between 2001 (Reidpath & Allotey, 2001) and the present, with Kennan & Markauskaite (2015), Peset *et al.* (2015) and Whitmire, Boock & Sutton (2015) being examples of more recent publications.

It is important to note that a caveat needs to be attached when investigating the date ranges of RDM studies. Although, as mentioned earlier, RDM studies were conducted as early as 2001, it appears that these studies were mostly interested in the data sharing practices of researchers. RDM as a practice involving many separate and distinct categories and topics, does not seem to have been investigated during these years. The study of Reidpath & Allotey (2001:125), examining the data sharing practices of medical researchers, is a case in point. A 2002 study (Campbell *et al.,* 2002) investigating the prevalence of data withholding in genetics, is another example of the leaning taken by studies at the start of the millennium.

Studies focussing solely on RDM, as opposed to the narrower investigations into data sharing only, only appear a few years after the first data sharing study. This researcher has not been able to trace reliable sources indicating when more elaborate RDM studies commenced, and has resorted to doing a bibliographic study in an effort to establish when the first detailed RDM study was published. Although earlier studies investigating research data issues did take place, examples being the 1976 Canadian report investigating access to research data (in Humphrey, 2012), or the 1996 report on a data policy and research data access (*ibid.*), this researcher is of the opinion that a 2001 Canadian needs assessment (Social Sciences and Humanities Research Council of Canada, 2001), and a 2003 study (Lord & Macdonald, 2004), prepared for the JISC Committee for the Support of Research, and reporting on a data audit that had been done in an effort to establish the data curation requirements, are a few of the first examples of more elaborate RDM studies. Another example of an earlier RDM study would be an exploratory investigation into the data practices of habitat ecologists (Borgman, Wallis & Enyedy, 2006). This researcher concedes that earlier studies than the ones mentioned here, might have been conducted, but has not as yet come across these when gathering published literature and study results.

Currently, new RDM studies are being added to the array of RDM-studies being published, on an almost quarterly basis. This researcher actually realised that many RDM studies, both locally and abroad, would be published during the writing of this dissertation, as well as after completion of the study. This study therefore possesses a double-edged sword: it attempts to make a valuable scientific contribution during a topic's prosperous period, while potentially useful similar studies are also being published, and many will be too late for inclusion into literature critique. Even so, participating in the thriving period of RDM studies is preferable to publishing later. Studies appear to often be used as a stepping stone to establishing RDM regimes at an institution and cannot be delayed for the benefit of a more complete or representative literature survey.

2.3.4 RDM studies: aim/goal of study

An examination of the literature available on RDM practices of researchers revealed that studies could be categorised into one of three groups:

- studies where only **one aspect** of RDM was examined. Examples of these are the studies of Reidpath & Allotey (2001:125), Campbell *et al., (*2002), and Blumenthal *et al., (*2006:137), which tended to focus solely on sharing of research data. These studies into sharing practices were conducted before the surge in more detailed RDM

studies, where RDM is seen as a practice involving many separate and distinct categories and topics, started.

- studies focussing **solely on RDM practices**. The majority of literature analysed by this researcher when writing this chapter, involved a thorough study of many aspects of RDM. Unlike earlier studies focusing mostly on data sharing, no specific sub-activity or practice forming part of RDM is given more prominence than other practices. Studies are concerned with investigating all or most activities forming part of RDM. Examples of detailed studies, where ascertaining the RDM practices of a group of scientists is the main aim, include the data management survey at the University of Iowa (Gu & Averkamp, 2012), the data management survey at the University of Hertfordshire, UK (Nassiri & Worthington, 2012), and the data management survey at the University of Nottingham (Parsons, Grimshaw & Williamson, 2013). This category of studies has the distinguishing feature, and added advantage, of always posting results and findings in detailed and categorised form. Proof of this point is the fact that the study by Gu and Averkamp's study (2012:4) reported on 13 separate RDM aspects, while Parsons, Grimshaw & Willaimson (2013:1) separated their findings into 15 distinct RDM categories.
- studies where the RDM investigation is a **sub-division of the study**; the aim of the study was to establish the state of e-research practices and skills of researchers. Examples of such e-research-focussed studies include Bradbury & Borchert (2010) investigating the practices of researchers at the Queensland University of Technology (QUT), and the Research Information Network (RIN) case study providing a detailed analysis of how humanities researchers discover, use, create and manage their information resources (Research Information Network, 2011). This type of study, almost hidden within a bigger study, was often difficult to trace, and many times stumbled upon by pure chance.

To summarise, one can say that RDM studies are conducted when researchers are interested in RDM practices only, or when RDM forms part of a bigger set of behaviours, such as e-research practices. Alternatively, only one RDM aspect, most often researchers' data sharing practices, would be the focus of a study.

In addition to RDM studies having different aims (as mentioned above), it is also possible to distinguish between RDM studies being conducted for the purpose of LIS academic research, or those carried out to inform institutional policy or service development. In many instances, the two may overlap, as is the case with the current study. With regards to RDM studies conducted to gain information informing RDM policies and services, one can expect

the response analyses to be less rigorous than the academic RDM studies, as these studies would not be aiming to make claims about the population concerned. With policy-related studies, the objective is to demonstrate that a reasonable cross-section of the target population have supplied information on relevant topics.

2.3.5 RDM studies: publication types and sources

Of particular interest to this researcher when finding, collecting and analysing literature related to RDM study findings, was the use of different publication types by authors, when making their publications available. A standout feature of this discovery is the fact that popular scholarly online search engines and portals have proven to be insufficient when tracing related literature, as a big part of RDM study publications are not found inside this realm. Grey literature would seem to be a common realm of RDM studies. The implications of many RDM studies occupying the grey literature domain can be said to not only result in RDM studies being difficult or even impossible to trace, but that the added value of such information/data sharing, for both author and reader, is lost.

Additional search techniques such as studying existing bibliographies, or using mainstream search engines e.g. Google, as opposed to Google Scholar, proved to be fruitful when gathering literature applicable to this study. Formal commercial resources, including ScienceDirect and Dialog, did not deliver additional results.

RDM study literature was found to end up, or be published in any of a variety of formats. A very prevalent way of portraying study results, turned out to be in the form of a report published on the university library website of the university being surveyed, or as a report deposited in the institutional repository/digital conservancy of the organisation being surveyed. Examples of such reports include the study into RDM practices at the University of Minnesota (Marcus *et al.,* 2007:1), the QUT (Bradbury & Borchert, 2010), Purdue University (Carlson *et al.,* 2011), and among University of Central Florida researchers (Beile, 2014). These two publication types just mentioned: a report published on university library/university RDM websites, or a report published in the university repository, were the most common literature sources and account for about 45% of literature sources traced in this study. Included in this group would be study results being published on a tertiary institute's RDM-specific website, as was used for the University of Nottingham study (Parsons, Grimshaw & Williamson, 2013). Discussing study results on an RDM blog, and providing study links on an RDM blog, as was done by authors involved in Oxford University studies (Patrick, 2012 also Wilson, 2013).

This researcher also traced several reports which had been published as a project report on a website created specifically for a project; the project often being strongly connected to RDM. Reports were also found to be published on the funder's website, or website/repository of a larger body, such as a study published on the website of the UK Data Information Specialists Committee (Gibbs, 2009), or the report on the Edinburgh Data Audit being published on the JISC repository (Ekmekcioglu & Rice, 2009).

A popular publication type, nearly as prevalent as study reports on RDM, is the journal article. This category made up about 20% of publication types traced. Journals used by researchers could be divided into two groups: library and information science journals, or discipline-specific (discipline of study respondents) journals. Examples of the former are *Ariadne* (Westra, 2010), *Science & Technology Libraries* (Peters & Dryden, 2011), *Journal of Web Librarianship* (Bardyn, Resnick & Camina, 2012), and *International Journal of Digital Curation* (Halbert, 2013), while the latter category includes *PLoS Medicine* (Piwowar *et al.,* 2008), *PLoSONE* (Tenopir *et al.,* 2011), and *Journal of Agricultural & Food Information* (Diekmann, 2012). Sometimes, an overlapping journal, which contains aspects pertaining to the respondents' discipline as well as the researcher's profession, is used, as is the case with *Journal of Professional Issues in Engineering Education and Practice* (Johnston & Jeffryes, 2014:1).

Another publication category used was conference proceedings, meaning that a published conference paper was the final deliverable. Papers presented at the *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (Wynholds *et al.,* 2011), the *2012 IFLA Conference*, as well as the *2013 Association of College & Research Libraries Conference* (Doucette & Fyfe, 2013) portray the use of this practice.

This researcher concedes that many authors, reporting on RDM findings, might indeed use more than one publication type/format/source when publishing his work. This practice is common and expected in the case of publications being stored in an institutional repository: it would often have been published somewhere else, too. In a similar vein, findings might have been augmented and as a result of this, presented at more than one conference, and shared on other platforms too. The University of Pretoria's study results being shared on Slideshare (Pienaar: 2011), and the University of Pretoria's institutional repository (Pienaar, 2010) as well as at several conferences (Pienaar, 2015:10), exemplify this theme. The current flourish of RDM studies, as well as the interest in study findings, makes this a common occurrence within this subject area.

2.3.6 RDM studies: institutes/disciplines/groups surveyed

As was expected, the majority of studies investigated by this researcher when analysing RDM studies, were interested in discovering the state of RDM within a specific institute. These institutes could be universities, such as the studies into the RDM behaviours at the University of Oxford in the UK (Martinez-Uribe, 2008), Cornell University, USA (Steinhart *et al.,* 2008) and Emory University, USA (Akers & Doty, 2013). Similar tertiary institutes, although semantically different in title, include the QUT (Bradbury & Borchert, 2010), Georgia Tech, USA (Parham, Bodnar & Fuchs, 2012), and the London School of Economics and Political Science (Raggett, 2012a). Studies mentioned here are classed together as they were all interested in investigating the RDM behaviour of researchers at a specific single tertiary institute.

For the purposes of this heading's categorisation, it is important to further distinguish between studies where more than one tertiary institute were involved, and the previous group. Studies investigating the RDM practices of researchers at the universities of Queensland, Melbourne as well as the QUT (Henty *et al.,* 2008:1), the universities of Cambridge and Glasgow (Freiman *et al.*, 2010:2), and the universities of Edinburgh, Oxfordshire, Bath and Bristol (Lyon *et al.*, 2010), belong in this group.

Under this heading, studies can further be subdivided into those that involved as many as possible faculties, disciplines, or subjects, and those that were strictly discipline-specific. Examples of studies including all faculties, or as many faculties and departments as possible, would be the study examining the RDM behaviours of university researchers in Canada, all subject areas included (Mowers, Humphrey & Perry, 2013), and the University of Central Florida study (Beile, 2014). In contrast to these all-encompassing studies, studies that limited the discipline scope to a single discipline only, form a large part of RDM literature. Examples of these include a study limited to agricultural scientists (Diekmann, 2012:14), a study looking at social sciences only (Jahnke & Asher, 2012:3), or a study involving geography researchers only (Tam, Fry & Probets, 2014:721). Although a reason for subject limitation is often not supplied, Westra (2010), when studying the RDM practices of researchers in natural sciences only, said that this action provided this researcher with a more manageable scope.

When looking at studies incorporating more than one organisation, it was seen that not all studies limit their investigation to a similar type of institution. Put another way: not all studies including more than one institute, limit affiliation to universities only, or research centres only. These studies would audit RDM practices of researchers, scientists of graduate students, irrespective of affiliation. A few examples of these would be an investigation into the RDM

practices of humanities researchers from five countries (RIN, 2011), a study of Canadian graduate students (Doucette & Fyfe, 2013), and a study investigating stakeholders worldwide (Sayogo & Pardo, 2013:S23). A study by Diekmann (2012), intent on establishing the RDM trends among agricultural researchers anywhere in the USA, is another study belonging to this non-affiliated group.

In summary, RDM studies investigated by this researcher revealed a wide range of groups investigated. Single faculties at a university, single universities involving many faculties, researchers in a certain subject discipline worldwide, different types of tertiary institutes or research centres, be it continent-specific, collaboration-specific or globally : the permutations were varied and numerous, and added to the plethora of RDM practices discovered during this literature survey.

2.3.7 RDM studies: respondents

As could be expected (since studies analysed here were investigating research data), the overwhelming majority of studies analysed by this researcher exclusively investigated the RDM practices of researchers and scientists. Exceptions to this rule were found: support staff were surveyed during an RDM study at the QUT (Bradbury & Borchert, 2010), an RDM study at the University of Oregon incorporated the responses of a maps librarian (Westra, 2010), an RDM study at the universities of Glasgow and Cambridge involved computing officers (Ward *et al.,* 2011), and administrative staff as well as librarians formed part of an RDM study sent out mainly using listserves connected to the American Library Association (Halbert, 2013:116). In addition to these studies, Mossink & Bijsterbosch (2013) included research funders and publishers when examining research practices in mainly Europe, respondents of the RDM study at the University of Central Florida included administrative staff (Beile, 2014), and a preliminary investigation into RDM practices in Malaysia included administrative personnel as well as IT staff (Johare, 2014).

It should be noted that the above-mentioned non-researcher groups were involved in RDM studies in addition to the respondent component consisting of researchers, also being investigated. It would seem as if most RDM studies, even those also involving non-researchers, included researchers by default. The crux of what was being investigated (i.e. the management of data generated during research) determined this obvious study trait. Studies traced by this researcher have shown that the inclusion, percentage-wise, of non-research staff varied between studies. Examples of this variance include Halbert's RDM study (2013:116), where 76% of respondents were classified as librarians, and Beile's study (2014), where the administrative component formed 10% of the respondents. It needs to be

mentioned that this researcher has focussed her attention on RDM studies where researcher behaviour, rather than the behaviour of LIS personnel or other support services, were investigated. The reason behind this decision is the fact that the role and function of US and European support services, specifically with regards to RDM, is currently too different from the South African research support scenario.

Researchers investigating RDM themes are often able to draw respondents from a range and variety of positions, responsibility and level of experience. This feature is especially prevalent with studies investigating RDM behaviours at tertiary institutes. Although too numerous to list them all, a study at the University of California Los Angeles (Borgman, Wallis & Enyedy, 2006) involving scientists, research partners, graduate students, post-doctoral fellows and research staff, a study investigating RDM themes at three Australian tertiary institutions involving academic staff, postgraduate students, and even emeritus (Henty *et al.,* 2008:2), the RDM investigation at the University of Exeter (Open Exeter Project Team, 2012:3) involving postgraduate students and academic staff, a study at the London School of Economics and Political Science (LSE) involving research staff and research students (Raggett, 2012a), and a study at the University of Nottingham (Parsons, Grimshaw & Williamson, 2013:6) involving career researchers and postgraduate researchers, are examples of studies incorporating research-involved respondents from different university positions, and differing levels of research experience.

Of particular relevance to this study are RDM studies showing the differences in RDM behaviour between different faculty ranks or level of research experience. Although RDM differences, according to research level, or experience level, is often not indicated in results, this researcher aimed at establishing similarities and differences between emerging CSIR researchers and emerging researchers, as well as experienced researchers, respectively as well as combined.

2.3.8 RDM studies: sample size

Considerable variance between studies could be seen when sample size was examined. This seems to be a natural outcome of target group and aim of study, as well as survey tool used: a study investigating the RDM practices of researchers, worldwide, by means of an online questionnaire or bibliometric methods, would invariably involve and demand a bigger sample size than a study interested in a comprehensive investigation of a specific unit at a research centre, using in-depth interviews to obtain information.

Examples of studies having large sample sizes, include the study of Campbell *et al.* having a sample of 1849 (2002:473), Kuipers & van der Hoeven having a sample of 1840 (2009:4),

and Tenopir, making use of 1329 respondents (2011:1). It should be mentioned that all of these large-sample studies, with the exception of the Kuipers study, investigated sharing practices only, and had no interest in other RDM behaviours.

In sharp contrast, several studies making use of in-depth RDM surveys, being interested in obtaining comprehensive data from a smaller group, and usually, but not always, limited to a specific unit, specific institute, or a specific discipline. The in-depth case studies done at the universities of Edinburgh, Oxfordshire, Bath and Bristol (Lyon *et al.,* 2010:6), the study investigating geography departments at universities in the United Kingdom (Tam, Fry & Probets, 2014:721), and a study into the RDM practices at Malaysian research institutes (Johare, 2014), all featured sample sizes of ten or fewer researchers.

In short, sample sizes investigated by this researcher during the course of the literature survey, revealed it to be a characteristic prone to variance. Sample sizes ranged from single figures, to a study consisting of many thousands of subjects.

2.3.9 RDM studies: survey tool used

Upon analysing survey tools used in previous RDM studies, it was found that a range of tools were used across the spectrum of studies looked at. Online questionnaires/web surveys, interviews (one on one, as well as group), supplementary desk research/bibliographic studies, questionnaires distributed via mail and via email, focus groups, ethnographic observations/ attending team meetings, as well as self-administered tick-lists handed to researchers to be used as they were working, were all methods implemented when investigating RDM practices.

Researchers could make use of any number of survey instruments; this researcher has found the use of more than one survey instrument to be just as prevalent as the use of a single tool only. Examples of a single instrument survey include the investigation into RDM practices, using an online survey, at three Australian universities (Henty *et al.,* 2008), Diekmann's study using interviews only (2012:17), and Campbell *et al.* (2002:473) making use of mailed surveys when investigating the data sharing practices of life science researchers.

Conversely, several studies made use of more than one survey instrument. The survey into RDM practices at the QUT (Bradbury & Borchert, 2010), using an online survey and following it up with the implementation of focus groups, the survey on RDM practices, done throughout the USA (respondents would be members of American Library Association list-

serves), making use of an online survey as well as focus groups (Halbert, 2013:116), and the survey at the University of Bath using an online survey, followed by interviews (Pink *et al.,* 2013), are three examples of studies using one data gathering tool, and supplementing findings by using an additional data gathering instrument.

This researcher aimed at establishing which study instrument was most commonly used. As it turned out, there are two survey tools being used far more frequently than other instruments; the two tools are the online survey and the personal interview/use of an interview schedule. This researcher found these two tools to be used equally often. When available RDM studies were gathered, close to 100 studies containing information about the survey instrument used, were collected by the researcher. Upon analysis of these studies, including those that investigated sharing practices only, it was discovered that an interview was the data gathering method in roughly half of all cases, while online surveys were used in the other half of instances. Looking at available literature, these two data gathering tools (the interview and the online survey), outnumber all other data collecting tools used when RDM surveys are being conducted.

A quick perusal of specific online survey tools used shows the plethora of software options available. Bristol Online Surveys (Alexogiannopoulos, McKenney & Pickton, 2010:35), Google Forms (Raggett, 2012b:5), and Qualtrics (Doty *et al.,* 2013:6), are just three of the many online survey tools used in previous RDM surveys. It might prove beneficial to examine the stated benefits and shortcomings of these applications in the methodology chapter of this study; a case in point being QuestionMark Perception (Gibbs, 2009:3) which was used during an RDM study and then found not fit for this purpose, and would not be reused by the same researcher. On the other hand, Google Forms was chosen as a suitable program (Raggett, 2012b:5) as it provided live statistics as well as results that could be readily downloaded.

Interviews could either be done on a one-on-one basis, or by doing group interviews. Examples of personal interviews include a study into RDM practices at the University of Edinburgh (Ekmekcioglu & Rice, 2009:7) as well as a study at the University of Northampton (Alexogiannopoulos, McKenney & Pickton, 2010:11); group interviews, involving focus groups, were used at QUT (Bradbury & Borchert, 2010) as well as at the UCLA (Bardyn, Resnick & Camina, 2012:274). A few studies made use of both types of interviews when gathering data: studies conducted at Purdue University (Witt, 2009) and a survey investigating RDM practices at two UK universities (Ward *et al.,* 2011) fall into this group. The RDM studies analysed by this researcher did not reveal a predilection for one over the

other. Furthermore, both interview types could be used as an only survey tool, or were sometimes used in conjunction with other tools.

Mailed questionnaires (Campbell *et al.,* 2002:473), ethnographic observations (Borgman, Wallis & Enyedy, 2006), self-administered probes to be completed by researchers (Research Information Network, 2009:11), and bibliometric studies (Piwowar, 2011:1 also Williams, 2012:312), although detected by this researcher, tend to be rarely used. The usage for each of these methods, with the exception of bibliographic studies, was found to be lower than 3% in all RDM studies analysed. Bibliometric surveys were slightly more common, used in 7% of studies analysed. The caveat in this instance would be the subject area investigated: bibliographic methods used as survey tool tend to be used only half as much when investigating RDM, as opposed to its usage when investigating data sharing practices. As a data gathering tool it would seem to be particularly useful when analysing a single behavioural aspect and wanting to make use of a large number of publications: Piwowar (2011:1) used a total of 11603 articles when analysing the prevalence of data sharing, worldwide.

In summary: examining the method of data gathering, or survey tool used by researchers investigating RDM practices, has revealed the use of online surveys, and personal interviews (one-on-one as well as group) to be the tools most frequently used. These two tools are used equally often, and were seen to be used as only survey tool in a study, or in conjunction with another tool.

2.3.10 RDM studies: survey framework used

According to Jones, Ross & Ruusalepp (2009), the Data Audit Framework (DAF), previously known as the Data Asset Framework, is

> '*a framework developed by the JISC-funded DAFD project to identify data assets held within higher and further educational institutions and to explore how they are managed. The framework is structured around audit at departmental or unit level with results being amassed to obtain an institutional or national perspective*'.

A standout feature, when investigating frameworks used in RDM studies, is that DAF is the most popular framework in post-2008 studies. It outranks all other frameworks and is used as a foundation on which to base the survey instrument in close to 50% of studies examined by this researcher.

The first recorded instance of DAF being used as framework at research institutions, was during the implementation phase of the framework, when projects at the University of Edinburgh, Imperial College, King's College and University College, London, were funded to test the toolkit and to promote its uptake.

Following on from implementation projects, the DAF was soon to be used at other universities based in the United Kingdom. Studies examining the RDM behaviours at the University of Oxford (Martinez-Uribe, 2008), the University of Northampton (Alexogiannopoulos, McKenney & Pickton, 2010), and the London School of Hygiene and Tropical Medicine (Knight, 2013:5), the University of Hertfordshire (Nassiri & Worthington, 2012), the University of Exeter (Open Exeter Project Team, 2012:3), and the University of Nottingham (Parsons, Grimshaw & Williamson, 2013:3), are examples of the framework gaining immediate popularity after its publication.

Outside of the United Kingdom, the DAF was seen to be a popular framework as well. In South Africa, Pienaar (2011) as well as Patterton (2014a) made use of a survey tool based on the DAF-based tool used by Martinez-Uribe (2008). In the USA, Westra (2010) based his University of Oregon survey framework primarily on the DAF, a study at the University of Houston based the survey framework on the DAF, (Peters & Dryden, 2011:391), and a study investigating RDM at Georgia Tech (Parham, Bodnar & Fuchs, 2012:11) used the DAF as well.

Despite its popularity, usefulness and accessibility, several high-profile RDM studies adhered to their own framework, or did not mention using the DAF as framework. A study into data curation behaviours at California Polytechnic State University (Scaramozzino, Ramirez & McGaughey, 2012:353), RDM themes at Canadian Universities (Mowers, Humphrey & Perry, 2013), and RDM practices of geography researchers (Tam, Fry & Probets, 2014:721) revealed the use of survey frameworks not related to the DAF. In a similar vein, the study into RDM practices throughout Europe (Mossink & Bijsterbosch, 2013:5) describes the structure of their questionnaire as 'based on the target key areas identified by the JISC Managing Research Data Programme', but do not explicitly state making use of the DAF as reference framework.

In conclusion: researchers are able to make use of various frameworks when designing the structure of their surveys. This researcher is of the opinion that its widespread use is due to its proven use by leading tertiary institutes, aggressive marketing of the DAF, its readily and online availability, lack of costs involved, ease of use, added advantage of comparing with other studies using the DAF, plus the range of survey samples to choose from. At least one

author has stated that the DAF as a survey tool contains the most comprehensive set of questions (Westra, 2010), when compared with other survey tools. The use of the DAF in a previous RDM study by this researcher (Patterton, 2014a) will serve as a stepping stone for its use in the intended study, albeit in an adapted format.

2.4 RDM-related literature

2.4.1 Introduction

The previous section of the literature analysis focused on aspects of the literature that dealt with study-related aspects: topics such as survey framework, sample size, date range and institutes surveyed were analysed and discussed. This section will deal will topics related to RDM practices only: researcher behaviour as revealed by RDM studies were examined and reviewed.

2.4.2 Overview

This part of the literature critique will be perusing the RDM-related actions of researchers and scientists, working at institutes of higher education or research institutes, as uncovered by RDM studies and published in mainly articles, research reports and project reports. This researcher attempted to establish patterns and themes in data-related behaviour, take note of outlying results and serious discrepancies in findings, and will seek to explain discrepancies and outliers.

Variance, seen in a very broad context, would appear to exist when looking at subjects studied, as well as survey frameworks used. Despite this, and because of the heterogeneous characteristics of studies, when examining studies on a more precise trait level (disciplines, institution, research culture, funding, funders requirements, to name but a few identifying traits), this researcher expected the establishing of general behavioural themes to not be an easy task. In addition, generalising the literature survey findings to researchers, research groups, disciplines or centres was thought by this researcher to be unwise and scientifically unsound. Academic researchers and scientists tend to be made up of a diverse, multifarious group of individuals: possessing different levels of experience, working in different setups, and belonging to different disciplines. Scientists within a single institute might belong to different age groups, have different experience levels, have different funders, have a different research culture, work under different principal investigators, deal with different data formats, different data software, different data confidentiality types, and have different budgets. This makes it very hard to generalise about their behaviour, and to use RDM findings of one study to make predictions about the RDM practices of another

group. A RIN-inspired study, for example, detected and commented on diverse behaviour not only between different disciplines, but also between teams in the same discipline, as well as between members of the same research team (University of Edinburgh, 2009:61).

Notwithstanding the dangers of generalisation forthcoming from this literature chapter, this researcher aimed at establishing some identifiable patterns. These patterns as well as findings on the periphery, were noted, and where possible or required, an explanation was put forward.

RDM findings are discussed under the following headings:

- state of RDM in organisation/institute,
- data formats/data size,
- data storage/data backups,
- data preservation,
- data sharing,
- use of metadata,
- data management plans,
- data training undergone,
- services required/recommendations, and
- limitations.

2.4.3 State of RDM in organisation/institute

This section of the chapter tried to establish the current state of the art of RDM at the respondent's institution: is an RDM policy in place, have procedures been drawn up, and does the institute provide assistance in the form of training, guidelines and funding?

What needs to be mentioned first and foremost is that this topic was only investigated by a few studies. Furthermore, studies were mostly conducted before institutional RDM was implemented at examined institutions; low levels of institutional RDM can be expected to be revealed. Also, it was not an easy task deciding whether the relatively small number of studies dealing with the issue of institutional involvement and guidance could be used to extrapolate to the situation befalling other researchers. As such, this researcher did not presume that findings mentioned here are generalizable, typical or to be expected.

Studies investigating the issue of institutional maturity reported some variance with regards to this institutional feature, with the mention of '…formalised policies largely absent' at the University of Oregon (Westra, 2010), and the 45% of respondents in a study involving university researchers USA-wide, stating having designated scientific data curation and

support units on their campus (Soehner, Steeves & Ward, 2010:14), being the two extremes displayed. Even within the same institution, as shown at the University of Oxford, variable degrees of RDM maturity can be found (Martinez-Uribe, 2008:6). It is interesting and important to note that no RDM study revealed the level of institutional RDM involvement as complete, fully implemented or highly structured.

RDM policies and procedures were generally found to be not implemented, while RDM as part of research was not supported, and not funded. The findings of a study into RDM behaviour of Malaysian researchers, showed that only 41% of institutes have preservation policies, and that no Malaysian researchers are expected to adhere to a collection policy (Johare, 2014). While 38% of scientists surveyed in a worldwide study (Tenopir, 2011:5) admitted to having a formal process in place, 47% of researchers in the same study revealed that there is no formal process, either in their projects or at their institution, for storing data beyond the life of the project. The same study also showed that 48% of institutions do not provide funds to support RDM during the life of the project, 59% provide no RDM training, and 59% supply no funding to support RDM beyond the life of the project. Likewise, while 45% of respondents, from various USA institutes, indicated some degree of institutional RDM involvement, the majority stated that their institution had no designated units to provide data curation and support of RDM (Soehner, Steeves & Ward, 2010:14). These findings are supported by researchers at the University of Bath, where it was shown that their RDM practices are guided by intuition rather than good practice (Jones K, 2011:1), and a study in the USA, involving mainly librarians, revealing that although 87% of respondents agreed that an institution-wide RDM policy is valuable, in practice 72% of respondents mentioned that there is no RDM policy in place at their institute (Halbert, 2013:117).

Additional evidence of low institutional involvement with RDM can be found when examining findings of studies involving various USA institutions (Keralis *et al.,* 2012:23); 72% of respondents have no institutional RDM policy, while a further 19% revealed to not be aware of the existence of any such policy. At the University of Southampton (Takeda *et al.,* 2010:4), RDM guidance and advice was shown to be limited, research knowledge of capability and resources was lacking, and researchers tend to resort to own best RDM efforts in many cases.

This researcher was of the opinion that two interesting findings, emanating from the studies, needed to be mentioned in this section. The first is that there is a possibility that researchers, when embarking on RDM, might tend more to discipline-specific developments than to developments arranged or implemented by their institute (Patrick, 2012). The second interesting finding is the comment made by Westra (2010), when investigating RDM

practices at the University of Oregon, that the survey actually prompted basic record keeping and file management practices after the study. The implication of these two observations is that RDM, or in this context specifically, institutional or organisational RDM support, would benefit from being more visible, marketing itself, and providing not only policies and RDM guidelines, but explaining the rationale and benefits of good RDM practice.

2.4.4 Data formats, file formats, and data size

Although data formats and data size form part of RDM, they are seen by this author to be more of an RDM 'feature' than an RDM 'activity'. Despite differentiating between RDM 'activities' and RDM 'features', this researcher is of the opinion that gathering information about a researcher's data formats and data volumes creates context for RDM activities such as storage, backups, preservation and archiving. It also provides a setting for software requirements, infrastructure needs, and other technological support needed. In other words: the ability to collect/generate data, access data, use data, and store data is highly dependent on infrastructure available, as well as software at a researcher's disposal. As such, it is vital to investigate the data formats, file formats and data size in order to inform institutional infrastructure and software needs and limitations. Acquiring information about data formats and data size is vital as this information would inform the ability or capacity to collect, access, use and store research data.

Study findings reveal a wide range of formats and dataset volumes to be prevalent in a researcher's data arsenal. Apart from these two variables, diversity in data types were also found when investigating values, or characteristics such as data collection method/data creation technique (Martinez-Uribe, 2008:8; Ekmekcioglu & Rice, 2009:18-20, also Pienaar, 2011:15), technologies used to create data (Raggett, 2012a:2) data type as pertaining to digital/non digital data (Henty *et al.,* 2008:3; Parsons, Grimshaw & Williamson, 2013:9; Johare, 2014), software tools (Henty *et al.,* 2008:6 also Gibson & Gross, 2013:12), proprietary and open source formats (Martinez-Uribe, 2008:8 also Gibson & Gross, 2013:12) as well as the differences that exist between static and dynamic datasets (Lord & Macdonald, 2003:34; Witt, 2009). An important issue emanating from these findings would be the fact that researchers often do not recognise the implications of software choices and later access to data (Henty *et al.,* 2008:41).

Henty *et al.* (2008:4-5), Martinez-Uribe (2008:8), Ekmekcioglu & Rice (2009: 18, 20), Gu & Averkamp (2012:1), Nassiri & Worthington (2012), Gibson & Gross (2013:12), Averkamp, Gu & Rogers (2014:7), Beile (2014), and Sewerin *et al.* (2015:5) all found their survey respondents to be using a wide range of data formats. Furthermore, studies have also

revealed a variety of data types used across institutions investigated, as seen in the findings of Averkamp, Gu & Rogers (2014:7), Buys & Shaw (2015:9), Kennan & Markauskaite (2015:69) as well as Whitmore, Boock & Sutton (2015:382). This researcher has also found researchers at a South African research council to be making use of a wide range of data formats (Patterton, 2014a). A numerically-precise portrayal of the diversity data formats is mentioned in the findings of a study involving UK researchers: 25 different data formats were mentioned by 34 survey respondents (Lord & Macdonald, 2003:32). Similarly, 150 different file formats were mentioned by 148 survey respondents at the California University Boulder (Rankin *et al.,* 2012:6). It is important to note that this research data multiformity was found in recent as well as older studies, in different subject disciplines, in different continents, and at universities as well as research centres.

Besides the diversity in data formats displayed, the use of different file types within the same data format was also revealed. An example of this would be Beile's study, showing that JPEG as well as TIF could be used as file types for images, PDF as well as DOC could be used for text document file types, and MPG as well as MOV could be used for digital video data (Beile, 2014). This finding is also supported by looking at the RDM behaviour of University of Northampton researchers: a great variety in file types was found to exist for database data, audio data and video data (Alexogiannopoulos, McKenney & Pickton, 2010:19-21). This trend can be said to have serious implications for preservation planning.

Apart from determining whether data formats used were uniform or diverse, several studies attempted to establish which data types were more prevalent within their group of respondents. Even though different institutions and subject disciplines were involved, studies investigating data type prevalence showed that there are a handful of data types which tend to be used more often than others. These most common types were found to be textual data, numerical data, and tabulated data. In some studies, textual data was referred to as documents, while tabulated could be seen in the same vein as spread sheets. Databases were often found to be a 'top three' data type as well. Studies highlighting the prevalence of these data types include a study involving three Australian universities showing spread sheets and database data to be used by two thirds of respondents (Henty *et al.,* 2008:4), a worldwide study involving 1840 respondents (Kuipers & van der Hoeven, 2009:30), where text documents/office documents reigned supreme, a study at the University College London (Polydoratou, 2009:299) indicating numerical, textual, and database data as the most widely used types, a University of Nottingham study (Alexogiannopoulos, McKenney & Pickton, 2010:18-21) finding that most researchers were using documents and spread sheets as data types, a study involving Dutch researchers where spread sheets were found to be most common, and used by 69% of respondents (Dillo & Doorn, 2011:14), a study at the

University of Bath, where the most popular types were text files and spread sheets (Jones K, 2011), a worldwide study of biodiversity researchers showing the prevalence of word documents (Enke *et al.,* 2012:29), a survey at the University of Iowa (Gu & Averkamp, 2012:7) revealing spread sheets, database and textual data to be the most common types, 67% of Georgia Tech researchers using data in text format (Parham, Bodnar & Fuchs, 2012:12), a study at the University of Oxford showing that the data in numerical format is used by 65% of researchers (Patrick, 2012), and a University of Central Florida study showing spread sheets and tabulated data to be most commonly used (Beile, 2014). Similarly, Buys & Shaw's study (2015:9) reveal spreadsheets, structured data, text and images to be the most common types, Sewerin *et al.* (2015:5) mention numerical data and text data, while Whitmire, Boock & Sutton (2015:388) mention spreadsheets and digital text. This researcher's study involving researchers at the CSIR support the above findings: the two most common data types used were revealed to be textual data and spread sheets (Patterton, 2014a).

Besides often showing the diversity of data types within the same study, or within the same group, studies have also highlighted the diversity of data types used by individual researchers. A study involving 1329 scientists worldwide found that many researchers tend to use more than one data type (Tenopir *et al.,* 2011:7), University of Iowa researchers said that they use two or more types (Gu & Averkamp, 2012:1), while all 148 respondents taking part in the California University Boulder study attested to be using more than four different data types (Rankin *et al.,* 2012:6). To summarise: data type diversity is a realistic component of the data arsenal of each individual scientist. This diversity in data types can also be extrapolated to individual disciplines, a case in point being the variety of data formats, software and dataset sizes used by agricultural scientists throughout the USA (Diekmann, 2012:21-23) as well as the heterogeneity of biodiversity researchers' data formats, worldwide (Enke *et al.,* 2012:28-29).

Listing all popular data types was an exhausting task, but this researcher feels obliged to mention the following types as well, as they were shown to be commonly-used data types in many studies examined: digital images (Martinez-Uribe, 2008:8, 23; Gibson & Gross, 2013:12 also Patterton, 2014a), digital video (Martinez-Uribe, 2008:8; Alexogiannopoulos, McKenney & Pickton, 2010:19; Patrick, 2012; Gibson & Gross, 2013:12, as well as Patterton, 2014a), digital audio (Martinez-Uribe, 2008:8; Alexogiannopoulos, McKenney & Pickton, 2010:19; Patrick, 2012; Gibson & Gross, 2013:12, as well as Patterton, 2014a), scanned documents (Parham, Bodnar & Fuchs, 2012:12), and geospatial data (Patrick, 2012, Patterton, 2014a).

Several studies confirm that intra-survey data set size variance among respondents (Akers & Doty, 2013:8; Van Tuyl & Michalek, 2015:3; Whitmire, Boock & Sutton, 2015:387). Furthermore, Akers & Doty (2013:8), Buys & Shaw (2015:13) as well as Whitmire, Boock & Sutton (2015:387) have found that some researchers are unaware of how much data they have stored.

Disciplinary differences with regards to data formats and data set sizes were shown to exist: a study at the University of Oxford, found that humanities scientists would usually only work with a few megabytes of data, while activities such as medical science simulations would run into several terabytes of data (Martinez-Uribe, 2008:8). This finding is supported by a study at Emory University, where researchers in basic sciences had larger datasets, and arts and humanities researchers were found to not only work with smaller datasets, but often be unaware of how much data they were storing (Akers & Doty, 2013:8). Similarly, researchers working with climatic forecasts and climatic simulation at the CSIR were found to work with datasets in excess of several terabytes, with researchers working with text data (e.g. interview data in the natural sciences, interview data in the social sciences) producing significantly smaller datasets (Patterton, 2014a). With regards to data formats, it was found that spread sheets and instrument data were used by 69% and 61% of researchers in the exact sciences respectively, compared to statistical data and spread sheets used by 68% and 53% of researchers in the humanities and social sciences (Dillo & Doorn, 2011:14). Furthermore, some humanities researchers at the University of Oxford (Wilson et *al.,* 2010) did not see themselves as working with data at all; Carlson (2011: 6) makes a similar statement about humanities researchers not seeing themselves as working with data, and adds that the term ‘data’ is often seen as numerical or tabular data by default. Findings of a study involving Canadian researchers probably capture the essence of disciplinary trends relating to data formats the best: 51% of respondents indicated storing their research data in a format standard in their field (Mowers, Humprey & Perry, 2013:5). It might be the case, worldwide, that researchers generally store most of their data in a standard format familiar within their discipline.

Before concluding the section on data formats and data set sizes, it is interesting to take note of some findings revealed in various singular studies, i.e. findings unique to that study as the topic or question was not really broached in other studies. Obsolescence, or outdated formats, was a topic examined at the University of Northampton (Alexogiannopoulos, McKenney & Pickton, 2010:30), where it was revealed that many researchers were still collecting data in outdated formats. Furthermore, it was discovered that researchers often do not recognise the implications of software choices and later access to data (Henty *et al.,* 2008:41).

In conclusion: data types, formats and dataset sizes portray a great deal of diversity. This diversity exists not only between studies, but can also vary between different groups within the same study. Furthermore, depending on aspects such as project, and data gathering tool/method, mentioned diversity also exists when the data of individual researchers are investigated.

2.4.5 Data storage and data backups

A high percentage of studies included questions about the storage of research data. In most instances, respondents were asked to indicate where research data were stored. Additionally, several studies also asked whether the available storage space met their storage needs, what their storage needs were, and what the data backup practices were. In some instances, differences between groups, whether on grounds of subject discipline, research intensity, or faculty rank, were also investigated.

The most common primary storage place for research data, as indicated by study respondents, proved to be locally, on the researcher's personal computer or laptop. This was found across all disciplines, institutions, geographical areas, and levels of research experience, as will be shown in the next few paragraphs.

Studies conducted in the United States, specifically the University of Houston (Peters & Dryden, 2011:393), University of Iowa (Gu & Averkamp, 2012:14), California Polytechnic State University (Scaramozzino, Ramirez & McGaughey, 2012:361), Emory University (Akers & Doty, 2013:9), University of Central Florida (Beile, 2014), Northwestern University (Buys & Shaw, 2015:12) as well as Oregon State University (Whitmire, Boock & Sutton, 2015: 388) revealed the primary place of storage to be on their personal computer or laptop. In the United Kingdom, respondents from the University of Oxford (Martinez-Uribe, 2008:8), the University of Bath (Jones K, 2011:7), the University of Hertfordshire (Nassiri & Worthington, 2012), and the University of Nottingham (Parsons, Grimshaw & Williamson, 2013:10) indicated similar practices. Moreover, a survey into RDM practices of Dutch researchers showed that more than 70% of respondents store their research data on their own computer (Dillo & Doorn, 2011:13). Findings of RDM investigations using Canadian researchers (Mowers, Humprey & Perry, 2013:9, as well as Sewerin *et al.,* 2015:5), as well as a study involving South African researchers (Patterton, 2014a), also support this trend. Marchionini's statement (2012: 12) is an apt summary of data storage practices: researchers are relying on themselves to store data.

However, investigations into storage of research data do not end with the indication of a primary storage location. It was found that respondents mostly stored their research data in

more than one place. An example of this trend is the conclusion reached that 52% of researchers at the London School of Hygiene and Tropical Medicine (LSHTM) store their data in two to three different locations, while 37% of respondents make use of four to eight locations (Knight, 2013:4). Similarly, Scaramozzino, Ramirez & McGaughey (2012:358) found that although 94% of respondents store a primary copy of a dataset on their personal computer, 30% of researchers would also make use of other computers and portable devices to store data. Correspondingly, while the personal computer was shown to be the most common storage place, Peters & Dryden (2011:393) state that numerous storage places and methods were being used by everyone interviewed. In agreement with this, Parsons, Grimshaw & Williamson (2013:11) mentions that the most common answer, when asked in how many places data was stored, turned out to be 'five'. A survey involving Australian researchers is in agreement with this; most researchers use more than one place when storing data (Henty *et al.,* 2008:9).

Additional storage locations indicated, as well as storage locations indicated by those that do not use the personal computer/laptop as primary location, are plentiful. After scrutinising available studies, this researcher is able to conclude that researchers make use of the following when storing data:

- a network server (Martinez-Uribe, 2008:8; Gu & Averkamp, 2012:15; Beile, 2014; Patterton, 2014a),
- a shared drive or shared server (Akers & Doty, 2013:9; Patterton, 2014a),
- laboratory computers (Akers & Doty, 2013:9; Mowers, Humprey & Perry, 2013:9),
- a home computer/personal computer/personal laptop (Wilson & Patrick, 2010:19; Enke *et al.,* 2012:29),
- disks such as CD-ROMs or DVD's (Mowers, Humprey & Perry, 2013),
- external hard drives (Dillo & Doorn, 2011:13; Enke *et al.,* 2012:29; Nassiri & Worthington),
- memory sticks/flash drives (Henty *et al.,* 2008:9),
- web-based storage, e.g. Dropbox and Google Drive (Raggett, 2012a:2; Pink *et al.,* 2013:13),
- journals (Marcus *et al.,* 2007:19), and
- paper or lab books (Marcus *et al.,* 2007:17-20; Parsons, Grimshaw & Williamson, 2013:9-10).

What can be deduced by looking at the range of storage locations used, is that research data, although mostly in digital format, is often also still being stored in paper format/non-digital format. Furthermore, as mentioned by Raggett (2012a:5-7), it would seem as if

researchers are voting with their feet, by using storage options not necessarily supplied or encouraged by their organisation. The use of cloud-based services for data maintenance, data sharing and data backups, while not being recommended and endorsed by the LSE, was cited as a case in point (Raggett, 2012:5).

A characteristic feature of the storage-practices as revealed through cited studies, is that an organised and systematic way of storing data was only found in the minority of studies conducted. Many authors commented on the inconsistency and unplanned way in which respondents are conducting research data storage. A study into RDM practices of University of Edinburgh researchers described their storage practices as '…ad hoc' (Ekmekcioglu & Rice, 2009:22), with little chance of retrieval. Pienaar (2011:16) also makes use of the exact term ('ad hoc') when describing the data storage practices at the University of Pretoria. Further evidence of the disorganised data storage practices prevalent are evident in the summarising adjectives and phrases used by authors: '…messy' (Marcus *et al.,* 2007:10), '…store data in haphazard manner' (Griffiths, 2009:48), '…much confusion' (Peters & Dryden, 2011:394), and '…a highly fragmented activity….each researcher has own strategy' (Diekmann, 2012:23).

These disorganised and random practices were often found to have negative repercussions for the researchers involved, or would reveal the potential of future undesired consequences. The survey conducted at the University of Exeter (Open Exeter Project Team, 2012:4) found storage practices to reveal varying degrees of information security measures. Similarly, Australian researchers indicated using storage devices and media that are unreliable and short-lived (Henty *et al.,* 2008:41), while two studies combining the RDM practices of two UK universities revealed that researchers often resort to seeking cheap solutions to their storage predicaments, without realising the risks or benefits of each (Freiman *et al.,* 2010:4; Ward *et al.,* 2011:267).

The issue of data storage does not come without its concerns and challenges. Space allocation as a worrisome issue was mentioned by University of Exeter researchers (Open Exeter Project Team, 2012:4), where the requirement of many researchers exceeding the standard 20 GB network space allocation. Respondents at the Universities of Cambridge and Glasgow (Ward *et al.,* 2011:267) tend to support this sentiment, believing that there is not enough institutional storage space available, while another survey at the same two institutions revealed that researchers find the available network server storage insufficient and slow (Freiman *et al.,* 2010:4). Australian researchers working in a collaborative setting expressed their concerns at the space limitations on the available server (Gibson & Gross, 2013:13), and a similar opinion was held at the University of Bath (Jones K, 2011), where

insufficient storage space was indicated to be the biggest data management issue. Most of the respondents taking part in a RDM survey at the University of Northampton mentioned that they had experienced problems due to lack of storage space.

An interesting theme was detected at the University of Northampton; it would seem as if storage needs and behaviour often tend to vary throughout the research lifecycle (Alexogiannopoulos, McKenney & Pickton, 2010:5). This would mean that different storage devices tend to be prominent at the data collection, data analysis and project completion stages of research.

Although this researcher expected dataset size to play a role when deciding on storage options, this issue was not commonly included in survey frameworks, nor mentioned when respondents were probed for RDM concerns. This researcher suspects that dataset size, and the problems associated with it, tend to fall under the umbrella-issue of 'lack of sufficient data storage space'. Nevertheless, dataset size (as well as the data held by data intensive groups) was found to be a concern for some. Australian researchers working in collaborative settings tend to make more use of storage alternatives than their less data-intensive counterparts, leading to the use of external hard drives, laptops and commercial data storing options (Gibson & Gross, 2013:13).

As a few studies were interested in the RDM differences between faculty ranks, as well as disciplinary differences, this issue deserves to be mentioned as well. It was discovered that there are no faculty rank differences in the amount of research data stored, or methods of storing, when researchers at Emory University were questioned about storage practices (Akers & Doty, 2012:18-19). Although not explicitly stated as a disciplinary difference, two studies at the University of Oxford seem to support the notion that humanities researchers tend to be making less use of institutional servers than other disciplines, and would appear to have little awareness of existing central services, as these servers were found to be used by only humanities researchers possessing considerable experience on narrow data projects, where IT staff were directly engaged (Wilson & Patrick, 2010:19, Wilson *et al.,* 2011). This finding is in agreement with a study at Emory university, where it was found that researchers in the basic sciences were more likely to use external hard drives, university servers, and the hard drive of the laboratory instrument, while researchers in the arts and humanities were more likely to use computer hard drives and the internet when storing data (Akers & Doty, 2013:9).

While not all studies investigating data storage included findings on data backups, many studies did indeed probe this separate aspect of data storage. In general, backup of research data is an activity most researchers complete regularly. This statement is

supported by the findings into RDM behaviours of researchers at the University of Northampton (Alexogiannopoulos, McKenney & Pickton, 2010:23), California Polytechnic State University researchers (Scaramozzino, Ramirez & McGaughey, 2012:358), and researchers at the CSIR, South Africa (Patterton, 2014a). Backup percentages of more than 85% were mentioned in all three studies conducted. Likewise, almost all researchers from Oxford University have backup strategies (Martinez-Uribe, 2008:8), and all researchers at the LSE indicated backing up their data (Raggett, 2012:30). In addition, respondents from the Universities of Cambridge and Glasgow (Freiman *et al.,* 2010:4) stated that all their data were being backed up by IT.

As was the case with data storage, the localities of data backups were found to be diverse and subject to individual practices. In other words, not only were findings between studies quite varied, but backup location differences were also discovered to vary in the same study. Authors investigating researcher backup behaviour describe it as follows: '…responses varied widely' (Peters & Dryden, 2011:394), '…highly individualised' (Diekmann, 2012:23), '…a mixed approach' (Raggett, 2012a:13), as well as '…sporadic' (Johnston & Jeffryes, 2014:9).

While findings cannot be generalised, storage options such as the following were indicated by respondents:

- the office computer (Scaramozzino, Ramirez & McGaughey, 2012:358),
- work server, networked storage or services offered by the university (Ekmekcioglu & Rice, 2009:22; Martinez-Uribe, 2008:8; Pink *et al.,* 2013:13; Van Tuyl & Michalek, 2015:3),
- external devices in the form of memory sticks/flashdrives and external hard drives (Beile, 2014; Scaramozzino, Ramirez & McGaughey, 2012:358),
- a cloud server (Van Tuyl & Michalek, 2015:3),
- CD's and DVD's (Martinez-Uribe, 2008:8),
- commercial services (Diekmann, 2012:24), and
- '…personally owned machines' (Wilson and Patrick, 2010:5; Wilson *et al.,* 2011).

This researcher is of the opinion that this last category is probably synonymous with personal computers and laptops, mentioned earlier.

Frequency of data backups, as with the activity of backing up data, as well as backup media used, is an RDM activity portraying diverse behaviour. Furthermore, backup frequency was an area not often included in studies, making generalisations regarding this aspect of RDM behaviour problematic. While some studies found the most common backup frequency to be

daily (Jones K, 2011 also Patterton, 2014a), other studies noted backup frequency to be most commonly weekly (Scaramozzino, Ramirez & McGaughey, 2012:358), or even non-specifically described as '…regularly by IT' (Freiman *et al.,* 2010:4).

Alexogiannopoulos, McKenney & Pickton (2010:23), Scaramozzino, Ramirez & McGaughey (2012:358), Averkamp, Gu & Rogers (2014:12) as well as Van Tuyl & Michalek (2015:3) have established similar backup behaviours in their respective studies, with all studies reporting backup frequencies of 85% or higher. Backup of all research data was reported by Martinez-Uribe (2008:8) as well as Freiman *et al.* (2010:4).

Group differences (differences being disciplinary, faculty rank, or researcher experience) were not commonly investigated, or reported. A lone study specifically interested in RDM behaviour as it relates to university faculty rank, found no differences with regards to backup behaviour (Akers & Doty, 2012:17).

Backup of research data is an RDM activity not without its share of issues, problems and concerns. Study respondents have indicated not using, or being being fully aware of existing central services, able to aid with data backups, offered by their institution (Wilson *et al.,* 2011). A study involving Dutch researchers revealed the frequency of data backup loss due to the absence of systematic RDM, and the absence of adequate means for archiving (Dillo & Doorn, 2011:21). Similarly, agricultural researchers in the USA expressed their concerns about this RDM activity, by mentioning that there is a need for a more standardised and centralized way to do backups (Diekmann, 2012:24). Furthermore, researchers at the University of Houston were found to meet the question of data backups with '…great uncertainty and discomfort' (Peters & Dryden, 2011:394), gave responses that varied widely and generally portrayed much confusion when the topic of backups was broached. Another issue indicated as cause for concern was the over-reliance on University of Bath researchers on external hard drives for backups (Jones K, 2011).

In recapitulating the main behavioural themes relating to data storage and data backups, a standout feature is the diversity of practices. This diversity was found to exist between studies, and even within the same study. Both data storage and data backups were discovered to include many different formats, with most researchers storing data in more than one location, and on more than one type of device. Familiarity with, and application of organised, structured and consistent storage and backup practices, were also discovered to be a diverse trait and activity. Both data storage and data backup activities can be said to benefit from training, marketing and more prominent institutional support and infrastructure.

2.4.6 Data preservation

Data preservation in the context of RDM refers to the process of maintaining research data over time, ensuring that the data can be '…found, understood, accessed, and used in the future' (University of Cambridge, 2012). Investigating the extent to which this component of RDM was practiced, or applied to the research data of respondents, formed a part of many studies analysed by this researcher.

Findings displayed considerable variance, and varied between '…not sanguine about institutional ability to support long-term preservation of data' (ECAR, 2009:119), '…sometimes neglected once project is complete' (Alexogiannopoulos, McKenney & Pickton, 2010:29), '…few think long-term preservation' (Jahnke & Asher, 2012:3), a '…general absence of practices' (Johare, 2014) on the one end, and a study indicating that 68% of respondents take measures to preserve data (Beile, 2014). Not only did findings vary between studies, but it was also revealed that preservation practices within the same institute, e.g. University of Minnesota, weren't always uniform (Marcus *et al.,* 2007:10, 19). Here, data preservation practices were described as '…idiosyncratic', '…haphazard', '…in great need of attention', and varying between '…preserving very little', and '…preserving everything' (*ibid.*); in other words, data preservation practices were left to individual whim. Inconsistency of preservation practices were also revealed at the University of Southampton (Takeda *et al.,* 2010) and the CSIR (Patterton, 2014a).

In more recent studies, Kennan & Markauskaite (2015:84) report that only 8% of Australian academics surveyed make use of an internal data storage facility or an institutional data service after project completion, and report that a third of respondents experienced serious data preservation issues. Sewerin *et al*., in a study involving Canadian researchers (2015: 3), state that a long-term data preservation platform, or institutional repository, was indicated by their survey's respondents to be one of their requirements.

Apart from preservation practices between studies, or within the same study, showing great differences, attitudes toward the practice, as well as knowledge pertaining to it, were found to portray considerable disagreement. While more than 85% of University of Hertfordshire researchers felt that research data should be preserved (Nassiri & Worthington, 2012), researchers at the universities of Cambridge and Glasgow (Ward *et al.,* 2011:267), showed that researchers believe that making backups is equivalent to preservation. The same study found that researchers were also unsure which formats were better for preserving data. As could be expected, Freiman's study at the same institute (2010:4) revealed the same theme: researchers were often unsure about which formats/media are best when it comes to long-

term data preservation. Malaysian researchers were found to lack preservation knowledge, and a lack of understanding of many technical terms (Johare, 2014). Similarly, a much earlier study involving UK researchers showed that awareness of issues, especially longevity issues, was generally low and that nearly half of respondents had never received curation guidance (Lord & Macdonald, 2003:28). Here, the respondents who had received guidance described the guidance offered as '…not helpful', being too much of a '…motherhood statement' and not able to solve the practical and financial problems of long-term archiving (*ibid.)*

Several studies probed respondents specifically on the issue of preservation period for preserved data. 42% of University of Hertfordshire researchers felt that their data should be preserved for more than 10 years (Nassiri & Worthington, 2012), Australian researchers indicated that their valuable data should have value for more than 10 years (Henty *et al.,* 2008:16), most researchers at the University of Oregon (Westra, 2010) as well as the CSIR (Patterton, 2014a) felt that their data should be preserved indefinitely; this inclination towards an indefinite period of preservation was also shown to exist in a collaborative Australian research environment (Gibson & Gross, 2013:13), and the University of Houston (Peters & Dryden, 2011:394). Although these findings might indicate a preference for indefinite preservation, this was not seen to be the case at all institutes: although many University of Southampton researchers want to keep their data forever, it was established that, in practice, the mean was five years (Takeda *et al.,* 2010).

Challenges, obstacles and other issues contributing towards data preservation being seen as a problematic area, were plentiful, and included infrastructure, finances, knowledge, experience and support. In particular, the following challenges were mentioned:

- no clear written preservation policies (Johare, 2014),
- lack of policies and guidelines (Nassiri & Worthington, 2012),
- insufficient infrastructure is in place (Nassiri & Worthington, 2012),
- lack of facilities (Nassiri & Worthington, 2012),
- a need for preservation training (Knight, 2013:19),
- difficulty managing influx of new data including time constraints (Westra, 2010),
- time and effort required to deposit and preserve data (Freiman *et al.,* 2010:4; Williams, 2012:318),
- concern about platform/software obsolescence (Freiman *et al.,* 2010:4),
- concern about data loss (Freiman *et al.,* 2010:4),
- file migration issues (Johnston & Jeffryes, 2014:14),

- lack of clarity about who is responsible for data preservation (Johnston & Jeffryes, 2014:14),
- lack of understanding of legislative and funding body requirements with regards to long-term preservation of data (Bradbury & Borchert, 2010),
- lack of institutional support offered (Tenopir *et al.,* 2011:19),
- lack of assistance (Marcus *et al.,* 2007:10), no centralised storage available/storage problems (Peters & Dryden, 2011:394; Johnston & Jeffryes, 2014:13), and
- funding mechanisms (ECAR, 2009:119).

An early study investigating the state of RDM in the UK (Lord & Macdonald, 2003:32) had revealed some alarming trends: 42% of respondents made use of software whose longevity was questionable, while a high percentage of researchers indicated that they had even written their own software. This researcher is of the opinion that recent advances in UK RDM, as well as marketing of the activity, may have led to preservation improvements in the meantime.

The requirements for making use of available preservation facilities were aspects considered in some studies: a study examining researchers worldwide (Tenopir *et al.,* 2011:9) found that repositories need to accommodate for varying levels of security, and should apply access restrictions, while the ability to edit data, delete data, data creator involvement, as well as the assurance that the database will be maintained for the long term, was reported in a study investigating biodiversity researchers (Enke *et al.,* 2012:29). Another study involving researchers worldwide showed that the possibility of reanalysis of data to be the most important driver for the preservation of research data (Kuipers & van der Hoeven, 2009:4). Here, it was felt that infrastructure used for long-term preservation should safeguard against lack of sustainable hardware, software or support, as these were seen as serious threats to digital preservation.

Adding to the above, disciplinary differences, as well as preservation differences related to research experience, were found to exist when researchers preserve data; this trend will be discussed in section 2.5.12, titled ‘Group differences in RDM: discipline, faculty rank/research experience’.

In summary: although the importance of long-term preservation of research data is understood by many, the practical application thereof is random and arbitrary. Considerable variance between studies, as well as within studies, was displayed. Several factors contribute to the haphazardness and underuse of long term preservation as an RDM activity; these factors revolve around issues of infrastructure, training, funding, and support.

Similarly, requirements and conditions contributing to the use of long-term preservation facilities by researchers, were mentioned and should be implemented in institutional RDM regimes.

2.4.7 Data sharing

The inclusion of a data sharing question in a RDM study, was probably the most common mutually-held characteristic in studies examined. Apart from incorporating this aspect into comprehensive RDM studies, several studies, interested in data sharing practices exclusively, have also been conducted. This section of the literature review, rather than reporting on all sharing-related practices, provides a brief overview of main trends. Possible reasons or explanations for these trends are also forward.

In general, and as with most RDM behavioural aspects discussed up to now, data sharing is an aspect displaying much variance between studies, and between different groups studied. In some instances, such as at the University of Central Florida (Beile, 2014), Unisa (Bezuidenhout & Macanda, 2014) as well as the LSTHM (Knight, 2013:4) researchers indicated sharing willingness, with 69%, 70% and 70% of researchers respectively indicating their sharing prevalence. Similarly, a study examining Dutch researchers (Dillo & Doorn, 2011:5) discovered data sharing to be quite common. These tendencies are also supported by Buys & Shaw, 2015:14 as well as Sewerin *et al.* (2015:6). These incidences of high sharing prevalence is in stark contrast to low percentages of sharing among European researchers (Kuipers & van der Hoeven, 2009:33), a survey at the University of Hertfordshire (Nassiri & Worthington, 2012), where sharing was shown to be an uncommon event, and an Emory University survey (Akers & Doty, 2013:9) as well as ten universities in Australia (Kennan & Markauskaite, 2015:69), where most researchers do not share data outside their group, and displayed the other end of the data sharing spectrum.

Several studies have shown that the method used when requesting data could also influence data sharing. Buys & Shaw (2015:15), Kennan & Markauskaite (2015:81) as well as Sewerin et al. (2015:6) have reported that researchers show a preference for personal requests and private negotiation when it comes to sharing data.

In many instances, respondents were questioned about their reasons for not sharing, and these revelations, being plentiful, diverse, and providing insight into the sharing mentality of researchers, are worth mentioning here. The list of reasons for not sharing research data include:

- fear of data misuse (Kuipers & van der Hoeven, 2009:4; Dillo & Doorn, 2011:7; Wynholds *et al.,* 2011:384; Hall, 2013:383; Mossink & Bijsterbosch, 2013:15; Sayogo & Pardo, 2013:S27; Peset, 2014:1; Wiley, 2014:1),
- fear of data misinterpretation (Dillo & Doorn, 2011:7; Peters & Dryden, 2011:395; Tenopir *et al.,* 2011:18; Wynholds *et al.,* 2011:384; Diekmann, 2012:28; Enke *et al.,* 2012:27; Hall, 2013:383; Mossink & Bijsterbosch, 2013:15; Sayogo & Pardo, 2013:S27),
- legal concerns (Griffiths, 2009:51; Kuipers & van der Hoeven, 2009:4; Enke *et al.,* 2012:27; Patrick, 2012; Mossink & Bijsterbosch, 2013:15; Sayogo & Pardo, 2013:S28; Peset, 2014:1; Buys & Shaw, 2015:5; Kennan & Markauskaite, 2015:82; Van Tuyl & Michalek, 2015:19),
- ethical concerns/confidentiality issues (Griffiths, 2009:51; Dillo & Doorn, 2011:11; Peters & Dryden, 2011:395; Enke *et al.,* 2012:27; Patrick, 2012; Raggett, 2012a:15; Akers & Doty, 2013:10-11; Hall, 2013:383),
- IP rights/commercial concerns (Peters & Dryden, 2011:395; Bardyn, Resnick & Camina, 2012: 282; Diekmann, 2012:28; Patrick, 2012; Raggett, 2012a:15),
- no reason to share (Raggett, 2012a:15),
- being unable to anonymise data (Raggett, 2012a:15),
- data sharing being a time-consuming activity (Griffiths, 2009:51; Freiman *et al.,* 2010:4; Peters & Dryden, 2011:395; Tenopir *et al.,* 2011:20; Enke *et al.,* 2012:27; Williams, 2012:318; Averkamp, Gu & Rogers, 2014:15; Sewerin *et al.,* 2015:6; Van Tuyl & Michalek, 2015:19),
- not having funds to enable data sharing (Freiman *et al.,* 2010:4; Tenopir *et al.,* 2011:20; Enke *et al.,* 2012:30),
- sharing not required by funders (Martinez-Uribe, 2008:10; Wiley, 2014:1),
- data sharing being too work-intensive (Campbell *et al.,* 2002:479; Martinez-Uribe, 2008:10; Peters & Dryden, 2011:395; Diekmann, 2012:29; Hall, 2013:383; Williams, 2012:318),
- lack of sharing experience/lack of sharing skills (Marcus *et al.,* 2007:18; Griffiths, 2009:51; Diekmann, 2012:27),
- data quality issues (Wilson & Patrick, 2010:40; Diekmann, 2012:28; Sewerin *et al.,* 2015:6),
- increased competitiveness of academic research in their fields, competitive research advantage, fear of data being scooped (Diekmann, 2012:28; Averkamp, Gu & Rogers, 2014:1; Wiley, 2014:1; Kennan & Markauskaite, 2015:82; Sewerin *et al.,* 2015:6; Van Tuyl & Michalek, 2015:19),

- too much work had been put into creating the data (Wynholds *et al.,* 2011:385),
- lack of professional award for sharing data (Griffiths, 2009:51; Enke *et al.,* 2012:30; Akers & Doty, 2013:11),
- loss of control over own data (Enke *et al.,* 2012:27),
- protecting own ability to publish (Campbell *et al.,* 2002:479; Griffiths, 2009:51),
- protecting colleague's ability to publish (Campbell *et al.,* 2002:479),
- lack of data sharing resources (Griffiths, 2009:51; Kennan & Markauskaite, 2015:82; Sewerin *et al.,* 2015:6), and
- viewing data as of a researcher's 'intellectual capital'/strong sense of ownership (RIN, 2009:39; Dillo & Doorn, 2011:7).

Looking at this list, this researcher is in agreement with Raggett (2012a:9), who states that many of the non-sharing reasons supplied are based in fear, and that training can overcome this and encourage an open frame of mind. It is therefore recommended that training materials (*ibid.*) should also be designed to include this culture change element.

Whereas the above list mainly portrays attitudes, fears and concerns, additional hindrances to data sharing, more related to infrastructure and technology, were also mentioned by researchers. Infrastructure-related concerns making sharing problematic were stated to be:

- access restrictions on their IT infrastructure (Gibson & Gross, 2013:13),
- difficulty in using shared drives and accessing the commercial partner's VPN (Pink *et al.,* 2013:19),
- difficulties with data size and software being used when sharing (Bradbury & Borchert, 2010), and
- incompatibility of data types (Kuipers & van der Hoeven, 2009:33).

Looking at non-sharing reasons supplied, it is important to note that these reasons emanate from a variety of subject disciplines, research groups, and research levels, and are supplied here to give an overview of possible reasons why researchers are unwilling to share data. Interdisciplinary differences in data sharing do exist: at the University of Oxford, it was discovered that researchers in the mathematical, physical and life sciences were the most willing and open with regards to data sharing, with 56% saying that they would share all or most of their data. Researchers in the medical sciences, with a sharing prevalence of only 9%, displayed the most reluctance towards this activity (Wilson, 2013). Here, the issue of ethics and privacy concerns would explain the glaring difference. At Emory university, disciplinary differences were also discovered (Akers & Doty, 2013:10): basic science researchers were more willing to share than other disciplines, while the arts and humanities were least likely to

share with other researchers, even when working on the same project. Once again, medical researchers were least likely to share outside their project; social science researchers were just as unwilling to share. Interestingly enough, arts and humanities researchers were more willing to share data with the public, than with other researchers. Even more interesting, is that about half of all respondents were not willing to share their data with project funders. Looking at these results, and probing for reasons of data withholding, the authors found that issues related to data confidentiality (specifically pertaining to medical data), lack of recognition when sharing data (arts/humanities) and possibility of data misinterpretation/data misuse, were the most common reasons for not sharing data (*ibid.*).

Although not doing interdisciplinary comparisons, a few studies also focussed on one discipline exclusively, resulting in the following conclusions:

- Nearly all environmental studies researchers see the value in data sharing (Hall, 2013:383).
- Life sciences researchers are reluctant to share data making up their 'intellectual capital' (RIN, 2009:7).
- Life science researchers readily share data not adding value (RIN, 2009:39).
- Many life science researchers are not willing to use data from other researchers; too many differences in experimental design and data collection practice can influence relevance, applicability and usefulness (*ibid.*).
- Humanities researchers generally only share upon request; data banks are rare (RIN, 2011).
- Humanities seem willing to share, but in practice do not; a sense of data ownership features strongly in this discipline (Wilson & Patrick, 2010:5,10).

Looking at these findings, the role of several data determinants are highlighted once again: keen sense of data ownership, lessening of sense of data ownership when data are deemed as less valuable, avoidance of data re-use when differences in experimental design or data collection practices are discovered, and availability of data sharing facilities. Although studies above were discipline-specific, these clinchers are common to all disciplines and provide an insightful glimpse into the sharing mind set and sharing attitude of researchers.

A few studies made possible the determination of factors to be present, or data sharing conditions to be improved, before researchers would be willing to share data. Researchers indicated that data sharing, or depositing data in a shareable data bank, would be more likely if:

- a written explanation on how to use data, could be included (Patrick, 2012),

- data have already been published (Borgman, Wallis & Enyedy, 2006:170; RIN, 2009:39; Enke *et al.,* 2012:29),
- there is a formal data citation (Tenopir *et al.,* 2011:19),
- a data repository has sharing restrictions (Tenopir *et al.,* 2011:19),
- some form of compensation (e.g. monetary, co-authorship, or acknowledgement) is received (Reidpath & Allotey, 2001:133; Cragin *et al.,* 2010:4031),
- the data creator is informed who the requestor is, and for what purpose data will be used (Pryor, 2009:78; Research Information Network 2009:39),
- formal application has to be made before data are reused (Pryor, 2009:78),
- there is sufficient time to complete analysis (Pryor, 2009:80), and
- there is sufficient time to explore IP rights (RIN, 2009:39).

Conditions listed here once again underline the importance of data-sharing determinants as mentioned earlier: Patrick's findings reveal the fear of data misinterpretation, Tenopir's findings reveal the need for acknowledgement, Pryor findings reveal the need for data quality, and so forth.

When sharing data, methods most commonly used included:

- emails (Pryor, 2009:80; Alexogiannopoulos, McKenney & Pickton, 2010:24; Peters & Dryden, 2011:396; Nassiri & Worthington, 2012; Raggett, 2012a:21; Akers & Doty, 2013:9; Pink *et al.,* 2013:20; Patterton, 2014a; Buys & Shaw, 2015:15; Kennan & Markauskaite, 2015:81),
- ftp (Nassiri & Worthington, 2012:24),
- an internet service (Diekmann, 2012:27; Raggett, 2012a:21; Pink *et al.,* 2013:20),
- collaborative web space (Peters & Dryden, 2011:396),
- portable media/external devices (Martinez-Uribe, 2008:9; Alexogiannopoulos, McKenney & Pickton, 2010:24; Peters & Dryden, 2011:396; Pink *et al.,* 2013:20),
- data centres, either departmental, national, international (Martinez-Uribe, 2008:9; Wiley, 2015:1),
- supplementary files on journal website (Diekmann, 2012:27; Akers & Doty, 2013:9; Buys & Shaw, 2015:15; Wiley, 2015),
- meetings (Pryor, 2009:80), presentations (Pryor, 2009:80; Diekmann, 2012:27), and
- shared folders within the group (RIN, 2009:41).

Sharing methods were often chosen for a specific reason, or under specific circumstances. Alternatively, certain sharing methods present their own set of concerns:

- use of emails can be problematic when encrypted data are shared (Nassiri & Worthington, 2012),
- emails present an added advantage of not having to meet the recipient (Alexogiannopoulos, McKenney & Pickton, 2010:24),
- researchers often indicated being in need of better sharing methods than are currently available/used (Henty *et al.,* 2008:13-15),
- very big datasets cannot be emailed (Martinez-Uribe, 2008:9, Alexogiannopoulos, McKenney & Pickton, 2010:24),
- emails tend to be a difficult sharing method when researchers in a group are working on and updating the same file (Alexogiannopoulos, McKenney & Pickton, 2010:24),
- USB sticks are a preferred sharing method when datasets are very big (Alexogiannopoulos, McKenney & Pickton, 2010:24), and
- researchers displayed concern about the ownership and maintenance of departmental data banks (Martinez-Uribe, 2008:9).

These concerns and usage-trends, pertaining to data sharing methods, are just a smattering of aspects researchers need to take into account when deciding on a data sharing method.

In summary: data sharing among researchers is an often-studied topic, resulting in a considerable amount of literature to be analysed by this researcher, and presenting an area that in itself is interesting, complicated and diversified enough to warrant a separate study. Data sharing by researchers display variance in behaviour, and its application or implementation is influenced by the presence of many determining factors. While many reasons for non-sharing can be seen as valid, the suggestion was put forward (Raggett, 2012a:9) that lessening of sharing-related fear, and changes in sharing-related research culture, could be brought about through purposeful training and training materials.

2.4.8 Use of metadata

The word ‘metadata’ means ‘data about data’ (Dublin core Metadata Initiative, 2014). Metadata can be seen to articulate the context, conditions and circumstances for an object of interest; seen in an RDM environment it would often refer to a ‘highly structured, machine-readable subset of data documentation that may be indexed and stored within a database (University of Sheffield, 2014). By adding this information to a dataset, the dataset’s relevance and context can be easily found when needed.

Most survey frameworks, but not all, included at least one question on the use of metadata when managing research data. Studies interested in discovering more detailed use of metadata would include not only a question about the use or non-use of metadata, but would also ask respondents about their adherence to a metadata standard, their satisfaction with tools available for creating metadata, and their satisfaction with the quality, or re-use potential of the metadata.

Responses (percentage-wise) on this aspect of RDM vary widely between studies; a generalised conclusion regarding the exact use of metadata cannot be made. Most studies are in agreement that metadata are not used by the majority of researchers. Some studies, such as the survey at the University of Colorado Boulder, reported that 63% of respondents did not make use of metadata (Task Force, 2012:7). Other studies indicating such a high figure include the survey at Emory University (Akers & Doty, 2013:12), researchers in many faculties at the University of Nottingham (Parsons, Grimshaw and Williamson, 2013:5), and a study investigating practices in 50 departments at the University of Central Florida (Beile, 2014). In all these instances, it was indicated that more than 50% of users do not add metadata to their datasets.

Not quite at the opposite end of the spectrum, but providing a more positive picture, are a few studies where the use of metadata was seen to be just below 50%, or just above 50%. In other words, these studies do not indicate a majority non-use of metadata, but do indicate that it is an area still in need of training and advancement. A national survey of Canadian researchers in all disciplines has found that half (53%) of researchers do add metadata to datasets (Mowers, Humphrey & Perry, 2013). In a similar vein, metadata usage by Malaysian researchers (Johare, 2014) was found to be 42%, while the study of Whitmire, Boock & Sutton (2015: 390) found that 53% of respondents made use of metadata.

Several studies found that metadata were being added by the majority of researchers. Examples of these are the survey of researchers at the QUT (Bradbury & Borchert, 2010), where 72% of users add metadata, and the University of Hertfordshire (Nassiri & Worthington, 2012), where 80% of researchers do add metadata to their datasets.

The conclusion drawn from these results indicates that the use of metadata is subject to diverse practices. It further portrays that even when taking into account institutes where metadata usage is seen to be flourishing, current metadata application is an area indicating room for improvement at all institutes surveyed.

It was interesting to note that a few studies mentioned that the concept of ‘metadata’ was not understood by all study participants. It was found that 20% of survey respondents at the

University of Hertfordshire had either no idea what metadata were, or were not creating metadata (Nassiri & Worthington, 2012). A similar situation was found to exist at the University of Nottingham, where researchers described the term 'metadata' incorrectly (Parsons, Grimshaw & Williamson, 2013:18). The majority of scientists and researchers at the University of Minnesota (Johnston & Jeffryes, 2014:10) also had no idea what the term 'metadata' meant. Similarly, students and researchers at Purdue University (Carlson *et al.,* 2011:11) indicated that 'metadata' as a term is not often used. These findings are in agreement with a study by this researcher (Patterton, 2014a) who discovered that many of the respondents, when asked about their metadata usage, responded by asking the interviewer what was meant by the term 'metadata'. In addition, Van Tuyl & Michalek (2015: 18) report that respondents felt that they do not know enough about metadata standards to be creating metadata. Jahnke & Asher (2012:3) state that metadata are mostly added by university researchers and graduate students in the USA, if this activity helps the researcher complete his work.

Questioning respondents about their adherence to a specific metadata standard was another topic included in many studies. When looking at the findings across studies, the non-adherence to any standard when creating metadata, is a reality which cannot be disputed. Studies were clear on this trend: responses varied from 'none use formal data standards' (Peters & Dryden, 2011:395), to a finding of 88% not using a metadata standard (Mowers, Humphrey & Perry, 2013), to an indication that '…very few are aware of existing data standards' (Martinez-Uribe, 2008:9).

Further studies supporting these findings were the worldwide study of scientists (Tenopir *et al.,* 2011:9), where 56% of respondents do not use a standard, and 22% have created their own standard, the University of Nottingham, where only 18% of respondents were found to stick to a metadata standard (Parsons, Grimshaw & Williamson, 2013:18), and the University of Central Florida where the majority of researchers do not make use of a standard (Beile, 2014:11). Similarly, Yeumo (2014:0) mentions that the use of metadata standards is uncommon (used by 23% of respondents only), and that most respondents do not see how standards could be beneficial to research, while Whitmire, Boock & Sutton (2015:390) report that 74% of respondents do not make use of a metadata standard, or are making use of a standard devised in the laboratory. These mentioned studies are just a few examples of RDM studies portraying the general absence of metadata standards when managing research data. Bearing in mind that the majority of studies found metadata standards users to be in the minority, and with this minority being part of another minority (respondents adding metadata), it can be said with certainty that the low usage of metadata standards seems to prevail in science.

The repercussions of not adding metadata, or not being metadata-knowledgeable, were mentioned in a study on researchers in the United Kingdom: it was found that the need for metadata could be a barrier to depositing in archives, and data repositories (Ward *et al.,* 2011:267). Conversely, and according to the DCC (2015), researchers making use of a metadata standard are ensuring the creation of 'rich, consistent metadata which will support the long-term discovery, use and integrity of digital resources'.

A few studies were interested in the possible differences that might exist between different groups of metadata users. No difference in metadata use was found to exist between PhD students and non-PhD students (Raggett, 2012a:17). The same author was also unable to establish that research experience was an indicator of metadata use (*ibid*.). Another study found that high quality metadata seems to be more prevalent in big projects (Martinez-Uribe, 2008:9).

When looking at various disciplines, no differences were found to exist with regards to metadata use in different fields of study at Emory University (Akers & Doty, 2013:12). These findings differ from a Research Information Network study (Griffiths, 2009:49) where disciplinary metadata practices were influenced by a number of factors including data centres that encouraged certain standards, for example in metadata for astronomy, crystallography and the arts and humanities.

In conclusion, and despite some diverse findings, a few general trends with regards to the use of metadata use among researchers were apparent. In the words of Johnston & Jeffryes (2014:10), the use of metadata seems to be '…ad hoc and varied'. Use of metadata vs non-use of metadata varied, and no hard and fast rule can be applied to the prevalence of this practice. Metadata standard adherence tends to be low; most studies reported that more than 50% of respondents do not stick to any standard. Although reasons for these trends are not supplied, it would seem that the conclusion reached, after studying Warwick University researchers (Delasalle, 2013), might explain this phenomenon somewhat. It was stated that researchers are not prioritising metadata as they are sceptical of the role of the metadata; this statement indicates an area which at many higher education institutes could benefit from metadata training and additional institutional assistance in the form of metadata tools/software and supportive infrastructure.

2.4.9 Data management plans

A data management plan (DMP) typically states '…what data will be created and how, it outlines the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied' (DCC, 2014). Apart from being

increasingly required by funders, the use of a DMP may help ensure that data are organised well and better annotated, and would also play a role in increased research efficiency. By taking the benefits of a DMP into account, the inclusion of questions pertaining to the use of such a plan is a natural step when studying the RDM practices of researchers.

Considerable variance in answers was found across all studies investigating DMP usage. Findings ranged from the majority not using such a plan, as was the case with 80% of Australian researchers (Henty *et al.,* 2008:8), 61% of respondents in a worldwide study (Enke *et al.,* 2012:28), 80% of researchers at the LSE (Raggett, 2012a:2), 76% of researchers at the University of Colorado Boulder (Rankin *et al.,* 2012:33), 66% of researchers at the University of Nottingham (Parsons, Grimshaw & Williamson, 2013:20-21), and 81% of researchers at the University of Bath (Pink *et al.,* 2013:11). In addition, Buys & Shaw's study revealed that only 45% of respondents used DMPs (2015:15), Kennan & Markauskaite (2015:69) state that only a small portion of their respondents used DMPs, and Van Tuyl & Michalek (2015: 1) reported that only 44% of respondents had to submit DMPs.

Of further interest was the discovery by some authors that many researchers are not only unaware of the existence of a DMP within their group, or not, but were also not familiar with the concept of a DMP. It was found that 47% of researchers at Georgia Tech indicated that they do not know enough about DMP's to be using them (Parham, Bodnar & Fuchs, 2012:12), while agricultural scientists in the USA were found to have very little experience with the use of a DMP (Diekmann, 2012:25-26). At the University of Nottingham (Parsons, Grimshaw & Williamson, 2013:21), it was not clear to the authors whether all research group members were made aware of the existence of a DMP.

Several studies probed deeper into the use of DMP's, and requested reasons for the non-use of DMP's. Similarly, a few studies also questioned researchers about their perceptions regarding the benefits of a DMP, or why they were using it. Also investigated in some studies was whether researchers being surveyed were required by funders to submit a DMP.

As could be expected, funders and institutional policies are some of the main drivers of researchers implementing the use of a DMP. This was mentioned in the study into RDM practices of European researchers (Mossink & Bijsterbosch, 2013:5,19), where a quarter of the funders were found to require a DMP, at the time of the study. 30% of respondents at the University of Iowa (Gu & Averkamp, 2012:10) mentioned that a DMP is a requirement. In similar fashion, a study predating the NSF requirement of a DMP found that 55% of scientists do not make use of a DMP, as it is not required. An interesting finding related to DMP's at the University of Nottingham (Parsons, Grimshaw & Williamson, 2013:21), was that only 30% of DMP-users admitted to be making use of these plans due to it being a

funding requirement; and 63% of DMP-users made use of such a plan and stated that funder-requirement was not the reason for it.

Additional reasons or arguments in favour of the use of DMP were showcased in several studies. Although only 19% of University of Bath researchers indicated submitting a DMP, the reasons for the plan's use was understood, and cited as assisting with ownership, storage and archive issues at a later date (Pink *et al.,* 2013:11). Without going into specifics, 75% of respondents in a worldwide biodiversity study regard a DMP as a good idea; 78% feel it is an extremely important part of research (Enke *et al.,* 2012:28). Although DMP-usage at the California Polytechnic State University was found to be low, researchers do recognise the need for such a plan (Scaramozzino, Ramirez & McGaughey, 2012:359). Researchers at Georgia Tech (Parham, Bodnar & Fuchs, 2012:12), although indicating their lack of knowledge regarding DMP's, felt that research norms, as well as convenience when using research data in future, were reasons for using a DMP.

An interesting finding, indicating lack of insight into the need and benefits of a DMP, was the common belief held by agricultural scientists (Diekmann, 2012:25) that having a preliminary dataset prior to requesting funding constituted proof of data management. At Georgia Tech (Parham, Bodnar & Fuchs, 2012:12) researchers not only revealed their lack of knowledge relating to DMP's, but 40% of respondents mentioned that such a plan was unnecessary. As this study was conducted before the DMP requirements of the National Science Foundation in the USA came into being, these findings were indicated as hardly surprising. Other reasons indicated as contributors towards the non-use of a DMP include the lack of information available about DMP's, and the extra demands it made on time (Peters & Dryden, 2011:395).

Several studies succeeded in revealing a discrepancy between DMP belief and DMP practice. A good example of this is the study investigating practices at the Polytechnic State University (Scaramozzino, Ramirez & McGaughey, 2012:359), where 84% of researchers felt that it is important to have a data preservation plan. In practice, fewer than 15% of respondents use a DMP. In addition, fewer than 30% of those indicating that DMP's are valuable, admitted to be adhering to good data management practices. What is important in this study is noting that the term 'data preservation plan' and not 'data management plan', was used.

Although not many studies investigated possible disciplinary differences, faculty-ranking differences, or researcher-experience differences when it comes to the use of DMP's, a couple of studies touched on the topic. Even though it was expected that basic science researchers would be more likely to make use of DMP's, this was in fact not found to be the

case among Australian researchers (Henty *et al.,* 2008:8). Here, the largest proportion of researchers with DMP's were social scientists, followed by researchers in the fields of medicine and health, with humanities and arts, as well as law researchers, the groups making least use of DMP's. Henty *et al.* also report that researchers in the science disciplines were found to be more attuned to the need for a DMP, than researchers in the humanities and creative arts (2008:8). Within the field of biodiversity itself, disciplinary differences seem to exist (Enke *et al.,* 2012:28). When looking at faculty-rank/research-level differences, it was discovered that University of Nottingham research fellows, career researchers and other researchers (professors, assistant professors, research assistants, research associates, research officers, technicians and managers) were more likely than post-doctoral researchers or PhD researchers to have developed a DMP (Parsons, Grimshaw & Williamson, 2013:21). This possibly indicates that researchers at principal investigator level, in essence: researchers who are tasked with funding applications, tend to work more with DMP's than other researchers. Moreover, creators of highly sensitive data tend to make more use of DMP's than creators of less sensitive data (Martinez-Uribe, 2008:7).

In summary, when taking available studies investigating the use of DMP's into account, this researcher finds the use of DMP to be an activity revealing diverse use and perceptions. Its main drive would be policies, but its implementation and use can also be inspired by knowledge and support. In spite of this, aspects such as the time required creating such a plan, or the perception that a DMP might not be necessary, can hinder its implementation.

2.4.10 RDM training undergone by researchers

Ascertaining whether RDM training had been received by researchers, as well as the type of training that had been received (training sources, training complexity, training topics) were included as survey aspect in several studies.

Section 2.4.3., where this researcher reported on the state of RDM within studies institutes, provided a glimpse into what may be expected when level of researcher training was investigated. With the conclusion reached in mentioned section that RDM was found, in general, not be implemented at the time of the studies, and the conclusion reached that RDM tools as well as RDM as good research practice should be promoted/marketed, this researcher suspected that this section too would reveal a lack of RDM training in studied samples.

Very little to no formal RDM training received by the University of Oxford researchers as well as university researchers all over the USA (Patrick, 2012; Jahnke & Asher, 2012:3), the prevalence of scattershot training at the University of Minnesota (Johnston & Jeffryes, 2014:5), a mentioning of the varied training of graduate students at Purdue University (Carlson *et al.,* 2011:9), as well as the finding that most students at Northampton University had not received any formal RDM training (Alexogiannopoulos, McKenney & Pickton, 2010). These mentioned findings reveal an expected yet worrying issue: at the time of the studies, most respondents had not received any formal RDM training.

This finding does not mean that no training was received, or that no researcher had managed to obtain RDM-knowledge or experience throughout the course of their careers/studies. Several studies attempted to establish the sources of training supplied. Patrick (2012) reports on the investigation into the anecdotal evidence that Oxford University researchers might receive 'on the job' informal training, and found this to be true: formal training was not prevalent at all, with only one out of 11 RDM tasks revealing more than 25% formal training, while five of the 11 tasks showed at least 25% informal training. Still, no RDM task was shown to be more than half-trained, either way. Tasks most trained were day to day RDM activities (backing up, file-naming, etc.) while RDM-activities related to post-project tasks (preservation, sharing, IP, etc.) revealed the least training.

At the University of Minnesota (Johnston & Jeffryes, 2014:5), it was established that principal investigators (PI's) were expected to complete web-based modules in RDM. Unfortunately, modules were found to be in need of updating and displayed weaknesses with regards to data organisation, data sharing, and preservation information. While it was mentioned that the RDM skills of graduate students at this institute could be addressed elsewhere, concerns were also expressed that the timing of RDM training might be too early in careers, and were probably not constant or available at the point of need. A needs assessment showed that students had indeed not received formal training, but instead relied on peers, family, and previous experience for RDM direction.

Considerable variance in the training and training methods were detected when Purdue University researchers were studied (Carlson *et al.,* 2011:9). This haphazardness of RDM training had serious implications for university research: faculties were often not able to use data after students graduate. Although faculty researchers felt that some form of data literacy for students was needed, they expressed a reluctance and uncertainty in teaching them, and were concerned about getting too involved, or making the students' work too difficult. It could also not be articulated precisely what RDM skills were needed by students.

Different training sources were mentioned when Canadian university researchers were questioned about their RDM training received: 21% had done a research methods course, 22% had done another course where RDM was discussed, and 15% had attended a workshop where RDM was discussed (Doucette & Fyfe, 2013:168). No difference in training scores between masters and doctoral students could be established. Informal training sources were indicated as well: 56% of respondents educated themselves, while 33% indicated making use of self-directed learning. Informal RDM discussions with a colleague, used by 93% of respondents, can also be seen as a popular method of obtaining RDM information, in an informal setting.

When investigating the educational needs and wants of California Polytechnic State University researchers, it was found that 50% of respondents were not confident in their data preservation abilities, and as a whole, 20% of respondents engage in self-education (Scaramozzino, Ramirez & McGaughey, 2012:360). Of particular interest was the finding that only 7% of the non-confident group indicated educating themselves. In this non-confident group, (while not seeking to train themselves), a high figure of 70% indicated a strong desire for more guidance and education on best practices. Only 12.5% of researchers, who admitted to regularly educating themselves, expressed a desire for more guidance and training. The deduction made here is that not only do researchers differ in their training behaviour, but their subsequent requirements as a result of such behaviour are also influenced, and decidedly different.

In summary, studies analysed by this researcher indicated low levels of prevalence of formal RDM training received. Sources of such training were found to be workshops and courses, and could be either RDM-specific or include RDM as a sub-topic. In spite of, or perhaps, because of lack of availability of formal training, resorting to autodidactic training is an avenue followed by many researchers. Training preferences as well as RDM-information seeking behaviour were found to be individualistic traits, and not related to discipline, research group or institute (see more about this in section 2.4.11.2: Training tools/formats). These particular and distinctive behavioural aspects, related to training, are a feature of all RDM practices, and summed up impeccably by the finding of Gibson & Gross (2013:14), stating that even if training opportunities were offered, there can be no guarantee of their uptake.

2.4.11 RDM recommendations and requirements

Asking researchers about services required, and enquiring about RDM-related recommendations to be implemented, resulting in a better institutional RDM regime as well as improved personal RDM practices, were found to be a part of many RDM studies.

RDM is a complex activity, made up of various subsections (see all earlier headings in this chapter), and dependent on training, funding, infrastructure and support. As a result of this, suggested services were indicated as being essential and much-needed, plentiful, and varied. As could be expected, with studies being conducted prior to, or at the start of organisationally-supported RDM services, the majority of recommendations and service requirements were actually a request for training. The relatively low prevalence of RDM-specific training undergone by most respondents, and discussed under a previous heading, was a further reason why training and guidelines feature strongly when researchers were asked to suggest services needed.

This section consists of two subsections: services required/RDM recommendations, and training formats. This researcher is of the opinion that it is important to state that not all recommendations and suggested services can be mentioned here; instead an attempt will be made at highlighting the services generally shown to be in high demand, while at the same time taking note of services seemingly in low demand. The section on training formats is thought by this researcher to be not only required, due to training featured strongly as a requirement when services were suggested, but also necessary in that it reveals the diversity of training formats required, as well as preferred/not preferred, by different researchers and through different studies. Non-training-related requests and recommendations, such as requesting more time for RDM, as well as better RDM funding (Mowers, Humphrey & Perry, 2013) were few and far between.

2.4.11.1 Services required/recommendations put forward

In general, it would seem that relatively few RDM services, low levels of institutional RDM support, and equally low levels of researcher awareness of services and support mechanisms, were in existence when studies were conducted. Low levels of organisational training and support (see section 2.4.3, 2.4.10) as well as awareness of services which may indeed exist, were detected through analysing the literature. Stating that the University of Northampton researchers are '…ill or misinformed' about available services, and were worried about the lack of RDM guidelines (Alexogiannopoulos, McKenney & Pickton, 2010), that researchers at the universities of Cambridge and Glasgow experience difficulty in finding

relevant guidance when needed (Ward *et al.,* 2011:268), that there is lack of awareness of university RDM services (Nassiri & Worthington, 2012), the reliance of University of Minnesota researchers on peers, family and previous experience for direction (Johnston and Jeffryes, 2014:9), that University of Oxford researchers felt that little or no RDM support was available (Martinez-Uribe, 2008:10) and that there is a need for RDM '…advocacy and training' (Pink *et al.,* 2013:3) support the notion that at the time of the studies, support and awareness of support, tended to be problematic areas.

Suggestions for required services included training in basic, as well as technical or more advanced RDM aspects. A good example of this range in requirements can be found in the findings of a survey at the University of Oxford (Wilson & Patrick, 2010:32); researchers felt that they needed training in both broad data skills, as well as narrow RDM skills. Training in basic practical RDM issues, such as data storage and data creation were mentioned at many institutions. Researchers at several USA universities (Jahnke & Asher, 2012), the LSHTM (Knight, 2013:4) as well as Georgia Tech (Parham, Bodnar & Fuchs, 2012:12) formed part of this group. Similarly, requests for data management training (Beile, 2014), guidance on best RDM practice (Ekmekcioglu & Rice, 2009:5, Buys & Shaw, 2015:15), RDM assistance (Rankin *et al.,* 2012:32), and training in general RDM practices (Kouper *et al.,* 2013) accentuate the need for training in basic RDM.

As could be expected, assistance and guidance around network issues, as well as software and hardware problems (Beile, 2014), also featured when analysing requirements.

Storage and preservation services, as well as the guidance around those topics, would seem to be sorely needed. While pleas for adequate storage space were quite common (Jahnke & Asher, 2012:3; Parham, Bodnar & Fuchs, 2012:12; Gibson & Gross, 2013:13,15; Parsons, Grimshaw & Williamson, 2013:35), university researchers at Pretoria (Pienaar, 2011), Hertfordshire (Nassiri & Worthington, 2012), Nottingham (Parsons, Grimshaw and Williamson, 2013:35), LSHTM (Knight, 2013:4), Georgia Tech (Parham, Bodnar & Fuchs, 2012:12), Oxford (Patrick, 2012), as well as Carnegie Mellon (Van Tuyl & Michalek, 2015:1) reveal the need for data preservation aspects such as the need for training in data curation, and the need for options for secure preservation.

Adding on to storage and preservation services, assistance related to the RDM infrastructure (Ekmekcioglu & Rice, 2009:5), or a request for better infrastructure (Westra, 2010) were mentioned by respondents.

More specific infrastructural requests, mentioned as storage/preservation/data access-related services required, included:

- collaboration tools (Jahnke & Asher, 2012:3),
- sharing tools (Parham, Bodnar & Fuchs, 2012:12),
- data sharing services taking into account internal as well as external partners (Jones K, 2011),
- assistance with data sharing obligations and data sharing agreements (Knight, 2013:4) as well as Buys & Shaw (2015:15),
- a central university server or repository (Pienaar, 2011),
- backing up guidelines (Wilson *et al.,* 2011),
- online spaces able to accommodate large volumes of data and provide access control (Jahnke & Asher, 2012:3),
- facilities promoting open sharing and reuse (Open Exeter project, 2012:4),
- a facility able to showcase final data sets (Parsons, Grimshaw and Williamson, 2013:33),
- a facility providing one-stop data access across disciplines/collective data facilities (Mowers, Humprey & Perry, 2013),
- collective data facilities (Nassiri & Worthington, 2012), and
- an infrastructure able to allow the 'integration of multiple data sources' (Wynholds *et al.,* 2011:386).

In addition to these requests, a request was made at the University of Oxford for sustainable infrastructure allowing publication of data for those disciplines not currently served by domain-specific services (Martinez-Uribe, 2008:11). Looking at diversity of services required in this domain, it immediately springs to mind that data access, storage, preservation and its associated RDM aspects are issues of concern to many researchers.

Adding on to the requirement for collaboration and sharing tools, studies revealed the desire for training in areas such as:

- data anonymization (Delasalle, 2013 also Mowers, Humprey & Perry, 2013),
- issues surrounding data sensitivity (Nassiri & Worthington, 2012; Parsons, Grimshaw & Williamson, 2013:33),
- data security (Marcus *et al.,* 2007:18, Delasalle, 2013),
- copyright and IP licensing (JISC, 2014:8; Simukovic, Kindling & Schirmbacher, 2014:5),
- IP rights when publishing data (Martinez-Uribe, 2008:10, Knight, 2013:4), and
- data ownership and rights (Pink *et al.,* 2013:24, Wilson & Patrick, 2010:38).

Furthermore, the following access-related issues and requests were shown to be troublesome:

- data access in the researcher's own discipline (Mowers, Humprey & Perry, 2013),
- access to the activities of others (Mowers, Humprey & Perry, 2013), and
- being worried about privacy access control (Jahnke & Asher, 2012:3).

With the increasing need for a DMP and the relatively low usage of DMP's (at the time of studies, and as discussed in section 3.5.11), requests for DMP assistance and guidance were shown to be a required service at many institutes. Included in this group are researchers from three Australian universities (Henty *et al.,* 2008:8), the University of Oxford (Martinez-Uribe, 2008:11), the LSHTM (Knight, 2013:21), Georgia Tech (Parham, Bodnar & Fuchs, 2012:12), the Colorado University Boulder (Rankin *et al.,* 2012:32), Emory University (Akers & Doty, 2013:12 also Kouper *et al.,* 2013), Northwestern University (Buys & Shaw, 2015:12), the University of Toronto (Sewerin *et al.,* 2015:7), as well at Carnegie Mellon University (Van Tuyl & Michalek, 2015:1).

An issue closely related to the completion and understanding of a DMP, would be meeting the requirements of funders. With the previous paragraph indicating a definite need in the field of DMP's, the desire for assistance in understanding funders' requirements (Knight, 2013:21, Buys & Shaw, 2015:15 as well as Sewerin *et al.,* 2015:3), and in meeting funders' requirements (Parham, Bodnar & Fuchs, 2012:12) is to be expected.

This researcher discussed the relatively low prevalence of metadata usage in an earlier section in this chapter, and as expected, services related to metadata featured prominently on the wish-lists of respondents. Metadata training (Beile, 2014; Rankin *et al.*, 2012:32; Carlson *et al.,* 2011:11 and Van Tuyl & Michalek (2015:1), guidelines for creating metadata (Mowers, Humphrey & Perry, 2013), assistance with metadata description and discovery (Bardyn, Resmick & Camina, 2012:282), tools for managing metadata (Parham, Bodnar & Fuchs, 2012:12), and assistance with discipline specific metadata standards (Mowers, Humphrey & Perry, 2013) were indicated as areas in which there was a need for services and support.

Contrary to what was expected, was, that apart from USA researchers indicating their need for ethics guidance (Jahnke & Asher, 2012:13), and ethics training mentioned as being needed at Purdue University (Carlson *et al.,* 2011:14), ethics training as an RDM-area did not feature strongly. This researcher is of the opinion that this might be due to research ethics forming part of general research methods training at most of the institutes studied.

Non-training-related requests and recommendations, such as requesting more time for RDM, as well as better RDM funding (Mowers, Humphrey & Perry, 2013) were few and far between. Still, funding was in several instances indicated to be lacking. Study findings indicated that funding for a narrower concept such as data storage (Parham, Bodnar & Fuchs, 2012:12), as well as a broader issue such as more funds for RDM, including but not limited to RDM funding supplied by grants (Mowers, Humprey & Perry, 2013;Research Information Network [RIN], 2009:44; Fearon *et al.,* 2013), were areas in need of support.

Other requirements not related to training, not yet mentioned, and more directed towards infrastructural, procedural and organisational implementation/involvement, were the indication of a need for visible ethics policies (Mowers, Humphrey & Perry, 2013), the desire for other stakeholders, such as journals, publishers and data centres to be brought in (Mossink & Bijsterbosch, 2013:6), a suggestion that standardised documents and agreements might be a way of putting measures in place to manage industry expectations (Pink *et al.,* 2013:25).

Although data citation was not found to feature prominently as a researcher training requirement, indication of data citation training as a training requirement was found by Akers & Doty (2013:13), as well as Van Tuyl & Michalek (2015:1).

2.4.11.2 Training tools/formats

Throughout available studies perused, a variety of training tools and training formats were put forward. If anything, training tools and training formats suggested by survey respondents reveal the diversity that exists between researchers. This researcher, after trying to make sense of the variety of training formats and training tools, was not able to find a link between type of institute, subject discipline, and type of data dealt with. Having said that, this does not mean there are no correlations between type of training preferred and other researcher characteristics; this researcher just was not able to determine such a link as yet.

While researchers at the University of Edinburgh were requesting ‘expert support staff’ (Ekmekcioglu & Rice, 2009:5), LSE researchers were taking it one step further, saying that a discipline-trained data specialist at the end of their research project is a high priority, followed by the requirement of the same specialist at the start of the project (Raggett, 2012a:28). The need for helpdesk- style assistance (face to face, rather than telephonic), was also highly ranked, but lower than the need for a discipline-trained data specialist mentioned here. Researchers at the University of Oxford were in agreement with this, stating a high need for a RDM specialist to be embedded in research groups, and adding that such a person should have relevant experience in the research area (Martinez-Uribe, 2008:10). In

a similar vein, respondents from the Universities of Cambridge and Oxford felt that appointing a local RDM champion (someone with sufficient RDM knowhow) in the different research units, would be a positive step (Freiman *et al.,* 2010:5); an activity also planned and mentioned at the University of Hertfordshire (Nassiri & Worthington, 2012).

The need or preference to make use of online training tools proved to be another common finding. Researchers at the University of Edinburgh indicated their liking for RDM web pages (Ekmekcioglu, 2009), while an RDM website was mentioned by University of Nottingham researchers (Parsons, Grimshaw & Williamson, 2013:33). LSE researchers felt that web-based FAQs would be the most-preferred training tool (Raggett, 2012a:28), and while not indicating FAQs as their preferred training tool, researchers at the Universities of Cambridge and Glasgow showed an interest in FAQs, newsfeeds, email alerts, factsheets, crib sheets, diagrams and checklists, while also mentioning their preference for online resources such as online tutorials, video, and interactive learning (Ward *et al.,* 2011:269). The preference for online or electronic training tools is also required by Glasgow and Cambridge researchers; stating that simple, accessible visual guidance is required (Freiman *et al.,* 2010:6).

Specific requirements relating to training materials should be taken note of:

- use of jargon-free language is required (Freiman *et al.,* 2010:5),
- sources must be easily accessible (Freiman *et al.,* 2010:5),
- support has to be tailored, and relevant to the researcher (Ward *et al.,* 2011:269),
- advice must be brief and practical; long policy documents are not needed (Ward *et al.,* 2011:269),
- examples used and guidelines should be discipline-specific (Freiman *et al.,* 2010:5, Ekmekcioglu & Rice, 2009:5),
- training needs should be based upon actual research data problems commonly faced, and not promoted as generic RDM training (Wilson & Patrick, 2010:), and
- training materials should be customisable and relevant to specific faculties (Wilson & Patrick, 2010:38). The need for 'pragmatic assistance' requested by University of Edinburgh researchers (Ekmekcioglu & Rice, 2009:5) supports this idea that training should be to the point, practical, relevant, and discipline specific.

As far as requirements and preferences for non-online training materials go, somewhat of a variance in findings was shown:

- While a study of European researchers showed their need for support and training, it was stated that flyers, meetings, seminars and presentations were not regarded as

effective RDM training tools (Mossink & Bijsterbosch, 2013:48). Here, researchers were probably more inclined to prefer online training sources.

- Other studies showed a more positive stance towards non-online training: Emory University researchers mentioned that they were definitely in favour of faculty workshops on RDM (Doty *et al.,* 2013).
- Researchers at the Universities of Cambridge & Glasgow (Freiman *et al.,* 2010:6) showed a positive attitude towards making future use of one on one advice/training.
- Humanities researchers at Oxford, on the other hand, indicated a preference for face-to- face courses, but stated that it is necessary to include supplementary online content (Wilson & Patrick, 2010:43).
- Oxford humanities researchers also stated that it is important that an advisory service is able to deal one on one technical advice (Wilson & Patrick, 2010:43).
- Although LSE researchers, as stated in an earlier paragraph, indicated their first training choice to be web-based FAQs, their positive point of view on email/phones support, as well as data specialists assisting with RDM, needs to be mentioned in this section as well (Raggett, 2012a:28).

Regarding the ambiance and feel of training, there would seem to be an overall preference towards non-structured training. A good example would be the findings at LSE, where an aversion to all forms of formal training, except a 'one-hour training course', was revealed (Raggett, 2012a:28). LSE researchers further showed a clear preference towards training methods that are informal, ad hoc, and targeted at the researcher and their individual case.

Support for training sessions not being long, was found at the University Of Edinburgh (Ekmekcioglu & Rice, 2009:5), where respondents felt that training opportunities should be short and focussed. This sentiment was supported by researchers at LSE, who responded negatively to one day training (this group preferred one hour training), and training being part of a PhD course.

Some studies were interested in determining when RDM training should be given, and although these studies are not prolific, a few interesting observations were made. At the University of Exeter, the need for thorough, guided RDM training throughout the full project life cycle was desired (Open Exeter Project Team, 2012:4). Correspondingly, researchers at Hertfordshire stated to be in need of help in the whole project lifecycle (Nassiri & Worthington, 2012). In contrast, University of Oregon researchers stated that their greatest need is for assistance with RDM at the very beginning, or at the very end, of the data life cycle (Westra, 2010). LSE researchers however, showed a slightly bigger need for a data

specialist near the end of their research, than at the start of the project, while rating the need for both to be rather high (Raggett, 2012a:28).

When summarising the above sections (recommended services, as well as service tools/formats), a standout feature is the variety of services needed. Basic as well as more detailed, aspect-specific RDM services, with RDM training in particular, is required. Available RDM tools and services should be marketed and promoted. With regards to training tools and training formats, although variance was also indicated, it would seem as if researchers prefer to receive autodidactic training, want to be in control of training (at a time and pace suited to them, e.g. online sources), want training tools to be user-friendly, focussed, discipline-specific, and available at point of need.

2.4.12 Group differences in RDM: discipline, faculty rank/research experience

This study, while mainly interested in the RDM behaviours of emerging researchers, also aimed at establishing differences in RDM behaviours between different groups of researchers. This would be done by making use of data gathered during this study, data gathered during an earlier RDM study, as well as findings of studies found via this literature analysis.

This section of the literature analysis briefly discusses the findings of studies interested in similar differences. As the determination of group differences was most commonly not the main aim of RDM studies, detailed analysis of differences and findings will not be done here. Instead, RDM activities showing differences/no differences will be mentioned, as will the studies that did have as its purpose the ascertainment of RDM group differences.

A 2012 study at the Emory University showed an explicit interest in the differences among faculty ranks in their views on research data management (Akers & Doty, 2012:17-18). Findings can be summarised briefly as stating that there were no differences in the amount of research data stored, methods of storing, as well as backups made. Furthermore, faculty ranks could not predict sharing willingness, or method of sharing. However, faculty rank was found to predict reasons for not sharing: '…full and associate professors were more likely than assistant-professors and non-tenure track faculty members to state that it takes too much time or effort to share their research data' (2012:18). It would appear that junior faculty members are more inclined to viewing the sharing of research data to be an expected part of the research process; it would then not be viewed as imposing on their time.

Faculty rank differences pertaining to other data sharing reasons, data repository usage, and familiarity with data documentation and metadata, could not be established.

More faculty rank differences were revealed when respondents were questioned about their familiarity with federal funding agency requirements (2012:17): most non-tenure track faculty members displayed less familiarity than tenure track staff, and the reason for this might be found in their faculty positions focusing more on teaching as opposed to research grant reliance.

A study into the RDM behaviour of researchers across Europe (Kuipers & van der Hoeven, 2009) compared their sample's experienced respondents (more than 20 years' experience) to the group of respondents that stated having less than ten years research experience. Apart from a few small differences, only two big discrepancies were found between these two groups. First, novice researchers showed more eagerness in using the research output from other disciplines than experienced researchers, and second, (although usage is uncommon in both groups) novice researchers showed a more positive attitude towards online collaborative platforms than experienced researchers (2009:22).

A survey investigating the RDM practices at the University of Northampton (Alexogiannopoulos, McKenney & Pickton, 2010:16) divided researchers into three groups: research student, independent researcher, and group researcher/collaborator. Research students displayed less RDM experience than other researchers, and while more aware of technological developments than other researchers, more comfortable with changes in computer software/hardware, were in general more blasé about the need to look after research data. Furthermore, more established researchers tended to show greater RDM awareness than researchers with less experience, and were to be more willing to adapt to new RDM regimes if the benefits of such behavioural changes were demonstrated clearly. However, it was not clear if lack of research experience, or belonging to 'Generation Y', was the main contributing factor for these differences.

A single RDM experience-related behavioural difference was established when investigating LSE researchers, when it was found that researchers with more than nine years' experience would be more likely to have used data archiving services than those with less experience (Raggett, 2012a:26).

Reporting on disciplinary RDM differences seem quite prevalent in RDM literature, with many studies reporting on differences between discipline extremes such as basic sciences versus the humanities, and others indicating discipline differences after studying many different faculties and subject areas. Disciplinary differences were found in several areas, including RDM implementation and sharing, and will be discussed briefly.

In general, and broadly speaking, it would seem as if authors found better implemented RDM practices, understanding and awareness in the basic, natural sciences than in the humanities. This phenomenon was seen to exist at the University of Exeter (Open Exeter Project Team, 2012:4), as well across the Netherlands (Dillo & Doorn, 2011:4, 23). Furthermore, several studies, although not comparing the humanities with other disciplines directly, reported on the low levels of RDM uptake in this discipline. At the University of Oxford (Wilson & Patrick, 2010:5), humanities researchers were found to be mostly unaware of policies relating to RDM, and displayed little awareness of existing central services (Wilson *et al.,* 2011). A study involving five countries and investigating RDM in the humanities, found that this discipline displays a limited uptake of advanced tools for RDM and sharing, across all humanities disciplines (Research Information Network, 2011). At the University of Oxford (Wilson *et al.,* 2011), this discipline was also mentioned as not seeing themselves as working with data, as a large percentage of humanities researchers tend to only see numerical or tabular data as research data. The extent of this RDM misunderstanding is accentuated further by the fact that the Oxford survey had to adopt a broader definition of research data in order to deal with this dilemma.

Further RDM differences between disciplines, but not related to RDM implementation or state of the art, should also be mentioned. A disciplinary difference detected at Emory University relating to knowledge of data set size (basic sciences vs. arts and humanities), was discussed in section 2.4.4. (Akers & Doty, 2013:8). Further humanities research data characteristics also touched on, but not related to RDM progress, include the tendency of humanities data, more than research data in other disciplines, not to depreciate over time (Wilson & Patrick, 2010:5), the diversity of data and sources of data, and the fact that collections of humanities data are frequently incomplete/inconsistent due to the inconsistent nature of sources (Wilson & Patrick, 2010:10). Patrick (2012) adds a further disciplinary difference to the equation: 80% of humanities researchers do individual research, while teams are the more common trend in the natural sciences.

Basic science researchers at Emory University seem more likely to use repositories or data banks to deposit their data, than other Emory researchers (Akers & Doty, 2013:11). Furthermore, medical researchers showed a very high interest in starting to deposit their data. However, the same authors mention that fewer arts and humanities repositories exist; this difference in facilities and services is also mentioned in a report investigating researchers' data sharing and data publication activities (Research Information Network, 2011). Adding on to this, it was found that researchers in the atmospheric sciences, environmental sciences, and ecology were more likely to report having RDM support, and a

formal established process during the life of their projects (Tenopir *et al.,* 2011:13). Here, social science researchers reported the least support and implementation of a process.

Tenopir *et al.* (2011:13) were also interested in the disciplinary differences with regards to RDM support after the life of the project, funding, and data access. Findings show that half of atmospheric scientists have a formal post-project process, compared to 33% of social scientists, with computer science and engineering, biology and medical science also reporting a low prevalence rate. Only 23% of social scientists, compared to 42% of atmospheric scientists, feel that they have the necessary funds; for long-term funding the social scientists indication was 15% compared to the 22-27% stated in other disciplines. As regards data access, 80% of social sciences agreed that lack of access to research data is a major research impediment.

Sharing of research data was found to be related to discipline: medicine and social science researchers worldwide were less likely to share electronically, while 90% of atmospheric science researchers, and 85% of biologists, would share their data (Tenopir *et al.,* 2011:14). Adding on to this, a study investigating UK researchers (Key Perspectives Ltd., 2010:5) and looking at RDM traits of the arts and humanities, the social sciences, the life sciences, and the physical sciences, mention that arts and humanities researchers do not publish a great deal of research data, although the number of sharing researchers in these fields, are growing. A troublesome concern, and also indicated in an earlier paragraph, is that there is no abundance of data repositories in these fields; this study mentions the disaggregation of data repositories with the withdrawal of Arts and Humanities Research Council in 2008, to be a major factor in the low prevalence of data sharing and data publishing in these fields.

An important issue, relevant to data sharing in the social sciences, relate to the possible fallibility of anonymization techniques (Key Perspectives Ltd., 2010:6). Confidentiality in the social sciences tends to be an issue, and can be a barrier to sharing. Damaging of relationships during longitudinal studies, no data access control, insufficient data protection, and complexity of datasets are hindrances to data sharing, while extracting as much value as is possible from a dataset is a common social sciences data practice.

In contrast to the social sciences findings in the Key Perspectives study (*ibid*.), sharing in other disciplines (life sciences, physical sciences) was found to be more common. Having said that, disciplines within the life sciences, such as neuropsychology, where de-identification of data is required, present sharing problems similar to the social sciences. Similarly, a physical science discipline such as astronomy showed advanced and formalised sharing structures, while a discipline such as design and manufacturing revealed data sharing paucity (*ibid.*). Funding was seen to be a problem in the life sciences, while data set

sizes and the need for compression, often leading to loss of data as it needs to be compressed 20 times, caused concern. Recently, good lossless compression, resulting in data compressed only three times, have been developed. In summary: the Key Perspectives study (*ibid*.) showed that disciplinary differences with regards to RDM behaviour do exist, and have portrayed these interdisciplinary differences by looking at the arts and humanities, the social sciences, the life sciences and the physical sciences, and discussing their particular findings under the headings of sharing, data discovery and data preservation.

Interesting findings emanating from a study into visual arts research (Garrett *et al.,* 2012:8), is that the term 'research data' is often not understood by visual researchers, as artistic research is a relatively new research field. As a new research field, displaying low levels of RDM awareness and uptake, and being part of a bigger discipline (i.e. humanities) already seen as the black sheep of RDM disciplines, the visual art can be seen as an area in need of RDM implementation, and this author predicts great progress and RDM advances in subsequent RDM studies involving visual arts researchers.

In summary: discipline-specific as well as research experience/faculty rank-specific differences, related to RDM behaviour, were found to exist. Differences were more pronounced and varied in the latter than in the former. Although humanities disciplines generally displayed low levels of RDM uptake, RDM practices in other disciplines were also prone to variance. Nevertheless, disciplinary RDM differences are part and parcel of the RDM situation of most multi-disciplinary institutes, and the reasons for these differences are often complex and research-culture-based. When trying to explain these differences, it might prove worthwhile to consider aspects such as heritage and practices of niche research communities, type and quantity of data, uniqueness of data, policies of funding bodies, as well as the provision of storage infrastructure, to name but a few possible contributing factors (Key Perspectives Ltd., 2010:25-26).

2.5 Limitations and concerns

This researcher has made use of a wide variety of literature sources, with sources spanning more than a decade, and covering a wide range of researchers, institutes, disciplines, and geographic localities. This diversity in studies has provided this researcher with insight and understanding into the broader RDM practices of researchers, problems faced by them, and services required in order to improve an organisation's RGM setup.

It needs to be mentioned that **specificity** of studies should always be taken into account. This specificity pertains to study sample, survey tool used, organisational support, funding available, research discipline, and funder requirements, to name but a few variables.

Apart from these variables, behavioural aspects such as research culture, individual preferences, personality and research style of group leader, could also influence RDM displayed within a sample.

These aspects are all relevant when **generalising** about RDM behaviour, or trying to establish trends within a specific research group, discipline, or institute:

A second concern held by this researcher, relates to **multifaceted nature of RDM.** The RDM studies analysed by this researcher has shown that diversity was a common feature found, and that due to the nature of questions fielded and topics broached, **direct comparisons** are not always possible. Therefore, in practice, whereas many studies would be interested in backups, some would enquire whether backups were made, others would be interested in regularity of backups, or storage space needed for backups, or storage media used when making backups. While several different studies might have investigated RDM backups, it is often the case that the complexities surrounding backups, and the various sub-activities contributing to the concept of 'data backups', result in survey questions differing between studies. In effect, this will contribute to intra-study inadequacies and deficiencies. Nevertheless, these intricacies, although not always directly comparable, do contribute to the reader's knowledge, awareness and perception of the topic.

2.6 Chapter summary and conclusion

This chapter aimed at analysing and discussing scholarly studies relevant to the subject of RDM practices as displayed by researchers. Standout features of literature surveyed, spread out over more than a decade, revealed wide-ranging findings: RDM practices could be non-existent, or still in its infancy, or showing a good deal of institutional support and involvement. This description can be used for all RDM behavioural aspects, be it backups, support, training supplied, adherence to DMP's, use of metadata, or data sharing practices.

Establishing trends is not an easy task; generalisations are to be steered away from when study samples, survey tools, and survey schedules (i.e. RDM topics investigated) are as diverse as this literature collection turned out to be. Each of the separate topics examined, be it behavioural in nature (e.g. use of metadata, data sharing practices, use of a DMP), or topics related to non-behavioural aspects (e.g. sample size, study date, geographic locality, disciplines included) managed to show that RDM practices cannot be forecast, generalised

or pigeonholed. Adding onto this: variables playing a role in RDM behaviour are numerous and real: organisational support and infrastructure, institutional policies, funder's requirements and, to a lesser extent, RDM training, are all factors contributing to the probability of a researcher applying RDM, or adhering to good RDM practices.

What can be said with certainty, is that most researchers realise the value of good RDM practices, will benefit from RDM training and RDM training materials, and would welcome institutional support and infrastructure. Having said that, it is almost certain that 'time and effort required', as well as 'funding required', mentioned as a data sharing constraint, would feature as an RDM constraint too. It therefore seems quite plausible to foresee that the implementation of an institutional RDM policy, and adherence to it, would benefit from RDM tools, infrastructure and materials being funded, user-friendly, and streamlined within the research process.

With regards to her own RDM study, this researcher was in two minds as to what to expect when surveying the RDM practices of emerging CSIR researchers. As the CSIR currently has no official RDM regime, meaning that there is no CSIR-wide policy, procedures, funding, infrastructure or training materials, it could be that RDM practices displayed by these researchers will not be adhering to best practices, not be consistent, and may in fact be totally absent.

On the other hand, with emerging researchers working in different subject disciplines, in different research groups, in different CSIR units, having studied at different universities, working under different group leaders or study supervisors, and displaying different academic personalities and practices, it could be that many emerging researchers are already adhering to good RDM practices.

It is expected that the sample to be studied by this researcher, as was the case when different studies were compared, will reveal RDM behaviours portraying variety and extremes, and for good measure, some interesting and peculiar habits and practices.

The next chapter will focus on giving a detailed and thorough description of the methodological steps used by the researcher during this study.

3. Chapter 3: Methodology

3.1 Introduction

This chapter deals with the methodology implemented during this study. It was this researcher's intention that after studying this chapter, the reader will have been given a clear indication of the means by which this researcher achieved her research aims.

To achieve this goal, this researcher has described and elaborated on the research design, and the methods and procedures used in this study. Each section or topic within this chapter needed to contain enough detail, or portray enough information, to enable a reader of the complete chapter, to repeat the study. As such, this chapter is a detailed description of the methods and procedures used, with its desired outcome the understanding of the research design, methods, procedures followed, as well as the reasons for implementing those strategies. In short: this chapter is a detailed account of 'how' this researcher investigated that which she said the dissertation would focus on.

Aspects covered in this chapter range from broad theoretical discussions about research approaches, to very narrow explorations of questionnaire content and phrasing. Chronologically, this chapter starts with a discussion of the research approach used in this study, and then moves towards discussions of the research method used, followed by a section on the data collection tool used. Thereafter, the contents or makeup of the data collection tool are discussed. Data tool administration involving aspects such as recruitment, communications with respondents as well as questionnaire dynamics, also feature in the methodology.

Besides the above, this chapter covers topics such as the target population involved, leading to a discourse on the sample chosen. In addition, methodological aspects such as ethical considerations, as well as the obtaining of ethical clearance, are addressed.

A methodology chapter cannot be considered complete without revealing how data will be analysed, and the ways in which it will be introduced to the reader. As a result, the latter part of this chapter explains the data analysis methods followed, as well as the data presentation methods and options used.

As many diverse methodological aspects are covered in this chapter, this researcher hoped to provide a clarifying and unifying view of the methodology followed by concluding the chapter with a chronological diagram.

Readers of this chapter will notice that there are frequent references to appendices; these attachments are meant to assist the reader when a topic is discussed, by showing a copy of the document being referred to. An example of this is found in section 3.6.3, which focusses on the questions asked and RDM topics included in the online questionnaire. Within that section, reference to Appendix 2 is made as that attachment contains a copy of the online questionnaire, albeit in Microsoft Word format.

3.2 Research approaches

A common practice or tendency found by this researcher when browsing through theses and dissertations was the discussion of the main research approaches in the research discipline, and its applicability to the study in question. Keeping in step with tradition, this researcher embarked on a similar undertaking. By looking at the main features of main research approaches, but specifically at the ways in which they overlap, she attempted at demonstrating that research approach pedantry is often not a desired trait when discussing research methodology.

As is most commonly done, the three main approaches namely quantitative research, qualitative research, and mixed-methods research were looked at, and their main features noted. Thereafter, this researcher correlated these features with the planned features of her study approach/design, in an effort to determine the applicable research approach. Reasons for the choice, as well as problems faced when making the decision, are supplied.

3.2.1 Quantitative approach

When looking at the short descriptions of quantitative research, in order to create a practical, abbreviated version containing the stand-out features of the approach, this researcher noticed the following standout features:

- Quantitative methods are those employed by study researchers and experimenters who rely on **numbers and statistical methods** (Singleton & Straits, 2005:75).
- It is also referred to as the **'scientific' method**, focussed on measuring and summarizing the characteristics of entities being studied (Wallace and Van Fleet, 2012:25).
- Quantitative research assumes the **objective reality** of social facts (Pickard, 2013: 13).
- The quantitative process is far more linear than the qualitative approach. A hypothesis, although common, is only required when true experimental research is

chosen as the method. In the absence of a hypothesis, **research aims and objectives** will suffice (Pickard, 2013:18).

- **Statistical methods** are used as analytical tool. **Data reduction** (a technique used to transform raw data into useful data) is essential to quantitative research (Wallace and Van Fleet, 2012:25).
- **Results** are presented in a **fixed** way, and are vital in quantitative research. This is in stark contrast to the 'why' and 'what' focus of qualitative research (Lichtmann, 2014:13).
- The researcher is seen as an **objective, neutral observer**; the statement is made that any other researcher would view the phenomenon the same way (Wallace and Van Fleet, 2012:25, as well as Lichtmann, 2014:13).

The points listed above can be seen as an abbreviated account of what constitutes the typical quantitative approach. However, this is by no means an exhaustive list of the characteristics of quantitative research.

3.2.2 Qualitative approach

When viewing brief descriptions of the qualitative research approach, the researcher encountered the following short explanations:

- Researchers typically make use of **intensive interviews, participant observation, and depth analyses** of historical materials (Singleton & Straits, 2005:75).
- Study focus is on the **particular**, rather than generalizable (Wallace & Van Fleet, 2012:27)
- The approach is often termed **'naturalistic'** (Wallace & Van Fleet, 2012:25).
- This approach emphasizes **depth and complexity**, rather than quantity (Wallace & Van Fleet, 2012:27).
- This approach tends to be more **exploratory** than conclusive in nature (Wallace & Van Fleet, 2012:27).
- Often, there is **no detailed plan** before research begins; the study has an **iterative** nature (Pickard, 2013:14).
- The **'why' and 'what'** of human behaviour feature strongly in qualitative research (Lichtmann, 2014:12). This is in stark contrast to the results-orientated outlook of quantitative research. **Themes**, and not stark data results, are sought.
- This approach encourages the presentation of findings in **alternative styles** (Lichtmann, 2014:13).

- **Data reduction is often not necessary** nor encouraged (Wallace and Van Fleet, 2012:25).

Looking back at a previous CSIR RDM investigation (Patterton, 2014a), this researcher recognised this approach as the one used in the previous study. Detailed interviews were used, the interviewer often did not stick to the interview schedule but adapted and adjusted, depending on interview progression, topics covered, interviewee answer complexity as well as topic dwindling.

3.2.3 Approach of this study: 'quantitative, but not at the end of the spectrum'

This researcher encountered numerous sources where researchers refer to or create tables showcasing the differences between quantitative and qualitative research, examples being the comparative tables supplied by Schurink (2000:242-243), and Lichtmann (2014:17). These tables are used to demonstrate the obvious differences between the two main approaches, and serve as proof that these two major research and evaluation paradigms are, to quote Wallace and Van Fleet (2012:25) 'competing for dominance'. This researcher was not intent on repeating mentioned tables here; these tables are readily available, easily understood and often end up being quite comprehensive and detailed in size and content. For the purposes of this study then: while this researcher was aware that she had resorted to very short descriptions of the two major research approaches, she was hopeful that the descriptions supplied will enable the reader to distinguish a typical quantitative approach from a typical qualitative one.

While not creating an elaborate contrasting table here, this researcher regarded it as essential to acknowledge and emphasize that differences between these two approaches do exist. This was found to be especially true in extreme cases, for instance where a quantitative study, having a hypothesis, and making use of a true and rigid experimental method, is contrasted against a qualitative study where intensive interviewing, open to adaptation and adjusting, is interested in exploring a new phenomenon. While these differences are real and obvious, the two research approaches do not often display this level of disparity. Indeed, some studies would incorporate elements belonging to both approaches and term their approach a 'mixed method', while others design their study in such a way that it is not possible to tell at first glance whether a quantitative, qualitative or mixed research approach was followed.

Wallace & Van Fleet (2012:26) state that the quantitative and qualitative approach are often treated as being opposites, and states that this viewpoint resembles 'opposing armed camps' (2012:26). In agreement with this, Pickard (2013:13) is of the opinion that they are

viewed as two conflicting paradigms. While this is the outlook of many, Wallace & Van Fleet (2012:26) further mention that others view all research approaches to be places on a continuum, with the quantitative approach at the one end and qualitative at the other. According to these two authors (2012:26), it might be more accurate to view different research approaches as displaying different perspectives, and state that it is indeed possible that any given phenomenon can be examined from either a qualitative or quantitative perspective, or even a combination of both. Pickard (2013:13) is in accord with this, and indicates that for many, methodological dualism (i.e. mixed methods approach) is seen as the only pragmatic option. Thus, the mixed methods approach is often not regarded as a separate methodology (2013:14), but rather as a method for combining the existing quantitative and qualitative methodologies in various ways to address various research questions.

Lichtmann, when describing the mixed methods approach (2014:83), and the various ways in which different approaches can be combined, state that some qualitative researchers even combine two or more qualitative approaches, thereby creating a new research approach. She attempts to summarise the various ways in which qualitative researchers can choose an approach, and as examples mention that researchers can now choose between either a philosophical/theoretical approach as opposed to a research approach, can choose a particular qualitative approach, can select a combination of qualitative approaches, can select a combination of qualitative and quantitative approaches, or can use a generic approach, not selecting any particular research approach. The mixed methods approach is described by Pickard (2013:18) to be an approach taking on many forms: there is no fixed mixed-method design; this approach has many and varied combinations.

This researcher, while contemplating the topics and issues to be included in the yet-to-be-written methodology chapter, at first felt compelled to ascribe a particular research approach category to this study. It was clear to this researcher that a previous RDM study at the same institute (Patterton, 2014a), conducted by this researcher also but involving a different target population (established researchers versus the current study's focus on emerging researchers), could be seen as being a qualitative approach. In-depth interviews were the data collection tools, and questions were open-ended in nature. Furthermore, the interviewer did not stick to a rigid interview method, or question order, but adapted and adjusted as was needed. Often, respondents would misinterpret a question and it would be rephrased, or explained by way of examples. At other times, respondents would answer a question before it was asked, rendering the asking of a question to be redundant. Analysing the answers, coding the answers, and assigning categories to the answers were activities performed by

this researcher. These mentioned tasks, activities and features are all very typical of the qualitative approach.

Still on the topic of the previous RDM study (Patterton, 2014a): it would be hardly realistic to assume that all researchers would have performed this study similarly, viewed the phenomenon in the same way, would have received identical responses and data, or ended up with the same categories, codes, results and findings. As stated in section 3.2.1, a key feature of quantitative research entails the objective, removed stance of the researcher: the study was conducted in such a way that all researchers repeating the study would have an identical viewpoint of the phenomenon. A study relying on one-on-one interviews, dependent on interviewer prompts and actions, and with data analysis and results probably influenced by investigator bias and interpretation, can in no way be seen to be part of quantitative research. Therefore, the previous CSIR RDM study portrayed many features of qualitative research, no features of quantitative research, with the result that categorisation of the study approach is a straightforward task.

The current study, once again a study into RDM practices of CSIR researchers, differs from the previous study on various levels. These differences have resulted not only from the researcher's dissatisfaction with methodological aspects related to the previous study's design and approach, but can also be attributed to current time constraints, and increased accessibility as well as improved functionality of other data collection tools. Although these methodological aspects are discussed in more detail later in the chapter, the following aspects are briefly mentioned with the aim of portraying the research approach specific to this study:

- This study, although not having a hypothesis, had **research objectives and aims.**
- This study investigated the behaviours of a **particular** group; results were **not to be generalized.**
- This study made use of a **detailed research plan**. A sample was identified, a data collection tool was designed (online questionnaire), and the data collection technique and process were structured and undertaken systematically. All respondents were asked the same questions in exactly the same order.
- Although data analysis involved numbers as well as words, the representation of data by means of graphs in effect reduced the **data to numbers**. **Data reduction** was critical to this study. Having said that, several **quotes**, supplied by respondents in response to an open-ended question, were used as is in the results chapter.
- This study, while focussing on the **'what' of RDM,** was not concerned with the 'why' of RDM. Detailed philosophical and theoretical discussions did not form part of this

study. Having stated this: the fact that the questionnaire contained open-ended questions, and that answers supplied in fact added a qualitative slant to the survey, cannot be ignored.

- The researcher was **mostly, but not entirely, an objective bystander**. It was assumed that, apart from the open-ended questions, where the researcher might have assigned categories and codes to the quotes given, she had interpreted the questionnaire answers and the phenomenon exactly like another researcher would have done. Conversely, the structured research design and tool, plus the questionnaire's closed-ended questions, assumed the viewing of the phenomenon by other researchers to be similar to this view obtained by this researcher. The percentage closed-ended questions content to open-ended question content was estimated to be 90:10.

Taking the above into account, this researcher could not, as with the previous study, refer to this study as being purely qualitative in its approach. At times, while the chapter was being written, this researcher viewed this study as more quantitative than qualitative. This ambivalence in thought pattern stemmed mainly from the fact that the research was mostly structured, the researcher was mostly an objective bystander, and the data were mostly reduced to numbers. While these factors are most often associated with quantitative research, this researcher was also not convinced that this study qualified to be termed as purely 'quantitative' in nature. She adopted a pragmatic approach instead, and adopted the following viewpoint: this study could be seen as mixed-methods research, as it had elements of both quantitative and qualitative research research within the same study.

To summarise: the features and characteristics of this study showed it to have qualitative as well as quantitative research elements. The overlapping of features, the absence of clear-cut categories and the emergence of the mixed methods approach has led to a situation where the labelling of this study was not only a difficult task, but this researcher was also not convinced that the labelling of was indeed required. Therefore, while happy to label her research as qualitative to meet requirements, she regarded the activity of research approach 'choice' to be not vital to the methodology chapter. In a research environment and field characterised by dynamic changes in data collection tools and data collection methods (to name but two methodological aspects), it is unrealistic to expect research designs to neatly fall into a specific category. As was seen in this study, the need to place the study into a research approach category actually had this researcher categorising her own study as quantitative, mixed methods, as well as qualitative during different phases of this chapter writing. She regarded the activity as being confusing, and marred by too many ambivalent options.

3.3 Research method: case study vs. survey

While the data collection tool used in this study was the questionnaire (see sections 3.4 and 3.5), the specific research method to be used was an issue given serious thought by this researcher. A case study, being either a quantitative or qualitative way of studying 'the particular within context' (Pickard, 2013:101), or 'the specific and detailed study of a case' (Lichtman, 2014:119) at first seemed a method ideally suited to this topic. A longer definition of the method was also useful: Wallace & Van Fleet (2012:217) describe it as a method used by the researcher, when, instead of studying many instances of some phenomenon to reach an inductive conclusion about the phenomenon in general, she would study one instance/case in great detail as a means of explicating the phenomenon. Fouche & De Vos (2000:125) mention that any type of research method can be used in cases studies, but state that data are typically obtained through observation, in interviews with key informants, and from available documents. According to Wallace & Van Fleet (2012:217), activities commonly included when the case study is used in library and information studies, would be direct contact with the case entity, visits to the case, interviews, examination of the facility, an analysis of routines, determination of trends, and other useful activities designed to build a complete picture of the case. As stated by Fouche & De Vos (2000:125) it is presumed that a thorough investigation and description of the case unit (being an individual case, group or community) would enable a researcher to develop insights, ideas, questions and hypotheses for further study.

Such a case study, if applied to this study's topic, would entail the researcher focusing on one emerging researcher (or a very small research unit, at most), having direct contact with the researcher while observing their daily activities and data management activities, having in-depth interviews, physically examining their data documentation and data metadata, inspecting their data archives or repositories, and sitting down with them to get a picture of the step-by-step procedures followed when dealing with a dataset. These activities would be in agreement with Pickard (2013:109), who is of the opinion that focused exploration, by means of interviews, observation, and content analysis of documentation are important means of obtaining data during a typical case study.

In spite of the many advantages of a case study, including level of detail and the potential for serendipity (Pickard, 2013:218), this researcher was not convinced that using the case study as research method would be ideal for this study. Possible negatives, such as the time required for data collection as well as data analysis, possible softness of data, the potential for bias, low potential for reliability and the uncertain potential for validity (Pickard, 2013: 219) played a role in making this researcher realise that other methods should be

investigated. The chosen research method should be able to fit into this researcher's planned data collection timing slot (six weeks at most), be a method more valid and reliable than the case study, allow for the researchers to take part in the study without feeling intruded upon, and not be a method where case bias or investigator bias would have influenced the study. When taking these requirements into account, this researcher made a decision to use the survey as research method in this study.

To summarise what has been discussed so far: while the case study and survey are often seen to be very similar, there are definite stark differences between the two methods. Pickard (2013:117) is of the opinion that the biggest difference between the two is found in in consistency and structure: with surveys, questioning must be consistent, meaning that the entire sample must be asked the same questions in the same way, and usually restrict their potential responses within pre-coded parameters. In other words, surveys are designed to produce a generalization within a target population, while the case study is concerned with individual perceptions, beliefs and emotions.

Before providing more details on the survey as research method, it is important to state that while this study did not use the case study as research method, this study may indeed still be seen by readers to be a case study. Several reasons can be given for this perception: details of how to conduct a case study are often not spelled out, and a case study could indeed involve more than one data gathering method. It could also include just one data gathering method, meaning that this study, using a questionnaire to collect data, could be seen as an investigation into a 'particular person, program, curriculum or technique'; in other words, a case study. While agreeing with this in principle, this researcher was of the opinion that when limiting the data collection tool in this study to a questionnaire only, and not engaging in typical in-person interviews, observation and direct contact, it should be categorised as closer to a survey than a case study. While the study indeed has some characteristics of a case study, it was viewed by this researcher as more typical of a survey.

Survey research exist in many forms and formats (Hank, Jordan & Wildemuth, 2009:256), and supports the collection of a variety of data, including the beliefs, opinions, attributes and behaviours of respondents. Wallace & Van Fleet (2012:146) define a survey as a 'descriptive study' of a 'sample', conducted as an approach to understanding the 'qualities of a population'. In agreement with this, Pickard (2013:111) states that the purpose of survey research is to gather and analyse information by questioning individuals who are either representative of the research population, or are the entire research population. With these two descriptions in mind, this researcher, interested in gathering and analysing information on the RDM practices of the CSIR's emerging researchers by questioning individuals who

are representative of this target population, or by questioning the entire population, viewed her research method as being typical of a survey.

Survey research enables researchers to 'estimate the distribution of characteristics in a population' (Dillman, 2007:9); this description, when paraphrased and applied to this study, would indicate that using survey research would enable this researcher to estimate the distribution of RDM practices and behaviours among emerging CSIR researchers, by making use of qualitative research and also involving aspects of quantitative research.

Several characteristics of survey research resonate with this researcher:

- Survey research can include qualitative and quantitative research (Pickard, 2013:111). As discussed in section 3.2.3, this study has elements of both quantitative and qualitative research. Survey research suits the dichotomous feature of this study to a tee.
- There are options for data collection within a survey; questionnaires and interviews are two examples of such tools (Singleton & Straits, 2005:219; Pickard, 2013:111). This researcher made use of a questionnaire as data collection tool, and decided on this practice after using interviews in a previous survey, and evaluating its suitability to this study. The questionnaire as data collection tool is discussed in more detail in section 3.4.
- Singleton & Straits (2005:221) identify different types of surveys: unstructured, structured and semi-structured. Looking at the types listed, this researcher described the survey technique in question as structured, as all objectives were very specific, all questions were written beforehand, and all questions were presented in the same order to all respondents. Furthermore, the sections comprising the introduction, the closing as well as the bridging sections were structured and similar for all questionnaires and respondents. These features are discussed in more detail in section 3.6.
- Surveys may further be categorised according to their aims; Pickard (2013:111) distinguishes between descriptive surveys and explanatory surveys. This study was concerned with the description of trends and patterns within the sample group (as is the case with descriptive surveys), rather than establishing a cause and effect relationship without experimental manipulation (as is the case with explanatory surveys). This approach does not lend itself to sophisticated statistical analyses, nor is it the survey type's purpose (Pickard, 2013:112). This researcher, using a descriptive survey, was more interested in an interpretation of the results and in establishing trends and patterns, based on the facts gathered via the questionnaires.

- Probability sampling is vital when generalising to the target population; when using non-probability sampling care must be taken when generalising to the wider population (Pickard, 2013:111; Singleton & Straits, 2005:219). This researcher made use of a non-probability sampling technique; this concept is discussed in more detail in section 3.7.1.
- Answers are numerically coded and analysed (Singleton & Straits, 2005:219). This study made use of basic numerical coding and analyses; methods are discussed in section 3.8.

When discussing surveys, Singleton & Straits mention that exceptions to the general survey rule can be found for each survey feature (2005:219). A good example of such an exception is found within this study: as seen in the section discussing sampling (section 3.7.1), and in contrast with the typical sampling method stated in the fifth bullet above, this researcher made use of non-probability sampling technique. The rationale behind this is discussed in section 3.7.1 as well.

Hank, Jordan & Wildemuth state that although survey research is a commonplace method, surveys are not always simple to design and administer (2009:256-257). Attention to a series of critical components ensuring the effective collection of data, and its implementation is deemed vital by stated authors. Aspects such as the survey design, pretesting and pilot testing (with revisions if necessary), survey administration and data analysis are stated to be aspects necessitating due attention and scrutiny. These aspects are all covered within this methodology chapter.

The survey process follows definite steps; these steps, according to Pickard (2013:114-116) occur in the following order:

- identification of a general topic area: this researcher, after attending a CSIR Emerging Researcher Symposium, felt that the RDM practices of this group of researchers were worthy of investigating,
- investigation of the literature: this has been done; an expansive literature review can be found in Chapter 2 of this dissertation,
- establishing a hypothesis or aims and objectives: this has been done, and is discussed in detail in Chapter 1 of this dissertation,
- identification of a suitable population: see section 3.7.1,
- application of an appropriate sampling technique: see section 3.7.1,
- selection and design of a data collection instrument: see sections 3.5 and 3.6,
- piloting the data collection instrument: see section 3.6.6,

- collection of data: see section 3.7.4,
- analysis of data: see section 3.8 as well as Chapter 5, and
- presenting findings and conclusions: see Chapters 5 and 6.

As can be seen from the above list, the chronological steps followed in this methodology chapter, as well as the chapters within this dissertation, closely mimic the survey steps outlined above. For this reason, the survey steps are not discussed in more detail here, as the reader is referred to the relevant detailed discussion of the stage elsewhere in the dissertation.

This researcher was well aware of possible limitations present when implementing the survey approach. Aspects mentioned by Singleton & Straits (2005:227), such as the survey's inability to establish a cause-and-effect relationship, or its inability to enable the researcher to change the course of research after the study has begun, have been taken note of. She was of the opinion that this study's non-interest in cause-and-effect, its viewing of association between variables as of secondary importance to this study, and her previous experience with questions found to be important when determining RDM behaviour, should alleviate the dangers of stated limitations. A further limitation, namely possible lack of respondent truthfulness, was an aspect beyond control of the researcher, and hopefully lessened by emphasizing in the questionnaire introduction that confidentiality would be guaranteed (see Appendix 7). Singleton & Straits (2005:227) also mention the danger of question misinterpretation; this researcher addressed this issue by having her study supervisors evaluate and revise the questionnaire by means of a pre-test, and submitted the evaluated/revised questionnaire to two CSIR researchers during a structured pilot-test. The pre-test and pilot tests are discussed in more detail in section 3.6.6.

3.4 Data gathering tool: the questionnaire

With the research method of this study established as a survey, the next step entailed determining which data gathering technique would be used. While studying survey-related methodology, this researcher noticed the existence of a multitude of data gathering techniques. Surveys may indeed be administered in a variety of ways; the oldest and most highly regarded method of survey research being, according to Singleton & Straits (2005: 236), the face-to-face interview. Hank, Jordan & Wildemuth (2009:259) mention that other popular administering techniques include mailing the survey, emailing the survey, using an online survey, using a synchronous online chat session, telephonic survey, interactive voice response survey, and other variations on these approaches. Wallace & Van Fleet (2012:180) consider the most popular techniques to be the in-person interview, telephone interviews,

questionnaires, direct observation, and case study. Each of these methods has its own advantages and disadvantages, and it was up to this study's researcher to decide on the method best suited to her study.

This study made use of an online survey questionnaire to collect data. While the two terms 'survey' and 'questionnaire' are often seen to be exchangeable, this researcher uses the term 'survey' when discussing the data gathering method in a broader sense, while 'questionnaire' as well as 'online questionnaire', 'electronic questionnaire' and 'web-based questionnaire' are used when discussing the data gathering technique, or tool, in finer detail (see sections 3.5 and 3.6 for questionnaire discussion). This is in line with Pickard emphasizing that although the terms 'survey' and 'questionnaire' are often used interchangeably, they are not synonymous. Pickard (2013:117) states that 'survey' is a research method, while 'questionnaire' is a specific data collection technique.

Deciding on choice of survey technique to be used during this study was not an easy task. Many options were available, and each was seen to be able to contribute to the expansion of information on the RDM practices of young researchers in its own unique, incomparable manner. After considering the options deemed practical and suited to this study, this researcher narrowed her choices down to two techniques. Having used the in-person interview during a previous CSIR survey (Patterton, 2014a), but seeing the benefits and advantages of the questionnaire as a data collection tool, it was considered appropriate and necessary to dedicate a section of this chapter to briefly comparing the interview method with the questionnaire method. In particular, the disadvantages of the interview technique, as experienced by this researcher during the previous CSIR RDM survey (Patterton, 2014a), as well as the advantages and limitations of the questionnaire as data collection technique, were be examined.

3.4.1 Interview disadvantages

This researcher, after using the in-person interviews exclusively during an earlier CSIR RDM survey (Patterton, 2014a), felt the need to consider alternative ways of collecting data during this study. While conducting in-person interviews, interview transcriptions and data analysis during the 2014 survey, several of the in-person interview's disadvantages were a source of concern for this researcher, and prompted her towards thinking that a future survey should use a more suitable method. In particular, the following disadvantages, also mentioned in survey literature, were found to be worrisome:

- Absence of anonymity (Wallace & Van Fleet, 2012:181): although confidentiality in survey results could be guaranteed, the potential for respondent anonymity during

the interviews conducted in the 2014 survey was non-existent. Thus, the danger of respondent bias, stress and pressure to respond in a certain manner throughout the interview was an ever-present danger.

- Unpredictability (Wallace & Van Fleet, 2012:181): it often happened that due to long-winded and off-topic answers, this researcher ended up with a long recording, filled with irrelevant data. This unpredictable trait of interviews, often making it difficult to get the session back on track, and requiring post-interview time and effort in order to obtain a valid transcription, was a disadvantage in need of eliminating in future surveys.
- Potential for interviewer influence (Wallace & Van Fleet, 2012:181): unconscious non-verbal cues from the study researcher might have had the potential to influence the respondents' answers (e.g. study researcher nodding, frowning, smiling, or writing furiously).
- Potential for respondent stress (Wallace & Van Fleet, 2012:181): although not intended, an unexperienced interviewer such as this researcher interviewing researchers during 2013, might have acted in an incompetent manner, or could have intimidated or near-insulted an already-stressed respondent. In addition, and in contrast to the previous bullet, it might be that even a non-emotional, objective, professional and aloof interviewer could have heightened the stress levels of respondents. As stated by Wallace & Van Fleet (2012:181): an 'ill-at-ease' respondent is often not a 'good source of information'.
- Intrusion of recording processes (Wallace & Van Fleet, 2012:181): during the previous CSIR RDM survey, this researcher made use of an audio recorder to capture the respondents' replies. Certain risks are associated with their use, including increased nervousness by interviewed persons, termination of interview as soon as a recorder is mentioned, or regeneration of nervousness when the recorder has to be adjusted during the interview.

Despite their clear advantages, including the possibility to interact and explain, as well as the depth and detail of information, this researcher was of the opinion that a less intrusive, more compact/time-efficient method would be more suited to future surveys. After studying the advantages and disadvantages of various survey methods, the feasibility of using a questionnaire in this study, and in particular an online questionnaire, was investigated.

3.4.2 Advantages of questionnaires

Based on Wallace & Van Fleet (2012:186-187), the advantages of questionnaires, when compared with in-person interviews, direct observation, case study or focus groups, and placed in the context of this study, are the following:

- Low cost: Costs of administering a questionnaire, be it electronic, mailed, or web-based, would be less expensive than having the interviewer travel to emerging CSIR researchers, as members of the target population are stationed across the country.
- Low pressure on respondents: according to Wallace and Van Fleet (2012:186), requesting questionnaire participation produces a lower sense of pressure to respond than a request for an in-person interview. This researcher, being involved in a new CSIR area (RDM), yet to be implemented at the CSIR, saw the benefit of not making researchers associate RDM-activities with heightened stress experienced.
- Ease of preparation and administration: the design and preparation of this study's questionnaire was relatively easy, as it drew from experience gained during a previous RDM survey (Patterton, 2014a) as well as the analysis of other surveys. Administration, involving the identification of target population, sample, and distribution of questionnaires were simpler and less demanding than would have been the case with alternative data collection methods.
- Ease of tabulation of responses: the researcher selected and constructed questions in such a way that the majority could be automatically recorded in a database.
- Potential for confidentiality or anonymity: a questionnaire, especially an online questionnaire devoid of personal questions, is able to convey and guarantee a greater sense of anonymity than a personal interview.
- Ease of quantification: questionnaire data analysis is much simpler, and can be performed in less time, than analysing data gathered via interviews.
- Potential for large response: whereas other techniques are limited by factors such as time and expense, the use of a questionnaire theoretically made it possible to use a very large sample. The in-person interview technique used during a previous CSIR RDM study could only include respondents who worked within the immediate vicinity (maximum of an hour's drive) of the researcher.
- Limited variation in responses: the possibility of using closed-ended questions made the questionnaire a preferred tool for the purposes of this study; this is in stark contrast with the previous CSIR RDM survey (Patterton, 2014a) where the use of open-ended questions only resulted in information overload and extreme difficulty when transcribing interviews, as well as analysing results.

- High potential for reliability: it was hoped that by making use of a pre-tested questionnaire, and minimising negative behavioural aspects such as performance pressure, embarrassing answers and invasion of privacy, the tool could be said to produce stable and consistent results. According to Wallace & Van Fleet, questionnaire research carries a high potential for replication, and therefore reliability, as the questionnaire is the tangible record of the research and evaluation design (2012:187).
- Questionnaires may be distributed in various ways; mailings, using electronic media, using the telephone, or even placing questionnaires in places such as hotel rooms, are examples of popular distribution methods used by researchers. This factor left this researcher with a variety of distribution methods to choose from, enabling her to select one best suited to her research needs.

3.4.3 Disadvantages/limitations of questionnaires

The questionnaire has some serious drawbacks too. These disadvantages, as stated by Wallace & Van Fleet (2012:187-188), needed to be considered, and this researcher addressed them in the following manner:

- Absence of personal contact: while this is seen as a drawback by some, this researcher aimed at overcoming this issue ensuring that questionnaire instructions were clear and unambiguous. Pretesting and a pilot run of the questionnaire also played a role in eliminating the need for personal contact. A previous CSIR RDM survey (Patterton, 2014a) actually demonstrated that not all CSIR researchers regarded personal contact as a positive experience; at least one researcher declined to be interviewed and on obtaining the interview questions, insisted on sending responses via email, rather than being interviewed.
- Low response rate: this researcher attempted to maximise the survey response by studying survey literature and implementing measures suggested. These methods, discussed in more details in section 3.7.4, included contacting candidates on four occasions, and adhering to the contact characteristics and requirements as described and stipulated in survey literature as far as possible.
- Self-selecting sample: Wallace & Van Fleet admit that regardless of response rate, respondents constitute a self-selecting sample (2012:188). This feature is one that the researcher could not control, nor change: she was not in a position to accurately know what differences between the two groups of emerging researchers motivated one to respond, and the other not to.

- Flaws in questionnaire design: the questionnaire was subjected to various pre-testing measures, including a mock survey to test the questionnaire software, a pre-test involving the study supervisors, and a pilot study involving CSIR researchers. Feedback received after these events indicated that the layout of the online questionnaire was not ideal; as a result of this feedback, this researcher adjusted the survey margins. Furthermore, the estimated time of completion as stated in the cover letter was changed from 20 minutes to a more realistic 30 minutes.
- Misinterpretation of questions: the questionnaire was subjected to various pre-testing measures, including a pre-test involving the study supervisors, and a pilot study involving CSIR researchers. No feedback regarding survey instructions, cover letter content, or survey questions being unclear were received after these pre-survey events. Even so, the cover letter sent to actual survey respondents (i.e. CSIR emerging researchers) still encouraged candidates to report difficulties and uncertainties.
- Differentiation between fact and opinion: while this is a danger in questionnaires making extensive use of open-ended questions (for more information, see section 3.6.2), this researcher attempted at minimizing this phenomenon by making mostly use of closed-ended questions, and emphasizing that truthfulness, albeit it a revelation of bad RDM practices, was crucial in this study. In addition, sufficient opportunity for opinions and perceptions, rather than facts, was included with several questions forming part of this questionnaire (see section 3.6.3).
- Misinterpretation of responses: although this was a real danger, this researcher aimed at minimising this hazard by making use of multiple choice questions where possible, with answers able to meet the demands of each question for each respondent. Furthermore, most questions featured an ‘other’ box where researchers could elaborate on answers. It was hoped that by using this option, misinterpretation by the researcher would be minimised.
- Deliberate sabotage: this was a danger that was real, and was anticipated. CSIR personnel are subjected to many surveys annually and this might have led to survey fatigue and deliberate untruthfulness in responses. This researcher attempted at minimizing the danger of deliberate sabotage by communicating to candidates that the completion of the questionnaire, as well as each of the questions, was not compulsory. Four contacts were made with each respondent; during these activities this researcher was cognizant of the fact that the timing of contacts, wording used and quality of relations with candidates were vital to the standard, reliability and

validity of questionnaire replies. Candidates were at no point be badgered into questionnaire completion.

After taking into account the advantages and limitations of both the 'in-person' interview as well as the questionnaire, and looking at ways in which the limitations of a questionnaire might be lessened, this researcher made the decision to choose the questionnaire as data collection tool for this study. The next section deals with the type of questionnaire implemented, as various formats of the questionnaire could be used during survey research.

3.5 The online questionnaire

Questionnaires may be distributed in various ways: Wallace & Van Fleet (2012:186) regard the two most popular formats to be mailed questionnaires and electronically-distributed questionnaires. Singleton and Straits (2005:243) go one step further and mention that electronic questionnaires could be administered via email, the internet (i.e. an online/ web-based questionnaire), Interactive Voice Response, or computerized self-administered questionnaires. All of these methods have their advantages and disadvantages, as well as scenarios and target population they are more suited to, or study designs not being a good match with a particular format.

This particular study, already discussed as being survey research involving the questionnaire as data collection tool, was deemed by this researcher to be ideally suited to an online questionnaire. Study features such as the geographical spread of the target population, availability and user-friendliness of online questionnaire tools, and the fact that all members of the sample group have access to a personal computer as well as excellent internet connections via the CSIR network, convinced this researcher to make use of the online questionnaire in this study.

As a result of this choice, the following sections look at the characteristics and distinguishing features of the online survey, the advantages offered when making use of them, and possible disadvantages or limitations that the researcher might have come across. Ways in which this researcher tried to minimise the negative aspects often accompanying online questionnaires, are also addressed.

3.5.1 Characteristics of the online questionnaire

According to Dillman (2007:352), the collection of survey data through self-administered surveys by the World Wide Web is just as significant a development as the ground-breaking introduction of random sampling in the 1940s, and the introduction of telephonic interviews in the 1970s. This type of data gathering tool, also referred to as the self-administered

electronic web survey (Singleton & Straits, 2005:243), an online questionnaire (Wallace & Van Fleet, 2012:202) or a web-based survey (Hank, Jordan & Wildemuth, 2009:258) coincided with the widespread adoption of the World Wide Web in the 1990s (Hank, Jordan & Wildemuth, 2009:258). As stated by Dillman (2007:352), a result of this global technological phenomenon has been a transformation in the way in which most major surveys are currently done.

The online survey has some unique features: it involves computer-to-computer transmission of a questionnaire placed on specially designed web pages (Singleton & Straits, 2005:244), and the major costs are often internet service provider fees as opposed to travel costs, paper and telephonic expenses, and transcription costs of other methods (Singleton & Straits, 2005:244, Hank, Jordan & Wildemuth, 2009:260). In addition, features such as scroll bars, pop-up boxes, drop-down lists and simple animation are usually employed (Dillmann, 2007:373).

3.5.2 Advantages of the online questionnaire

Dillmann (2007:352) is of the opinion that the major benefit of online surveys lies in their potential to bring about cost efficiencies. Comparisons with the cost-saving benefits of earlier survey-related breakthroughs are made: he mentions that the past advances in survey design were typically motivated by cost considerations, and in a similar fashion, the electronic survey has the potential to bring about efficiencies of comparable importance. Other advantages, over and above the financial savings, according to Wallace & Van Fleet (2012:202) are its:

- global reach,
- appeal in commercial applications,
- flexibility,
- speed and timeliness,
- technological enhancement,
- convenience,
- ease of data entry and analysis,
- ability to use a diverse range of question structures,
- relatively low administration costs,
- ease of follow-up,
- ease of sampling,
- ability to work with very large samples,
- ease of constructing controlled path sequences,

- ability to require that questions be answered, and
- ease of determining differences between respondents and non-respondents.

Of the factors mentioned above, the following are of particular importance and relevance to the current study, and were regarded as the reasons this researcher made the decision to implement an online questionnaire:

- Global reach: using an online survey enabled the involvement of emerging CSIR researchers not based on a CSIR campus close to this researcher. This was in stark contrast to the previous CSIR RDM survey (Patterton, 2014a), where in-person interviews were done and only candidates within driving distance of this researcher could be included in the study.
- Speed and timeliness: a previous CSIR survey (Patterton, 2014a) done by this researcher, and using paper-based face-to-face interviews, required, on average, about 120 minutes of combined interview and travel time from this researcher, per interview. Transcription and summary of transcription into usable qualitative format required about two days per interview. When taking required time into account, making use of a web survey to be completed in the candidate's own time was a speedier option.
- Convenience: a web survey, when compared to a face-to-face interview or telephonic interview, eliminates the need for personal contact, a set time slot and a venue suiting both parties. Respondents are free to complete the survey in their own time and at venue suited to them. An only requirement for completion can be said to be web access. Further conveniences include the nearly complete elimination of paper and postage.
- Ease of data entry and analysis: a previous CSIR survey (Patterton, 2014a) done by this researcher, and using face-to-face interviews, made use of open-ended questions only. Recorded and often long recorded responses had to be transcribed and summarised before data analysis could commence; an online survey using closed-ended questions would eliminate those stages. Data analysis with the online questionnaire can also be seen to be faster than the case would be with other types of questionnaires.
- Ability to use a diverse range of question structures: a previous CSIR survey (Patterton, 2014a) done by this researcher, and using paper-based face-to-face interviews, made use of open-ended questions only. This often resulted in an overload of information, as well as an overly long interview session. In this respect, making use of a web survey and limiting the inclusion of open-ended questions, while

making full use of a variety of closed-ended questions, is a preferred method. Question types included in this online questionnaire are quite diverse and are discussed in more detail in section 3.6.2.

- Relatively low administration costs: many online survey tools are free of charge. Although making use of a telephonic survey, a face-to-face survey on the CSIR Pretoria campus, an online chat survey or a telephonic survey would be relatively inexpensive to conduct, the fact that free online survey tools exist made it an appealing alternative when its other advantages are included in the equation.
- Ease of follow-up: many online survey tools automatically send reminders to respondents who had not completed the survey; apart form it being a necessary function it is also a step up from the previous CSIR survey (Patterton, 2014a) where this researcher had to manually re-schedule missed interviews.

3.5.3 Disadvantages of the online questionnaire

Disadvantages should be taken note of too; Wallace & Van Fleet (2012:202) mention the following online survey aspects to be important considerations:

- tendency of participants to view invitation email as spam,
- respondent's lack of expertise/experience with online tools,
- technological inconsistency,
- necessity for extreme clarity in instructions,
- impersonal nature,
- privacy and security issues, and
- general low response rates.

However, this researcher was in agreement with Singleton & Straits (2005:245) who state that it is possible to address the problems of online surveys. Thus, within this study, the obvious online survey weaknesses were treated as follows:

- Tendency of participants to view invitation email as spam: this researcher, whose CSIR email address is visible to respondents, sent out a pre-survey email to respondents. This pre-survey email, also termed the 'first contact', explained the aim and scope of the project, study and survey. Proof of ethical approval obtained from the University of Pretoria, as well as a document indicating approval for a request to conduct research at the CSIR, was be relayed to candidates. These features are discussed in more detail in section 3.7.4.

- Respondents' lack of expertise/experience with online tools: candidates were all in possession of a doctorate, or busy with doctoral level studies, and were employed at the CSIR. This researcher was confident that a high level of computer literacy among all target population members could be safely assumed.
- Technological inconsistency: internet downtime at the CSIR is a rare occurrence. Furthermore, all employees had access to a computer featuring the necessary software and programmes able to access the internet-based survey this researcher had used.
- Necessity for extreme clarity in instructions: the online survey was pre-tested by study supervisors, and pilot-tested by two CSIR researchers. Feedback indicated the need for more clarity or removal of ambiguity in instructions/questions. Contact details of this researcher were also supplied in the introduction and after-note of the final version of the online survey.
- Impersonal nature: this researcher was of the opinion that this was not a valid concern within this particular study. Candidates were all full-time CSIR employees, and had their own deadlines to meet and schedules to adhere too. In fact, it became clear to this researcher, when conducting the previous CSIR survey (Patterton, 2014a), that some candidates actually preferred impersonal contact to an invasion of office space, privacy and time.
- Privacy and security issues: although this survey did not gather data that could be classified as being of a sensitive nature, respondents were in no way required to supply their names or contact details; the completion of these two questions was not compulsory. Furthermore, survey answers were not accessible to the general online public, and where required, de-identification was used prior to publication of study results.
- Generally low response rates: to compensate for low rates, this researcher decided to make use of the total population as sample group. Two reminders were sent out via email, at approximately one week and two weeks after the first email was sent to candidates. The timing and wording of the reminders was in line with suggested survey reminders discussed in survey literature.

After making a final decision to use the online questionnaire as survey tool, this researcher had to progress to the next step in the methodology process. This process – the planning of the online questionnaire – is discussed in the next section.

3.6 Questionnaire planning and design

Wallace & Van Fleet (2012:189) state that while a questionnaire is quite easy to design, prepare and administer, a 'good' questionnaire requires expertise, care and planning. Various aspects needed to be considered when planning this online questionnaire: aspects such as the online survey tool/software to be used, questionnaire design and layout, RDM areas to be probed, choice of questions, and question structure and phrasing needed to be mulled over and implemented in a suitable manner. Pickard (2013:208) warns that because the researcher is not present when the questionnaire is being completed, it is important that the respondent understands the questionnaire, and that no additional information outside the questionnaire instructions needs to be provided to respondents. Also, to encourage responses, a questionnaire needs to look good and read well, and instructions should be clear and plausible (Pickard, 2013:208).

Pickard (2013:208) provides a helpful chart, chronologically displaying the steps that need to be taken when designing a questionnaire. These steps – reviewing the data requirements of the research question, developing a list of questions, evaluating each question, determining the form of the question, constructing the wording of a question, organising the structure of the questionnaire, evaluating and piloting the questionnaire, making amendments – have been found by this researcher to be a helpful tool during the questionnaire's design phase, and were frequently be used as a guiding reference when designing the questionnaire.

The steps taken when designing the final data collection tool, are discussed under separate headings.

3.6.1 Aspects to consider/Data requirements

When deciding on the RDM topics and questions to be included in the questionnaire, two important considerations faced this researcher: what were the data requirements of the study's research question, and what RDM topics were included in other surveys and analysed by this researcher?

Firstly, it was critical that the RDM topics and questions included in the questionnaire, should result in the gathering of data of such a nature to allow for the attainment of the following study goals, mentioned in Chapter 1:

- establish what the international developments with regards to RDM practices are,
- assist all stakeholders in understanding the types of data held by emerging researchers in the CSIR,

- determine the current RDM practices of emerging researchers at the CSIR. Practices pertaining to data sharing, data annotation, data storage, data backups as well as the use of metadata will be ascertained,
- identify the RDM needs and requirements of emerging researchers at the CSIR,
- compare the RDM practices, needs as well as challenges of emerging researchers in the various CSIR operating units,
- compare the RDM practices, needs as well as challenges of emerging researchers with those of established scientific groups in the same science research institute,
- compare the RDM practices, needs as well as challenges of emerging CSIR researchers with those of researchers/scientists in other parts of the world, and
- demonstrate best practices with regards to the management of the research data collected during this study: develop a RDM plan, add metadata, collect context giving information, promote the visibility and re-use of the data set, secure the data and ensuring that longer term preservation activities take place.

Furthermore, it was imperative that the data gathered in this study, the RDM topics examined, and even the research method and data gathering tool used be of sure a nature as to make a comparison with results, studies and surveys mentioned in Chapter 2, possible. As stated by Hank, Jordan & Wildemuth (2009:257), the researcher has to examine the literature in her area of interest, and consult those studies that may have posed similar questions. Furthermore, it is stated by them that the researcher should consider using or revising the surveys used in those studies, and even contact the authors of previous studies to obtain more details, or clarifying information, on their study procedures.

This researcher has indeed done a thorough analysis of surveys used in other RDM studies; the results being shared in Chapter 2. When revisiting Chapter 2, it was seen that the main RDM areas, together with their sub-categories, were:

- state of the RDM within the organisation,
- data formats, file formats and data size,
- data storage and data backups,
- data preservation,
- data sharing, sharing tools used,
- use of metadata, metadata standards,
- data management plans,
- RDM training undergone by researchers,

- RDM recommendations and requirements, including RDM services, RDM training and RDM tools, and
- group differences in RDM.

It was then crucial that these RDM areas also be included in the questionnaire to be used in this study.

Apart from keeping the two previously-mentioned major aspects in mind (study goals, survey results comparison), this researcher was made aware of additional questionnaire considerations when studying literature pertaining to the topic. Pickard (2013:208) states that because the researcher is not present to give respondents additional or clarifying information, and because questionnaire design and format has an effect on questionnaire response, designing a questionnaire is 'serious business'. With this warning in mind, this researcher ensured that the questionnaire adhered to the following design suggestions:

- questions should be clear and not ambiguous (Pickard, 2013:209),
- complete sentences should be used to minimise the risk of misinterpretation (Hank, Jordan & Wildemuth, 2009:257),
- ask only what can be answered and what is necessary to satisfy the research objectives (Hank, Jordan & Wildemuth, 2009:257),
- use neutral language; beware of bias, offence and subordination (Hank, Jordan & Wildemuth, 2009:257, also Pickard, 2013:209),
- avoid personal questions ((Wallace & Van Fleet, 2012:197),
- consideration to question specificity is important (Hank, Jordan & Wildemuth, 2009:257),
- avoid double-barrelled questions (Hank, Jordan & Wildemuth, 2009:257),
- participation is voluntary, as is the completion of each question (Hank, Jordan & Wildemuth, 2009:257),
- questionnaire should be short enough to be completed in a reasonable time (Pickard, 2013:209),
- lengthy instructions and definitions should be avoided (Wallace & Van Fleet, 2012:199), and
- a cluttered appearance should be avoided (Pickard, 2013:209).

3.6.2 Types of questions

A questionnaire consists of different types of questions; according to Wallace & Van Fleet (2012:189) questions can be categorised and understood in two major ways: by purpose and by structure.

3.6.2.1 Questions by structure

When categorizing by structure, Wallace & Van Fleet (2012:191) state that the two fundamental structural categories of questions are open-ended and closed-ended. Singleton & Straits (2005:266-267) mention that the terms 'free-response questions' and 'fixed-choice questions' are often used when referring to these two categories of questions. Open-ended questions are free from parameters restricting the respondent, and according to Pickard (2013:219) tend to be essentially descriptive questions requiring a more detailed and personal response. Closed questions, on the other hand, deliberately limit the respondent's range of possible answers (Wallace & Van Fleet, 2012:192). Each of these structural types has its own aim and purpose, as well as advantages and disadvantages, and will be discussed in due course. In addition to these considerations, question structure can be further subdivided: examples of these sub-categories would be Pickard placing closed questions in one of three categories; these categories being the dichotomous question, the multiple dichotomous question, and the rank order question (2013:211). Wallace and Van Fleet (2012:192-196) are of the opinion that closed-ended fall into five prominent categories, and mention the multiple response list question, the single response force question, the ranked list question, the true/false (dichotomous) question, and the Likert scale question.

When deciding on type of question based on structure, to be used in this study, this researcher took into consideration the following aspects:

- aim/purpose of question type,
- advantages of question type,
- disadvantages of question type,
- applicability of question type,
- previous experience of question type in an earlier CSIR survey by the same researcher (Patterton, 2014a); whether it worked well or is best avoided,
- question type being included as question option in online survey tool, and
- question types used in consulted and analysed RDM surveys (see Chapter 2: literature analysis).

As a result of these considerations, the survey questionnaire used in this study made use of the following question types, categorised by structure:

- **Open-ended questions:** this type of question gives the respondent the opportunity of writing any answer in the space provided (Fouche, 2000:160). Apart from an obvious aspect such as name and contact details, the open question has advantages when a variable is unexplored or not well-known to the researcher (Fouche, 2000:160). Pickard also states that an open-ended question is able to add detail to a closed question, can bring a totally new perspective to an issue, even one not considered by the researcher (2013:219). It is, however, important to guard against too many open questions, as completion can be time-consuming and lead to incomplete replies. In addition, data-processing time might be lengthened, and the questionnaire runs the risk of being more liable to error (Fouche, 2000:160). Pickard (2013:219) warns against using open-ended questions as the only data collection method in a study, and does not recommend relying on participants to fill out 'vast empty spaces' on a questionnaire. Hank, Jordan & Wildemuth (2009:258) cautions against open-ended question instructions not being clear, and suggest the use of parameters in helping respondents devise answers that are appropriate in length, content and detail.

  Bearing these aspects in mind, the open-ended questions used in this questionnaire asked for:

  - CSIR unit of respondent,
  - more details about 'other' when an 'other' option in a multiple choice question had been chosen,
  - more details about RMD training received had respondents indicated 'Yes' to a dichotomous question about RDM training received,
  - more details about a data repository currently being used to share data,
  - more details about RDM services deemed important by the respondent, and not already covered in a previous question about RDM services, and
  - more details about any other RDM-related concern, issue or problem, not already covered in the survey.

The format and content of the open-ended questions used in the questionnaire can be seen in Appendix 2, and are also discussed in section 3.6.3.

- **Closed-ended questions:** this type of question requires the respondent to choose a response from responses provided. According to Fouche (2000:160), closed-ended

questions are typically used when a substantial amount of information about a subject exists and the responses are generally well-known. Closed-ended questions have the advantages of being easier on the respondent as it requires 'less effort' and 'less facility with words' (Singleton & Straits, 2005:268). Furthermore, closed-ended questions enhance standardization, require less work and training to administer, and allow for shorter time expenditure (Singleton & Straits, 2005:268). Despite these advantages, Fouche (2000:161) states that closed-ended questions can never completely provide for the response variety which may exist on any particular topic. In a similar vein, Hank, Jordan & Wildemuth (2009:258) warn against providing too many response options, but add that researchers should also make sure all possible responses are identified. To compensate for this ambivalent limitation, Fouche (2000:161) as well as Hank, Jordan & Wildemuth (2009:258) suggest the inclusion of an 'other' category as a perpetual accompaniment to closed-ended questions.

Bearing these aspects in mind, the following categories of closed-ended questions were included in this questionnaire:

- **Dichotomous questions:** this type of question allows the respondent to choose from two possible responses, a third neutral response such as 'don't know' could also be added. Dichotomous questions were used when asking:
  - whether the respondent had developed a research data management plan,
  - whether the respondent was aware of any policy or requirements from their funders with regards to research data management,
  - whether metadata had been created,
  - whether a metadata standard was adhered to,
  - whether their data was subject to confidentiality/security measures, and
  - whether they had received RDM training.

- **Multiple response list questions:** this type of question is accompanied by the instruction '*Select all that apply'*. The respondent is allowed to identify more than one answer to the question, and is under no obligation to rank his selection of answers. Multiple response questions were used when asking:
  - what types of research data respondents work with,
  - which software applications were used when analysing or manipulating research data,

- where research data were stored,
- where research data were backed up,
- what steps were taken to ensure privacy of research data,
- with whom research data were shared,
- which methods/infrastructure were used when sharing research data,
- which research data management tasks are usually performed, and
- which RDM areas they would like to receive training in.

- **Single response forced questions:** this type of question requires the respondent to select one and only one answer from a list. Wallace and Van Fleet (2012:193) mention that this type of question might create problems of interpretation, and could also make choosing an answer difficult when a respondent finds two or more categories to be equally true as an answer. Single response questions were used when asking respondents to stipulate:
  - the academic discipline their PhD is part of,
  - the volume of research data across all of their CSIR work,
  - frequency of data backup,
  - the owner of IP rights for their research data,
  - frequency of sharing request from other researchers for respondent's research data,
  - ability to supply research data to others,
  - frequency of sharing request from respondent for other researchers' research data, and
  - where research data were preserved after results were published.

- **Likert scale questions:** According to Wallace & Van Fleet (2012:194) these are also known as rating scale questions, and require the respondent to indicate a response on an explicit scale. Pickard states that a Likert scale allows a respondent to demonstrate their level of agreement with a statement (2013:213), while not indicating the interval measure between each choice. It is often used when gathering opinion behavioural data, or as mentioned by Singleton & Straits (2005:273) it measures the 'strength and intensity of a respondent's feelings'. Rating scale questions were used when asking respondents to indicate:
  - the importance they attach to several listed RDM services in improving research in their discipline, and

- the importance they attach to several listed RDM standards, policies, principles and practices in helping CSIR researchers manage their research data.

- Other categories of closed-ended questions mentioned in the literature, such as ranked list questions, matrix-type questions, statements, or follow-up questions, are not included in this questionnaire.
- It should be noted that the open-ended questions mentioned in an earlier section, asking details about 'name' and 'email' address, are sometimes categorised as closed-ended questions. Fouche (2000:162) states that these closed-ended questions are '**completion questions**', and are used to collect data when there are too many response options to enable meaningful classification.

The format and content of the closed-ended questions used in the questionnaire can be seen in Appendix 2, and are also discussed in section 3.6.3.

3.6.2.2 Questions by purpose

According to Wallace & Van Fleet (2012:189) questions can also be categorised by their 'primary intent', meaning that questions can be classified on their purpose of probing respondents' perception of 'facts, knowledge, behaviours, or opinions'.

- **Factual questions:** this category of questions assumes that anyone with access to required information would answer in an identical manner. Examples of factual questions included in this questionnaire were questions asking about:
  - research data types used,
  - research data volume gathered, and
  - the owner of IP rights for their research data.
- **Knowledge-eliciting questions:** these questions are designed to assess the knowledge base of a respondent, and not, as the previous category, their access to data. Examples of knowledge-eliciting questions included in this questionnaire were questions asking about the awareness of funder requirements.
- **Behavioural questions:** these type of questions address 'actual, perceived or projected' behaviours (Wallace & Van Fleet, 2012:190), and are very personal in nature. Examples of behavioural questions included in this questionnaire were questions asking about:
  - respondent's asking other researchers for research data (sharing needs), and

  - the use of metadata when managing research data.
- **Opinion questions:** this category of questions gathers data about opinions, and answers might be completely unrelated to the facts about a phenomenon. Examples of behavioural questions included in this questionnaire were questions asking about:
  - the importance a respondent attaches to pre-identified RDM services, and
  - the importance a respondent attaches to pre-identified RDM standards, policies, practices and principles.

A copy of the questionnaire to be used, and questions included, can be viewed in section 3.6.3, or in Appendix 2, or online at https://eSurv.org?s=LHIIMF_e2e40c29 .

3.6.3 List of questions

After considering the aspects mentioned in section 3.6.1, as well as structural considerations mentioned in 3.6.2, this researcher decided on using the following 31 questions in the questionnaire:

- **CSIR Unit (question 1):** respondents were asked to indicate which CSIR unit they were part of. The question is open-ended.
- **Academic discipline (question 2):** respondents were asked to indicate which academic discipline their PhD was part of. It is a closed-ended question, with four options supplied. In addition, an 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, is a fifth option available.
- **Data formats/types of data (question 3):** respondents were asked to indicate which data types they were currently making use of. It is a closed-ended question, with 15 options given. Respondents may select all applicable options. An 'other' option, which can be ticked and also add text-information to, is also supplied. The data formats included were based on the most common data formats indicated during a previous CSIR survey (Patterton, 2014a), as well as formats commonly indicated throughout surveys analysed in Chapter 2, section 2.4.4).
- **Data volume (question 4):** respondents were being asked to indicate the volume of research data across all of their CSIR work. It is a closed-ended question, with respondents having to choose one answer from nine options. All size ranges were included, as well as an 'I don't know' option for respondents not aware of the data volume they work with.
- **Software applications used (question 5):** respondents were asked to indicate the types of software used when analysing or manipulating data. It is a closed-ended

question, with 19 options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Software options included were based on the most common software tools indicated during a previous CSIR survey (Patterton, 2014a).

- **Research Data Management Plan (question 6):** respondents were asked to indicate whether they are making use of a data management plan. It is a closed-ended question, with respondents required to choose one applicable answer from 'yes', 'no' and 'I don't know'.
- **Awareness of funder's data policies/requirements (question 7):** respondents were asked to indicate whether they are aware of their funder having an RDM policy, or RDM requirements. It is a closed-ended question, with respondents required to choose one applicable answer from 'yes', 'no' and 'not applicable'.
- **Data storage (question 8):** respondents were asked to indicate where their research data were stored. It is a closed-ended question, with 13 options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in RDM surveys analysed (see Chapter 2, section 2.4.5), as well as storage locations indicated during a previous CSIR RDM survey (Patterton, 2014a).
- **Frequency of data backups (question 9):** respondents were being asked to indicate how often data backups are made. It is a closed-ended multiple choice question, with respondents having to choose only one option.
- **Location of data backups (question 10):** respondents were being asked to indicate where there data are backed up to. It is a closed-ended question, with eleven options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in a previous CSIR RDM survey (Patterton, 2014a).
- **Use of metadata (question 11):** respondents were being asked to indicate whether they are making use of metadata when managing their data. It is a closed-ended multiple choice question, with respondents having to choose only one option. The options are 'yes', 'no', 'sometimes' or 'I don't know'.
- **Use of metadata standards (question 12):** respondents were asked to indicate whether they are making use of a specific metadata standard when managing their data. It is a closed-ended multiple choice question, with respondents having to

choose only one option. The options were 'yes', 'no', 'sometimes', 'I don't know', or 'not applicable'. When choosing 'yes' or 'sometimes', the respondent was asked to supply clarifying/additional information about the standard used. An empty text box was supplied for this purpose.

- **IP ownership of data (question 13):** respondents were asked to indicate who the Intellectual Property Rights owner of their research data was. It is a closed-ended multiple choice question, with five options supplied, and respondents having to choose only one option. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options supplied are based on answers supplied in a previous CSIR RDM survey (Patterton, 2014a).
- **Data confidentiality/Data sensitivity (question 14):** respondents were asked to indicate whether their research data are subject to confidentiality/sensitivity matters. It is a closed-ended multiple choice question, with respondents having to choose only one option. The options were 'yes', 'no', or 'unsure'.
- **Data security measures (question 15):** respondents were asked to indicate the steps they were taking, or would be taking to ensure data security. It is a closed-ended question, with eight options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in a previous CSIR RDM survey (Patterton, 2014a).
- **Data sharing (question 16):** respondents were asked to indicate with whom they share their research data. It is a closed-ended question, with ten options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in RDM surveys analysed (see Chapter 2, section 2.4.7).
- **Frequency of data sharing requests received (question 17):** respondents were asked to indicate how often they have received requests for their data, during the last five years. It is a closed-ended question, with five options supplied. Only one answer was to be selected.
- **Frequency of data sharing requests fulfilled (question 18):** respondents were asked to indicate how often they had been able to give access to their research data. It is a closed-ended question, with four options supplied. Only one answer was to be selected.

- **Explanation of answer supplied in question 18 (question 19):** respondents were asked to explain the answer supplied in the previous question. It is an open-ended question, with text-box supplied.
- **Data sharing method (question 20):** respondents were asked to indicate the methods used when sharing data with others. It is a closed-ended question, with eight options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in RDM surveys analysed (see Chapter 2, section 2.4.7).
- **Details on use of curated digital data repository (question 21):** respondents who had indicated that they make use of a digital repository when sharing their data, were asked to provide more details about the repository. It is an open-ended question, comprising a text box allowing respondents to add a response in their own words.
- **Frequency of data sharing requests made (question 22):** respondents were asked to indicate how often they have requested research data from others, during the last five years. It is a closed-ended question, with five options supplied. Only one answer was to be selected.
- **Explanation of answer supplied in question 22 (question 23):** respondents who answered 'no, never' to question 22, were asked to explain why. A text box was supplied.
- **Data preservation (question 24):** respondents were asked to indicate what was done with their research data after the results had been published. It is a closed-ended question, with six options supplied. Only one answer was to be selected. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in RDM surveys analysed (see Chapter 2, section 2.4.6).
- **Other RDM tasks performed (question 25):** respondents were asked to indicate which other RDM tasks they generally perform. It is a closed-ended question, with 11 RDM tasks listed. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in RDM surveys analysed (see Chapter 2, section 2.4).
- **RDM training received (question 26):** respondents were asked to indicate whether they have received RDM training. It is a closed-ended question, options are 'yes', 'no'

and 'I cannot remember', with respondents asked to supply clarifying information should the 'yes' option be chosen.

- **RDM training required (question 27):** respondents were asked to indicate the RDM areas they would like to receive training/guidance in. It is a closed-ended question, with 12 options supplied. Respondents were able to select all applicable options. An 'other' option, which can be ticked and has an empty text box enabling the respondent to add clarifying information, was one of the options given. Options included were based on answers supplied in RDM surveys analysed (see Chapter 2, section 2.4.11).
- **RDM services (question 28):** respondents were asked to rate several RDM-related services according to its importance in improving research. It is a closed-ended question, with eight services listed. Five rating options were available: 'not important', 'somewhat important', 'important', 'very important' and 'not familiar with this services'. RDM services listed were based on options used in RDM surveys analysed (see Chapter 2, section 2.4.11).
- **RDM practices, policies, principles (question 29):** respondents were asked to rate several RDM policies, principles, practices and standards according to its importance in assisting researchers when managing their data. It is a closed-ended question, with 13 policies/practices listed. Five rating options were available: 'not important', 'somewhat important', 'important', 'very important' and 'not familiar with this services'. RDM services listed were based on options used in RDM surveys analysed (see Chapter 2, section 2.4.11).
- **Additional RDM services required (question 30):** respondents were asked to state any RDM service or policy deemed important to them, and not covered in the previous questions asking them to rank data services/policies. It is an open-ended question, comprising a text box allowing respondents to add a response in their own words.
- **Any RDM-related concern/suggestion (question 31):** respondents were asked to state any RDM-related concern, suggestion or complaint, not covered in any of the previous questions. It is an open-ended question, comprising a text box allowing respondents to add a response in their own words.

Looking at these questions, this researcher was confident that both aspects regarded as important when considering the RDM aspects to be included in this survey (comparison with other survey results, study goals discussed in Chapter 1) had been considered and addressed. An aspect not included in the questionnaire, namely requesting the respondent to describe the state of RDM within the CSIR, was considered to be unnecessary. This

researcher, as a permanent employee of the CSIR, responsible for RDM in the CSIR, and having conducted a previous CSIR RDM survey, was in an informed and knowledgeable position to describe the RDM status quo at the CSIR.

Another RDM topic, addressed in a few studied surveys but not explicitly probed within this study, was the question of group differences with regards to RDM behaviour. It was not one of the main aims of this study to discover group differences in RDM behaviour. However, as a result of data collected, this researcher found herself in a position able to comment on trends noticed among the following:

- RDM behaviours of emerging CSIR researchers could be compared with RDM behaviours of established CSIR researchers, and
- RDM practices of emerging CSIR researchers could be compared with those of researchers/scientists in other parts of the world.

However, regardless of data collected, group differences such as gender, higher education institute attended, or research position at the CSIR, was not be investigated during this study.

3.6.4 Choice of survey software

A large selection of online survey tools was available to this researcher for use, and deciding on one of them proved to not be an easy task. In the end, eSurv.org, accessible from http://esurv.org/docs/?Welcome was chosen, as the platform proved to be user-friendly, free of charge, and provided the survey user with an editable professional looking survey. The survey tool also provided for an unlimited number of questions and responses. In addition, time restrictions were not part of this software, and an unlimited number of email invitations could be created. Furthermore, a wide variety of question options, required by this researcher for this particular study, was available, and its email centre was able to send email invitations and reminders to participants, as well as create links to a survey and survey reports (eSurv, 2015).

3.6.5 Additional online questionnaire considerations

Hank, Jordan & Wildemuth (2009:258) state that due consideration should be paid towards the physical appearance of a survey. With web-based surveys, different browsers may alter the intended display of the survey; surveys to be accessed via the web need to be designed with these renderability concerns in mind. Furthermore, these authors also suggest the use of a progress bar to be included so respondents will be made aware of their progress as they move through the survey. With these suggestions in mind, this researcher tested the online

questionnaire's appearance on different browsers, and was satisfied that its intended appearance and functionality would not be altered. A progress bar was also made part of the online questionnaire to be used in this study. With feedback received after the pilot-run of the questionnaire, and no requests for changes in its physical aspects or functionality forthcoming, this researcher was confident that the appearance of this web-based questionnaire was satisfactory.

3.6.6 Questionnaire analysis: pre-evaluation

Once the questionnaire had been drafted, and before administering the tool, this researcher felt it important to implement a variety of measures whereby the quality of the questionnaire could be ascertained. The survey tool used in this this study had been subjected to two types of pre-evaluation methods, namely pretesting and pilot testing. In addition to these two tests, an informal survey using the eSurv software, and incorporating question types to be used in the formal, final survey, was also tested. This last method – a mock survey – was done prior to questionnaire design and is discussed prior to the other two tests.

Pre-survey administration activities are discussed extensively in questionnaire-related literature; this researcher quickly realised that these activities form part and parcel of any study using a newly-designed questionnaire and cannot be omitted. Different terms are used to discuss these preliminary activities: Singleton & Straits refer to it as 'field pretesting' (2005:298), Wallace and Van Fleet use the terms 'expert advice and pretesting' (2012:203), while Hank, Jordan & Wildemuth describe the activities as 'pretesting and pilot testing' (2009:259). Dillman (2007:140-147) warns that while pre-testing has always been a highly touted part of questionnaire design, in practice it is often done haphazardly. To combat this, he has divided the pre-testing phase into four distinct phases; the detailed discussion of the pretesting phase by Dillman was seen as useful by this researcher, and consulted frequently during the pre-test phase.

The various pre-test measures implemented by this researcher, are now discussed in more detail.

3.6.6.1 Mock survey

This was a test done to ascertain and evaluate the functionality of eSurv as a reliable online tool. The mock survey, consisting of arbitrary non-RDM questions, was created by this researcher and distributed to colleagues (librarians, information specialists and library managers) within her CSIR unit. Six different question types, including open-ended questions and closed-ended question types such as the single response question, a Likert

scale question, a multiple choice question and a ranked response question, were included in the mock survey. 12 respondents completed the survey, and no problems with technology, software or answering of different questions types were reported to this researcher. The successful run of the mock eSurv survey, coupled with the features mentioned in the previous paragraph, enabled this researcher to make the decision to use the eSurv platform in the proposed study.

3.6.6.2 Pre-testing and pilot testing: background

After the preliminary question type, question content, and questionnaire layout had been completed, the survey tool was subjected to two types of pre-evaluation methods. These methods are commonly known as pretesting and pilot testing. Pretesting is the review of the survey instrument by experts or members of the target audience, while pilot testing refers to the realistic administration of the survey to a sample from the survey audience (Hank, Jordan & Wildemuth, 2009:259). While Wallace & Van Fleet (2012:203) state the importance of similar activities, they refer to the actions as 'expert advice' and 'pretesting', respectively. So often, these terms are used interchangeably, another example being Singleton & Straits (2005:248) referring to a pre-test as 'the trying out of the survey on a small sample of persons having characteristics similar to those of the target group of respondents'.

While terminology may differ, the purpose of pre-test activities is viewed similarly by all sources. In brief: the purpose of pre-survey testing is to determine whether the survey instrument is able to serve the purpose for which it was designed, or if there is a need for further refinement, revision or changes (Singleton & Straits, 2005:248). Hank, Jordan & Wildemuth (2009:259) encapsulate this as evaluating the reliability and validity of the survey, and also see it as a way to uncover problems, as well as correct misspellings and grammatical errors. Similarly, pretesting, as stated by Wallace & Van Fleet, assesses the quality of the questionnaire prior to questionnaire administration (2012:203). Dillman (2007:140) state that pre-testing has always been a highly touted part of questionnaire design, but in practice is often done haphazardly. To combat this, he has divided the pre-testing phase into four distinct phases; these four phases will be touched on in the section following.

3.6.6.3 Pre-testing

Pre-testing**,** or having the instrument reviewed by experts (Hank, Jordan & Wildemuth, 2009:259) was done by having this researcher's supervisors examine and critique the instrument schedule. Wallace & Van Fleet state that these experts may be individuals with greater questionnaire expertise than the researcher, and are expected to provide

constructive comments on the questionnaire structure, form, wording and other technical matters (2012:203). They may also be individuals with greater subject expertise than the researcher and may be called upon to advise on the subject matter, use of jargon, and comprehensibility of the survey. As a result of this, the questionnaire's structure, questions used, RDM areas included, as well as web functionality, were also inspected by them. Comments and suggestions regarding these mentioned survey aspects were requested by this researcher from the experts, provided to this researcher, and where deemed necessary, these changes were incorporated into the survey.

Dillman (2007:140-141) states that this first pre-testing activity, being a reviewing by knowledgeable colleagues and analysts, should be used to determine whether all necessary questions have been included, whether some questions can be eliminated, whether comparisons with other studies are possible, and if there are merits to modernising the question categories used. Dillman (2007:141) also supports the idea that knowledgeable people would include data analysts (who would know whether question responses would be usable), survey-experienced persons (who would notice an unbalanced answer scale or missing answer option), or people with knowledge of policies (some answers might have no consequences for company decisions). In short, Dillman (2007:141) supports the idea of the first pre-test phase involving persons with diverse expertise; these persons could be as few as two or as many as dozens of individuals.

The above activity can be said to have as its goal the finalization of the 'substantive content' of the questionnaire (Dillman, 2007:141). A next pre-testing phase, evaluating the comprehensibility of the wording, the similarity of question interpretation between candidates, the availability of an answer for every question by every respondent, and the impression created by the appearance of the questionnaire, is an advised activity. This researcher, while not embarking on a series of cognitive interviews as discussed by Dillman (2007:142), was confident that the evaluation of the questionnaire by her two study leaders should address shortcomings present in those mentioned areas.

3.6.6.4 Pilot study

A pilot-run of the study, described by Hank, Jordan & Wildemuth (2009:259) as following a pre-test, entailing a replication of the administration of the full-scale survey design and involving a small sample from the target audience, was felt by this researcher to be appropriate. In slight contrast with the requirement that the pilot-study sample include members of the target population, this researcher decided to make use of two established CSIR researchers when pilot-testing the survey. In other words: the pilot study made use of two established researchers, older than 35, whose feedback regarding the survey was vital

even though their question responses did not qualify as survey research data. The rationale behind this decision was that making use of emerging CSIR researchers would decrease the sample size to be used in this eventual real study; this researcher was not in a position to predict the eventual response rate of the actual survey. Furthermore, as this researcher launched the pilot study as a means to gain an idea of the workability of the online questionnaire, as well as the modes of contact, the content of the answers were deemed to be less important during the pilot-run. What mattered first and foremost, was that modes of contact (web-based, e-mail) were effective and well-received, all pilot study respondents were able to view to access the online questionnaire, understand every question, and be able to choose an applicable answer for every question asked. As such, the pilot test involved all administration activities, as well as follow-up activities of the real survey. Furthermore, as in the real study, pilot survey candidates were contacted on four occasions. These activities are discussed in more detail later when the administering activities are covered; for now, it is sufficient to say that the pilot study included the following activities:

- first contact: pre-questionnaire notice informing the candidates of the questionnaire due to arrive, distributed via email,
- second contact: cover letter, informed consent letter and link to the online questionnaire,
- third contact: thank you note/reminder via email, and
- final contact: email sent to all candidates.

Initially, this researcher was indecisive about whether to make use of a pilot study during the pre-test phases of the questionnaire administration. The demands on time, effort, and resources were aspects to consider. The decision was made to go ahead with the pilot study: the possibility of getting questionnaire feedback/suggestions during the pilot phase was seen to be important and by making use of non-emerging CSIR researchers, the sample would not be in any way be decreased.

The following input was received from experts, as well as pilot study respondents:

- margins of the online survey were found to be ‘very broad’, and
- the survey took longer than was stated in the cover letter.

The following amendments were made:

- margins of the layout of the online survey were adjusted to be slightly narrower than before, and

- estimated completion time stated in the cover letter was changed from 20 minutes to 30 minutes.

3.7 Administering the survey

Administering the survey included various essential subtasks, and according to the researcher each of these tasks need to be explained and discussed. Defining and selecting the target population and sample, obtaining research consent and ethical clearance, recruiting candidates and distributing the questionnaire, deciding on methods and frequency of contacting candidates, and additional aspects related to the dynamics of survey administration, formed part of the survey administration process and are explained below.

3.7.1 Target population and sampling

This study was concerned with the RDM practices of emerging researchers in the CSIR. However, simply stating that 'emerging researchers working at the CSIR' was be the group being investigated, would not suffice; this researcher needed to precisely define the characteristics of survey candidates, how the members would be identified and chosen, and how big the survey group would be. As such, the concepts known as 'target population', 'sample' and 'sampling' need to be taken note of, applied in practice, and discussed in this section.

3.7.1.1 Target population

The target population, also referred to as the 'research population' or the 'population', is described by Pickard to be the entire set of individuals about which inference will be made (2013:60). Wallace & Van Fleet (2012:49) refer to the population as 'all the entities that are by definition of interest to the researcher'. They also state (2012:145) that the population is defined by the research question. Furthermore, taking note of the emphasis on 'all' is critical (2012:145). This study viewed the population to be all emerging researchers employed by the CSIR, but this broad population definition needed to be described in more detail, and its subcomponents clarified in order for readers to fully grasp the characteristics and limitations of this particular research population.

According to Singleton & Straits (2005:115) defining the population is a two-step process: the first step involves a clear identification of the target population, and a specification of criteria for determining which cases are included in the population, and which case are excluded. The second step is termed the 'sampling frame', and denotes the set of cases from which the eventual sample will be selected. Two ways of constructing a sampling frame

exist; the researcher could list all cases, or could provide a rule defining membership (Singleton & Straits, 2005:116). With this in mind, a more detailed description of population characteristics in this study, are as follows:

- **'Emerging'**[3]: for the purposes of this study, 'emerging' was defined as a person aged 35 years or younger, at the time the list of names was created and supplied by CSIR Human Resources to this researcher. Researchers meeting all other requirements, but being 36 years or older, were not included.
- '**CSIR researcher**': for the purposes of this study, a CSIR researcher was defined as a permanent CSIR employee, working in a research capacity, and currently busy with, or in possession of a PhD-degree. To be seen as 'busy with' a PhD-degree, a researcher would have to be registered at a tertiary institute, for a PhD-degree, for the 2015 academic year. The following criteria were not regarded as valid for population inclusion:
    - unregistered students without PhD-degrees,
    - students not in possession of a PhD and registered for a degree that was non-PhD, and
    - students having as highest qualification a master's degree, or currently registered for a master's degree as highest qualification. This researcher made the decision to insist on PhD level, to ensure that the researcher had dealt with research data. At master's level, there is the possibility that researchers might only have completed assignments or modules, and not gathered or dealt with research data before.

Part-time employees, contract employees, visiting researchers, and interns were not included in the research population.

Now that the criteria important when describing the target population have been identified, the population of this study can be stated to be all persons meeting all requirements as described above. An emerging CSIR researcher is a fulltime employee of the CSIR, aged 35 years or younger, and busy with or in possession of a PhD-degree. Acquiring a list of population members involved this researcher approaching the CSIR Human Resources division, and requesting a list of people meeting the criteria described earlier, while stating the intended use of the name list. After being supplied with an electronic list of names, this researcher was able to determine that the population had the following features:

[3] This definition is in line with the definition used by the CSIR when showcasing the work of the CSIR's 'emerging' researchers during its biennial Emerging Researcher Symposium. As a result of this practice, it is age, rather than years of research experience, which determines 'emerging' researcher eligibility within this study.

- the list comprised 179 persons, and
- all of the CSIR research units were involved; the members of the target population is shown in the table below:

Table 1: CSIR emerging researchers: target population

CSIR Research Unit	Number of researchers
Biosciences	28
Built Environment	5
Consulting and Analytical Services	1
Defence Peace Safety and Security	18
Materials Science and Manufacturing	43
Meraka Institute	21
Modelling and Digital Science	17
National Laser Centre	10
Natural Resources and the Environment	36
Total	**179**

As mentioned earlier, defining the population is seen to be a two-step process, with the first step (selection of criteria) just discussed. Step two of the process, which entails constructing the sampling frame, either by listing all cases or providing rules stipulating membership, would be the next step (Singleton & Straits, 2005:116). However, in this study, it would seem as if the criteria already listed could in fact be used to construct the sampling frame as well.

Singleton & Straits (2005:116) mention that ideally, the sampling frame and the target population should be identical, but that in practice it only happens for very small, geographically concentrated populations. Wallace & Van Fleet (2012:147) state that sampling frames are inherently imperfect representations of their corresponding populations, however, this study's researcher is confident that these two concepts (sampling frame and target population) are as identical as possible in this study. The size of the group, as well as the fact that a computerised CSIR personnel management system was able to capture all candidates precisely, meant that there would be no foreseeable omissions or differences between target population and sampling frame.

Having established the target population of the study, and being in possession of a valid, reliable and up-to-date document listing the names of all population members, this researcher was now in a position to regard this list as representing the sampling frame. The goal of having a sampling frame: having a source of information about a population used to make it possible to select a sample, could now be applied.

3.7.1.2 Sampling

The next step in the methodology entailed sampling: a procedure aimed at obtaining a sample to be surveyed that will be representative of the research population. Implemented in research for reasons of feasibility, it can take on a myriad of types, complexity and applicability, and will be discussed in the remainder of this section

A sample is defined as 'a subset of the population that is identified as being a useful representation of the population for purposes of a specific research endeavour' (Wallace & Van Fleet, 2012:49). The act of sampling is described as 'a process of selecting a subset of cases in order to draw conclusions about the entire set' (Singleton, 2005:146). Several aspects stood out to this researcher: it was a fragment of the population, this fragment was regarded as being representative of the target population, this fragment would be investigated, and conclusions about the fragment would be extrapolated to the target population. Taking this into account, this section, in brief, is a discussion of the fragment of the target population that was investigated, and touches upon its ability to represent the total population.

It is important to consider the reasons for sampling, as an understanding of sampling reason assisted the researcher when choosing an applicable sampling method. Wallace & Van Fleet (2012:49) state that sampling is first and foremost used as a tool of convenience, enabling researchers to make use of a portion, or a subset of the population, when trying to explain a facet of the population. The sample as a separate entity is never the focus of study: while the population is the true focus of the study, the use of a sample enables researchers to save resources. These resources could include time, money, and effort. Wallace & Van Fleet also state that sampling has additional benefits, such as the avoidance of waste, error reduction, as well as simplified data handling and analysis (2012:148). Because of these listed economy factors, Singleton & Straits (2005:146) are of the opinion that the use of a sample is unavoidable. These authors go one step further, and mention that the use of a sample may sometimes lead to more accurate conclusions than when using 'complete' enumerations (2005:146). Reasons given for sampling seem to indicate that it is seen to be convenient, often unavoidable, and in some cases, even used to deliver more accurate conclusions than when not sampling.

Apart from a sample having to meet requirements with regards to size, its composition needed to be such that it would qualify as a representative sample. Such a sample qualifies as 'a selected segment of a group that closely parallels the population as a whole in terms of the key variables and characteristics that are under mention' (Cherry, 2014). A

representative sample enables the researcher to generalise the results of the studies and experiments to the population as a whole. Key variables, deemed by this researcher important for this study and not to be disregarded when deciding on sampling, included disciplines, data types and file formats used, CSIR units and unit behaviour, as well as level of RDM training and experience. Sampling method used would need to be a method whereby the maximum number of members could be included, without making it too big to be managed by one researcher.

3.7.1.3 Sampling used in this study

When deciding on a sampling method to be used in this study, this researcher not only studied sampling literature, but also needed to consider the unique characteristics of the population, and the research questions this study tried to answer.

This researcher, after obtaining the list of names forming the target population, was of the opinion that the population was neither very big nor very small; in fact, this population with a total of 179 emerging CSIR researchers could probably be considered to be of medium and manageable size. With this population comprising researchers situated in nine CSIR operating units, involved with various research disciplines, and making use of a diverse range of data types, this researcher decided on investigating sampling options, tailor-made for small-to-medium groups, and aimed at involving as many of the members of the target population as possible.

Such a sampling design would indicate it belonging to the sampling techniques categorised under ‘purposive sampling’; according to Pickard (2013:64) the logic of this non-probability sampling method lies in the selection of ‘information-rich’ cases for study. This researcher, while aware that information-rich cases are those select individuals that can provide insight and a deep understanding regarding the topic of interest (Eliot, 2011), was also cognizant of the fact that almost each and every member of the population could be classified as an ‘information-rich’ case, due to their unit-specific and discipline-specific data. While ‘saturation’, or the point at which additional data fails to generate new information (Braun & Clarke, 2013:55) is stated to be a possible concern in studies, this researcher was wary of making use of a limiting sampling technique which might lead to the exclusion of information-rich cases. As such, she investigated whether a sampling technique, maximising the number of sample members, exists and whether it would be an acceptable method for this study. This line of reasoning was in line with the opinion of Pickard, who states that the choice of sampling technique is driven by the purpose of the research (2013:64). In this study, this researcher was interested in obtaining data about the RDM habits of emerging CSIR

researchers, preferably including participants from every CSIR operating unit, with responding researchers involved in as many CSIR disciplines or academic areas as possible. For this very reason, random sampling, where each member of the population has an equal chance of being selected (Strydom & De Vos, 2000:193) would not meet the expectations of this researcher.

Sampling in this study had to be purposive, in other words, it would involve a process of case selection other than random selection (Singleton & Straits, 2005:132). These authors also state that while purposive (also referred to as nonprobability sampling) sampling has some weaknesses, it is often a very viable means of case selection (2005:132). In one of the given examples, they describe studies with small populations (fewer than 100) as being a typical situation where each case should be studied in its own right in comparison with all others. This situation resonated highly with this researcher; she made an effort of establishing which sampling methods would involve the highest number of participants. If such a method did not exist, then she would have to choose between involving the whole population, or choose a purposive sampling method matching the sampling requirements of this researcher the closest.

Although not common in the literature, this researcher was able to trace several references mentioning the sampling method she had envisaged. Laerd (2012:1), in an online tutorial on sampling techniques, refers to a method called ‘total population sampling’. This is described as a purposive sampling technique that involves examining the total population that have a particular set of characteristics. In this instance, these particular characteristics would be their age (35 and younger), affiliation (permanent employee at the CSIR) as well as level of education (busy with, or in possession of a PhD degree). While the resource admits that it is not a frequently-used method, it is stated that it could be a preferred method when the population size is relatively small. The rationale behind this decision is that failing to include all candidates of a population with unique characteristics in the sample, could lead to missing a significant piece of information, which is an omission this researcher was trying to avoid.

Laerd (2012:2) also mentions a second instance where total population sampling could be implemented, stating that the sharing of uncommon characteristics by the population could persuade a researcher to make use of the total population sampling method. As mentioned earlier, this study’s researcher views the population used in this study – permanent CSIR employees, aged 35 and younger, busy with or in possession of a PhD – as a unique population, descriptively dissimilar to the characteristics of any other group of researchers, and meeting the requirements of this second category stated by Laerd (2012:2) too.

The creation of a total population sample, as could be expected, is quite straightforward. As described by Laerd (2012:2) it is a three-step process, entailing:

- defining the population characteristics,
- creating a list of the population, and
- contacting all members on the list.

During this study, these three steps were completed as follows:

- Characteristics of the population centred around three aspects: candidates had to be a permanent CSIR researcher, candidates had to be 35 years of age or younger, candidates had to be in possession of a PhD or busy with a PhD.
- Laerd (2012:3) mentions that a researcher often needs to make use of a 'gatekeeper' when trying to create a list of the population. In this study, this researcher made use of a CSIR Human Resources manager as gatekeeper, and requested a printout containing the names of all candidates possessing all three characteristics. After inspecting the list, and determining that it was valid (i.e. candidates not registered for the 2015 academic year should not be included, PhD-holders not involved with research should not be included), this researcher compiled a final population list.
- After completion of the various pre-test activities (described in section 3.6.5), and gaining ethical approval to conduct this study, the sample (in this case the target population) were contacted by this researcher. Although various methods of contact were used to communicate with the target sample throughout the study, the first contact was a pre-study notice in the form of an email, informing candidates of the study and the online questionnaire to be sent to them in due course. This method of contact, as well as other contact methods, is described in section 3.7.3.

As with any sampling method, this method has its advantages as well as disadvantages. Laerd (2012:3) states that the deep insight into the phenomenon investigated, as well as the reduced risk of missing potential insights from candidates not included, are two advantages of the method. On the other hand, aspects such as the total population sampling's requirement of a list and the resources required to draw up such a list, the possible difficulties experienced when trying to compile such a list, and possible difficulties in trying to contact all candidates on the list, could be seen as disadvantages of this method.

With regards to sample naming, additional sources used by this researcher when studying sampling techniques, such as the website of the Australian Bureau of Statistics (2014), as well as reference source on research in library and information science (Wallace & Van Fleet, 2012) provided her with an interesting conundrum. According to the Australian Bureau

of Statistics (ABS), a population can be studied using one of two approaches: taking a census, or selecting a sample. A census is described as a study of 'every unit, everyone or everything, in a population', while a sample is defined as a 'subset of units in a population', in other words, it is a partial enumeration. Similarly, Wallace & Van Fleet (2012:145) consider a descriptive study of an entire population to be a census. According to these definitions, this researcher did not make use of a sample, but instead, conducted a census, also termed as a complete enumeration of the population.

Wallace & Van Fleet (2012:145) state that there are only a few circumstances under which conducting a census is desirable or necessary, causing this researcher to study the pros and cons of using a census. ABS (2014:1), in agreement with Laerd (2012:3), mentions the advantages and disadvantages of using a census as opposed to using a sample. Absence of sampling error, as well as the possibility of obtaining detailed information about small sub-groups within the population, is stated as distinct advantages.

Disadvantages of the census method, as stated by ABS (2014:1) include the difficulty of enumerating all units of the population within the available time, higher costs, as well as the longer amount of time required to collect, process and release data. Wallace & Van Fleet (2012:146) further state that when a census involves a very large population, it adds to the complexity of this study. This researcher, taking the features of this study, its population and survey tool into account, was confident that the use of a free and web-based questionnaire should minimize the effect of disadvantages stated.

Another interesting literature source, comprising the discussion of various sampling techniques and presented as online training material for post-graduate students, was the web manual titled 'Non-probability sampling techniques' (Mugera, 2013). Here, total population sampling was stated to be a type of purposive sampling technique, and a method used by researchers when they choose to examine an entire population having a particular set of characteristics. Total population size of the population having this particular set of characteristics being small, is stated to often be a prerequisite for choosing this sampling technique.

To summarise: this researcher, when deciding on the sampling method to be used in this study, was faced with a myriad of choices. By taking into account the purpose of the study, the unique characteristics of the group, the size of the population as well as the survey tool being used, the decision was made to make use of a purposive sampling method called 'total population sampling'. Had one of these elements been slightly different, this method would not have been seen to be suitable or practical. An example of the study design leading to the choice of sampling method is the following: had this researcher decided on

using in-person interviews instead of an online questionnaire, using total population sampling would not have been the chosen sampling method. As this researcher was directly involved with a previous CSIR RDM survey (Patterton, 2014a), making use of in-person interviews, and has personal experience of the amounts of time, resources and efforts required to collect and process data using this method of data collection, the use of the total population sampling method would not have been a practical choice then. This study, using a free online questionnaire to collect data, and being less demanding with regards to time, resources and effort, was suited to the total population sampling technique.

3.7.2 Ethical concerns, ethical clearance and managerial permission

Research ethics, in layman's terms, is seen to be adhering to what is 'right and correct' (Strydom, 2000: 24). Put another way: ethics are concerned with treating people fairly, and not hurting anyone (Lichtman, 2014:56). When talking about research ethics in a more formal manner, it can be defined as the ethics of the 'planning, conduct and reporting of research' (Resources for Research Ethics Education, 2013). It can also be said to refer to the 'ethical norms, codes and regulation' which govern current research practices (Farrimond, 2013:13). It is primarily, but not exclusively, concerned with the protection of human and animal subjects.

This study, involving human subjects and their research behaviours, was subjected to ethical clearance before gathering of data could commence. Adding on to this: as a qualitative study it was particularly vulnerable to questions of ethical and political conduct (Glazier & Powell, 1992:201). Research aspects deemed important of investigating by the respective committees before ethical clearance was granted, are discussed in the sections below. Despite the fact that this study was subjected to ethical clearance and guidelines given by two Research Ethics Committees (CSIR as well as University of Pretoria), the final responsibility for ethical conduct lay with the researcher.

Adding on to the line of thought that final responsibility lies with the researcher, literature sources advise researchers to see research ethics as wider than the aspects stated on an ethics clearance form. According to Farrimond, ethics and research are indivisible, and is increasingly embedded in the research process itself (2013:58). Although ethical clearance was required before data collection could commence, this step was not be seen as separate from the research process, nor was adherence to the ethical guidelines supplied by both entities seen as the ultimate in ethical research. Braun & Clarke (2013:61) state that research ethics should be seen as something permeating the whole research practice, not as a separate stage or a 'hoop to jump through'. Ethical clearance obtained, as well as the

guidelines supplied, should be seen as the lowest level of ethics, and not the pinnacle (Braun & Clarke, 2013:61). So while micro-ethics, referring to subject consent, confidentiality, avoidance of subject harm, informing subjects (Brinkmann & Kvale, 2006:167) were evaluated by the Research Ethics Committees (RECs), it was also important to consider how the knowledge obtained would affect humans and society as it circulates in the wider culture. These considerations, termed macro-ethics (Brinkmann & Kvale, 2006:167), support the idea that research ethics is an omnipresent practice, and not limited to micro-ethics, referring to ethics at the level of research participants, only.

The implication of this line of thought for this researcher was that ethics formed part of all aspects of this research project. Put another way: while it was critical to obtain the go-ahead from an ethics committee and gain access to the field, it was equally and even critically important to maintain trust and a working relationship while research continued. According to Braun & Clarke (2013:61), such a relationship, characterised by trust and respect, and adhering to ethical principles, should even extend to the academic community as well as the wider world in which we conduct research practices. Farrimond refers to this ubiquity as a 'lifecycle' approach to research ethics: research ethics is not something complete by the writing of an ethics proposal; it is also not over once approval is received (2013:59). It includes ethical conduct when deciding on topic choice, design, data collection, analysis, dissemination, and avoiding all other potential transgressions.

This researcher, apart from adhering to obvious aspects such as anonymity, confidentiality, informed consent, and no harm to respondents, also needed to ethically uphold practices such as ethical data storage, and absence of value judgements pertaining to respondents. With regards to the release and publication of findings, it needed to be done in an accurate, objective, clear, unambiguous manner, containing all essential information. In addition to these, aspects mentioned by Farrimond (2013:69) such as 'data fishing' (the use of data mining methods to sample parts of a larger population data set that are/or may be too small for reliable statistical inferences to be made about the validity of any patterns discovered), manipulation of data, and the practice of excluding outliers, which should never form part of data analysis, needed to be avoided. An objective approach also needed to be used when informing respondents of the research results.

Adding on to these ethical aspects, Farrimond (2013:67) is of the opinion that conducting research with people known to the researcher, or at an institute where the researcher is already working, might present potential ethical issues. This researcher, knowing some of the respondents, and being a permanent employee of the institution where the research will be conducted (CSIR), took cognizance of this danger.

As mentioned earlier, this study, dealing with human subjects and their research behaviour, required ethical approval before data the study could commence. Being a full-time employee of the CSIR, as well as using this study for the purposes of obtaining a degree at the University of Pretoria (UP), meant that ethical clearance provided by the RECs of both these institutes, needed to be obtained. The remainder of this ethics section discusses the process as encountered at each of the two institutes. The CSIR Research Ethics Committee insisted on proof of ethical clearance from the University before considering ethical clearance of the study; as a result of this, ethical clearance obtained is discussed in the order in which it was obtained.

3.7.2.1 Ethics approval: University of Pretoria

In short, obtaining ethical clearance from the University of Pretoria entailed the following steps:

- Research may not be done without the prior written approval by an Ethics Committee.
- Each faculty has its own procedures to be followed.
- Ethical clearance submission was done online, by accessing the Ethical Clearance link on the University of Pretoria's webpage for the Faculty of Engineering, Built Environment and IT.
- The online ethics application form completed by this researcher, required information on the applicant, as well as research project details (problem, objectives, methods, materials, target group profile, as well as a copy of the questionnaire).
- Detailed information about the research subjects, including safety and health implications of participation, duration of participation, and handling of confidential information, needed to be supplied as well. In addition to this, information on the study's environmental impact, plus the dissemination of data, were also requested.
- The application was approved and an ethical clearance certificate was issued.

Upon receipt of the ethical clearance certificate from the University of Pretoria, this researcher proceeded with applying for ethical clearance from the CSIR Research Ethics Committee, discussed in the next section.

A copy of the wording of this ethical clearance certificate can be found in Appendix 3.

3.7.2.2 Ethics approval: CSIR

When perusing the CSIR's policy on Research Ethics, the term 'research ethics' was seen to be the 'moral principles guiding research, from its inception through to completion and publication of results and beyond' (Mathabe & Dlamini, 2010:2). This view is in agreement with statements earlier that ethical conduct should not be seen as separate from research, and that it needs to be implemented throughout the research cycle.

While the policy addresses issues such as its purpose, benefits, links to other policies and its regulatory framework, it is of more importance now to discuss the actual process of ethics approval, as experienced by the researcher. Using the policy, and taking into account an ethics approval submission to the CSIR Research Ethics Committee (REC) in 2013, the process happened in the following manner:

- This researcher, upon studying the CSIR's policy on Research Ethics, discovered that research proposals containing ethical issues involving/impacting human subjects/tissue, animals studies/aquatics, genetic manipulation of biological organisms, and research which might have a negative impact on the environment, need to be evaluated by the CSIR REC.
- As this study is free from ethical issues which might impact subjects and areas indicated in the first bullet point, this researcher was not required to submit an ethics clearance form per se. Feedback from the CSIR REC indicated that an expedited ethical procedure would suffice for this study.
- Instead, this researcher requested consent in proceeding with research. This involved downloading a two-page form from the CSIR intraweb, and electronically completing fields indicating details of the project, objectives of the study, critical research questions to be asked, and research approach to be used.
- In addition to the completed form, this researcher also needed to submit:
    - an outline of the online questionnaire,
    - a copy of the memorandum to be submitted to unit directors, requesting permission to contact the emerging researchers in their units,
    - a copy of the research proposal submitted to the University of Pretoria,
    - a copy of the proof of registration at the University of Pretoria for the period 2015, and
    - proof of ethical clearance from the University of Pretoria.
- The REC, after evaluating the application for ethical clearance, provided committee feedback to the researcher. Ethical concerns, areas/questions/activities in need of

amendment or removal, as well as recommendations, were conveyed to the researcher.

- Upon proof of implementation of recommendations, the researcher was given permission by the REC to commence with the proposed study.

A copy of the wording of this ethical clearance certificate can be found in Appendix 4.

3.7.2.3 Ethics summary

This study made use of human informants, and investigated their research data management practices. The inclusion of human informants meant that ethical approval from the institute where the research would be conducted, were to be obtained. In addition to this, this researcher also needed ethical clearance from the institute where this study would be submitted for postgraduate degree purposes. This researcher wished to emphasize that ethical research did not end once ethics approval was granted: this researcher endeavoured to adhere to ethical principles throughout the whole cycle of research, regardless of whether a particular practice was stipulated in the ethics application form.

3.7.2.4 Managerial permission

Apart from getting ethical clearance, this researcher also needed to gain the approval and official permission from relevant authorities at the CSIR. In the previous CSIR RDM survey, permission to conduct a survey entailing in-person interviews was obtained from the various CSIR units' strategic research managers (SRMs). This researcher, upon requesting requirements needed before obtaining research consent, was informed that CSIR research unit directors needed to be informed of the intended study, the online questionnaire to be used, and researchers to be involved. As a result of this, a memorandum was drafted, signed by this researcher's competency area manager at the CSIR, and emailed to each of the directors. In addition to this, copies of the cover letter, the questionnaire, as well as the research proposal, were also included in the email. A copy of this memorandum can be seen in Appendix 5.

Prior to circulation of the memorandum, it had been submitted as an accompanying attachment to the CSIR Research Ethics Committee (CSIR REC) when requesting consent to conduct research. Consent was given, and the memorandum evaluated by the CSIR REC as an acceptable way to request managerial permission.

All research units gave permission for the survey to proceed.

3.7.3 Questionnaire administration and recruitment

After receiving input from experts and pilot run candidates, and amendments being implemented, the final questionnaire was considered ready for distribution. Ethical clearance as well as CSIR unit managerial approval had already been obtained, meaning that this researcher was able to proceed with administering the final version of the survey to the targeted sample. Wallace & Van Fleet (2012:205) mention that questionnaire administration involves several chronological activities: development of a cover letter, actual distribution of the questionnaire, and follow-up activities.

3.7.3.1 Cover letter

Cover letters are considered vital to any questionnaire, and provides a 'persuasive introduction' to the survey, with the intent to motivate respondents to complete and return the questionnaire (Wallace & Van Fleet, 2012:205). In agreement with this, Singleton & Straits (2005:249) are of the opinion that a good cover letter enhances respondent cooperation.

A study of methodological literature shows that a good cover letter has distinctive traits; the following traits were applied to the questionnaire in question:

- A cover letter should be limited to one page (Dillman, 2007:158).
- Critical information include date, salutations, what the letter is about, why the study is useful and important, guarantee of confidentiality of answers, voluntary participation, contact details in case of questions, addition of a post-script, selection of a suitable mail-out date (Dillman, 2007:177).
- An explanation of how the sample was drawn (Singleton & Straits, 2005:249) is vital.
- An undertaking to supply respondents with the results of the study (Singleton & Straits, 2005:249), should be made.

While the inclusion of a financial incentive or token of appreciation is suggested by some (Dillman, 2007:167), this activity was not included in the current study. The reason for it not being considered is based on the fact that this study was not merely part of this researcher's master's degree studies; it was also a CSIR project and an activity done by a permanent employee of the CSIR and considered to be one of her daily work tasks.

An example of the cover letter used with this study's questionnaire, can be found in Appendix 7.

3.7.3.2 Distribution of the questionnaire

Hank, Jordan & Wildemuth (2009:261) mention that there are several ways to increase the response rate of a survey. One such method is to contact potential respondents multiple times. A five-contact framework to increase response rate is also suggested by Dillman (2007:150): these steps were studied by this researcher and considered to be in part worthy of implementing in this study. The five steps, outlined, are the following:

- step 1: a short, pre-notice contact to all members of the sample. This entails an email briefly explaining the project, study, target population, sample, data collection method, and data of survey commencement,
- step 2: distribution of the survey to study sample,
- step 3: a thank-you/reminder note a week after initial distribution,
- step 4: a second reminder about two weeks later, and
- step 5: a fifth follow-up contact using a different means of contact; those who had not completed the survey to be contacted by the researcher.

For the purposes of this study, this researcher decided on making use of four steps only, and did not include the fifth step. Implementing the first four steps and adhering as far as practically possible to its features and requirements, was deemed to be sufficient for this study while at the same time not causing the candidate to experience undue irritation, or pressure to participate. These steps, each different to the other and serving a different purpose, will be discussed in more detail in section 3.7.4.

The suggestion by Dillman (2007:152) that all contacts, regardless of type, should be personalized, and not generic in nature, was adhered to as far as the survey software allowed. An advantage of using eSurv as survey software was that automatically-generated emails could be tweaked to address each survey candidate by a title and name chosen by this researcher.

3.7.4 Contacts to be implemented

An elaborate description of the questionnaire administration processes is given by Dillman (2007:149-213), who refers to these steps as an ‘implementation process’ (2007:188). He mentions that the process should involve two definite points of contact: a first contact pre-notice letter, and a second contact questionnaire mailout including the cover letter. To improve response rates, he suggests a third contact involving a thank you note or a reminder, a fourth contact mentioning involving a more insistent tone and portraying to respondents a strong form of personalization as well as individual attention (Dillman,

2007:181). This contact reinforces the message that the respondent is important to the success of the survey. A fifth contact, which is a final effort to elicit a response (Dillman, 2007:184), will exhibit a greater overall intensity than the previous four contacts. The heightened intensity is not due to wording, as the fifth contact wording is softer than the preceding contact, but due to a different method of contact than the preceding contacts, being used.

This researcher, taking into account the reasons candidates might have for not completing the questionnaire (e.g. work schedules, not comfortable with revealing RDM behaviour, not being near a network computer due to fieldwork) and weighing it up against the need to receive as many as possible completed questionnaires, decided to use four contacts during the course of the study. These contact points were:

- initial pre-questionnaire email,
- questionnaire with cover letter,
- thank you note/reminder, and
- final email contact.

Each of these activities, forming part of the questionnaire administration phase, is discussed in more detail below.

3.7.4.1 Initial pre-questionnaire email

The purpose of this communication, according to Dillman (2007:156) is to provide the candidates with a 'positive and timely notice' of a future request to help with an important study/survey. He also states that this first contact needs to adhere to specific criteria: it needs to be brief, personalised, positively worded, and aimed at providing anticipation rather than details.

According to Clark, Dillman & Sinclair (1993:37) a pre-notice improves response to mail surveys; it is this researcher's wish that the same effect will be true for online questionnaires.

A copy of the pre-questionnaire notice can be found in Appendix 6.

This pre-questionnaire notice was sent out a week before the questionnaire was distributed, and was sent via email. Although brief, it included a date, inside address, a short description of what will happen (in other words, informing the candidate that she will receive a request to complete the online questionnaire), a short description of what it is about (mention will be made of the research questions, a short description of the usefulness of the study, a short thank you paragraph, and a signature.

3.7.4.2 Questionnaire distribution and cover letter (including informed consent form)

Approximately one week after the questionnaire pre-notice had been sent out, an email containing a link to the questionnaire, mounted on the World Wide Web, was sent to candidates. This email can also be labelled as the questionnaire's cover letter. The cover letter was discussed in detail in section 3.7.3.

A requirement of this study, namely that respondents should sign a consent form before proceeding with the questionnaire, was included with the cover letter. It appeared as a multiple choice question (consent was given or not given by answering yes or no to three consent-related statements), and was placed after the cover letter and before the first questionnaire question.

A copy of the cover letter used with this study's questionnaire, can be found in Appendix 2.

A copy of the informed consent wording forms part Appendix 2.

In addition to containing the wording of the cover letter as well as informed consent section, Appendix 2 also contains the remainder of the survey wording (i.e. questions, as well as answer options).

3.7.4.3 Reminder/thank you letter

Questionnaire follow up activities are done in an attempt to urge recipients to respond (Wallace & Van Fleet, 2012:206). When studying literature related to follow-up activities, various aspects are stated to be important considerations. This researcher implemented the most prominent of these considerations in the following manner:

- Wallace & Van Fleet suggest using a slightly different cover letter (should not look identical to the original letter), to be sent to non-responsive recipients (2012:206). Dillman (2007:179) agrees with this change in format, and mentions that when measuring questionnaire response rate, repeated stimuli are less effective than new stimuli. This contact's main function was to serve as a reminder to the recipient to complete the questionnaire. Once again, a link to the questionnaire being hosted on the internet, was supplied. All candidates, not only non-responsive candidates, will be contacted.
- Dillman (2007:179) mentions that the wording of a reminder letter should be given due consideration. For some people, due to missing the previous email, this could have been the first time they had learnt that a cover letter as well as a link to a questionnaire was sent to them. For this reason, the first few lines of the reminder

stated that a questionnaire link plus cover letter had been sent a fortnight ago, and also stated why it was sent. Dillman states that subsequent reminder letter information should include a statement of gratitude to recipients who had already completed the questionnaire, followed by a request to others to complete the survey 'today' (2007:180). In addition, a sentence amplifying the message of how important each recipient's contribution is to the study, was required. Furthermore, a link to the questionnaire was supplied with the reminder letter. For the purposes of this study, this inclusion was aimed at those who did not read or notice the previous email, or who might have deleted it.

- Although a postcard reminder as well as mailed reminders are suggested as ways to remind recipients (Wallace & Van Fleet, 2012:206), this researcher was of the opinion that these methods are outdated and less reliable than emailing a reminder.
- Wallace & Van Fleet (2012:206) also regard the timing of a reminder to be an important issue. It is stated that the researcher should be able to make an educated guess when the original questionnaire letter had been received, and how long it would have taken to complete the questionnaire. Dillman (2007:179) is of the opinion that one week is an appropriate interval of time for making an appeal that conveys a sense of importance, while at the same time does not sound unreasonable or impatient. The danger of a reminder being sent too close to the initial contact, is that a recipient might feel nagged, while a reminder sent too late could result in irrelevant or stale data (Wallace & Van Fleet, 2012:206). In this study, this researcher chose a period of one week to be the time lapse between the original questionnaire request and link sent out, and the mailing of a thank you note/reminder.
- Confidentiality and anonymity issues are also aspects to be considered important during the follow-up (Wallace & Van Fleet, 2012:207). Obviously, if a researcher is able to tell that a recipient has not completed a questionnaire, then he is clearly not anonymous and cannot be assured of confidentiality. In this study, as all candidates were contacted during the reminder/thank you phase, no additional assurances of confidentiality were required, nor were concerns raised by researchers.

A copy of the reminder letter used in this study can be found in Appendix 8.

3.7.4.4 The first 'replacement' questionnaire

Dillman (2007:181) is of the opinion that a fourth contact, which entails sending candidates a replacement questionnaire, and marked by a tone of insistence, should be made to remind recipients that their completed questionnaire has not been received. After careful consideration of the benefits and drawbacks of this contact, this researcher decided not to

implement this stage. The main reason for excluding this contact was that this researcher was wary of over-requesting a completed questionnaire from fulltime CSIR employees not receiving any kind of financial incentive for completion. In addition, it would have been extremely difficult to target those who have not yet responded as respondents are not obliged to provide their names. This researcher, for the sake of keeping good relations with future CSIR RDM-policy adherents, was wary of being seen as demanding, impatient or not understanding of heavy workloads.

3.7.4.5 Final contact: email to candidates

Dillman states that this contact is characterised by greater overall intensity than any of the previous contacts, not through its wording, but through the method used by the researcher when approaching non-responsive candidates (2007:184). He suggests that a different method of contact should be used; with this suggestion in mind, this researcher considered making use of telephonic contacts when contacting candidates for the last time. However, previous experience when trying to contact CSIR researchers telephonically, and discovering that they were often not working close to their office phones, or would make frequent use of voicemail, suggested that this might not be an ideal way to make final contact. A mailed letter was also considered, but not regarded by this researcher as being a suitable way of reaching out to candidates.

As a result of these stated limitations, this researcher again made use of email when contacting candidates for the last time. This final email contained the following information:

- The candidate was informed that a questionnaire and cover letter was sent to them previously.
- The letter explained that the contact was made to see if they have questions about the study.
- Candidates were encouraged to complete the survey.
- A link to the online questionnaire link was provided; previous emails might not have been accessible or might have been deleted.
- Candidates were thanked for their consideration, and ensured that they would not be contacted again.
- Wording was more relaxed than previous contacts.
- Usefulness and importance of each respondent was stated again, but different wording was used.
- It was important that the final contact had a different ‘look and feel’; wording, tone, message as well as overall appearance were not be copies of the previous contacts.

With regards to timing, Dillman (2007:188) advises that this contact should not be later than a week after the previous contact.

A copy of the outline of the final contact used by this researcher can be found in Appendix 9.

3.7.5 Dynamics of the questionnaire administration process

In an ideal research scenario, all sampled candidates would complete the distributed questionnaires, and no feedback from the researcher would be requested at all. In reality, researcher-involved activities such as handling returned incomplete questionnaires, and ad hoc questions from candidates directed towards the researcher need to be anticipated and treated in an ethical and professional manner so as to not influence the cooperation from emerging researchers or invalidate the findings.

According the Dillman (2007:189) the study researcher is bound to deal with dynamic and unpredictable respondent behaviour. During the course of this study, this researcher treated these behaviours in the following manner:

- Automated return e-mails stating that candidate is out of office: this researcher decided on a period of one month between the send out of the questionnaire, and the commencing of data analysis. No questionnaires were completed and returned after the one-month period had expired.

No ad-hoc and unique queries from candidates took place during the questionnaire administrating activities of the questionnaire.

3.8 Data analysis and data presentation

The final methodological aspects discussed in this chapter are procedures and activities implemented when gathered data were analysed. Data analysis methods used, as well as the method of presentation to the reason, are discussed. The reasons for making use of the mentioned procedures, as opposed to other methods, are also be touched on briefly. Key concepts of importance to the data analysis of this study, such as the use of spread sheets, the creation of data documentation, and the levels of measurement to be used in this study, are also be highlighted and expanded upon.

According to Wallace & Van Fleet the purpose of data analysis is twofold: it prepares the data for use (2012:265), and it makes sense of the data (2012:266). Applied here: this researcher analysed the data obtained via returned questionnaires in order to use it in

various ways, to understand, and enable readers to understand, what information is portrayed by the data.

Data analysis in this study was done by making use of the features offered by the online survey tool, Microsoft Excel as spreadsheet, and the calculative as well as graphical features offered by Microsoft Excel. While performing the calculations, analysis and data presentations herself, this researcher also made use of a CSIR statistician for advisory instructions.

3.8.1 Spreadsheet and data documentation

Data analysis commenced a week after the fourth contact (see section 3.7.4) was made. Data responses were recorded, displayed and summarised by the survey software. Esurv as a software tool records, displays and summarises the questionnaire responses in a simple, straightforward and user-friendly manner, and made the transfer to a spreadsheet an easy task. The creation of general purpose spreadsheets, consisting of questionnaire values and variables exported from the software to the spreadsheet, was the first data analysis step. Editing and cleaning of data was also performed. Prior to the creation of the spread sheets, this researcher decided on coding, as well as abbreviations and acronyms to be used in the spread sheet.

This researcher made use of Excel when creating spreadsheets. These simple spreadsheets formed the basis of all calculations done, made up the dataset connected to this study, and is also the format in which the data will be shared with interested parties. As such, it was vital that spreadsheets should make sense to the researcher as well as users of the data sets. With these aspects in mind, this researcher created the spreadsheets to be simple, user-friendly, easily understood, not prone to misinterpretation, and in a well-known, easily shareable format. Furthermore, data documentation accompanying the datasets and to provide additional information on the data collection tool, the data collection method, as well as the variables, and values forming part of the dataset, was compiled. Explanatory details on the coding, acronyms and abbreviations used within the spreadsheets also formed part of the data documentation. Data documentation connected to this data set was captured in text format (MS Word), amounted to 1 MB in size, and accompanied the dataset when added to a data repository or shared with interested parties. A copy of the data documentation forming part of this study can be viewed in Appendix 11.

This researcher aimed at keeping spreadsheet complexity to a minimum, and managed to succeed in using a single spreadsheet page per question. The total size of the dataset, in Excel format, came to 32.1 kilobytes.

During this stage of data analysis, while viewing the data created by the software tool, creating a spreadsheet, copying it to a spreadsheet, and coding the data or variables, this researcher inspected the data in order to determine the need for data cleaning. Data cleaning entailed activities such as checking for missing data, checking for duplication, as well as incorrect coding. This part of the data analysis process was also mentioned in the bullet points listed just before section 3.8.1. The process of data cleaning was considered by this researcher to be crucial to the data analysis, as ignoring it could have led to misleading research findings.

3.8.2 Univariate analysis

Univariate analysis, according to De Vos & Fouche (2000:204) is the simplest form of data analysis. This study made use of this type of data analysis. Univariate analysis entails the analysis of only one variable at a time, and is done with a view to describing that variable. In practical terms, this would mean that each variable, on its own, is discussed and displayed in the results chapter.

3.8.3 Levels of measurement

A variable can be classified according to its nature/type, and could be either a nominal, ordinal, interval or ratio variable in nature (Wallace & Van Fleet, 2012:137). A nominal variable, according to Pickard (2013:284) is one which is classified by categories which cannot be ordered; these categories will have names. This study, investigating behavioural trends and common denominators within the RDM habits of researchers, made mostly use of nominal data. An example of a nominal variable in this study was the issue of data formats used by emerging researchers. ‘Text’, ‘numerical’, ‘images’ and ‘video’ are examples of nominal variables as they cannot be ordered. Nominal variables, also known as categorical, qualitative or discrete data (Pickard, 2013:285) could then be said to possess different categories, but no obvious rank order. Categories, such as ‘text’, ‘numerical’, ‘images’ and ‘video’ were chosen by respondents in response to a question, but could not be quantified in a meaningful way, such as being subjected to arithmetic operations. This data could be coded, but this type of coding is only a numerical label used to identify the variable, and is not a meaningful numerical quantity.

Coding of nominal data, according to Singleton & Straits (2005:87), is only done for the researcher’s convenience when collecting and analysing data. When it comes to arithmetic operations, nominal data do not allow much in the way of mathematical relationships. For instance, this researcher was not able to state that ‘text data + numerical data = video data’, as all cases placed in the same category (i.e. data format) were equivalent.

Ordinal data, where the order of data points can be determined, but not the distance between points (Pickard, 2013:284), also formed part of the data collected, but to a lesser extent. As stated by Pickard (2013:285), ordinal data portray an inherent logical order in choices, even though the distances between points are arbitrary. Examples of ordinal data used on this study were the indications of frequency of data backups, or an estimation of the data storage size requirements. Ranking certain required RDM services would also be an example of an ordinal variable. According to Singleton and Straits (2005:88), ordinal data display the proper order of numbers, but do not indicate the distances between numbers. Thus, within this study: researchers might have indicated that the need for an RDM policy is greater than the need for a data repository, and that the need for a repository is greater than the need for a metadata scheme. Although a ranking could be indicated, the exact distance, or quantifiable level of need between services could not be determined.

Interval data are described as having the qualities of the nominal and ordinal variables, but also has the required equal distance or interval between numbers representing equal distances on the variable being measured (Singleton & Straits, 2005:89). In other words, as stated by Pickard (2013:84) both the order of data points as well as distance between data points can be determined. This type of variable did not feature in this study.

A fourth type of variable, called ratio data, is stated by Singleton & Straits (2005:89) to include the features of the previous data types mentioned, but would also have an absolute, non-arbitrary zero point. As is the case with interval data, ratio data did not feature in this study.

3.8.4 Measures of central tendency

Wildemuth (2009b:343) states that when reporting on data, even in simplistic terms, it is important that at least one measure of central tendency and one measure of dispersion be used. Wildemuth also mentions that the idea with the measurement of central tendency is to calculate or identify one number, or value, that describes an entire variable (2009:339). Sources on data analysis are unanimous in their listing and description of the types of central measurements can be found: the mode, the mean, and the median are the measures available. The mode is described as the value or category with the highest frequency, the mean is the mathematical average, while the median is the midpoint in the distribution, or the value of the middle response, where half of the responses are above it and half are below it (De Vos & Fouche, 2000:215, Pickard, 2013:288, Singleton & Straits, 2005:462).

The type of central tendency measurement to be used, would be dependent on the level of measurement used in the variable that one is trying to summarise. As stated by Wildemuth

(2009b:339): if one is making use of nominal data, then only the mode can be used. When making use of ordinal measurements, then the median is most commonly used, and to a lesser extent, the mode as well. In practical terms: this study made mostly use of the mode when analysing the data, but the median was also used. Two real examples were the following: when reporting on a central tendency using nominal data, such as a variable investigating data formats used, the category (data format) indicated to be most frequently used, and thus the mean, was supplied. With ordinal data, such as a ranking given to a specific service, the mode or the median was used. This would mean that the researcher could either have reported that the majority of respondents shared data most frequently on an annual basis (mode), or could have indicated which sharing frequency was the middle value.

Advantages and disadvantages of these two measures do exist. The mode is considered to be the weakest measurement, lacking the precision of other measures (Pickard, 2013:290), as well as not being a representative value when the majority of values cluster around a value that differs from the mode (Pickard, 2013:290). Yet, as stated by Wildemuth (2009b:341) it is the only measurement of central tendency that is appropriate when a researcher is working with nominal data. Its advantages include ease of obtaining ((Pickard, 2013:290) and not being affected by extreme scores in a skewed distribution, or outliers (Wildemuth, 2009b:341). With regards to the median, a disadvantage could be that the median value may not be characteristic of the dataset; on the other hand, being insensitive to extreme scores, it is a good value to report on when the distribution is skewed. As such, Pickard (2013:289) states that it is the most useful average when dealing with skewed data.

Use of the mean as measurement of central tendency was not used in this study. The reason for its exclusion is that this study was limited to nominal and ordinal level data only, and as stated by Pickard (2013:288), the mean can only be used when the researcher has interval or ratio level data, making a numerically meaningful interpretation possible.

A frequency distribution is another way of indicating measures of central tendency, and is seen as the most elementary summary and display of data collected on one variable. Pickard (2013:286) states that the frequency distribution is one of the first stages of analysing data, and a way for the reader to make sense of the results presented. The elements of a frequency distribution table may vary depending on the level of measurement used (Pickard, 2013:286), a frequency distribution table using ordinal data will be more simplistic than one making use of interval data. Within this study, and its presence of ordinal level data, a frequency table portraying frequency of data backups would include elements such as:

- column 1: value, or category of a variable being measured (e.g. never, daily, weekly, monthly, annually),
- column 2: frequency of response to each category of each variable,
- column 3: percentage of respondents from entire sample that belong to that category,
- column 4: valid percentage of respondents that belong to that category, and
- column 5: a cumulative percentage giving a rolling addition of the valid percentages of responses; this last column can only be used with ordinal and interval levels of measurement.

Although frequency distributions are often displayed in table format, they can also be displayed using a variety of graphic ways. De Vos & Fouche (2000:204) mention that bar graphs, histograms, frequency polygons, pie charts and pictograms are used to display frequency distributions. This researcher discusses the visualization of datasets later in this chapter.

3.8.5 Measures of dispersion

Wildemuth (2009b:341) states that it is often necessary to show how far the indicated values spread out around a central point, and that this is the goal of reporting on the measure of dispersion. The choice of measure of dispersion would depend on the measurement of central tendency used; according to Wildemuth (2009b:341) when one uses the mode or median, then the range, or interquartile range as measure of dispersion is used. The range would indicate the distance from the lowest to the highest score. An example of the practical application of the measure of dispersion in an RDM study would be the indication of the range that exists when participants indicate the size of a typical dataset.

3.8.6 Data correlation

Data correlation can be seen to be the relationship between two variables. As stated by Wildemuth (2009a:375): correlations assume that two variables studied are linearly related and that data are either ordinal, interval or ratio in measurement level. As interval and ratio data do not feature in this study at all, and ordinal data only used to a limited degree, this researcher did not make extensive use of correlations during data analysis.

3.8.7 Visualization of data

The aim of data visualization, or displaying the data, is to help both the researcher and audience to understand data. In this study, this researcher made use of tables as well as graphical content to assist herself and the reader in this regard. Tables, according to

Wallace & Van Fleet, (2012:274), are excellent tools when wanting to display a summary of data, as they are able to present a simplified view of a complex phenomenon. In this study, tables were used to report on data most effectively displayed when using text or numbers; examples of these variables could be tables displaying frequency distribution. In such a case, for instance when displaying the frequency distribution according to CSIR unit representation, a table making use of text as well as numbers, was used.

Data can also be represented graphically, or by using figures (De Vos & Fouche, 2000:209). While tables are able to provide a detailed summary of frequency distributions and other textually-oriented data, graphic representations may help the researcher and the reader to understand the data. Furthermore, it may also make the relationship between variables more understandable. In short, it could be said that graphs serve the same purpose as tables, but that graphs emphasize data relationships through visual cues. Wallace & Van Fleet (2012: 276) are of the opinion that graphs are a very effective tool when trying to present data, usually qualitative in nature, visually.

Wildemuth (2009c:350) mentions that when making use of graphs, the researcher should always aim at showing data clearly, at making large data sets coherent, and at avoiding data distortion. Wildemuth (2009c:353) further warns against deceiving with graphs, or using graphs in such a way as to not aid the reader in understanding the data. This researcher was aware that while graphical portrayal of data is a helpful and value-adding tool, she needed to guard against misrepresentation of data when using graphs. While these activities or misdemeanours are stated to occur largely unintentionally, it is worthwhile taking note that it is often found to occur when data are too complex, or when one of the axes has been manipulated leading to differences between categories being over- or understated. As this researcher made extensive use of graphs, the above-mentioned concerns and pitfalls were duly be kept in mind. Practical suggestions, such as Wildemuth's suggestion that a pie chart should display no more than six categories in order to avoid clutter (2009b:341), were applied. Thus: while this study lent itself to heavy use of graphs, this researcher was aware of scaling, perspective and aspect ratio issues which might have rendered the graph to a cluttered look, an unclear message, or misinterpretation.

With data measurement levels in this study confined to nominal and ordinal data, this researcher needed to ascertain which types of graphical representation would be more suited to this type of data. It would seem as if graphs such as pie charts and bar charts are useful for displaying nominal and ordinal data (Wildemuth, 2009c:351). Pie charts were most commonly used by this researcher to show how a particular variable is distributed, as this type of graphics tool illustrates the corresponding 'slice of pie' of each category.

Therefore, within this study, variables such as the use of metadata when managing data, or the amount of RDM training undergone by emerging researchers, were variables suited to portrayal via pie charts. Wildemuth (2009c:351) states that pie charts should not be used to show frequency within multiple choice answers where participants were able to indicate more than one applicable answer; a bar chart would be more suited to this type of question. In practice, this would have meant that the graphical representation of a question asking emerging researchers to indicate all the places their research data are being stored, would be a bar chart, and not a pie chart.

According to De Vos & Fouche (2000:209), bar charts are ideally suited to variables studied at the nominal level, while Pickard (2013:293) mentions that it is ideally used to show frequencies or percentages. Three forms are identified: simple bar charts, compound bar charts, and component bar charts (Pickard, 2013:293). Examples of the use of each within this study are as follows:

- simple bar charts: this type of graphic were used to indicate the frequency distribution of researchers' data storage requirements, or to indicate the CSIR unit the respondents belong to,
- compound bar charts: this type of graphic is used for a variable having more than one category, and encourages visual comparison (Pickard, 203:294). This researcher used this type of chart when showcasing the RDM training requirements of emerging researchers. Being a multiple choice question, with each respondent able to tick as many answers as are applicable, it lent itself ideally to such a graph, and
- component bar charts: these charts are used when displaying data that needs to demonstrate comparison, but instead of using multiple bars representing each piece of data connected with the group, all data elements are included in a single bar (Pickard, 2013:295). The same data variable as was discussed in the previous bullet point can be displayed using a component chart; the difference being that all categories of each variable would now be captured and visible within a single bar.

The flexibility of bar charts, being a tool able to represent a variety of different distributions as well as displaying variables of varying complexity, made it an ideal tool to be used in this study. This researcher was in agreement with Wildemuth (2009c:351) who emphasises that bar charts are particularly useful when trying to compare across categories, or for seeing trends across categories. This study, making use of a questionnaire where the majority of questions involve many categories per variable, made extensive use of bar charts in all its forms. As mentioned earlier, pie charts were also used; the latter being more ideal when illustrating six categories or less. Bar charts were used in instances where more complex

categories were illustrated, or where the use of pie charts would have resulted in a cluttered appearance.

3.9 Creation of a data management plan for this study (DMP)

Based on RDM best practices, this researcher has developed a data management plan for this study (see Appendix 1). As stated in Chapter 1, section 1.2, RDM is concerned with practices such as the planning, creating, storing, organising, accessing, sharing, describing, publishing and curating of data. Therefore, this researcher, in order to demonstrate RDM best practices, has followed the DMP guidelines supplied by the Digital Curation Centre (Jones S, 2011), and demonstrated that:

- a DMP has been created,
- data have been created/generated/collected using approved, valid and reliable methods and tools,
- data have been stored making use of more than one location,
- backups have been made,
- storage as well as backup locations are secure,
- data have been organised by making use of a personalised naming convention, enabling recognition and identification of the dataset,
- data access is well-managed, with confidentiality, sensitivity and ownership concerns addressed and applicable access measurements in place,
- data sharing issues, such as the use of a sharing agreement, sharing restrictions, and sharing timescale for release, are stated,
- expected difficulties when sharing data as well as measures to overcome these, are addressed,
- data documentation, as well as metadata pertaining to the data, was created,
- data documentation allows a fellow researcher to fully understand the data, and provides contextual details about how and why the data were created,
- metadata, as a subset of this documentation, describes the dataset in detail,
- adherence to Dublin Core as metadata standard was shown,
- data generated during this research are to be transferred to an accredited data repository, post-publication, and
- a Digital Object Identifier (DOI) still needs to be assigned.

Data curating practices have been considered, and will be implemented in due course. The dataset has been judged to be of value for at least ten years. As such, a decision still has to

be made regarding its archival format, and ways of assuring that data will remain accessible, still needs to be shown.

3.10 Chronology of research

An abbreviated version of the methodological steps followed, and described in this chapter, is supplied below:

Research proposal completed/Literature analysis completed/DMP created

Choice of research design made

Choice of data collection tool (questionnaire)

Choice of questionnaire (online questionnaire)

Mock survey to test online questionnaire functionality

Formulation of questions (question content, question type, wording)

Questionnaire appearance finalised

Pretesting (using experts, feedback received, amendments made, questionnaire finalised)

CSIR research consent requested & obtained

CSIR directors informed, permission to involve researchers obtained

University of Pretoria ethical consent requested & obtained

Sampling completed

First contact: Pre-survey notice distributed to survey sample

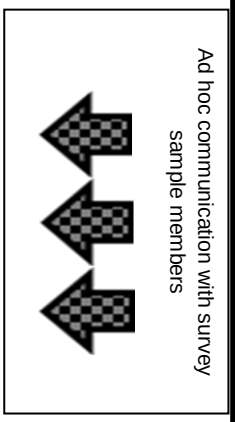
Second contact: Survey distributed to survey sample

Third contact: Thank you/reminder letter sent to survey sample

Fourth contact: Special procedures

Data analysis

Data presentation



3.11 Summary

This chapter aimed at describing the methodology used by this researcher when gathering data to be used to answer the study's research questions. The qualitative research approach was described, and the use of survey as research method was given due attention. Furthermore, the use of an online survey, together with its features, advantages and disadvantages, was discussed in detail.

In addition to the above, this researcher also described the questions used in the survey: the format, reasoning behind the format, and question wording. Following this, attention was paid to the administering of the survey, entailing the survey population and sample, obtaining ethical clearance, and the ways in which the survey sample would be contacted.

The methodology chapter was concluded with a section on the data analysis methods used, as well as describing how results would be displayed. Finally, the main chronological research steps followed by this researcher, was captured in an image.

With the methodology of this study described, the next chapter (Chapter 4: Results and discussion) deals with the portrayal of findings as revealed by the data gathered.

4. Chapter 4: Results and discussion

4.1 Introduction

This chapter portrays the research findings of the study as obtained via the submitted online questionnaires. Research results are reported on, and discussed. Charts, tables and graphs are used to illustrate and summarise survey responses. A narrative is used to best describe what this researcher considers the most relevant information contained in the survey responses.

When discussing study findings, results of this survey are, whenever possible, compared with results obtained during the earlier RDM study by the same researcher, involving experienced CSIR researchers (Patterton, 2014a). In addition to these comparisons, current survey findings were also compared with earlier RDM surveys done elsewhere, and mentioned in Chapter 2 (Literature Analysis). In particular, this researcher was interested in explaining how her findings confirm or diverge from those of previous studies, as well as her own earlier RDM survey findings. To summarise: the discussion looks at this survey's findings in relation to the theoretical framework introduced in the literature analysis.

The sequential order of reporting on results of the study mirrors the order of the 31 questions in the online survey. Each survey question is discussed under its own heading. Each separate question contains a table and/or graph portraying the question responses, followed by an explanation of said figure and/or table. Where possible or applicable, each question's findings is also analysed in terms of similarities and dissimilarities between this study and other RDM surveys, discussed in Chapter 2. Limitations of survey questions, as well as limitations of the survey tool used, also form part of this chapter.

In short: the goal of this chapter is therefor to address the results from the data collection and analysis, and to communicate what these findings mean for research data management behaviour of emerging CSIR researchers, and research data management at the CSIR. Where applicable, possible inadequacies of questions, be it question content, format or response analysis, with suggestions for improvement, are discussed.

4.2 Survey response: overview

The table below details the study population, the study sample used in this study, and number of survey respondents. Percentages are indicated where applicable.

Table 2: Respondents: overall view

Category	Category size	Percentage of population
Study population	179	-
Study sample	179	100%
Respondents	48	26.8%

The size of the study population was calculated by the researcher after obtaining a name list from the Human Resources division of the CSIR. The obtained list contained, as requested by this researcher, the names of CSIR researchers, permanently employed, age 35 years or younger, and who were either in possession of a doctorate, or currently registered for a PhD-degree. The total number of emerging CSIR researchers came to 179. This population number is higher than the previous CSIR RDM survey conducted by the same researcher (Patterton, 2014a), when the survey population comprised 98 senior research group leaders.

As stated in Chapter 3, this researcher decided on using total population sampling as sampling technique. As a result of this, the sample of this study consists of all members of the target population and totals 179 researchers.

By survey closing date, the total number of completed surveys came to 48. The number of completed surveys in this study (n=48) is slightly higher than the number of completed interviews (n=36) in the earlier CSIR RDM survey by Patterton (2014:3a).

The number of completed surveys for this study comprises 26.8% of the study population. With the survey population and survey sample in this study being the same number of emerging researchers, this would mean that completed surveys also make up 26.8% of the survey sample. This percentage, although slightly lower than the 40% attained in the earlier CSIR RDM study (Patterton, 2014a), still provided this researcher with data describing the RDM practices, needs and challenges of 48 emerging researchers at the CSIR.

4.3 Survey response: operating units

The table below indicates, per unit, the study population (column B), study sample (column C), and what percentage these numbers represent of the total population, and total sample size (column D). In addition, the table also illustrates the survey responses per unit (column

E), and the percentage these responses represent of the unit population and sample (column F). Furthermore, the respondent numbers per unit as a percentage of response totals, is shown in Column G.

The two charts below table 2 graphically illustrate the survey sample/population (figure 1), and the survey responses (figure 2). Discussion of these results follows the table and graphs.

Table 3: Respondents: unit-wise

A	B	C	D	E	F	G
Operating unit	**Population**	**Sample**	**% of total pop/sample**	**Respondent numbers**	**Response %**	**% of total responses**
Biosciences	28	28	15.6	10	35.7	20.8
Built Environment (BE)	5	5	2.8	0	0	0
Defence, Peace, Safety and Security (DPSS)	18	18	10.1	7	38.9	14.6
Implementation Unit (IU)	1	1	0.6	1	100	2.1
Modelling and Digital Science (MDS)	17	17	9.5	4	23.5	8.3
Meraka	21	21	11.7	7	33.3	14.6
Materials Science and Manufacturing (MSM)	41	41	22.9	5	12.2	10.4
National Laser Centre (NLC)	10	10	5.6	3	30	6.3
Natural Resources and the Environment (NRE)	38	38	21.2	11	28.9	22.9
Total	**179**	**179**	**100**	**48**	**n/a**	**100**

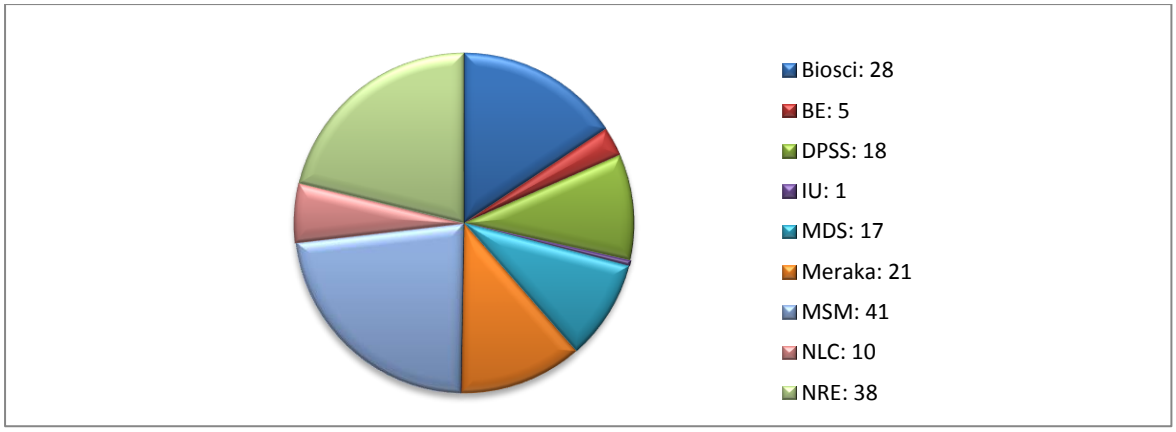


Figure 1: Emerging researcher population/ survey sample

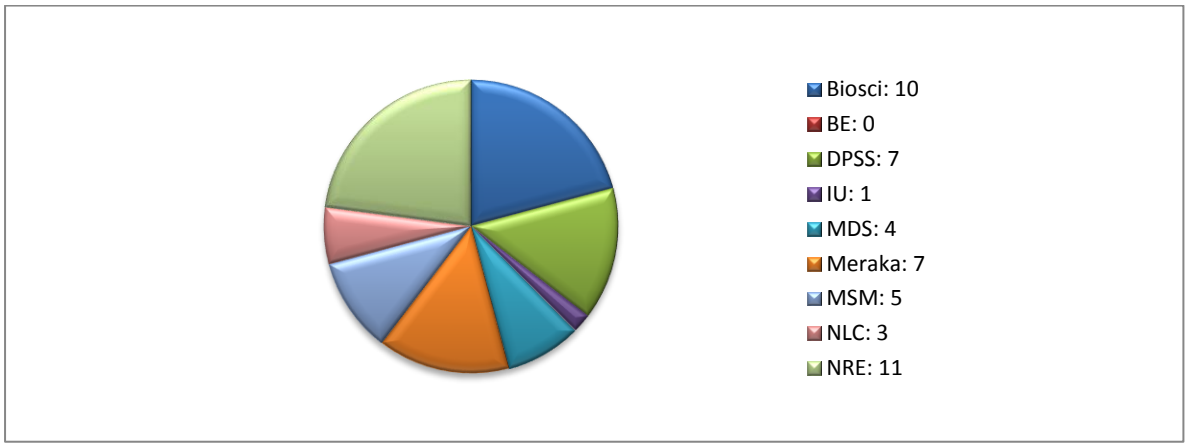


Figure 2: Survey respondents

Several observations can be made when looking at the CSIR emerging researcher **population**, as well as the **sample** (in this survey, the population equals the sample) as illustrated in table 2 and figure 1:

- At the time of the survey launch, 179 emerging CSIR researchers were identified.
- This population was spread over nine units.
- Not all operating units are equally represented.
- The three units with the highest representation are MSM (22.9%), NRE (21.25) and Biosciences (15.6%).
- At the other end of the scale, BE (2.8%) and Implementation Unit (0.6%) are the units contributing least to the emerging researcher population.
- As the sampling technique used in this study is the 'total population sampling' technique (see Chapter 3, section 3.7.1), the sample used in this study is an exact

replica of the survey population. This means that figure 1 can also be used to illustrate the survey sample composition.

- The survey sample's total number of emerging researchers, number of units involved, as well as number of emerging researchers per operating unit, is therefore exactly as described in the earlier bullets (as per survey population).

When looking at table 2 and figure 2, and focussing on **responses**, the following findings stand out:

- Survey respondents came to a total of 48.
- Respondents came from eight of the nine operating units forming part of the survey population/survey sample.
- Response rates between operating units showed variance.
- The Implementation Unit portrayed the highest response rate (100% of surveys completed), but it should be mentioned that this unit was found to only have one emerging researcher
- DPSS, with a response rate of 38.9%, Biosciences (35.7%) and Meraka (33.3%) were additional operating units displaying high response rates.
- No emerging researchers at Built Environment completed a survey. While the absence of any contribution by this unit can be said to be a limitation of this study, this researcher is of the opinion that the unit's non-participation should rather be seen as an unfortunate outcome, or less-than-ideal characteristic of the results, rather than a lacking methodology. Survey awareness, distribution and reminders to all sample members were identical in nature: a pre-survey survey information letter, a survey invitation containing a detailed cover letter, a survey reminder two weeks after the survey invitation, and a final survey reminder distributed a week after the first survey reminder, were the steps followed when administering the survey. Despite these steps, discussed in detail in Chapter 3, section 3.7, and seen to be characteristic of good survey administration practice (Dillman, 2007), no responses from the BE unit were forthcoming. This researcher was unwilling to place additional pressure on getting even only a single BE emerging researchers to submit a completed survey, as it was felt that such a move would jeopardise the respondent/researcher trust and nullify the anonymity assurance stipulated in the survey cover letter (see Appendix 7).
- It might be important to compare the figures obtained in column G (survey responses per unit as a percentage of the total responses), with column D (unit emerging researchers as a percentage of CSIR emerging researcher population). Such a

comparison would enable the researcher to determine whether the unit-related proportions of the population resemble the unit-related completed surveys in relation to all surveys completed. It is seen that two of the three units with the highest proportions (column D), namely MSM (22.9%), NRE (21.2%) and Biosciences (15.6%) could also be found occupying the three highest positions in column G. Moreover, the three units with the lowest representation of the population (NLC, BE, Implementation Unit) also occupied the three lowest spots in column G. This would seem to indicate that when general RDM trends are investigated, survey respondent composition closely mimics the survey population composition.

4.4 Survey response: academic discipline

The table below illustrates survey responses when asked to indicate which academic discipline their doctorate, or current PhD studies, were part of:

Table 4: Survey response (discipline)

Academic discipline	Number of responses	Response percentage
Humanities	0	0%
Social Sciences	0	0%
Natural Sciences	29	60.4%
Formal Sciences[4]	11	22.9%
Multidisciplinary/Other	8	16.7%
Total	**48**	**100.0%**

This researcher also deemed it important to gain clarifying information on the 'multidisciplinary/other' category; the results are indicated below:

Table 5: Survey response to 'other' disciplines

Academic discipline	Number of responses
BioEngineering	1
Biotechnology	1
Conservation ecology	1
Engineering	4
Health Sciences	1

[4] ***Formal sciences*** are disciplines concerned with ***formal*** systems, such as logic, mathematics, statistics, theoretical computer ***science***, information theory, game theory, systems theory, decision theory, and portions of linguistics

It is clear from these two tables that emerging researchers at the CSIR are all involved in SET research; the CSIR being an institute focussing on research related to science, engineering and technology. Put another way: as was expected, none of the respondents of this survey are involved in non-SET-based research, such as the humanities, or social sciences research. The implication of this is that the RDM-practices of CSIR researchers, being SET-based, should be discussed in this context, and not extrapolated to non-SET research. This would entail making recommendations (see Chapter 5, section 5.4) applicable to SET-research, and not to humanities or social sciences. It would also mean not blindly copying or duplicating the RDM-setups at research institutes not SET-based, but rather implementing an RDM-system that meets CSIR-specific needs, which is heavily invested in natural as well as formal sciences. The next section entails stepping away from demographic information and will be a portrayal and discussion of actual RDM behaviours, practices and perceptions displayed and held by the CSIR's emerging researchers, as revealed through survey results.

4.5 Types of research data

The graph below illustrates the responses given when emerging researchers were asked: 'What types of research data do you create or work with as part of your research? Select all that apply':

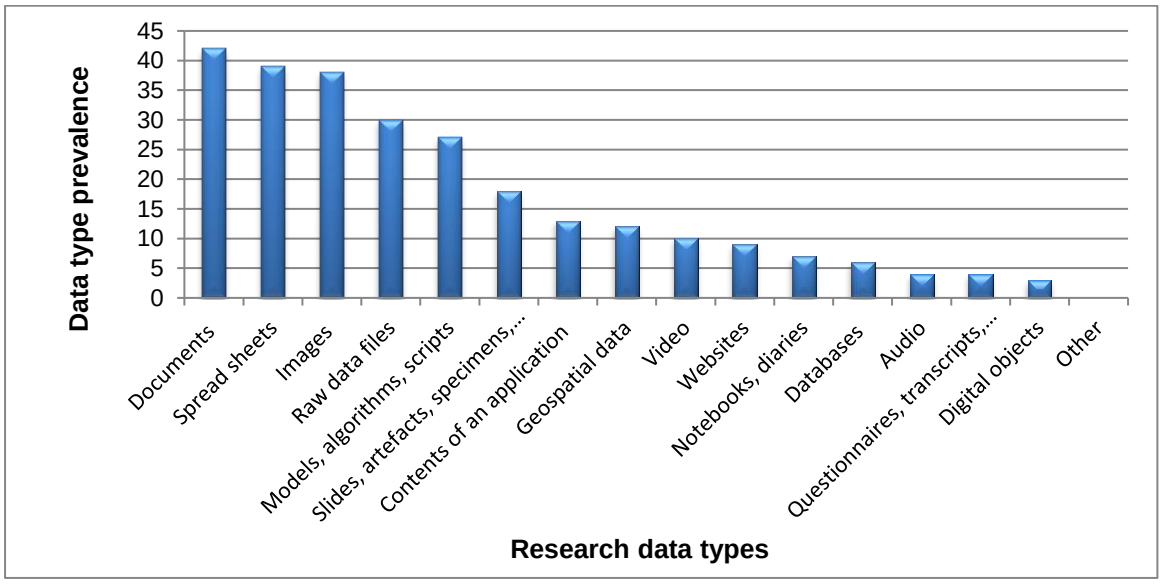


Figure 3: Research data types

In total, 15 different data types were seen to be used by emerging researchers. It has to be mentioned that the question format for this topic was in multiple choice style, with researchers being able to tick as many options as were applicable. Sixteen options were

supplied; and responses indicate that 15 data types were seen to be used across the CSIR. The results show that no emerging researcher clicked the 16th option given, namely 'other'. This researcher was surprised to find that only the 15 options listed in the question were chosen by respondents, and that no mention was made by any of the responding emerging researchers of other data types used during their research.

What can also be seen from the above chart is that emerging researchers make use of several data types. The 48 responding emerging researchers indicated a total of 262 data types: this amounts to an average of 5.5 different data types per researcher.

The most common datatypes were found to be text documents, spreadsheets, images, and raw data files, respectively used by 88%, 81%, 79% and 63% of emerging researchers. Datatypes not commonly used include audio (8%), questionnaires/transcripts (8%), and digital objects generated/acquired during data creation (6%).

The researcher was interested in comparing these results with a previous CSIR RDM survey conducted by her (Patterton, 2014a:4). The earlier study revealed the three most common data types to be spreadsheets, images, and text documents; a finding in close agreement with the current study. The findings with regards to datatypes not commonly used differ somewhat from the previous CSIR survey involving experienced researchers (Patterton, 2014a:4). The latter group indicated that audio data featured more prominently in research than the current group of respondents did. Furthermore, experienced researchers, although fewer in numbers in the previous study than the current group of respondents (36 versus 48) were shown to make use of 19 different data types, compared to the current study's 15 different data types. While reasons for these differences might be ascribed to the experienced researchers' longer research careers, as well as an experienced researchers more likely to be exposed to an entire research group's research activities, comparisons between the two CSIR RDM surveys for this RDM topic cannot really be made. The previous survey (Patterton, 2014a) made use of open-ended questions, while the current study investigated data types used by means of a multiple choice question.

Studies conducted elsewhere and reporting on data types commonly used, contain findings not dissimilar to this study's findings: Buys & Shaw's study (2015:9) reveal spreadsheets, structured data, text and images to be the most common types, Whitmire, Boock & Sutton (2015:388) mention spreadsheets and digital text, while Sewerin *et al.* (2015:5) mention numerical data and text data.

The use of more than one data format by every emerging researcher is a finding supported by previous studies: Rankin *et al.* (2012:6), Averkamp, Gu & Rogers (2014:7) as well as

Sewerin *et al.* (2015:5) found that researchers make use of more than one data format. Furthermore, studies have also revealed a variety of data types used across institutions investigated, as seen in the findings of Averkamp, Gu & Rogers (2014:7), Buys & Shaw (2015:9), Kennan & Markauskaite (2015:69) as well as Whitmore, Boock & Sutton (2015:382). As such, it can be said that the variety, type and number of data formats used by the CSIR's emerging researchers are in line with previous RDM study findings documented earlier, elsewhere on the globe.

Implications of the findings portrayed in figure 3 include the acceptance and acknowledgement of various formats by all RDM stakeholders, especially with regards to future data access, format obsolescence, data preservation, training requirements and compatibility issues.

Recommendations emanating from data format findings are addressed in Chapter 5, section 5.4.

4.6 Volume of research data

The two charts below illustrate the responses given when emerging researchers were asked to estimate the volume of their research data across all of their CSIR work. The first chart shows data volumes in order of size, while the second chart has data size arranged in order of prevalence.

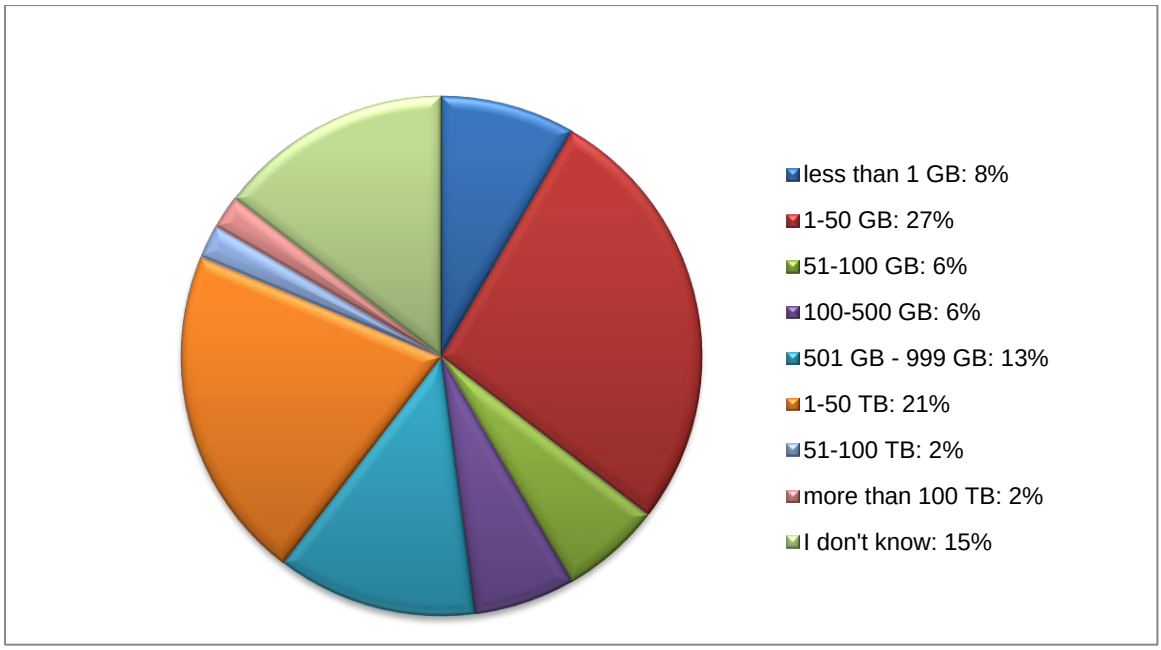


Figure 4: Research data volume (clockwise in size order)

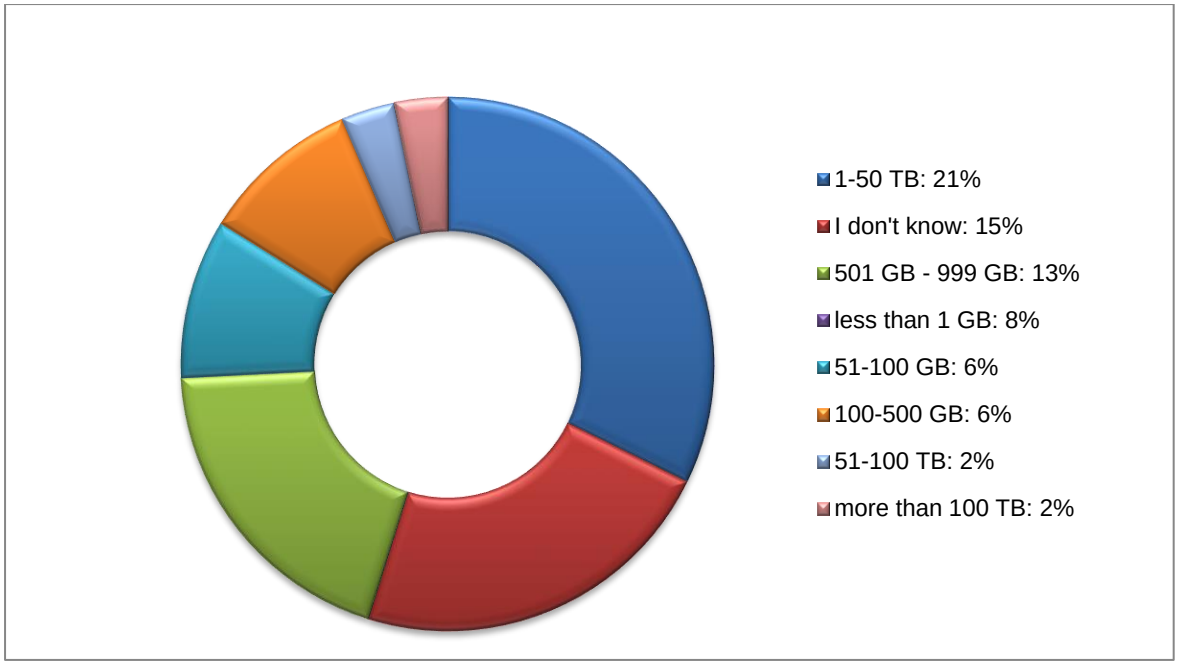


Figure 5: Research data volume (clockwise in prevalence order)

Responses indicate that emerging researchers across the CSIR hold data volumes belonging in all eight dataset size ranges included in this particular survey question. This means that some researchers' total data amount to less than 1 GB, while others have datasets at the opposite end of the size spectrum, i.e. more than 100 TB. In short: all size options forming part of the question were being used by the survey sample.

Data held by emerging researchers were most often found to be in the range of 1 to 50 TB (21% of respondents), and followed by data in the 501 GB to 999 GB range (13% of cases). It is also worth noting that 15% of respondents had no idea how much data they held. Data volumes least prevalent are seen to be the very large collections, with only 2% (i.e. one respondent) stating his data were in the 51 to 100 TB range. A similar response was found for data volumes bigger than 100 TB in size. It is interesting to note that in both instances (51-100TB, as well as data volumes bigger than 100TB) both respondents were seen to from the NRE unit.

When these results are compared with the previous CSIR survey, when Patterton (2014a: 5) investigated the data volumes held by experienced researchers, it is found not to be an easy task. This current study was interested in the total data held by emerging researchers, while the previous study asked experienced researchers about the size of a typical dataset. As a result of this, a direct comparison is not possible. It should, however, be mentioned that Patterton (2014a: 5) found that the biggest dataset size category were the least frequent dataset size used. Also, as is the case in the current emerging researcher study, it was

established that there is a subset of researchers who are not aware of the data volumes created.

Findings displayed above are in agreement with RDM studies done elsewhere: several studies confirm that data set size variance exists (Akers & Doty, 2013:8; Van Tuyl & Michalek, 2015:3; Whitmire, Boock & Sutton, 2015:387). Furthermore, Akers & Doty (2013:8), Buys & Shaw (2015:13) as well as Whitmire, Boock & Sutton (2015:387) have found that some researchers are unaware of how much data they have stored, supporting the graph above portraying that 15% of emerging researchers do not know how much data they have accumulated during their current research project.

Implications of findings displayed in figures 4 and 5 are mainly associated with storage issues of locations used, be it an emerging researcher's office computer, unit server, or data repository. Recommendations stemming from data volume findings are discussed in Chapter 5, section 5.4.

4.7 Software applications used for analysis/manipulation of data

The chart below illustrates all answers given when researchers were asked to indicate which software applications they were using for analysis or manipulation of their data:

Table 6: List of software applications

Access	Adobe Photoshop	ArcGIS	Discovery Studio
Design-Expert	ENVI	Eviews, ImageJ	Excel
FACSDIVA	FlowJo	Fortran	Galaxy
GAMMA	Gaussian	Grads	Julia
Labview	LaTeX	Matlab	Microsoft Visio
Microsoft Word	NCO	NVIVO	Python
QGIS	R	SAS	SigmaPlot
SPSS	Stata	VASP	

The above list contains all software applications indicated by survey respondents. In total, 31 different software applications were mentioned by respondents.

The ten software applications most frequently used are indicated in the chart below:

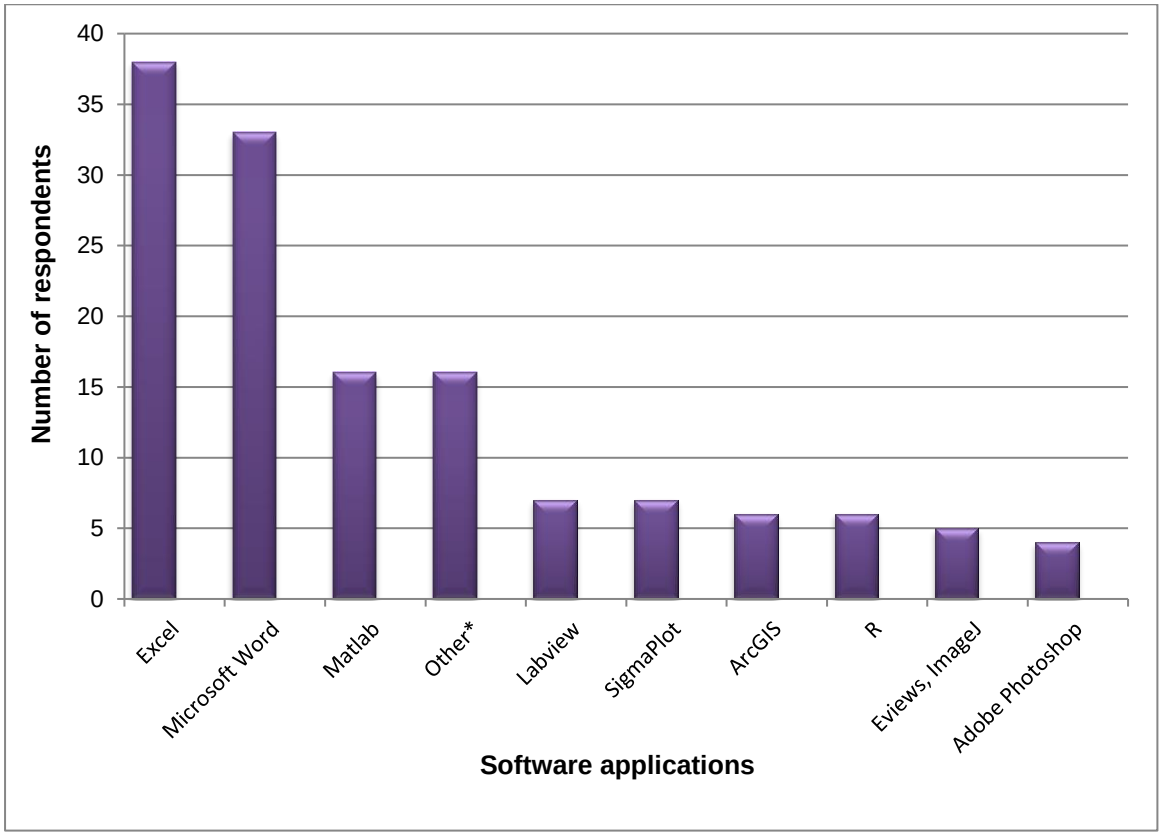


Figure 6: Software applications

Software applications most commonly used by emerging researchers were found to be Microsoft Excel, Microsoft Word, and Matlab. These findings are in agreement with the previous CSIR RDM survey (Patterton, 2014a:6) where Microsoft Office as software package, as well as Matlab, were the most commonly used software packages.

Although the 'other*' category is ranked equally to Matlab, an investigation of the clarifying details revealed that this category contained 15 different software applications (added to Table 5, see earlier). All applications in 'other*' were only mentioned once, except Python, with three mentions. Python did not make the list of ten most frequently-used software applications, and if placed in the figure above, would obtain a ranking just below 'Adobe Photoshop'.

This researcher is interested in comparing the software applications used by emerging researchers, with the software applications indicated by experienced CSIR researchers

(Patterton: 2014a). Quite a few similarities could be found: Microsoft Office was found to be the most widely used application, and Matlab was indicated to be a popular tool. The previous study, however, did not investigate the frequency of use, and as such, prevalence ranking could not be created.

An interesting difference between the two studies is found when looking at the number of software applications indicated by respondents. Although the earlier study by Patterton (2014a:3) involved fewer respondents than the current study (48 emerging researchers versus 36 experienced researchers), the 2014 study respondents listed more than double the number of software applications listed by emerging researchers. Experienced researchers indicated that they made use of 73 different software applications, while emerging researchers' responses came to a total of 31 software applications. This researcher is not able to tell the exact reason for this difference: possible explanations could be the following:

- Patterton's previous CSIR study (2014a:5) involved nine operating units; the current study incorporates eight operating units.
- It is possible for even one respondent, making use of many software applications, to make a significant difference to the survey total. This was indeed the case with the 2014 survey, as evidenced during a study interview when approximately ten software applications were listed by an experienced DPSS researcher.
- The possibility exists that emerging researchers might have interpreted this question as being more directed towards the clicking of options on the list, rather than a request for all software applications to be listed.
- It might be that experienced researchers were in fact more aware of all software applications used in their units, or were themselves using more applications than a typical emerging researcher.

Implications of emerging researchers making use of more than 30 types of software tools are numerous, and include the following:

- Varied and widespread use of software tools emphasize the need for data documentation/metadata added to datasets to be seen as a vital RDM activity.
- Data documentation/metadata added to datasets to include details on software tools used to create, visualize or view the data.
- Shared datasets to include necessary information on software tools used to create or view the data. Custom-designed software tools or those requiring subscriptions, or

not in common use, might have implications for ease of data sharing and data access.

- Software tools used need to be taken into account when data are preserved in the long term or archived. Software redundancy is a real threat and it is possibly a key concern when deciding on data preservation.

Recommendations forthcoming from this survey's software-related findings are elaborated on in Chapter 5, section 5.4.

4.8 Development of a research data management plan

The chart below illustrates the answers supplied when researchers were asked whether they have ever developed or submitted a research data management plan (DMP) for any of their projects:

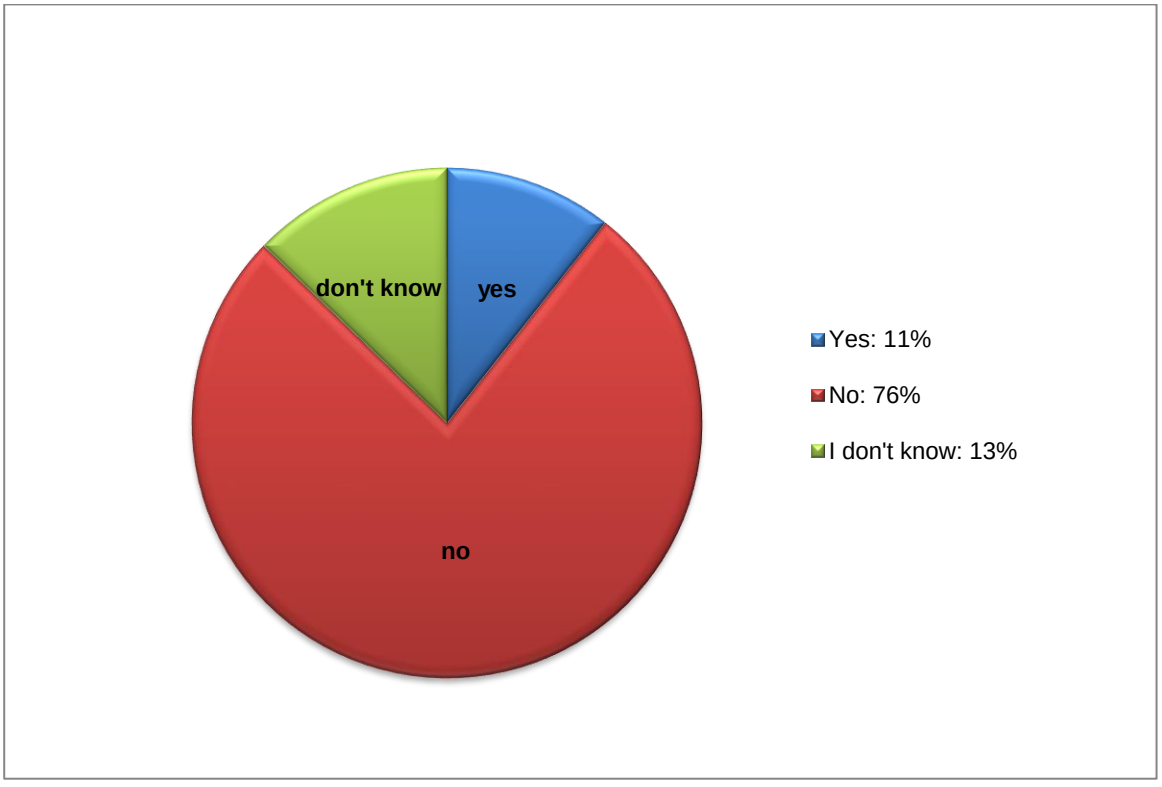


Figure 7: Use of DMPs

The majority (76%) of emerging researchers indicated that they had never developed/submitted a data management plan. Only a small minority (11%) admitted to having created such a plan. Thirteen percent of respondents were not sure; this researcher is of the opinion that the uncertainty indicated is probably due to collaborative research,

where the respondent is unsure whether other members of the project team had developed or created a data management plan.

The use of a data management plan was not an activity investigated in the previous CSIR RDM study; as such, a comparison cannot be made. However, looking at the responses given by experienced researchers when asked about the familiarity with research data management (Patterton, 2014a:7), some overlapping tendencies are noticed. It was found that when researchers were divided into three groups, namely a group applying RDM, a group that has heard of it, and a group that has never heard of it and never uses it, the group with the fewest experienced researchers were those applying RDM. Similarly, in the current study, the category with the fewest number of emerging researchers turned out to be those who submit data management plans. Although these two mentioned groups differ with regards to percentages of the total survey respondents (11% emerging researchers versus 28% experienced researchers) it would be safe to say that the majority of emerging researchers, as well as the majority of experienced researchers, are not familiar with RDM and as a result of this, do not submit DMPs.

This researcher was also interested in comparing the DMP-related findings of this study with those of other RDM studies. Studies are not unanimous in their findings and some diversity in results was discovered when this researcher studied earlier RDM studies. In general, findings showed that the majority of researchers do not make use of DMPs. Examples of these would be Pink *et al.*'s (2013:11) study showing that 81% of researchers did not made use of DMPs (2013:11), Buys & Shaw's study (2015:15) showing that only 45% of respondents used DMPs, Kennan & Markauskaite (2015:69) stating that only a small portion of their respondents used DMPs, and Van Tuyl & Michalek (2015:1) stating that only 44% of respondents had to submit DMPs. As was expected, funders and institutional policies are the main drivers of DMPs, a finding supported by Mossink & Bijsterbosch's study (2013:5) into the RDM practice of European researchers. These findings substantiate the results of the current study: DMPs are not being submitted by the CSIR's emerging researchers; only 11% of respondents stated that a DMP has ever been submitted by them. This low DMP-submission rate comes as no surprise, seeing that the CSIR does not have an institutional policy addressing the issue of data management plans. The previous CSIR study of Patterton (2014b:18) revealed, via survey interviews, that only one funder (Water Research Commission) required grant recipients to indicate, albeit in rudimentary terms, how they would be managing their data. It is encouraging to note that while not part of the application process at the time of the previous CSIR study, grant application forms for National Research Foundation funding now include a section titled 'Details of Research: Data Storage

and Dissemination', requiring applicants to complete the section in free-texted format (Pillay, 2016).

The benefits of creating a DMP are numerous and include being able to find and understand the data when needed, continuity is ensured should project members leave or new members join, and duplication (recollecting, reworking) of data is avoided. The repercussions of not having a DMP could result in the invalidation of stated benefits. With the majority of CSIR's emerging researchers currently neither expected to create DMPs, nor creating DMPs, Chapter 5 (section 5.4.4) contains recommendations on how to treat this current and serious CSIR RDM weakness.

4.9 Awareness of policy/funder requirements regarding research data management

The chart below illustrates the responses supplied when researchers were asked to **i**ndicate whether they were aware of any policies or requirements from their funder/s regarding research data management.

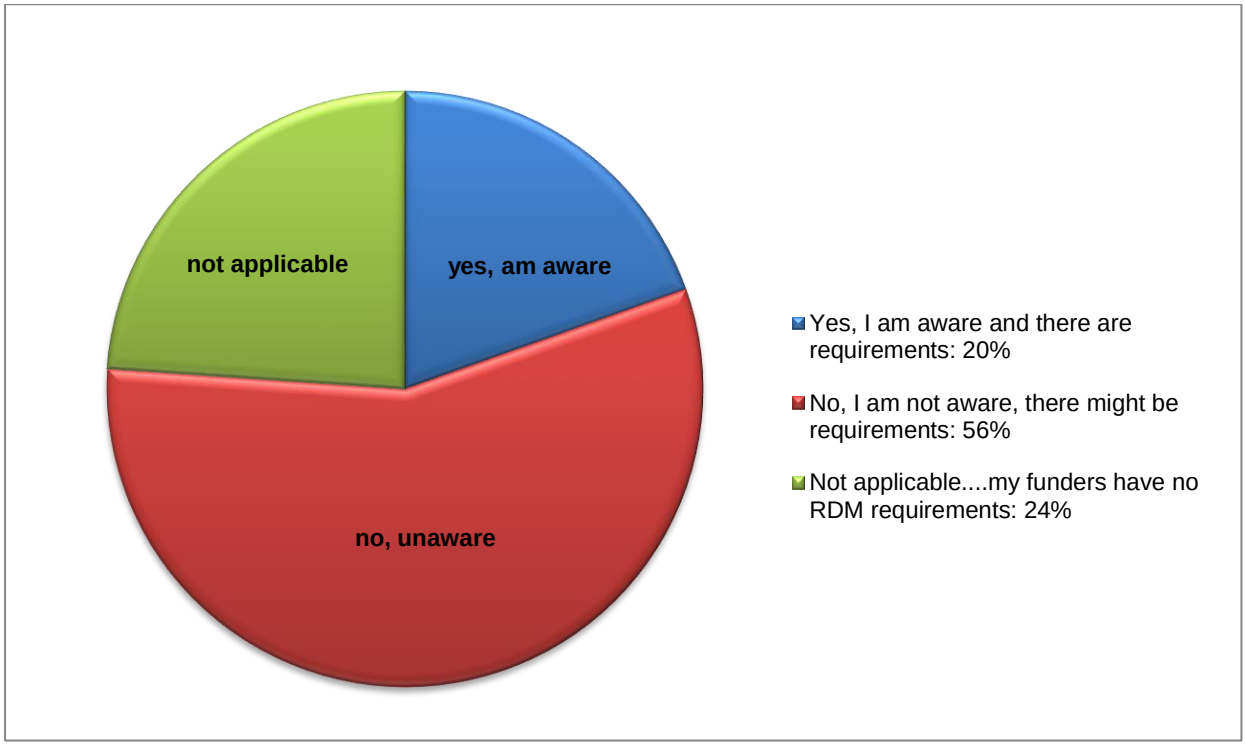


Figure 8: Funder requirements (awareness)

Responses indicate that the majority of emerging researchers (56%) are not aware of funder requirements. A fifth of emerging researchers are aware of funder requirements, while almost a quarter of respondents stated that their funders have no requirements.

These findings cannot be compared with the previous study as funder requirement awareness was not an aspect investigated. Studies done elsewhere, such as the study done by Knight (2013:21), although not reporting on 'awareness' of funder requirements, have found that researchers have a desire for assistance in understanding funders' requirements. A need for assistance in meeting funders' requirements, as expressed by researchers in a study by Parham, Bodnar & Fuchs (2012:12), is a possible related need when researchers battle with an awareness of funder requirements.

The conclusion drawn from the responses indicated in the chart above is that CSIR emerging researcher awareness of funder requirements is currently an RDM skill in dire need of improvement. With nearly six out of every ten emerging researchers being ignorant about funder requirements, the probable resulting non-adherence has implications for the quality management of research data, meeting funder requirements/satisfying funder expectations, and future grant awards, to name but a few possible consequences. Recommendations for addressing the low levels of awareness pertaining to funder requirements are put forward in Chapter 5, section 5.4.

4.10 Data storage location

In an effort to establish the locations used when storing research data, emerging researchers were given a multiple choice format question, with instructions to select all the locations/options used when storing data. As 'other' was also an available option, respondents were requested to supply clarifying information when choosing this option. The results of this question are illustrated in the graph below.

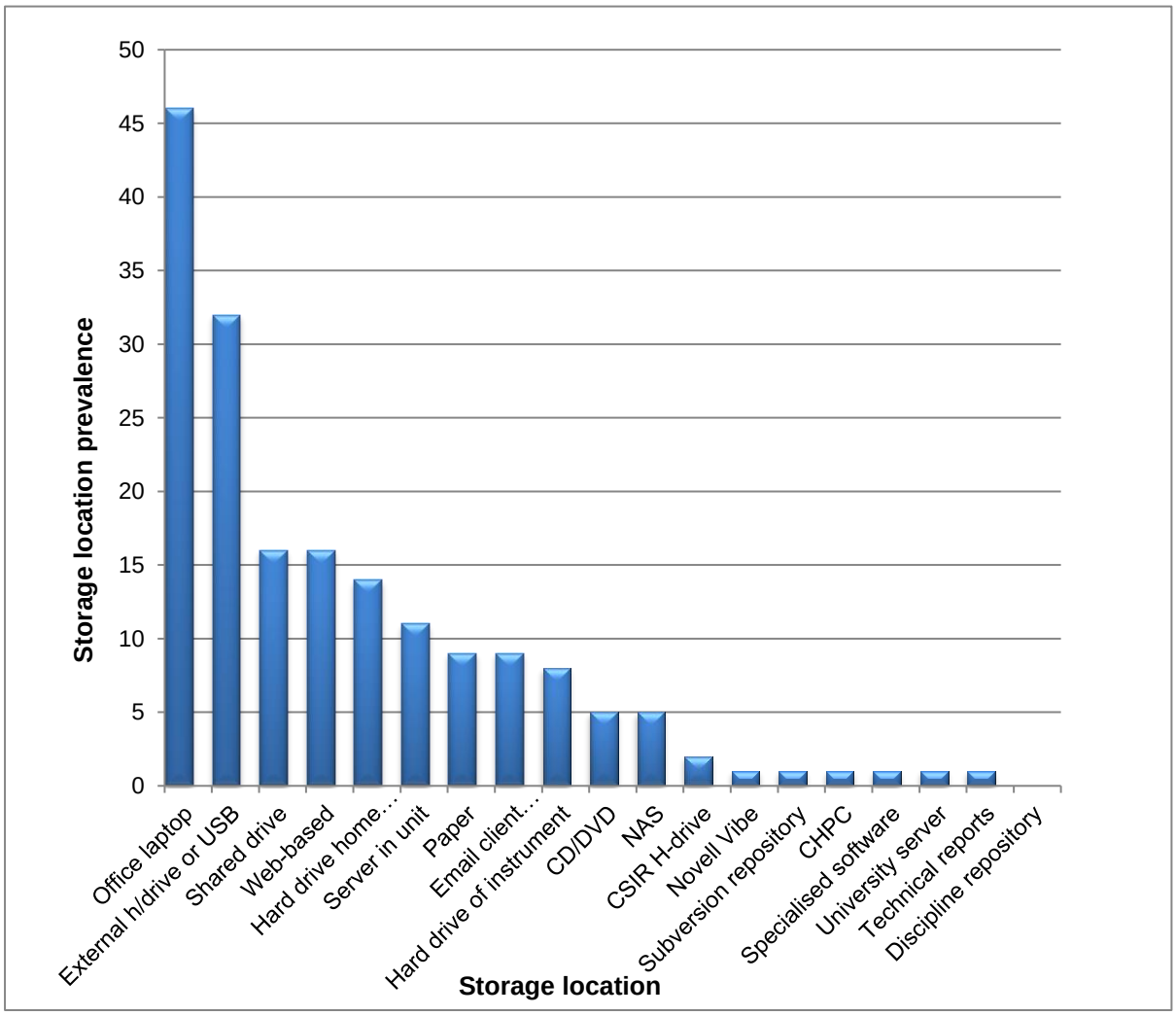


Figure 9: Data storage locations

As is seen from the graph above, a range of storage locations is used by emerging researchers when storing data. Furthermore, these locations are used in varying degrees of prevalence. The range, number and prevalence of locations are to be discussed in more detail in the remainder of this section.

This multiple-choice question featured 14 storage location options, with an additional 'other' option presenting an opportunity for elaborating on this storage choice. All-in-all, a total of 18

different storage locations were stated to be used. Respondents were requested to indicate as many storage locations as were applicable; this resulted in findings showing a total of 185 storage locations used by 48 emerging researchers: an average of 3.9 different locations per emerging researcher.

The overwhelming majority of respondents (96%) indicated that data were stored on the hard drive of their office personal computer, or office laptop. The second most commonly used location was found to be external hard drives/USB sticks, in other words, a small portable device, with 67% of respondents making use of this option. Other storage locations shown to be popular, were shared drives (a third of respondents), cloud services such as Dropbox or Microsoft OneDrive (a third of respondents), and the hard disk drive of their home computer/laptop/tablet (29% of respondents).

The graph above also shows that several location options were only being used by a tiny minority of respondents. The CSIR H-drive (used by two researchers), a subversion repository, the Centre for High Performance Computing, specialised software such as Torrent Suite, a university server, and using technical reports as a data storage location, were each indicated by one respondent each. The mentioning of technical reports as data storage location (as an answer in open-ended option 'other') probably implies that data were included with the report; it does, however, not indicate what the actual storage medium was.

An option supplied in this multiple choice question, namely discipline-specific data repositories, was not indicated as being a storage location for any of the respondents.

The findings of this investigated RDM area can be compared with those of the earlier RDM study (Patterton, 2014a:7) when experienced researchers were asked where they were storing research data. The results of the earlier study showed several similarities with the current study's results: emerging researchers made use of several different storage locations, and a personal computer or laptop was found to be the most commonly-used storage location. The CSIR shared drive, as well as external hard drives or portable storage devices, featured, just as is the case with the current study, in the top three of storage locations chosen.

This researcher was also interested in comparing the current study's storage findings with those of earlier findings on data storage practices, as documented in Chapter 2. When looking at general trends found across studies cited by this researcher, it was seen that the most common primary storage place for data proved to be locally, on the researcher's personal computer or laptop. This general trend was found to be true across disciplines, institutions, geographical areas, and levels of research experience. Beile (2014:10), Buys &

Shaw (2015:12), Sewerin *et al*. (2015:5) as well as Whitmire, Boock & Sutton (2015:388) are a few examples of authors revealing the primary storage location of research data to be on personal computers or laptops. Marchionini's statement (2012:12) is an apt summary of data storage practices: researchers are relying on themselves to store data. This trend seems to be in line with the results of the current study, as indicated in figure 8.

When comparing the number of storage locations used, on average by the CSIR's emerging researchers, with number of storage locations indicated in earlier studies, similarities are found. While this study shows that emerging researchers make use of approximately four storage locations, the study by Knight (2013:4) showed 52% of researchers to be using two to three locations, and 37% of respondents using four to eight locations. Similarly, Averkamp, Gu & Rogers (2014:9) found that most respondents made use of more than one data storage location. This study's finding of four locations to be the average number of locations used, also correlates with a study conducted by Parsons, Grimshaw & Williamson (2013:11), where data were found to be most commonly stored in five different places.

It can be said then that the storage locations used by CSIR emerging researchers are in line with storage locations indicated by the institute's experienced researchers, as well as researchers elsewhere on the globe.

Implications of storage location findings entails CSIR RDM stakeholders taking note of the range and number of storage options preferred and used, and keeping these practices in mind when future RDM procedures are drawn up. Other implications revolve around researchers possibly making use of devices not considered to meet security standards, making data prone to loss, failure, theft or similar. Recommendations emanating from the results pertaining to storage locations are discussed in Chapter 5, section 5.4.

4.11 Research data backup: frequency

Emerging researchers were asked to indicate, from a multiple choice list, how often their research data were backed up. The chart below illustrates the responses supplied.

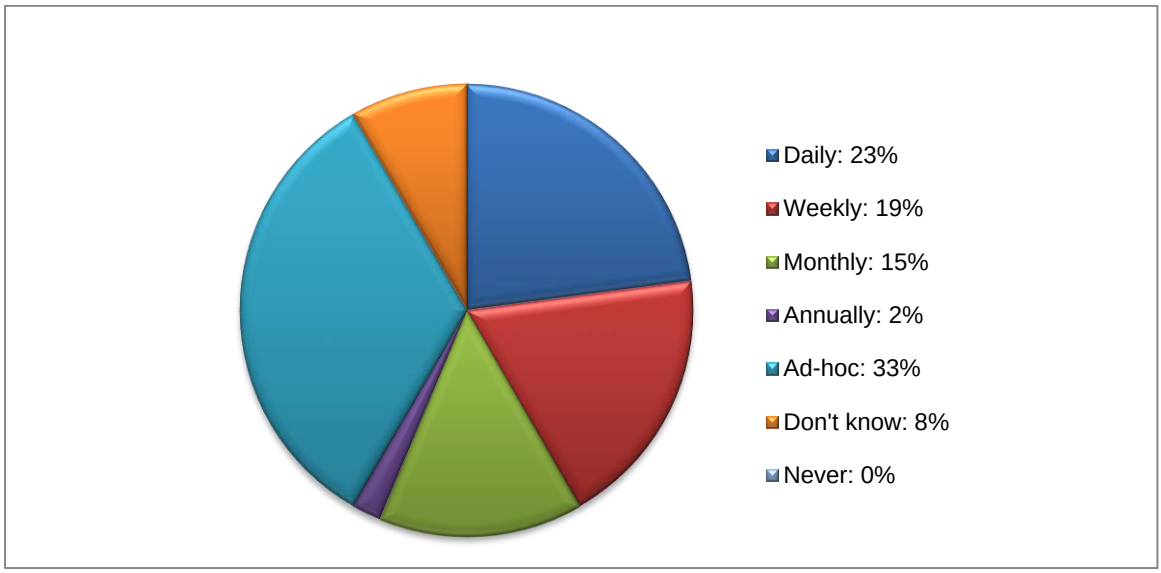


Figure 10: Data backups (frequency)

As can be seen from the graph above, emerging researchers' backup practices with regards to frequency was found to show variance. Apart from the category named 'never' (indicating that data are not backed up), all backup frequencies forming part of this question were seen to be used by emerging researchers. Prevalence of the different backup frequencies, as indicated in the size of the various graph segments, showed variance. A closer look at these frequency segments will be the focus point of this section.

Results indicate that the most prevalent backup frequency used by emerging researchers was on an ad-hoc basis, with a third of them falling into this category. The second most common backup frequency was backing up daily, with nearly a quarter of respondents indicating this option to be their backup frequency. One out five researchers made weekly backups. Two worrying findings came to light: 8% of researchers did not know how often their data are backed up, while 2% only backed up once a year.

When these results are compared with the previous RDM study by Patterton (2014a:10), it is seen that researchers from both studies place a high premium on daily backups. Furthermore, all respondents in both studies indicated backing up their data; it was nowhere found to be an RDM activity not performed by any researcher, either emerging or experienced. However, as is discussed in the previous paragraph as well as the paragraph

below, the fact that all CSIR researchers are seen to be backing up their data does not mean that their backup strategies cannot be questioned, or improved upon.

Another interesting finding displayed in the graph above, is that a third of respondents indicated their data backup frequency to be ad-hoc. This would mean that one in three emerging researchers back their data up without previous thought or preparation. In practice, it could mean that in a population of 179 emerging researchers, 60 researchers are making use of a backup strategy that is possibly lacking in justification, or to be used for the particular case at hand without consideration of wider application, or the implications of such a strategy. As ad-hoc backups were shown to be the back-up frequency/tendency forming the biggest part of the above chart, it is an RDM area flagged by this researcher to be addressed in planned RDM procedural documents.

This researcher is also interested in comparing emerging researcher backup frequency, with the findings of earlier studies, done both locally and abroad. Alexogiannopoulos, McKenney & Pickton (2010:23), Scaramozzino, Ramirez & McGaughey (2012:358), Averkamp, Gu & Rogers (2014:12) as well as Van Tuyl & Michalek (2015:3) have established similar backup behaviours in their respective studies, with all studies reporting backup frequencies of 85% or higher. It is encouraging to note that the CSIR's emerging researcher backup percentage of 100% exceeds these mentioned figures. This finding, namely that all researchers in the study were found to back up their data, is identical to the findings of Martinez-Uribe (2008:8) as well as Freiman *et al.* (2010:4), whose studies also established that all research data were backed up.

Backup frequency, as well as backup locations (see next section) can also be compared with studies further abroad. This study established that an 'ad-hoc' backup strategy was found to be the most prevalent backup trend; the findings of Peters & Dryden (2011:394), Raggett (2012a:13) as well as Johnston & Jeffryes (2014:9) respectively describing the backup behaviours identified by their studies to be 'responses varied widely', 'a mixed approach' and 'sporadic'. It is seen that the prevalence of the ad-hoc category displayed in the graph above, tends to be similar to backup trends established elsewhere.

Recommendations on improving backup practices related to frequency, are put forward in Chapter 5, section 5.4.

4.12 Research data backup: location

This researcher was interested in investigating backup locations used, and presented survey respondents with a multiple choice format question, where they were required to select all

relevant location options. Three additional options included being able to select 'don't know', 'not applicable, data are not backed up', as well as 'other'. For the latter, respondents were requested to supply clarifying details. The graph below shows responses given, as well as usage frequency of storage locations indicated. Responses indicated for survey option 'other' have been analysed and added to the graph.

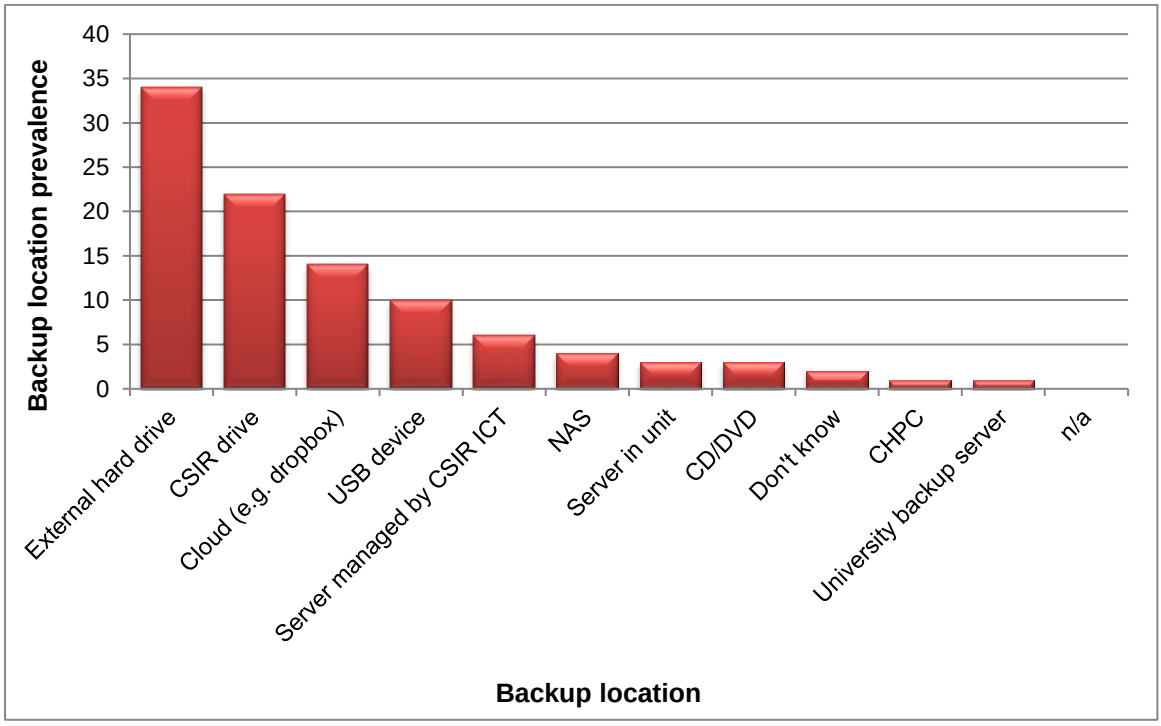


Figure 11: Backup locations

As is seen from the above graph, emerging researchers across the CSIR make use of many locations when backing up their data. The 48 survey respondents indicated a total of 100 backup locations, and even when correcting for the ambiguous category 'don't know', it amounts to approximately two backup locations per emerging researcher.

The most common backup location was found to be on an external drive, with 71% of respondents indicating that they made use of such a device. Other backup options commonly used for backups, were seen to be on a CSIR drive, and a cloud-based service, with 46% and 29% of respondents indicating that use was made of these backup options/locations.

Backup locations listed on this multiple choice format question, and not found to be commonly-used, include unit server (used by 6% of respondents), and CD/DVD (used by 6% of respondents). Backup locations listed when the 'other' category was elaborated on, were the Centre for High Performance Computing, as well as a university backup server, shown to be used by 6% of survey respondents each.

As was expected, and already established in section 4.11, none of the respondents indicated not backing up their data, i.e. all emerging researchers backup their data and made use of some sort of backup mechanism.

When these findings are compared with those of the earlier CSIR RDM study (Patterton, 2014a:10), quite a few commonalities in backup practices can be found. Emerging researchers favour external hard drives above all (71%), while experienced researchers indicated this type of device to be third most popular, and used by 39% of them. Making use of a CSIR drive was the most prevalent backup location for experienced researchers (47%), a figure closely matched albeit as third-most prevalent location for emerging researchers (46%).

Two marked differences are found between these two CSIR groups: experienced researchers did not seem to me making use of cloud-based storage as a backup location, while it is the third-most used location for emerging researchers, with nearly a third of them making use of it. Furthermore, only 3% of experienced researchers stated that they were using USB devices when backing up data, while these devices are much more prevalent with emerging researchers, used by 21%.

This researcher was also interested in comparing the findings of this study pertaining to backup location, with those of RDM studies done elsewhere. Many studies revealed backup strategies to be sporadic and ad-hoc (see section 4.11); the varied nature of study findings is a feature of RDM investigations. Van Tuyl & Michalek (2015:3) state that data are backed up to a departmental server, an external hard drive or to a cloud server, Scaramozzino, Ramirez & McGaughey (2012:358) report that backups were done on the office computer, Martinez-Uribe (2008:8) Ekmekcioglu & Rice (2009:22) and Pink *et al.* (2013:13) mention the use of the work server or network storage, Diekmann (2012:24) states that commercial services are used, and Martinez-Uribe (2008:8) refers to the use of CDs and DVDs as backup tools. These varied findings support the findings shown in figure 10: emerging researchers in the CSIR, as well as researchers across the globe, make use of a variety of locations or backup options when backing up their data.

This study found external hard drives to be the most common backup location used by emerging researchers; this practice mirrors the findings of an RDM study by Jones, K (2011:1). While this researcher takes cognizance of the fact that emerging researchers, on average, use more than one backup location, the over-reliance of this institute's young researchers on a device prone to theft, as well as disk crash, is a serious concern.

4.13 Documenting metadata

The chart below indicates the responses given by emerging researchers when asked: 'Do you document or record any metadata about your data?' A short definition of metadata was also included with the question, and was worded as follows: 'Metadata is data about research data, making it more meaningful or easier to search for'.

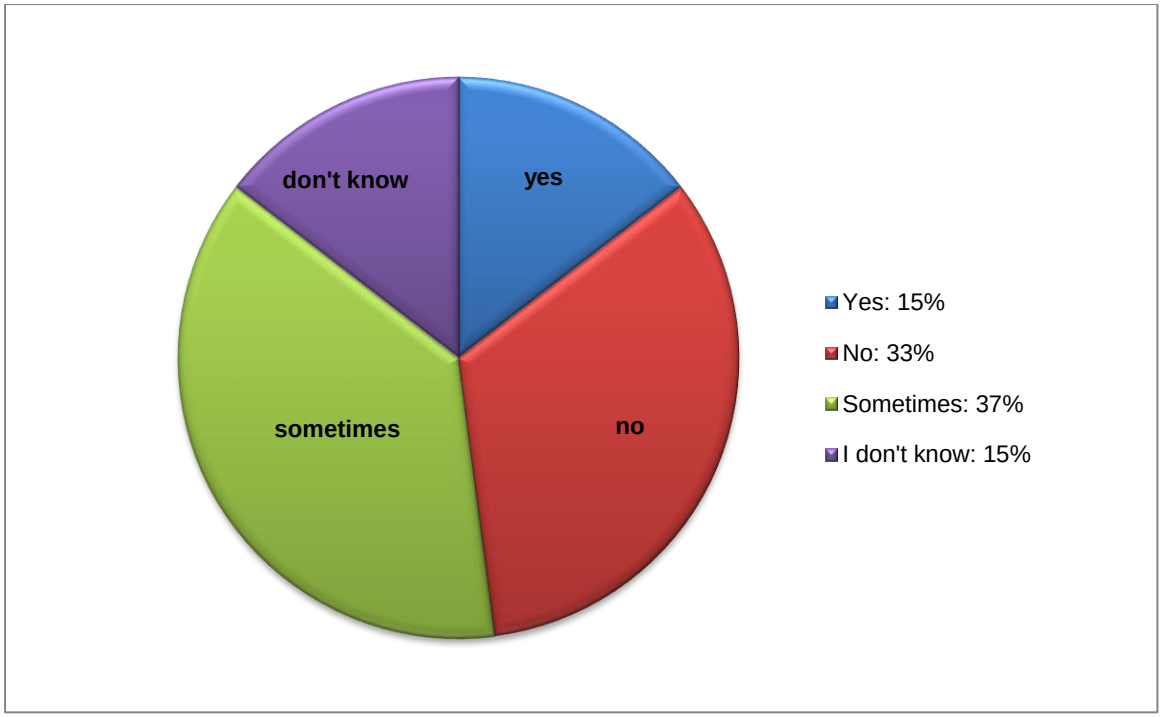


Figure 12: Metadata documentation

The use of metadata as an RDM practice showed a troublesome trend, with most of the respondents (37%) indicating that metadata were only sometimes added to research data. Furthermore, a third of respondents stated that metadata was never added to datasets. Only a small percentage (7 out of 48 responses) stated that they add metadata to datasets.

These findings seem to be in disagreement with the previous CSIR RDM study, in which Patterton (2014a:9) found that 42% of experienced CSIR researchers add metadata to datasets. This finding would seem to clash with the findings of an RDM study by Raggett (2012:17), who was not able to establish that research experience was an indicator of metadata use. Examining figure 12 further seems to support this statement, in that both CSIR studies revealed that 33% of researchers never add metadata to their datasets.

This researcher was also interested in comparing the metadata documentation practices of emerging researchers with findings of RDM studies done elsewhere. This was found to not be an easy task: as stated in Chapter 2, section 2.4.8, exact conclusions, after studying the

findings of several RDM studies, could not be made about the use of metadata. Nevertheless, some general trends support the findings of this study. As stated in section 2.4.8, most RDM studies show the majority of researchers not to be using metadata. While the study of Whitmire, Boock & Sutton (2015:390) found that 53% of respondents made use of metadata, Van Tuyl & Michalek (2015:18) report that respondents felt that they do not know enough about metadata standards to be creating metadata. Furthermore, a study at the University of Colorado Boulder reported that 63% of respondents did not make use of metadata (Task Force, 2012:7). Similarly, the studies of Akers & Doty (2013:12), Parsons, Grimshaw & Williamson (2013:5) as well as Beile (2014:14) all indicated that more than half of respondents did not add metadata to their datasets. Even though a few studies found the opposite to be true, i.e. metadata were being added by the majority of researchers, examples being Bradbury & Borchert (2010:6), and Nassiri & Worthington (2012:7), the findings reported by the majority of studies mimic those of this study. A small minority (15%) of emerging CSIR researchers always add metadata, and a third of them never add metadata to datasets.

Comparisons with both experienced researchers as well as researchers elsewhere in the world have led to the conclusion that the use/non-use of metadata is an activity varying widely between researchers and studies. As indicated via this study, metadata documentation is an activity currently under-utilized among the CSIR's emerging researchers. Repercussions of not making use of metadata could have far-reaching consequences for emerging researchers, as well as impact records management and research at the CSIR. Ramifications include being unable to reuse one's own data, or being unable to understand the data as a secondary user. This researcher is also in agreement with Ward *et al.* (2011:267), who state that not adding metadata could impact on being able to deposit in archives or data repositories. Low rates of metadata usage necessitate a discussion around metadata use; recommendations regarding implementation can be found in Chapter 5, section 5.4.

4.14 Use of metadata standards/guidelines

Emerging researchers were to indicate whether any standards or guidelines were adhered to when adding metadata to a dataset. The chart below details the responses given:

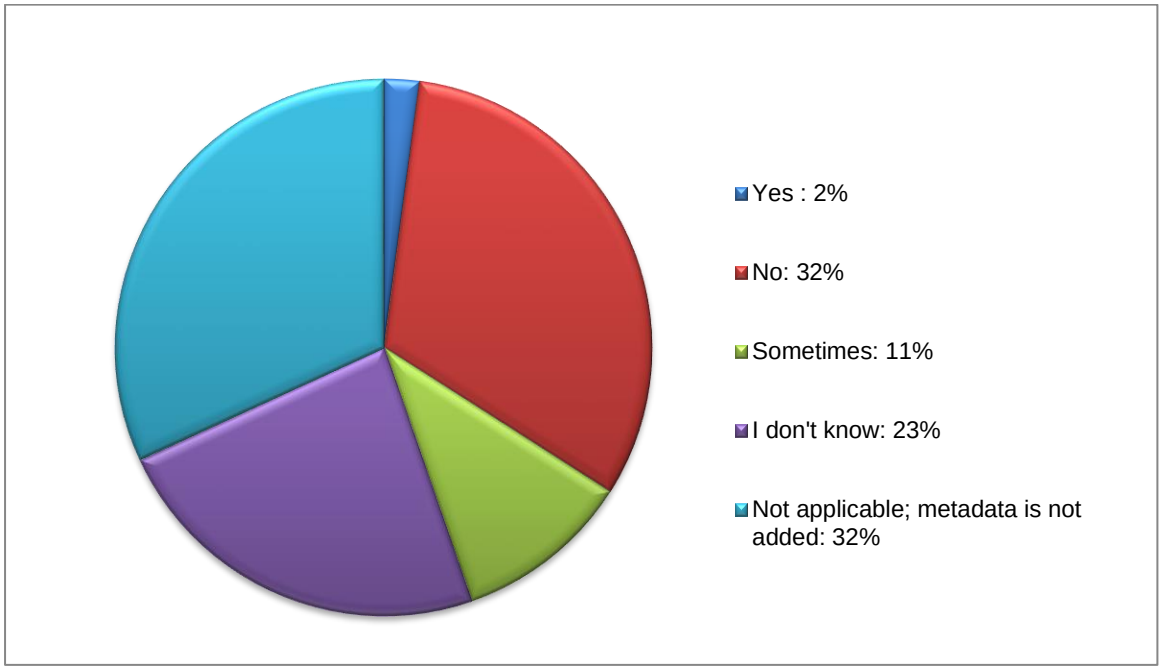


Figure 13: Metadata standards/guidelines

As can be deduced from the figure above, the use of metadata standards by emerging CSIR researchers is not a common RDM activity. Only one emerging researcher stated that adherence to metadata standards/guidelines is a practice always performed. A third of respondents do not add metadata, while another third of respondents do not make use of standards or guidelines when adding metadata. Furthermore: a quarter of respondents indicated not knowing whether a particular standard was being used; this researcher presumes that this could be due to metadata being added automatically by laboratory equipment, for example, or metadata being added by research team member/collaborators, resulting in not having hands-on knowledge of the exact nature of metadata usage pertaining to their research data.

When comparing the use of metadata standards in this survey with the metadata standards use of experienced CSIR researchers, a straightforward comparison is not possible. This is due to the fact that the current study did not exclude non-metadata users when enquiring about standards, while Patterton's previous study (2014a:9) focussed on the use of metadata standards among researchers adding metadata, only. Despite this, it is apparent

that not making use of metadata standards is prevalent with both groups. Respondents always making use of a standard when adding metadata were in both studies found to make up a small minority only.

Questions about adherence to a metadata standard are often included in RDM studies, enabling this researcher to compare the findings of this survey with previous studies. As stated in Chapter 2, non-adherence to a metadata standard is a trend found in most RDM studies. Whitmire, Boock & Sutton (2015:390) reported that 74% of respondents do not make use of a metadata standard, or are making use of a standard devised in the laboratory, while Peters & Dryden (2011:395) reported that no respondents make use of metadata standards. In addition, Mowers, Humphrey & Perry (2013:5) state that 88% of respondents do not use a metadata standard, while Martinez-Uribe (2008:9) claims that very few researchers are even aware of existing metadata standards. Yeumo (2014:0) mentions that the use of metadata standards is uncommon (used by 23% of respondents only), and that most respondents do not see how standards could be beneficial to research. These mentioned studies are just a few examples of RDM studies portraying the general absence of metadata standards when managing research data. Looking at previous studies, it is found that this aspect of emerging researcher's RDM behaviour is in full agreement with that of researchers around the world. Only 15% of respondents claimed to always add metadata to datasets, a further 37% stated that it is sometimes done, and in this group of researchers (roughly half of emerging researchers) only 2% always make use of a metadata standard.

According to the DCC (2015), researchers making use of a metadata standard are ensuring the creation of 'rich, consistent metadata which will support the long-term discovery, use and integrity of digital resources'. The current study has shown that emerging researchers do not adhere to a metadata standard, meaning that the benefits of metadata standard/guideline adherence will not be found to be part and parcel of research data managed by the CSIR's emerging researchers. The implications of not adhering to metadata standards, as well as suggestions on how to address this RDM gap, are discussed in Chapter 5, section 5.4.

4.15 Intellectual Property Rights ownership of research data

The chart below shows the responses given when researchers were asked to indicate the owner of the intellectual property rights (IPR) for their data.

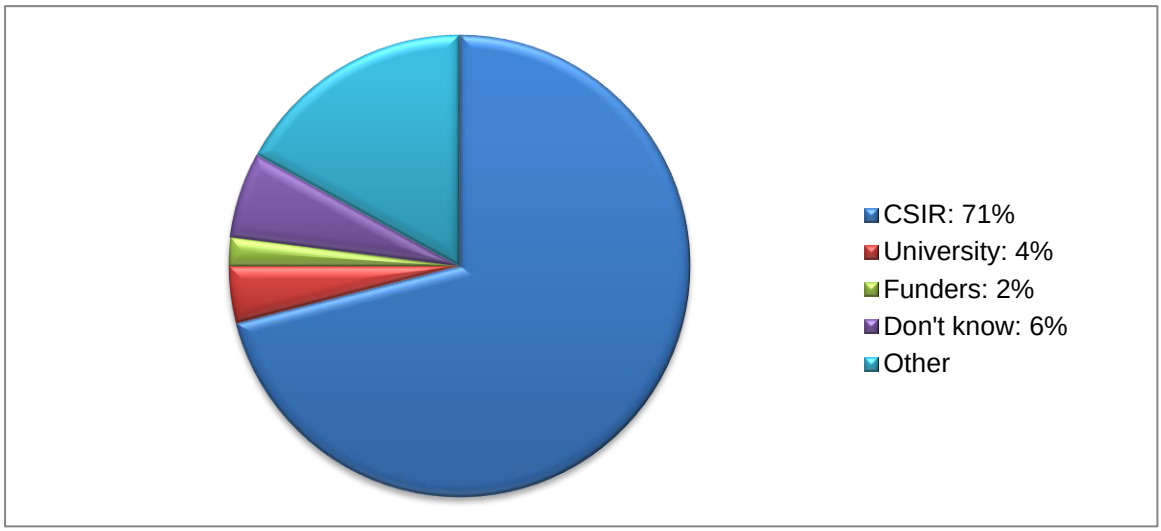


Figure 14: Research data ownership

As is seen from the above graph, 71% of emerging researchers stated that the CSIR owns the intellectual property right to their data. Other single stakeholders, such as the university (4%), or the funder (2%) were indicated to be IPR owners in a tiny minority of cases.

A category titled 'other', prompting respondents to explain why they had chosen that option, was picked by 16% of emerging researchers. Further investigation into this category revealed the most common 'other' option explanation to be joint custody by the CSIR and the university. Joint custody between the CSIR and the contract client was a further explanation given by one emerging researcher. Another single 'other' option response explained that IPR ownership would depend on the research performed; it would sometimes be the CSIR, sometimes the funder, and that some degree of uncertainty was held as to the role of the university.

An alarming finding shown in the graph above is the fact that 6% of respondents did not know who the IPR owner/s of their data is/are.

When comparing the IPR ownership statements of emerging researchers with the answers supplied by experienced researchers in the earlier CSIR RDM study (Patterton, 2014a:15), the results differ slightly. In both studies, the majority of respondents revealed that their data belonged to the CSIR. Emerging researchers indicated a higher percentage than the

experienced researchers; the current survey shows a figure of 71% while the previous study arrived at 39%.

The number of experienced researchers stating that their data belonged to the client/funder (28%), was markedly higher than the numbers indicated by emerging researchers (2%). Possible explanations put forward by this researcher for the discrepancy, are:

- Research data commonly evaluated by emerging researchers during this study, were probably different from the research group data evaluated by experienced researchers in the 2014 study. It is quite possible that emerging researchers only took their study data into account when giving a response. Adding on to this, experienced researchers were probably taking many years of dealing with research data into account, involving different funders, clients and projects. It might even be unwise to try and compare the IPR ownership rights of the two groups, with the survey sample, and in essence experience, type of data dealt with, and projects involved with, too different to overlap on this measurement aspect.
- It is hoped that the stated difference in IPR ownership is due to actual ownership variance, and not ignorance or guessing on the part of emerging researchers. It is a worrying fact that 6% of emerging researchers were not able to indicate whom their data belongs to.

4.16 Data confidentiality/sensitivity

Emerging researchers were asked whether their data was subject to confidentiality rules, or sensitive data restrictions. The graph below showcases responses given:

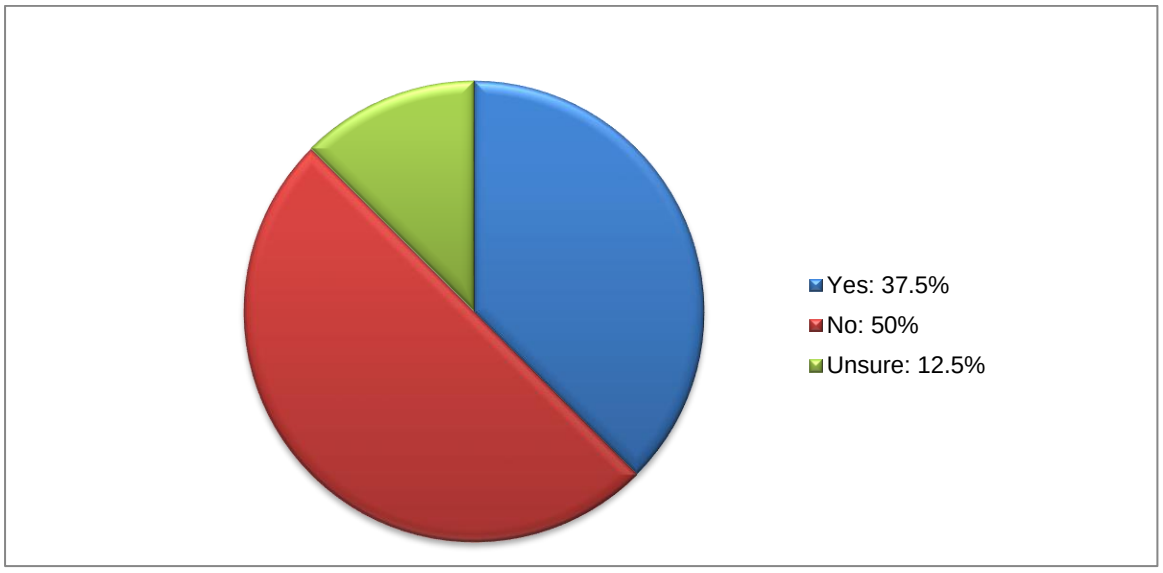


Figure 15: Data confidentiality/sensitivity

As can be seen from this graph, half of the data created by respondents in this study is not considered by them to be subject to confidentiality or sensitivity restrictions. However, 37.5% of emerging researchers indicated that their data are either confidential in nature, or are subject to sensitivity restrictions. A rather worrying aspect of the graph is that 12.5% of respondents were unsure how to classify their data.

When comparing these results with the 2014 CSIR RDM study by Patterton (2014a:16), results do not seem to be in agreement. Experienced researchers indicated that the majority (53%) of their data are confidential, and that only a quarter are not confidential. Possible explanations put forward by this researcher for the anomaly in research data classification at the same institute, are the following:

- PhD-data were being compared with CSIR research data used in the Research Group Leader's group.
- The current study included both the terms 'confidentiality' and 'sensitivity' in the survey question, whereas the interviewer in the previous study (Patterton, 2014a:16) mainly enquired about data confidentiality.
- When engaging in confidential research it is more likely that the organisation would allocate such work to experienced research staff.

Implications of findings related to this topic include uninformed emerging researchers not treating sensitive/confidential data accordingly, as well as the important role of data privacy practices, as 37.5% of data was described as being either confidential or sensitive in nature. The next section (section 4.17) deals with such practices.

4.17 Steps taken to ensure data privacy

Emerging researchers were asked to indicate the steps taken to ensure the privacy of their data. The graph below showcases the answers supplied:

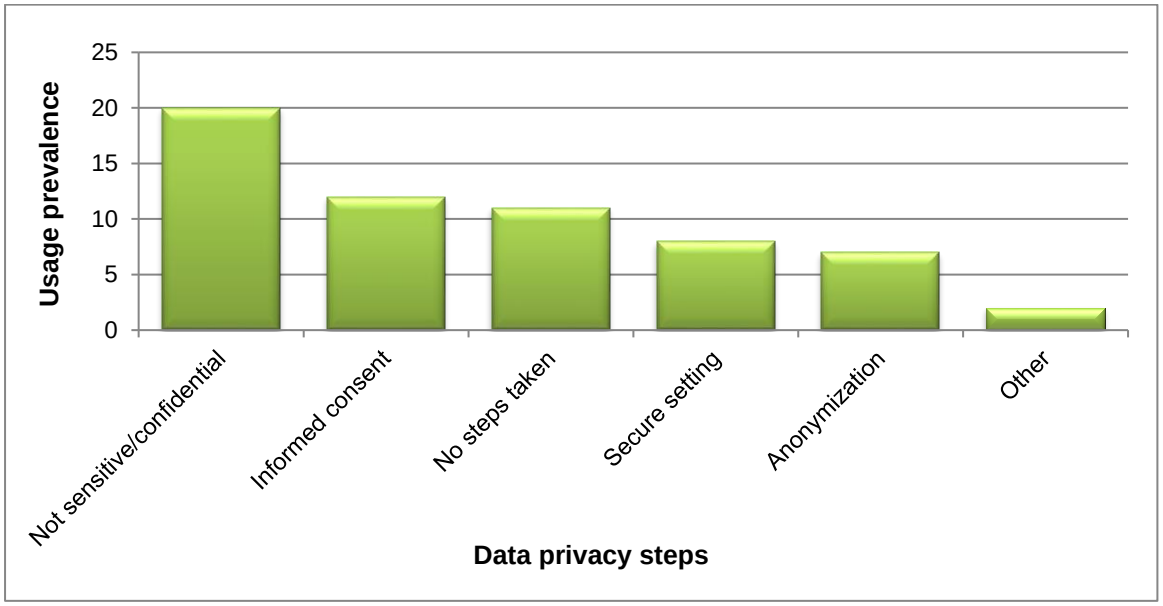


Figure 16: Data privacy steps

As is seen from the above chart, the most prevalent steps taken to protect the privacy of data included gaining informed consent, using a secure setting, and anonymising data. A worrisome aspect is the fact that 11 emerging researchers admitted to not taking any confidentiality steps; however, it is not clear whether this response is due to data not being confidential, or not treating confidential data accordingly.

This particular question was in multiple-choice format, with respondents requested to indicate all applicable options. Seven options were listed. An eighth option, being 'other', with the request to elaborate, was also given. It is interesting to note that no respondents indicated that they destroy their data within a short time of publishing research results, or destroy data within a specified time period. Furthermore, two respondents stipulated that they made use of 'other' methods to deal with data privacy issues, but did not provide adequate elaborating information. Only one respondent provided additional information, but on investigating the details it is seen that this is merely explaining that the respondent deals with confidential as well as non-confidential data.

The results of this question cannot readily be compared with the findings of the previous CSIR RDM study (Patterton, 2014a:12), as the earlier survey did not probe data privacy

steps. Instead, experienced researchers were asked to state the research data security measures taken by them. As such, responses given were more an indication of steps taken to ensure data did not get lost or damaged, than it was about making sure data are protected from unlawful access. Server access restrictions, believing it is ICT responsibility, and making use of backups, shown to be used by 28%, 17% and 17% of experienced researchers respectively, were the measures found to be most commonly used to ensure data are secure, not lost and not damaged (either accidently or maliciously). With both groups of researchers indicating the need for a secure setting for research data, either to protect data privacy or to ensure data are not lost, the importance of this aspect should be noted, and are discussed further in Chapter 5, section 5.4.

4.18 Data sharing: sharing parties

In order to establish with whom emerging researchers share their research data, respondents were asked the following question: 'Data sharing entails sharing informally with researchers, as well as more formal sharing such as data repositories, data banks and data centres, or submission to a journal to support publication. With whom do you share your data? Who can typically access the research data you are creating?' In addition, respondents were told that as many options as were applicable, could be chosen. The graph below displays all responses supplied.

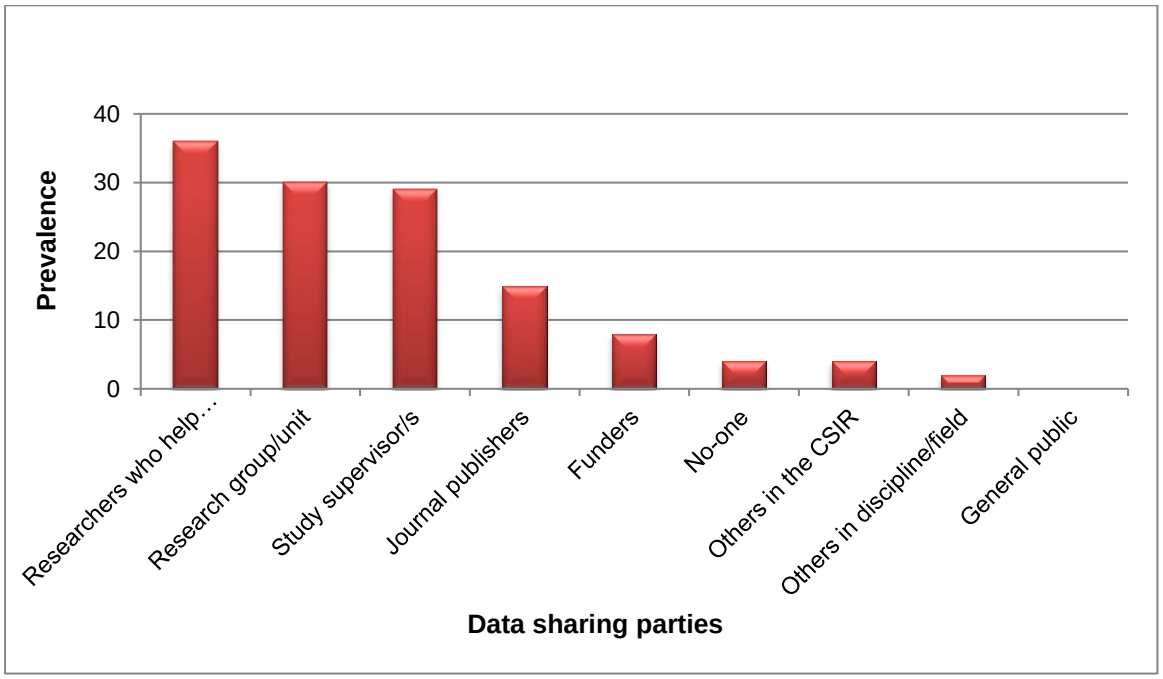


Figure 17: Data sharing parties

As can be seen from this graph, data are most often shared with researchers who helped create the data, fellow members of the research group or unit, as well as study supervisors. It is also interesting to note that no emerging researchers stated sharing data with the general public, and only two respondents were sharing data with others in the same discipline/field. Four respondents stated that data are never shared.

This researcher was interested in comparing the sharing practices of emerging researchers with her previous CSIR RDM study (Patterton, 2014a:15), in which experienced researchers at the CSIR were asked to indicate who the potential audiences for their research data were, or who might be interested in seeing it. Before making a comparison, this researcher felt obliged to emphasize the difference between the two studies when questioning respondents about sharing parties: the current survey clearly wanted researchers to state with whom data are shared, or who is able to access the data, while the previous study by Patterton (2014a:15) enquired about the 'potential audiences' for their research data. In short: the previous study investigated the 'potential interest' while the current study was more direct, in asking emerging researchers: 'With whom do you share your research data?'

Some degree of similarity is noticed, with the scientific fraternity a prevalent group in both studies. However, emerging researchers indicated not readily sharing with others in the research field, meaning those researchers not in the same CSIR unit, project or research team. 28% of experienced researchers stated that clients/funders would be interested in their data, while 16% of emerging researchers claimed to give access to the same group. 8% of experienced researchers felt that the public might be interested in their research data while none of the emerging researchers claimed to be sharing data with members of the public. While 12% of emerging researchers stated that they do not share their data with anyone, a closely-corresponding finding of 8% was found for experienced researchers.

This researcher was not fully able to compare this study's indicated data sharing recipients with similar surveys elsewhere, as study authors elsewhere were more interested in the link between disciplinary sharing differences pertaining to recipients of data. So while Akers & Doty (2013:10) established that medical researchers were least willing of all disciplines to share outside their project, and also report that arts and humanities researchers were more willing to share with the public than with other researchers, the findings of this study, not investigating disciplinary differences, cannot be compared with those of studies done elsewhere. A finding in agreement in both groups of studies (i.e. this study, and studies done elsewhere) is the non-sharing of with funders. Akers & Doty (2013:10) reports that half of all respondents were not willing to share with funders, while the percentage of emerging researchers indicated to be sharing data with funders, was found to be 16% only.

The implications of non-sharing of data would mean that sharing advantages, such as the increase in visibility of work, enhancement of a researcher's reputation, increase in citations, or the enabling of collaborations on related themes or even new topics, are lost. In addition to these advantages not being realised, the impact of not sharing data on research integrity is cause for concern. Failing to share data would affect the ability of others to scrutinize, validate or reproduce the research findings. Addressing the possible reasons why emerging researchers do not readily share data, as well as suggestions for implementing tools making sharing easier, are some of the issues discussed in Chapter 5, section 5.4.

4.19 Data sharing requests

Emerging researchers were asked whether, during the last five years, they had received requests from other researchers for access to their data. Five different options, with 'no, never' at the one end of the spectrum, and 'frequently: more than ten times', were included in this multiple choice question. The chart below illustrates the responses to the question:

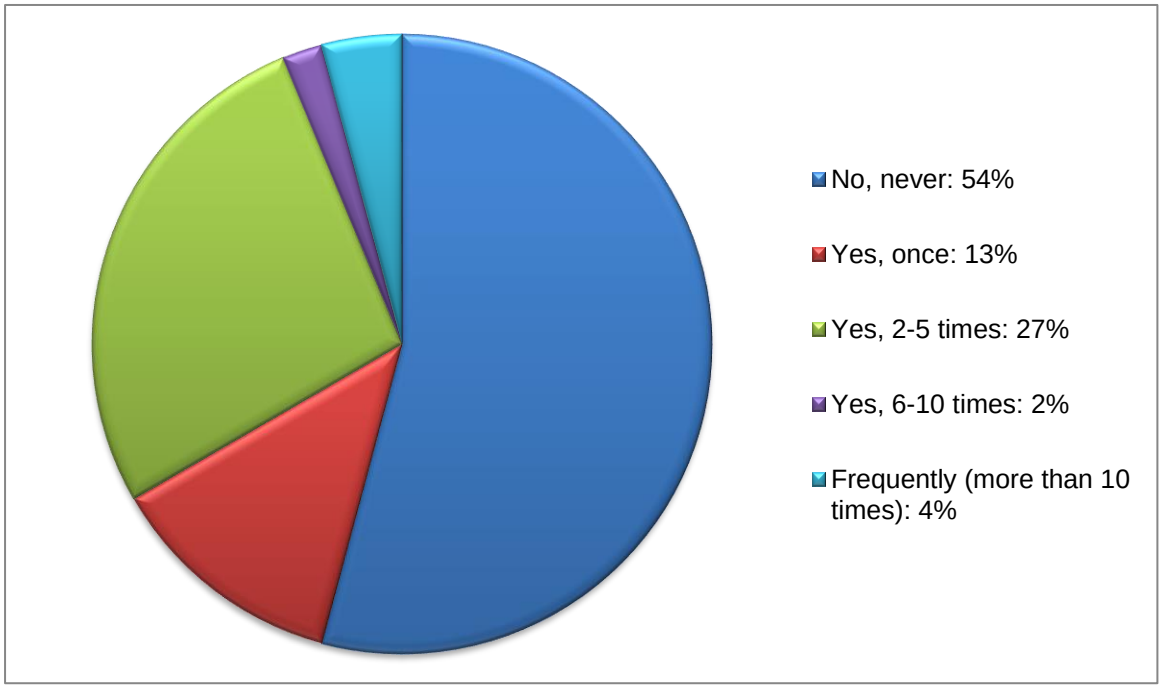


Figure 18: Data sharing requests

As can be seen from the chart above, data sharing requests directed towards CSIR emerging researchers cover the entire frequency range included in this survey question. The majority (54%) of emerging researchers stated that they had not received a request for their data during the last five years. About a quarter of respondents indicated that they had received between two and five requests for their data. Only in the minority of cases (2% of researchers had six to ten requests, 4% of researchers had more than ten requests) can

frequent/regular data sharing requests, from other parties, be seen to be a CSIR RDM occurrence.

Thus, while requests for data (requested from CSIR emerging researchers) are currently and within the last five years not a common occurrence, it does still occur.

While more than half of respondents indicated that they had never, in the last five years, been asked to share their data, it is expected that this figure might drop in future. The expected increase in data sharing awareness, expected increase in visibility of research data when deposited in data archives, as well as the fact that 46% of emerging researchers had received data sharing requests, makes this an RDM topic important to the survey population. Recommendations around data sharing are discussed in Chapter 5, section 5.4.

4.20 Providing access to own data

Emerging researchers were asked to indicate how often they have been able to provide access to their own data. In instances where researchers were never able to provide access, they were requested to provide additional, clarifying information. Responses are indicated below:

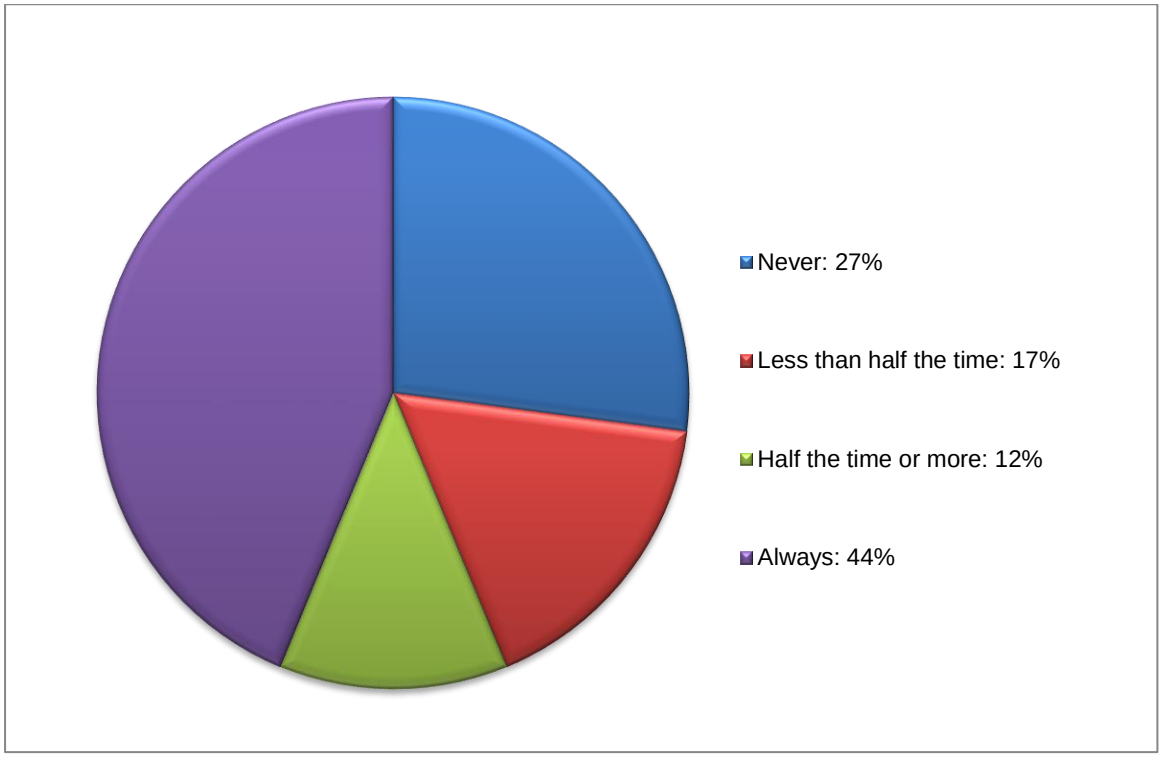


Figure 19: Access to own data

As is seen from the above chart, emerging researchers reported varying levels of success when asked about their ability to provide access to their research data. The group of emerging researchers indicating that they were always able to provide access to their own data, formed the largest group, at 44% of respondents. In total, 73% of researchers were able to provide access to their data, be it less than half the time, half the time or more, or always.

A worrying figure was the revelation that 27% of emerging researchers were never able to provide access to their data. This discrepancy with the results shown for an earlier data sharing question (section 4.19) can probably be ascribed to respondents interpreting the question (section 4.20) as meaning 'Would you have been able to share data if you had been asked?'. This unclarified discrepancy in findings could probably have been resolved through the use of focus group discussions; an issue that will be discussed in more detail in section 4.31 (Limitations of study).

Clarification of reasons for not providing access, were provided by 12 emerging researchers:

- data were not requested by other researchers (six responses),
- still more research to be published (using the data) for journal articles before releasing (two responses),
- IP laws; confidentiality agreements (two responses),
- 'I mostly work on development of systems' , and
- researcher still in early stages of PhD; no publishable results/data available yet.

When taking into account that the CSIR currently has no RDM policies or procedures in place, and therefore assuming that many researchers might not be aware of the advantages of data sharing, the reasons for not sharing appear valid and come as no surprise. No technology-related reasons are mentioned (such as 'requestor would need a specific software programme to view the data', or 'dataset was too big to transmit'), and it seems as if the majority of emerging researchers are not succumbing to non-sharing practices due to competitiveness, data quality issues or loss of control over own data, as was found in literature.

Willingness and ability to share research data is a topic found in many RDM studies; this researcher was able to compare the data sharing tendencies and non-sharing reasons with sharing-related findings published in earlier studies. As was discussed in Chapter 2, section 2.4.7, RDM literature concerned with data sharing practices reveal great variance. While some studies reveal a general willingness among researchers to share data (Knight, 2013:4; Buys & Shaw, 2015:14; Sewerin *et al.*, 2015:6), others show it to be an uncommon event

(Nassiri & Worthington, 2012; Kennan & Markauskaite, 2015:69). Thus, in agreement with studies across the globe showing data sharing practices to occur on a continuum, the data sharing tendencies of this study show emerging researchers to be a rather heterogeneous group.

When looking at previous RDM studies, it is noticed that the reasons given by respondents for not sharing data are numerous and diverse. Only a few reasons were given by this study's researchers for not sharing (as was discussed above); previous studies across the globe have stated many more reasons supplied by non-sharing researchers. Some of the reasons are the following:

- competitive research advantage (Kennan & Markauskaite, 2015:82; Van Tuyl & Michalek, 2015:19),
- privacy/subject protection/legal issues (Peset, 2014:1; Buys & Shaw, 2015:5; Kennan & Markauskaite, 2015:82; Van Tuyl & Michalek, 2015:19),
- fear of data misuse (Peset, 2014:1; Wiley, 2014:1),
- fear of data being scooped (Averkamp, Gu & Rogers, 2014:15; Wiley, 2014:1; Sewerin *et al.,* 2015:6),
- data sharing not required by funders or institution (Wiley, 2014:1),
- data sharing being a time-consuming activity, or not having funds or time to partake in data sharing (Averkamp, Gu & Rogers, 2014:15; Sewerin *et al.,* 2015:6; Van Tuyl & Michalek, 2015:19),
- data quality issues (Sewerin *et al.,* 2015:6), and
- lack of data sharing resources/technical difficulties (Kennan & Markauskaite, 2015:82; Sewerin *et al.,* 2015:6).

While many of these reasons are not yet listed by the CSIR's emerging researchers, this researcher is of the opinion that these concerns might surface in future, when their own research data are requested by others.

Carrying on from this, and in agreement with Raggett (2012a:9) who states that many of the non-sharing reasons are based in fear, this researcher addresses fear-related concerns, and ways to overcome this inhibiting RDM-related emotion, in Chapter 5, section 5.4.

4.21 Data sharing methods/infrastructures

The graph below illustrates the methods or infrastructures used by emerging researchers when sharing their data. As many options as possible could be indicated. Should researchers indicate making use of a curated digital data repository, they were requested to supply more details on its location, curator, characteristics, users and usage.

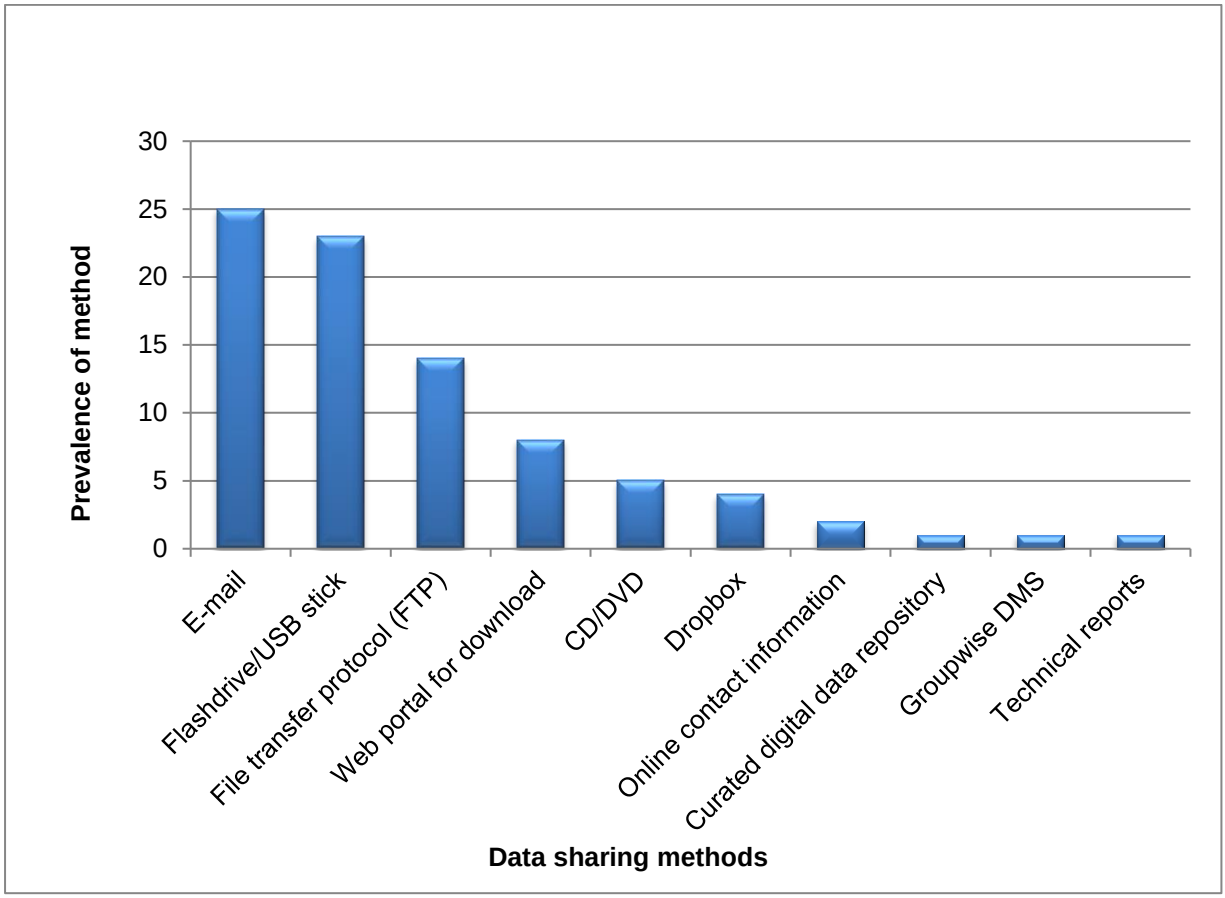


Figure 20: Data sharing methods/infrastructure

Looking at the above graph, it is noticed that in total, ten different data sharing infrastructures/methods are used by emerging researchers when sharing data. All in all, 44 emerging researchers responded to this question, indicating a total of 84 sharing methods used by all. This figure translates to 1.9 sharing methods per respondents; it can be assumed that, on average, an emerging researcher makes use of about two data sharing methods when sharing data. The number of respondents indicating that data are shared shows a discrepancy with sharing numbers indicated in an earlier sharing-related question; use of focus groups would have provided clarifying information and is a study limitation further discussed in section 4.31.

Two sharing methods are found to be more prevalent than the rest: email was stated to be used by 57% of respondents, while flash-drives are seen to be used by 52% of respondents.

Sharing infrastructures scoring low in prevalence were the provision of online contact information for data requests (indicated twice), making use of a curated digital data repository (indicated once), using GroupWise DMS, the CSIR document management system (indicated once) and placing data inside compulsory technical reports (indicated once).

The single emerging researcher making use of a curated digital data repository indicated that the repository was located on the K-drive of the Global Spatial Data Infrastructure server. It was also mentioned that their remote sensing data are stored there, and that strict sharing protocols are in place. The low prevalence of data sharing via a curated digital data repository, although cause for concern, is not altogether unexpected. The CSIR does not have a curated digital data repository as yet, and as it is seen as one of the hallmark features of an institution invested in RDM, it is an issue that is discussed in more detail in Chapter 5, sections 5.4.7 and 5.4.9.

During the previous CSIR RDM survey (Patterton, 2014a:17) experienced researchers had to indicate which types of data sharing methods they would use when sharing data. Results are in agreement with the current study: e-mailing was found to be a popular method, with 47% of experienced researchers opting to share data this way. Other sharing methods commonly used by experienced researchers, such as ftp (31%), data put on a disk (28%), the use of Dropbox (22%) and making use of flash drives (22%) further accentuate the commonality in practices when comparing emerging researchers' sharing methods with those of experienced researchers.

Questioning respondents about their data sharing practices is one of the topics most commonly found in RDM studies. As such, this researcher was able to make comparisons between this study's data sharing findings and findings of other RDM studies. Studies analysed by this researcher in Chapter 2 revealed common data sharing methods to be:

- emails (Akers & Doty, 2013:9; Pink *et al.,* 2013:20; Buys & Shaw, 2015:15; Kennan & Markauskaite, 2015:81),
- ftp (Nassiri & Worthington, 2012:24),
- internet services (Diekmann, 2012:27; Raggett, 2012a:21; Pink *et al.,* 2013:20),
- portable devices (Alexogiannopoulos, McKenney & Pickton, 2010:24; Peters & Dryden, 2011:396; Pink *et al.,* 2013:20),
- a collaborative web space (Peters & Dryden, 2011:396),

- data centres (Martinez-Uribe, 2008:9; Wiley, 2015),
- supplementary material in a journal (Diekmann, 2012:27; Akers & Doty, 2013:9; Buys & Shaw, 2015:15; Wiley, 2015),
- meetings (Pryor, 2009:80), and
- presentations and posters (Pryor, 2009:80; Diekmann, 2012:27).

These data sharing methods, seen to be used by researchers across the globe, are shown to be similar to methods used by emerging researchers in the current study.

Chapter 5, section 5.4 addresses implications of current sharing methods in slightly more detail, and this researcher suggests ways in which current under-utilized sharing tools, such as curated digital data repositories, can be included in the data management regime of emerging researchers.

4.22 Requesting access to secondary data

In an effort to establish the use of secondary data, emerging researchers were asked to indicate how often during the last five years they had asked other researchers to provide them with access to their research data. In the context of this study, secondary data are defined as other researchers' primary data. Should a respondent indicate that no data have ever been requested, they were asked to explain the reason why. The responses to the question are illustrated in the chart below:

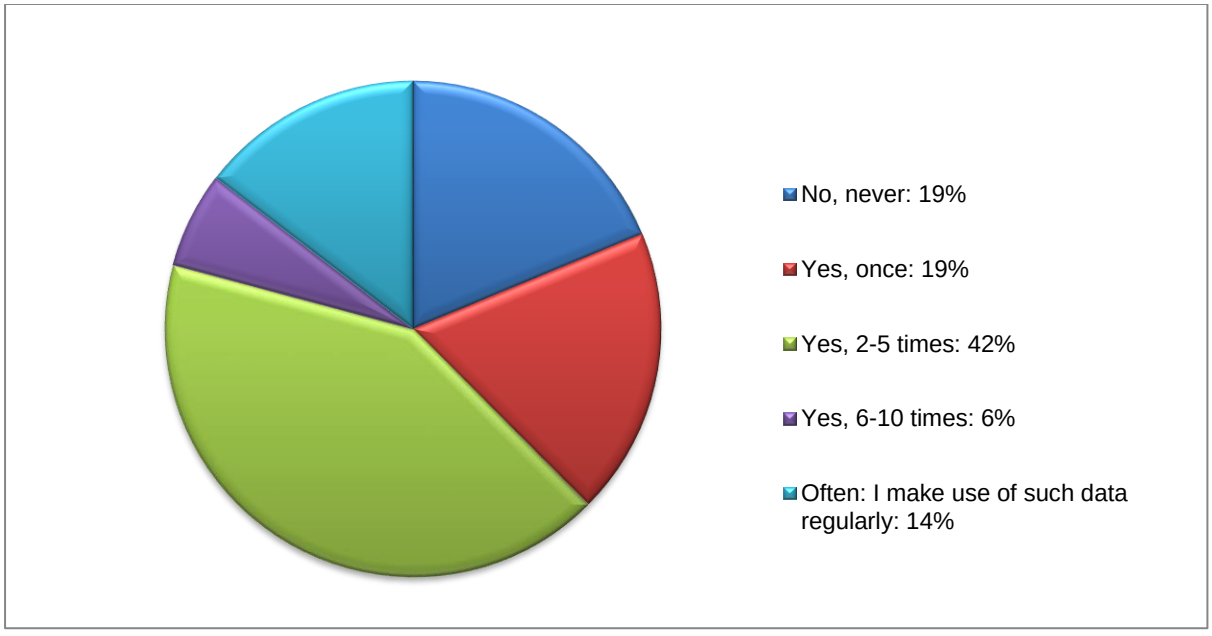


Figure 21: Requesting secondary data

As is seen from the above graph, requesting data from other researchers is a widely varying practice, with data requests ranging from something never done, to an activity often performed. Most commonly, nearly half of emerging researchers (42%) have requested data from others between two and five times during the last five years.

Almost one in five emerging researchers stated that data are never requested from others, with reasons for this behaviour listed to be:

- not required for the specific project (four responses),
- haven't had the need to ask; guidelines for data use almost always provided with available data,
- not required as all required information can be obtained from research articles (three responses),
- respondent is often provided with the data if involved in the study, and
- IPR constraints.

What can be gathered from these findings, is that emerging researchers do make use of secondary data. Regular, as well as infrequent secondary data use, amount to a total of 81% of respondents. One of the implications of this RDM activity would be ensuring that emerging researchers are knowledgeable about finding trustworthy secondary data, and are aware of proper data management practices. Recommendations in this regard are found in Chapter 5, section 5.4.

4.23 Data storage after publication

Emerging researchers were asked to indicate where their research data were stored or preserved, after their research results had been published. Respondents were given a multiple choice format question, containing five location options. A sixth option, 'other', asked emerging researchers elaborate when choosing this option. Findings for post-publication research data storage are shown below.

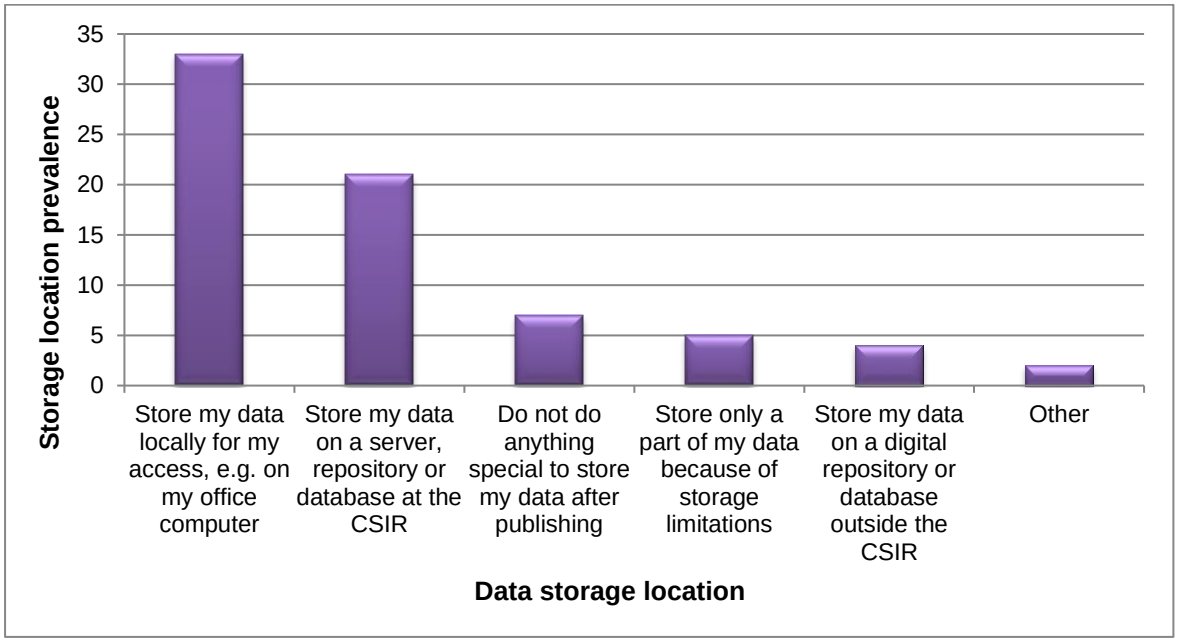


Figure 22: Post-publication data storage

As is seen from the graph above, several different storage locations are shown to be used by emerging CSIR researchers. All in all, a total of 72 different storage locations were indicated by 48 respondents, which roughly translates to 1.5 storage locations per respondents. In essence, it can be said that emerging researchers are making use of more than one location when storing data after publication.

All storage options listed in the question were shown to be used, to some extent, by emerging researchers. The storage option most commonly used, is storing research data locally for the researcher's own access; 69% of respondents indicated to be utilizing this location type. Also shown to be commonly used by emerging researchers post-publication, is the storage of data on a server, repository or database at the CSIR; roughly half of emerging researchers utilize this option. Storage options/locations not found to be commonly used, include the storing of only a part of data due to storage limitations (10%), and the storing of

data on a digital repository or database outside the CSIR (8%). The 'other' category, chosen by 4% of respondents, was found to contain the following clarifying details:

- post-publication research data were stored on a CD, home computer, or USB stick, and
- one respondent indicated that he had yet to publish his research findings.

It was interesting to note that 15% of researchers indicated not doing anything special to their data after publishing.

The previous CSIR RDM survey (Patterton, 2014a) did not specifically address the issue of post-publication data handling, but it did reveal inconsistency in preservation practices. Studies conducted elsewhere globally, and discussed in Chapter 2, section 2.4.6, revealed considerable variance when it comes to post-publication data treatment. Responses ranged from '…few think long-term preservation' (Jahnke & Asher, 2012:3) and '…sometimes neglected once project is complete' (Alexogionnopoulos, McKenney & Pickton, 2010:29) on the one end, to Beile's study (2014) revealing that 68% of respondents take measures to preserve data. Kennan & Markauskaite (2015:84) report that only 8% make use of an internal data storage facility or an institutional data service after project completion, and report that a third of respondents experienced serious data preservation issues, while Sewerin *et al.* (2015:3) state that a long-term data preservation platform, or institutional repository, is a requirement of their study's respondents. This variance in the findings of previous studies seems to be in agreement with this study's findings, displayed in the graph above.

Implications of post-publication data treatment as indicated in this survey, and in particular curated data repositories not being the preferred location or most prevalent location, could prove to be serious in the long term. Data being lost, data integrity being compromised, data formats becoming obsolete, and data not serving a secondary purpose are a few possible outcomes of haphazard post-publication data curation. Ways of addressing these dangers are found in Chapter 5, section 5.4.

4.24 RDM tasks performed

In an effort to establish what RDM tasks are currently being performed by emerging researchers, respondents were asked 'Which of the following data management tasks do you generally perform? Select all that apply'. The graph below illustrates the responses given:

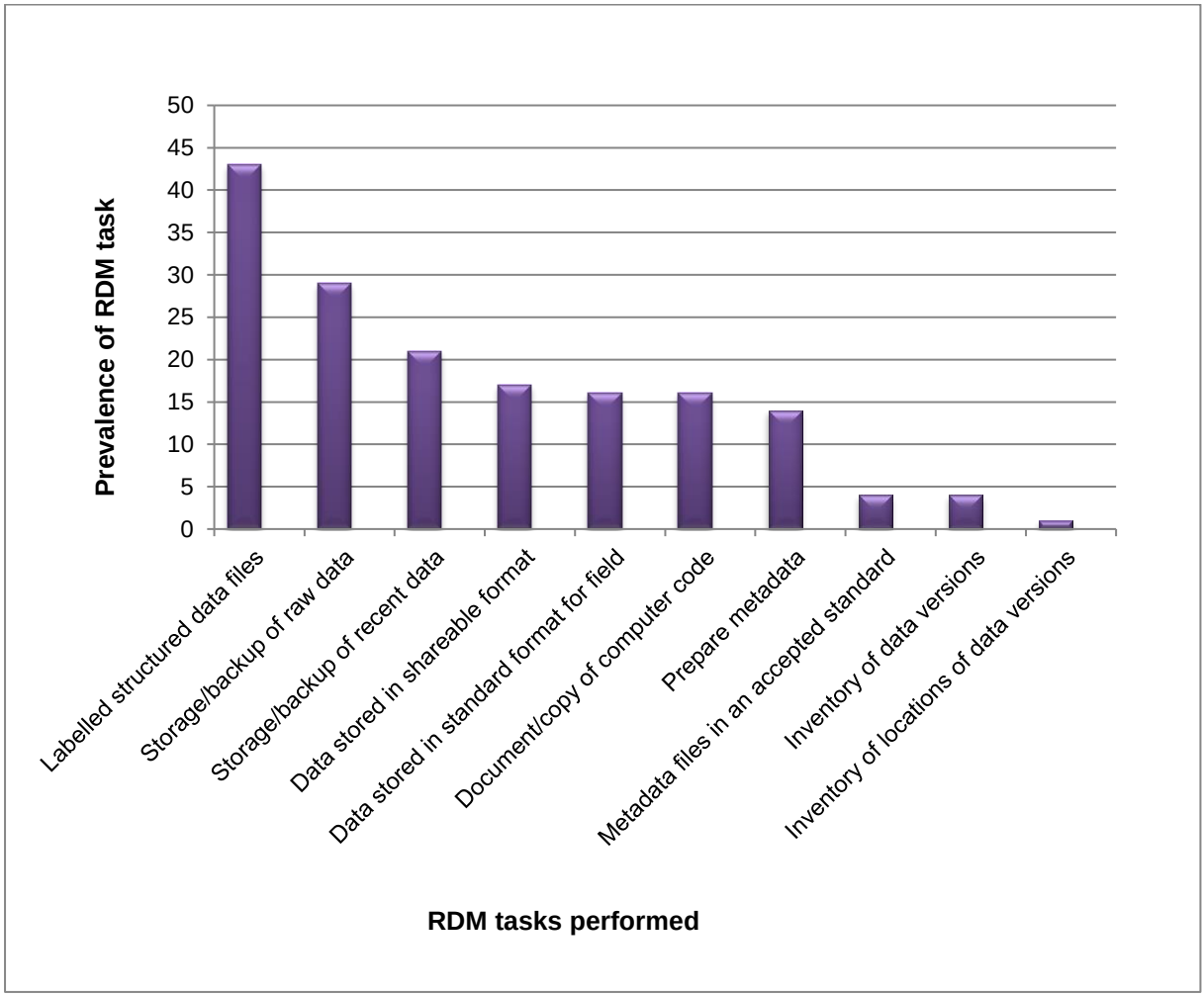


Figure 23: RDM tasks performed

As can be seen from the above chart, some RDM tasks are more prevalent among respondents than others. As such, organising data into structure data files with clear labelling is found to be the most common RDM activity (performed by 43 out of 48 respondents, i.e. 90%). Other prevailing RDM activities include ensuring the storage and backup of original raw data files/master version (performed by 29 out of 48 respondents, i.e.

60%), and ensuring that storage and backup of most recent data files (performed by 21 of the 48 respondents: 44%).

At the other end of the RDM task spectrum, it is noticed that three RDM activities in particular are generally not performed by emerging researchers. In both instances, making use of a metadata standard, and maintaining an inventory of data versions were found to be tasks executed by four respondents (8% of question respondents) only. The RDM activity discovered to be the least common with respondents, was maintaining an inventory of locations of data versions, with only one respondent indicating that it was an activity performed.

The remainder of RDM activities were all found to be carried out by roughly a third of respondents, with these activities showing prevalence scores between 30% and 36%.

Enquiring about the RDM activities usually performed was not part of the previous CSIR survey conducted by this researcher (Patterton, 2014a).

Implications of findings displayed above would suggest that while some RDM activities are already being implemented by emerging researchers, other RDM practices are not forming part of their research routine. The role of training, the importance of RDM funding, and ways to increase RDM awareness are discussed in Chapter 5, section 5.4.

4.25 Research data management training

Emerging researchers were asked whether they had ever received any RDM training. The chart below illustrates responses given:

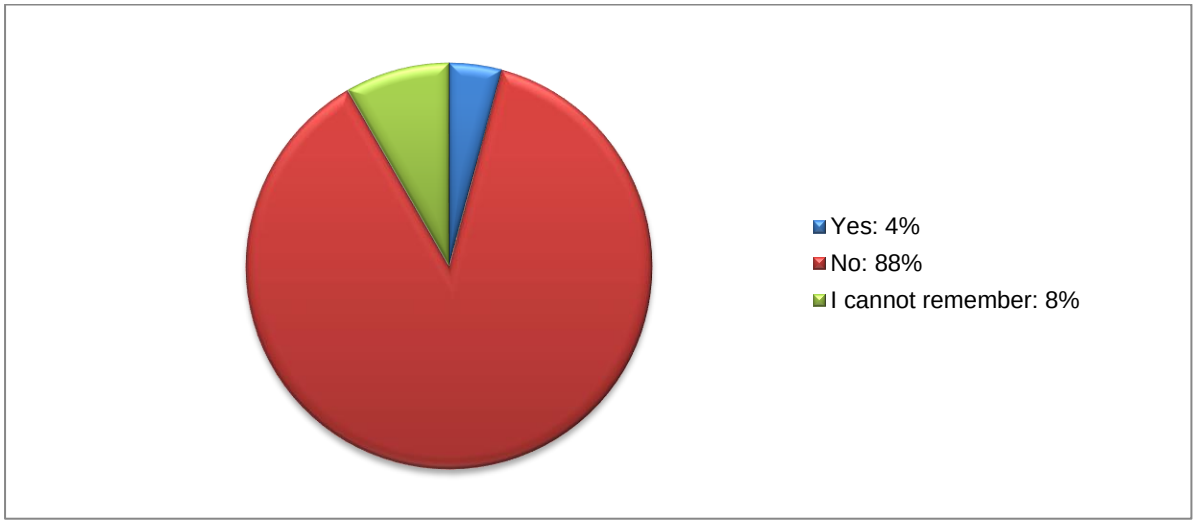


Figure 24: RDM training received

As is seen from the above chart, the majority of emerging researchers (42 of 48 respondents) indicated not receiving any RDM training ever. Only two respondents stated that they had undergone RDM training previously.

While a similar question was not posed to researchers during the earlier CSIR RDM survey (Patterton, 2014a:7), some conclusions as to the prevalence of RDM training received by experienced researchers can be drawn when looking at responses given when queried about their RDM awareness. The fact that only 28% of experienced researchers stated that RDM was a familiar practice and an applied one too, seems to correspond with the current survey findings; RDM training received by CSIR researchers is not a widely-experienced event.

This researcher was also interested in comparing the RDM training undergone by emerging researchers, with the training undergone by scientists/researchers as reported in previous studies. As reported in Chapter 2, very little to no formal training was received in all RDM studies consulted. The studies of Carlson *et al.* (2011:9), Jahnke & Asher (2012:3), Parsons, Grimshaw & Williamson (2013:21) and Patrick (2014) are examples of studies portraying the worrying fact that most respondents had not received any kind of formal RDM training, or researchers indicating a need for RDM training. Similar findings are shown in a JISC study analysing the results of several JISC studies (JISC, 2014). After analysing their survey results, a need for RDM training is also suggested by Knight (2013:4) as well as Rolando *et al.* (2013:27). The reason for the low level of RDM training shown in most studies, could be explained by the fact that RDM studies were usually performed prior to implementing RDM services at an institute, when no formalised training was on offer yet.

The findings of this survey pertaining to RDM training received convey a dire need for RDM training. With only 4% of respondents claiming to have received RDM training, it could be argued that RDM skills are lacking in most young researchers and that there is a dire need for RDM training. While it is possible that young researchers might have received informal training, or had taught themselves via online tools or similar, or had received RDM-knowledge during the course of their studies or short careers, it stands to reason that the numbers indicated in the chart above are not ideally suited to a leading scientific institution.

As is seen in later sections in this chapter (4.26 as well as 4.27), survey respondents regarded training materials and guidelines to be important RDM services. This service need, coupled with the finding that the majority of the CSIR's emerging researchers had never received formal RDM training, accentuates the fact that RDM training is one of the main RDM services that need to be addressed by all parties concerned. While online RDM

training tools such as web-based RDM courses, and webinars are on the increase, and institutes of higher education also starting to increase RDM awareness, this researcher acknowledges that RDM at the CSIR training at the CSIR is non-existent, leaving CSIR's emerging researchers to their own devices. Recommendations for addressing this state of affairs are put forward in Chapter 5, section 5.4.

4.26 RDM training: areas of interest

In an effort to establish RDM training requirements of emerging researchers, respondents were asked to indicate the RDM areas they were interested receiving RDM training in. Eleven RDM areas were added to this multiple-choice question, and emerging researchers could indicate all training areas they were interested in. In addition to the listed RDM areas, an 'other' option, prompting respondents to elaborate on the other RDM training options needed. Furthermore, an option enabling respondents to state that RDM training was not required, was included in the list of choices. The graph below illustrates the responses supplied:

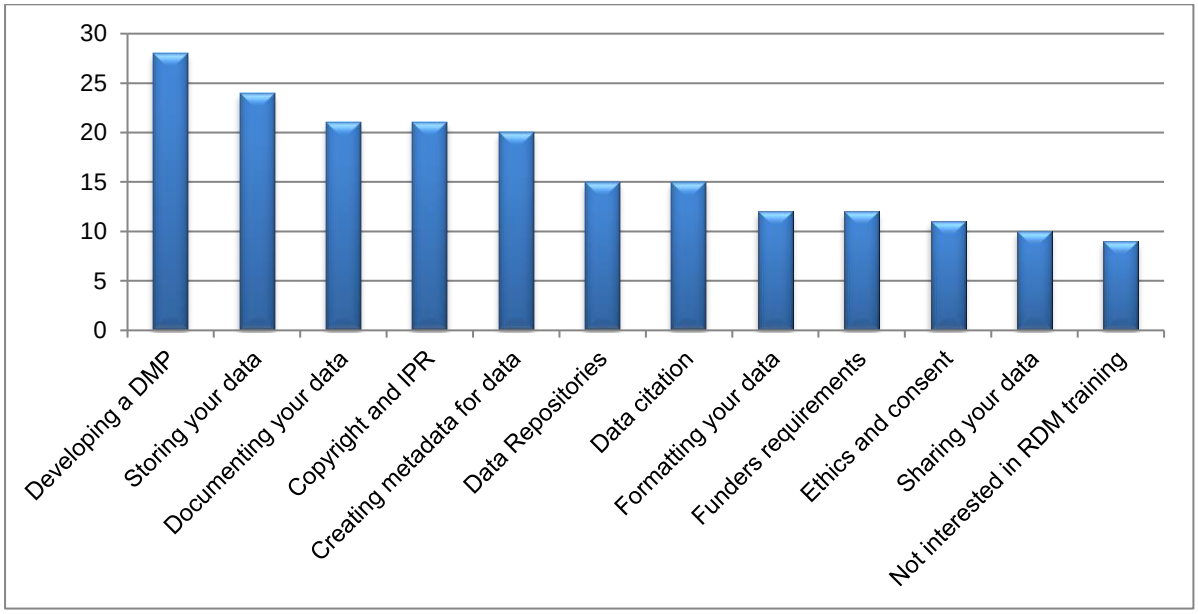


Figure 25: RDM training areas required

Looking at the chart above, it is seen that while varying degrees of interest were shown for the RDM areas forming part of the multiple choice question, most options revealed to hold some level of interest for emerging researchers. The RDM area seen to be most frequently ranked as training requirement, was the development of a data management plan, with 58% of emerging researchers indicating this as training requirement. The recent statement by the

NRF regarding proof of intended RDM (National Research Foundation, 2015) might be a contributing factor in this result.

In a similar vein, receiving training about data storage was seen to be a training requirement by half of respondents. Other areas seen to indicate high level of training interest, were data documentation (44% of respondents), copyright and IPR (44% of respondents), and metadata creation (42% of respondents).

Looking at the areas showing the least level of training interest, it is seen that training in the areas of ethics and consent (23%), and data sharing (21%) were spheres not seen to be highly regarded, training-wise. The low level of interest in ethics training might be due to the fact that ethics is only an issue in some disciplines, and it might be that those researchers in need of ethics training had already received adequate guidance and support.

It is interesting to note that although not the majority, an alarming 21% of respondents indicated that they were not interested in receiving RDM training. Whether this response was given because emerging researchers felt that already possessed adequate RDM skills, or had no time or desire to undergo RDM training, or felt that RDM was not an important part of a CSIR researcher's make-up, is not clear. However, as stated in section 4.25, the stated low levels of RDM training received is regarded as a red flag by this researcher. As such, the topic of RDM training is an area given due consideration in Chapter 5, section 5.4.5.

This researcher was also interested in comparing the training requirements as shown in this study, with those of researchers across the globe, as indicated by earlier studies. The main training requirements, as discussed in Chapter 2, section 2.4.11, can be recapped as follows:

- a need for training in broad data skills, as well as narrow RDM skills (Wilson & Patrick, 2010:32),
- a need for training in basic practical RDM issues, such as data storage and data creation (Parham, Bodnar & Fuchs, 2012:12 also Knight, 2013:4). Studies showing a need for training in 'general RDM practices' (Kouper *et al.*, 2013:376) and 'data management training' (Beile, 2014:18) as well as 'best data practices' (Buys & Shaw, 2015:15) add to this category of RDM training requirements,
- guidance around data storage and data preservation (Jahnke & Asher, 2012:3; Knight, 2013:4; Parham, Bodnar & Fuchs, 2012:12; Gibson & Gross, 2013:13-15; Parsons, Grimshaw & Williamson, 2013:35; Van Tuyl & Michalek, 2015:1),
- assistance with data sharing obligations and data sharing agreements (Knight, 2013: 4; Buys & Shaw, 2015:15),

- backing up guidelines (Wilson *et al.*, 2011:284),
- training in the area of data anonymization (Delasalle, 2013:102; Mowers, Humphrey & Perry, 2013:11),
- issues surrounding data sensitivity (Nassiri & Worthington, 2012:11; Parsons, Grimshaw & Williamson, 2013:33),
- data security training (Marcus *et al*., 2007:18 as well as Delasalle, 2013:102),
- training in copyright and IP licensing, also ownership and legal issues (Martinez-Uribe, 2008:10; Knight, 2013:4; Pink *et al.,* 2013:24; JISC, 2014: 8; Simukovic, 2014:5),
- training requirements around the topic of data management plans were shown by Martinez-Uribe (2008:11), Knight (2013: 21), Parham, Bodnar & Fuchs (2012:12), Rankin *et al.,* (2012:32), Buys & Shaw (2015:12), Sewerin *et al.* (2015:7), as well as Van Tuyl & Michalek (2015:1), to name but a few,
- training regarding funder requirements were mentioned by Knight (2013:21), Parham, Bodnar & Fuchs (2012:12), Buys & Shaw (2015:15) as well as Sewerin *et al.* (2015:3),
- training related to metadata was a prevalent requirement and reported by Carlson *et al.* (2011:11), Rankin *et al.* (2012:32), Beile (2014:18) as well as Van Tuyl & Michalek (2015:1), to name but a few studies,
- ethics training as a requirement was indicated by Carlson *et al.* (2011:14) as well as Jahnke & Asher (2012:13), and
- data citation training was reported on by Akers & Doty (2013:13) as well as Van Tuyl & Michalek (2015:1).

Looking at trends and training requirements as stated by researchers worldwide, much overlap is noticed between the training requirements indicated by the CSIR's emerging researchers, and training requirements of researchers elsewhere in the world.

Implications of findings related to RDM training required include the realisation that not all researchers are aware of RDM benefits, or of the existence of certain RDM activities. There is not a single RDM activity included in this question that is required by all researchers; in fact, the highest required percentage indicated for a single area was found to be 58%, with the rest of RDM areas included in the question scoring even lower. Recommendations are put forward in Chapter 5, section 5.4.5.

4.27 Importance of RDM-related services

In an effort to gauge the perception of emerging researchers towards RDM services, respondents were asked to rate the importance of eight RDM services. Adding on to this, an additional open-ended question, asking respondents to indicate any data-related service, data-related policy, or data-related practice, not mentioned in earlier questions, deemed important to them when managing research data. The responses of the question concerned with the RDM services are illustrated in the graph below, while the responses of the open-ended question (regarding RDM-related services) are discussed elsewhere in this section.

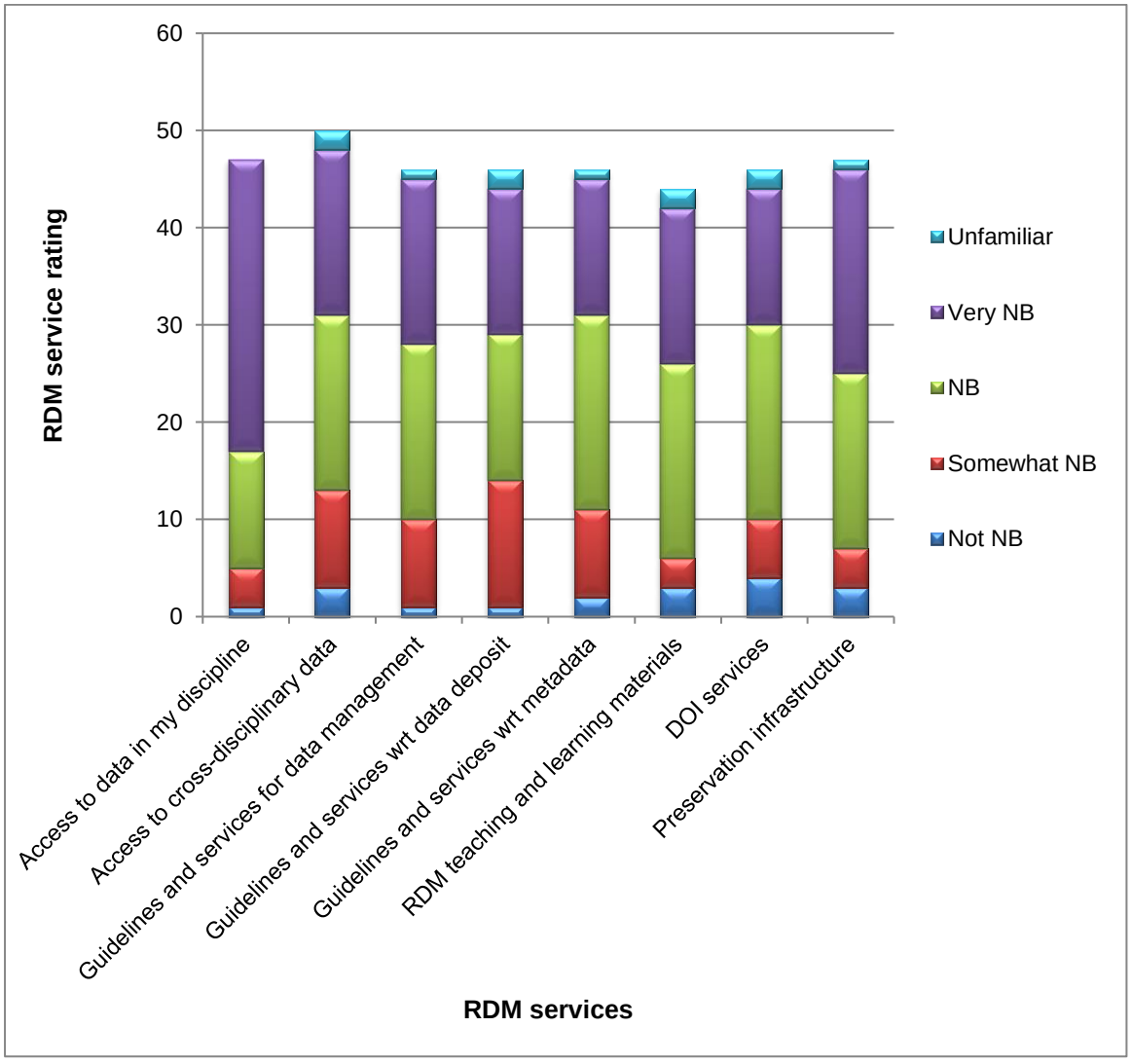


Figure 26: Importance of RDM-related services

As is seen from the graph above, emerging researchers evaluate the eight listed RDM services with varying degrees of importance. Furthermore, all of the services listed here were responses in all levels of evaluation, ranging between 'not important', 'somewhat

important', 'important', very important' and 'unfamiliar'. The only exception to this was in the case of the service listed as 'access to data in my discipline', where no respondent indicating that they were unfamiliar with the concept. A survey tool limitation was noticed when analysing results of this question, in that 50 responses by 47 respondents were shown for 'Access to cross-disciplinary data'. Although a closer examination of the results revealed this to be an isolated occurrence, it is nevertheless a survey tool limitation/error in need of further investigation, and will be discussed in section 4.30 (Limitations).

When looking at RDM services displaying the most responses in the evaluating category 'very important', it is seen that 'having access to data in their own discipline' (63% of responses for this question), and 'having the infrastructure at the CSIR to allow CSIR data to be preserved, and made available to others' (44% of responses for this question) had the highest number of responses. When adding up the results for the 'very important' category and the 'important' category, the RDM services described as 'having access to research data in my discipline' scored highest, with an accumulated total of 88% of responses for this service. Other RDM services scoring high when adding the total of the two mentioned categories, were shown to be 'having the infrastructure at the CSIR to allow CSIR data to be preserved, and made available to others ' (82% of respondents), and 'having teaching and learning materials so that researchers can work with data' (75% of respondents). Accumulated totals for the categories 'very important' and 'important' were never lower than 63% for any of the RDM services investigated in this question.

Looking at the RDM services evaluated as 'not important', the highest percentages were seen for the service described as 'having services necessary to assign a permanent DOI to my data', at 8%. The lowest scores for this evaluation category was found to be for 'having increased access to research data in my discipline' (2% of responses), which supports the previous paragraph stating the perceived importance of this RDM service. Additionally, two RDM services also revealed to have low scores in the 'not important' category, were 'guidelines and services supporting researchers in managing their research data' as well as 'guidelines and services supporting researchers in depositing their research data', with only 2% of respondents in both instances choosing this category.

There are implications of certain RDM services being rated highly, as well as being rated as either unimportant or a service researchers are not familiar with. This researcher acknowledges that marketing, training and guidance might be needed to address ignorance and unfamiliarity with issues. Furthermore, RDM services shown to be rated as important or very important reveal an insight and understanding of the importance of the rated RDM

service; in short: a positive finding. The ramifications of services rated on the extremes of the spectrum, are discussed in more detail in Chapter 5.

Although only completed by two respondents, responses to the open-ended question (see first paragraph of section 4.27 have revealed the following:

- the need for an e-research data infrastructure was expressed,
- the need for the development of a linked data infrastructure was mentioned, and
- the need for sufficient storage, able to accommodate the backing up of large volumes of research data, was indicated.

Although interested in comparing these findings with those of previous studies, either at the CSIR or internationally, such an action is not easily performed. The ratings requested from emerging researches in this study was not a question forming part of the previous CSIR study (Patterton, 2014a), and its exact counterpart is also not easily found in other RDM studies. As discussed in section 4.26, as well as in Chapter 2, section 2.4.11, most studies revealed training to be the most prevalent requirement, and attached a high rating to it. Nevertheless, the need for the following RDM-related services has been shown in earlier RDM studies:

- RDM guidelines (Alexogiannopoulos, McKenney & Pickton, 2010:33, as well as Ward *et al.*, 2011:268),
- preservation services and storage space (Jahnke & Asher, 2012:3; Parham, Bodnar & Fuchs, 2012:12; Gibson & Gross, 2013:13-15; Simukovic *et al.* 2014:5), RDM infrastructure (Westra, 2010:5), central repository (Pienaar, 2011:19),
- collaboration tools (Jahnke & Asher, 2012:3),
- sharing tools (Parham, Bodnar & Fuchs, 2012:12),
- data sharing services (Jones K, 2011:1),
- open sharing and reuse facilities (Open Exeter Project, 2012:4),
- a facility showcasing final data sets (Parsons, Grimshaw & Williamson, 2013:33),
- collective data sharing facilities (Nassiri & Worthington, 2012:14),
- one-stop data access across disciplines (Mowers, Humphrey & Perry, 2013:11), infrastructure allowing integration of multiple data sources (Wynholds *et al.,* 2011:386), and
- better RDM funding (Parham, Bodnar & Fuchs, 2012:12 as well as Mowers, Humphrey & Perry (2013:11).

While these unrated required RDM services listed cannot be directly compared with the findings illustrated in figure 25, they do indicate some degree of overlap between the current group of emerging researchers, and researchers surveyed in previous studies. Both groups have indicated a need for data access (in own discipline and across disciplines), RDM guidelines, preservation services, and a decent RDM infrastructure. As such, it can be safely assumed that the RDM needs of CSIR's emerging researchers are in line with global trends.

4.28 Importance of RDM-related standards, policies, principles and practices

In an effort to gauge the perception of emerging researchers towards RDM standards, policies, principles and practices, respondents were asked to rate the perceived importance of thirteen RDM areas fitting the categories mentioned. Five categories were available when evaluating RDM standards, namely 'not important', 'somewhat important', 'important', 'very important', and 'unfamiliar with concept'. The responses supplied are shown in the graph below.

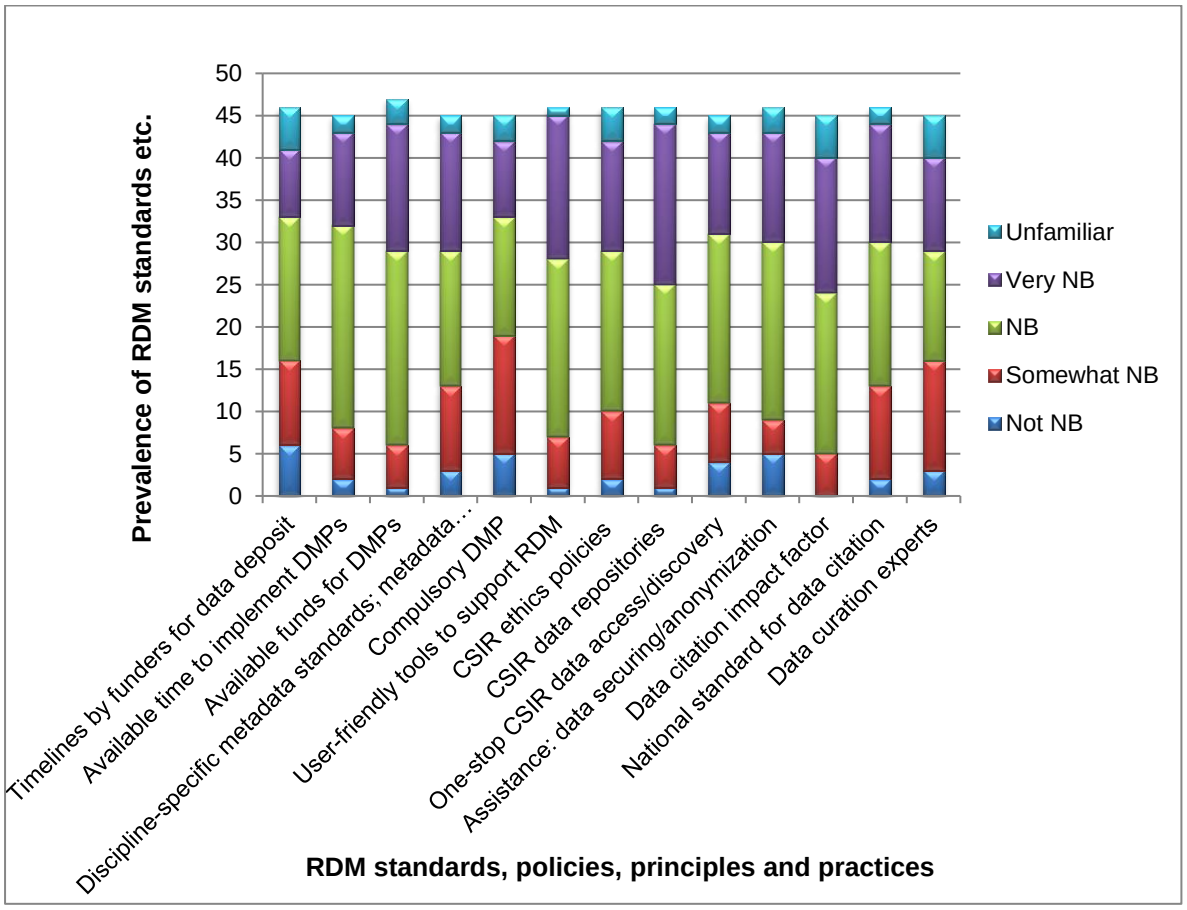


Figure 27: Importance of RDM standards, policies, principles and practices

RDM policies/practices forming part of this multiple choice question were all deemed to be an important part of an institute's RDM regime by this researcher. The responses to this question, as indicated in the chart above, would allow this researcher to establish:

- which practices and policies are rated highly be emerging researchers,
- which practices and policies are seen to be not important, and
- which practices and policies are unfamiliar to respondents.

When looking at all responses for this question, it is noticed that, with the exception of one practice/policy and one rating category, all options/ratings for all listed RDM practices have been indicated by respondents. The strength of the five levels of evaluation differ between RDM practices and it is in investigating these differences that a better idea of emerging researcher's perception towards listed RDM standards, policies, principles and practices can be formed.

RDM practices shown to have the highest percentage of votes in the 'very important' category, were the services described as 'CSIR data repositories where researchers can deposit data' (40% of responses), 'user-friendly tools to support RDM' (35%) and 'data citation impact factor/recognition when data are cited' (33%). When combining the 'very important' as well as the 'important' ratings, policies and practices seen to attain a high ranking are:

- CSIR data repositories where researchers can deposit data (a combined total of 79% of respondents indicated the two mentioned ratings),
- user-friendly tools/assistance to support RDM (79% of respondents),
- having available funds to implements data management plans, and manage data (79% of respondents). This practice was also elaborated on in the open-ended question discussed in section 4.29, where a respondent mentioned the challenges of funding RDM, and of making funders more aware of this necessity, and
- data citation impact factor or recognition (e.g. promotion, grants) when data are cited (73% of respondents).

Ideally, all options forming part of this multiple choice question should be rated as either important or very important. Four of the 13 options listed were deemed to be important by more than 75% of emerging researchers. All options were shown to be important to the majority of respondents (i.e. more than 50%); in other words, most emerging researchers view all RDM practices/policies listed as being vital to their execution of research data management.

RDM practices, policies, standards and principles shown to be more unfamiliar than the others, are:

- having dedicated data curation experts to curate research data (10% of respondents were not familiar with the concept),
- data citation impact factor or recognition (10% of respondents were not familiar with the concept),
- stated timelines by funding agencies for depositing data (10% of respondents were not familiar with the concept), and
- CSIR ethics policy pertaining to research data sharing and confidentiality (8% of respondents were not familiar with the concept).

Although these four RDM policies/practices were only unfamiliar to a small percentage of researchers, the fact that there is no understanding of what the practices entail, serves as a warning to this researcher that RDM marketing, awareness and training is lacking at the CSIR. Recommendations to address and rectify the situation are put forward in Chapter 5, section 5.4.

4.29 Additional RDM concerns, issues or problems

The final question included in the online questionnaire entailed asking respondents to indicate any other data-related concerns, issues or problems, not already covered in the survey. Four responses were supplied, and respondents mentioned the following:

- ‘Data generated comes from computer simulation software. Enormous amounts of data are generated, and only the interpreted data are important. The data itself are often of interest to the client only, and no one else. Backing up this data is often not possible due to demands on storage capacity. Therefore only the input files, along with the actual source code, are backed up. The source along with the input files is sufficient to regenerate the data. However, generating the data requires hundreds of hours of CPU time, and it would be nice to have a repository where such data can be backed up. Services such as Dropbox and Google Drive have space limitations.’
- ‘With online tools such as ResearchGate and OneDrive on the rise, I don’t see the need for a CSIR specific data management system.’
- ‘The importance of proper data curation is not currently widely acknowledged, and this can create challenges for finding and allocating funding for this task. Awareness needs to be raised among local funders and CSIR management to facilitate this more.’

- ‘This may fall under infrastructure, but besides storage capacity, transfer rates are important. Within each CSIR campus the rate is understandably fast, however, it is fractured between campuses. Another issue is ICT skills (or cooperation between ICT entities) in resolving network efficiency. We (CSIR PTA) are on SANREN. CHPC is on SANREN. Speeds between the two are much lower than expected.’

As is seen from the responses above, concerns in the areas of data storage, RDM funding, data transfer speeds and network efficiency were mentioned. The concerns listed here cannot be ignored by either this researcher or the CSIR research support environment: without adequate funding, storage space and network capability, CSIR RDM will battle to take off and will certainly not prosper. Recommendations stemming from these concerns are stipulated in Chapter 5, section 5.4.

In addition, the solitary comment stating that there is no need for a CSIR data management system conveys a perception held by at least one emerging researcher that the current web-based data storage tools are sufficient and trustworthy. Without going in too much detail about the advantages and disadvantages of online tools at this moment, this researcher has put forward recommendations in the next chapter, specifically addressing ways to increase awareness of web-based tools and the issues surrounding them. Furthermore, recommendations on how best to implement and market an envisaged CSIR RDM system are required, as some researcher negativity during the implementation phase could be expected.

4.30 Limitations of survey and survey questions

The advantages and disadvantages of online surveys were addressed in the previous chapter, section 3.5. However, this researcher wished to list issues specific to the delivery, format and content of this survey. It is hoped that by listing these issues, a better understanding of the nature and demands of RDM surveys can be obtained. Moreover, these limitations, as well as suggestions for survey improvements or modifications, could be of value to other researchers wanting to conduct similar RDM surveys in future.

Main limitations of this survey and survey instrument are seen to be the following:

- Only eight of the nine research units completed the survey (see discussion in section 4.3).
- It would be interesting to know the reasoning behind responses given. However, given that the survey comprised 31 questions, and some questions containing up to 19 response options, of which all could be selected, this would prove to be an

insurmountable task. Such an in-depth study of behavioural reasons would be better suited to an in-depth case study featuring fewer participants, or a focus group limited in size.

- The survey software used, although marketed as advertisement-free, did display advertisements to respondents. Advertisements were only removed after this researcher contacted eSurv and queried the status quo.
- This researcher came across a magnitude of different questions included in the RDM surveys analysed by her. While almost all of these questions were interesting, formed a part of RDM, and made additional RDM comparisons possible, not all were included in this survey. As mentioned in a previous bullet, including all possible RDM questions part of past RDM surveys is an impossible and impractical task.
- Discipline-specific RDM trends and inter-disciplinary comparisons were not done in this survey. The aim of this survey was to get a general overview of the RDM behaviours of the CSIR's emerging researchers. As many as possible different RDM activities or practices were included in the survey, aimed at rather addressing the multifaceted nature of RDM and its many components, rather than discipline-specific practices. Furthermore, inter-disciplinary comparison would require the participation and input from all CSIR's research units; a feat not accomplished within this study.
- In retrospect, it could also be stated that not having an RDM policy-related question included in the questionnaire could be a limitation of the study instrument. While there is no arguing this point, this researcher made the decision to limit the number of survey questions to 31, and rather have a questionnaire that would be completed within 30 minutes than one covering every possible RDM issue.

To summarise: a sufficient amount of RDM data was collected via an online-questionnaire completed by 48 emerging researchers. Discipline-specific RDM behaviour, as well as interdisciplinary comparisons, did not form part of this survey. A follow-up focus group study involving fewer respondents, and revealing in-depth RDM behaviour related to specific research disciplines, would add to the theoretical knowledge of CSIR emerging researcher behaviour.

4.31 Summary

Analysis of responses to 31 survey questions asked of the CSIR's emerging researchers has revealed the following main findings and RDM trends:

- **Population and respondents:** The CSIR has 179 emerging researchers. A total of 48 emerging researchers completed the survey, amounting to 26.8% of the survey population.
- **Discipline:** 60.4% of respondents work in the field of natural sciences, 22.9% in the formal sciences, and 16.7% are involved in a multidisciplinary field.
- **Data types:** Emerging CSIR researchers have indicated making use of 15 different data formats. Text documents, spreadsheets and images are the most commonly-used data types. Data types not used very often are audio, digital objects, and questionnaires/transcripts/codebooks.
- **Data volume:** Data volumes of 1-50GB were found to be most common, followed by volumes of 1-50 TB. 15% of respondents had no idea of the volume of their datasets.
- **Software tools:** A total of 31 different software tools were indicated by respondents. The most commonly-used tools were shown to be MS Excel and MS Word.
- **Data management plans:** The majority (76%) of the CSIR's emerging researchers have never created or submitted a data management plan.
- **Awareness of funder requirements:** The majority (56%) of the CSIR's emerging researchers are not aware of their funder's research data management requirements.
- **Data storage:** CSIR's emerging researchers have indicated using 18 different data storage locations. Data are most commonly stored on the hard drive of the office computer, on an external hard drive, on a shared drive, or on a web-based platform. Making use of a discipline-specific data repository is not a common RMD activity for this group.
- **Data backups:** All of the CSIR's emerging researchers back up their data. The most common backup frequency was show to be 'ad-hoc' (37%), with daily backups being the second most common backup frequency (33%).
- **Data backup location:** Emerging researchers most commonly back up their data to external hard drives, a CSIR drive, or to a cloud-based environment.
- **Metadata documentation:** No outright finding regarding the documentation of metadata could be made. Emerging researchers were found to most commonly add metadata sometimes (37% of respondents) or not at all (33% of respondents).
- **Use of metadata standards/guidelines:** Approximately a third of emerging researchers add metadata to their datasets, and only a further 2% of these researchers adhere to a metadata standard.
- **Intellectual Property Ownership of data:** The majority of emerging researchers (71%) stated that their data belong to the CSIR. University ownership was indicated

in 4% of responses, while only 2% of respondents mentioned their funders to be owners of their data.

- **Data confidentiality/sensitivity:** Half (50%) of respondents described their data as not being subject to confidentiality or sensitivity issues, while 37.5% indicated that it was. 12.5% of respondents were not able to classify their data.
- **Data privacy steps:** The most common step taken when data are prone to privacy issues, is obtaining informed consent.
- **Data sharing:** When data are shared by emerging researchers, it is most commonly shared with researchers who helped create the data, researchers within the same group, or with study supervisors. Data are least likely to be shared with the general public.
- **Data sharing requests:** The majority of emerging researchers (54%) indicated not receiving any requests for their data during the last five years. About a quarter of the respondents (27%) stated that between two and five requests for data access had been received during the last five years.
- **Data access:** 44% of respondents indicated that they were always able to provide access to their data. 27% of respondents stated that they were never able to provide access, with reasons being IP restrictions, being in the initial phases of a project, or wanting to make use of the data (publish) before sharing it. A discrepancy in dating sharing results was identified.
- **Data sharing methods:** The most common data sharing methods/tools were found to be email, flash drives/USB sticks, and FTP servers. Making use of a curated digital repository was a very uncommon data sharing method.
- **Requesting primary data from other researchers:** While responses were varied, 42% of respondents stated that they had requested access to other researchers' primary data during the last five years. 19% of researchers indicated never having asked for data access, with the most prevalent reasons being IPR restrictions, or not requiring access to the data.
- **Storage after publication:** Locations most commonly used for storing data post-publication, were found to be the office computer, or a server/repository/database belonging to the CSIR.
- **RDM tasks performed:** RDM tasks most commonly performed by emerging researchers were shown to be adding structured labels to data files, and backing up raw and recent data. Adding metadata to datasets, keeping an inventory of data versions, and keeping an inventory of locations of data versions, were RDM tasks not commonly performed.

- **RDM training received:** the overwhelming majority of emerging CSIR researchers (88%) had not received any RDM training.
- **Areas of interest for RDM training:** RDM training areas selected by most respondents include the creation of a data management plan, data storage, data documentation, copyright/intellectual property rights, and the creation of metadata.
- **Importance of RDM-related services:** Having access to data in their own discipline, and having infrastructures at the CSIR allowing for data to be preserved and made available to others, were rated by most respondents to be very important RDM-related services. DOI services received the highest number of responses amount for the 'not important' rating.
- **Importance of RDM-related standards, policies, principles and practices:** Having a CSIR data repository for data deposit, having user-friendly RDM tools, and the use of a data citation impact factor/recognition when data are being used, were rated by most respondents to be very important RDM-related practices.
- **Additional RDM concerns:** Having adequate storage capacity, adequate RDM funding, and decent data transfer rates between CSIR campuses.
- **Limitations of study:** Although a sufficient amount of RDM-related information was gained, the survey tool itself exposed several limitations. Another limitation could be seen to be the absence of focus group interviews to clarify discrepancies in survey responses. Furthermore, discipline-specific RDM behaviour, as well as interdisciplinary comparisons, did not form part of this survey.

This chapter reported on the results of the study. Demographic respondent information, as well as RDM behaviours, practices and perceptions were discussed, and compared with the earlier CSIR RDM survey as well as other surveys conducted elsewhere around the globe.

The next chapter entails establishing whether the study's research questions, as stipulated in Chapter 1, have been answered. Recommendations with regards to establishing research data management services at the CSIR, as well as any other recommendations forthcoming from this study's results, are put forward.

5. Chapter 5: Recommendations

5.1 Introduction

The previous chapter portrayed the results of the study as obtained via completed questionnaires. The responses to each question forming part of the web-based survey were analysed, documented in detail and discussed. In this chapter, this researcher has presented the study's research question and research sub-questions, and has portrayed how these questions have been addressed.

After presenting the study's research questions, the researcher documented the main findings of the study. Implications of these findings were stated and discussed. Following on this, a short conclusion was provided.

Based on the study findings and taking into account the main research question of the study, this researcher put forward several recommendations in the form of guidelines. Ideas for future research emanating from this study, were also stipulated.

This chapter is brought to an end by a study conclusion.

5.2 Research questions

This section comprises a main research question and several sub-questions. These questions, used to guide and centre the research, will be dealt with in reverse order. Reverse order was chosen as answering each of the sub-questions successfully would in turn lead to, and enable, the answering of the main question. Conversely, not addressing or being able to answer any of the sub-questions would lead to difficulty in answering the study's main research question.

A practical example of the relationship between the study's main research question and sub-questions is as follows: through identifying the data management practices of the CSIR's and thereby answering the research sub-question: '*What are the current RDM practices and trends among emerging researchers in the CSIR?'* this researcher was provided with findings which assisted in being able to give data management guidelines to the CSIR. In doing so, the researcher was able to answer the main research question:

How can an organisation like the CSIR ensure that future researchers apply best practices when managing the CSIR's research data?

The sub-questions, the main findings obtained when answering the sub-questions, as well as resulting implications, were addressed before moving over to the study's main research question.

5.2.1 Research sub-questions, findings and implications

This study aimed at answering five research sub-questions; each is discussed in turn.

5.2.1.1 Sub-question 1: What are the international RDM requirements, standards, best practices and expectations that are being developed?

The findings relating to this research question were detailed fully in Chapter 2 (Literature Analysis), and discussed in Chapter 4 (Results and discussion). Furthermore, international trends were once again briefly stated in section 5.2.1.1 (Sub-question 1: What are the international RDM requirements, standards, best practices and expectations that are being developed), when this researcher investigated possible deviations between the CSIR's emerging researchers, the CSIR's experienced researchers, and researchers worldwide. To summarize, the following international RDM trends, practices, and needs were found:

- **Data types:** Most researchers make use of several data types. Data types found to be most common are textual data, images, and spreadsheets.
- **Data volume:** Studies reveal great variance in data set sizes; and it can be deducted that a 'typical' dataset size does not exist. Another volume-related finding revealed researchers to often be unaware of how much data they have created.
- **Data management plans:** Studies reveal international diversity with regards to the use of data management plans. In general, findings showed that the majority of researchers do not yet make use of data management plans. Funders and institutional policies were seen to be the main drivers of DMPs; DMP usage is therefore a dynamic activity prone to sudden change once institutional or funder policies change.
- **Data storage location:** Researchers are relying on themselves to store data. The most common primary data storage location is locally, on a researcher's personal computer or laptop. This general trend was found to be true across disciplines, institutions, geographical areas, and levels of research experience. In addition to a primary data storage location, researchers also tend to make use of more than location when storing data.
- **Data backups:** All studies reveal that the overwhelming majority of researchers back up their data. With regards to backup frequency, it was shown that data are most

commonly backed up on an ad-hoc basis. Researchers make use of a variety of locations or backup options when backing up their data.

- **Use of metadata:** Most studies indicate that the majority of researchers do not create metadata for their datasets. Studies do reveal that the use/non-use of metadata is an activity varying widely between researchers and studies.
- **Use of metadata standards:** The use of a metadata standard is an uncommon activity amongst researchers adding metadata to datasets.
- **Data sharing:** This RDM activity varies widely between studies and researchers. Global sharing practices are currently on a continuum, with some studies showing a willingness to share and others indicating that it is an uncommon event. Common reasons put forward for being reluctant to share, include the need to retain a competitive advantage, legal issues/confidentiality, data sharing being a time-consuming activity, and fear of data being misused. Lack of data sharing resources, or experiencing technical difficulties when sharing, are also seen to be aspects limiting data sharing.
- **Data sharing methods:** Common data sharing methods and tools include email, ftp, internet services, portable devices, data centres, supplementary material in a journal, as well as meetings, presentations, and posters.
- **Post-publication data treatment:** Studies revealed variance when it comes to post-publication data treatment. Practices ranged from not thinking about long-term preservation at all, to making use of a long-term preservation platform.
- **RDM training received:** Very little to no formal RDM training received was found in all studies consulted.
- **RDM training requirements:** Guidance around data storage, backing up, creating a DMP, creating metadata, copyright/legal issues, ethics, data citation, and assistance with data sharing agreements is required.
- **Importance of RDM services:** Apart from RDM training, researchers worldwide place a high premium on data storage space, preservation services (RDM infrastructure), need for collaboration and sharing tools, one stop data access, and better RDM funding.

Findings obtained via RDM studies, and analysed by this researcher, reveal that RDM is a new and dynamic discipline and currently prone to many intra-study and well as inter-study differences. Furthermore, findings obtained via a particular study might only be relevant for a short period, as the implementation of funder requirements and institutional policies has a determining role in establishing RDM practices, as well as adherence to best practices.

5.2.1.2 Sub-question 2: What data practices need more formalised support: at CSIR, nationally, internationally?

Research data management practices of various groups were investigated by this researcher, with RDM practices of researchers nationally and internationally (i.e. non-CSIR researchers) documented and analysed in Chapter 2: Literature Analysis. In addition, RDM practices of the CSIR's experienced researchers were investigated in a previous RDM study (Patterton, 2014a), and also incorporated into Chapter 2: Literature Analysis. RDM comparisons between these groups were done and documented in Chapter 4: Results and Discussion.

In general, RDM practices of the CSIR's emerging researchers were found to be no different to the rest of the world. In addition, no major deviations were found when comparing the RDM practices of the CSIR's emerging researchers with the RDM practices of the CSIR's experienced researchers.

Main RDM trends and practices found to be mostly universal and similar in nature, include the following:

- **Data formats/types:** Researchers worldwide, experienced CSIR researchers as well as emerging CSIR researchers all make use of a variety of data formats when doing research. Most popular formats are textual data, spreadsheets, and images.
- **Data volume:** Studies investigating RDM practices of researchers worldwide, experienced CSIR researchers as well as emerging CSIR researchers report that there is not a typical dataset size. Dataset sizes vary from being very small text data to very large climatic datasets, for example.
- **Software applications:** Researchers worldwide, experienced CSIR researchers as well as emerging CSIR researchers have been found to make use of a variety of software tools. Experienced CSIR researchers, in total, were found to make use of more software applications than emerging CSIR researchers.
- **Data management plans (DMPs):** The use of DMPs was seen to be dependent on funder requirements. The majority of CSIR researchers (both the experienced as well as the emerging group) do not create and submit DMPs. These findings are in agreement with findings of studies done elsewhere; findings of worldwide studies were compared and analysed in section 4.8 (Development of a data management plan).
- **Data storage location:** Studies investigating RDM practices of researchers worldwide, experienced CSIR researchers as well as emerging CSIR researchers

report that data are generally stored locally, on the researcher's computer or laptop. This global general trend was found to be true across disciplines, institutions, geographical areas, and levels of research experience. In addition, studies also revealed that most researchers use more than one place to store their data.

- **Data backups:** Studies investigating RDM practices of researchers worldwide, experienced CSIR researchers as well as emerging CSIR researchers report that the majority of researchers back up their data. Data backup frequencies displayed similarities as well as variation, in that CSIR emerging researchers, as well as researchers worldwide, tend to back up on an ad-hoc basis. Differences will be stipulated in the next list, in the 'Deviations from current global practices' section.
- **Data backup locations:** Studies investigating RDM practices of researchers worldwide, experienced CSIR researchers as well as emerging CSIR researchers report that researchers make use of a variety of locations or backup options when backing up their data. The use of an external hard-drive as the most common backup location is a practice mirrored in at least one other study internationally, and was also seen to be one of the locations favoured by experienced CSIR researchers, albeit it not the most common backup location with last-mentioned group. A further data backup location difference will be mentioned in the next section ('Deviations from current global practices').
- **Use of metadata standards:** Whereas the use of metadata shown in this study showed some dissimilarity with other studies, adhering to a metadata standard has shown universal results. The majority of the CSIR's emerging researchers, the CSIR's experienced researchers as well as studies worldwide have indicated that a metadata standard is not used.
- **Data ownership:** The majority of the CSIR's emerging researchers as well as experienced researchers have indicated that their data belong to the CSIR. The percentage of emerging researchers indicating that data belong to funders, is markedly higher than was indicated by experienced researchers. No comparison can be made with international studies; this RDM aspect is not often included in RDM studies.
- **Data sharing:** Differences in question phrasing/wording have made comparisons of data sharing practices difficult. Data sharing among the CSIR's emerging researchers is seen to be a heterogeneous practice; this finding agrees with findings into sharing practices worldwide. In general, it can be said that sharing with the CSIR scientific fraternity was found to be a common activity, for CSIR emerging researchers as well as CSIR experienced researchers. Low prevalence of data

sharing with the general public was also found for both groups. Non-sharing of data was not found to be prevalent in either of the groups. Reasons for data sharing reluctance showed universal commonalities: retaining a competitive research advantage, privacy/legal issues, and data-sharing being a time-consuming activity were sharing-related concerns. Email was seen to be the most prevalent data sharing method for CSIR emerging researchers, CSIR experienced researchers, and researchers worldwide. Other common data sharing methods include the use of ftp, use of a CD/DVD, making use of Dropbox, and using a USB device.

- **Data storage after publication:** Great variance in post-publication data handling, as well as lack of uniformity relating to data preservation, was shown in this study as well as studies worldwide.
- **RDM training received:** The majority of the CSIR's emerging researchers, as well as researchers investigated worldwide, had received little or no RDM training.
- **RDM training interest:** Much overlap is noticed when comparing the RDM training requirements of emerging CSIR researchers with those of researchers worldwide. DMP training, data storage training, data documentation, training in copyright issues, and metadata training feature strongly in both groups.
- **Importance of RDM-related services:** Although making study comparisons proved to be difficult due to differences in survey question phrasing/wording, the CSIR's emerging researchers as well as researchers worldwide, place a high premium on storage space and preservation services (data infrastructure), RDM guidelines, and access to data required.

Deviations from current global practices were found for the following RDM activities:

- **Data backup frequency:** CSIR emerging researchers most frequently do backups on an ad-hoc basis. This finding differs from the previous CSIR RDM study, where the majority of CSIR experienced researchers reported to be making daily data backups.
- **Use of metadata:** The majority of the CSIR's emerging researchers do not, as a rule, add metadata to datasets; this findings clashes with the previous study in which it was shown that adding metadata is a more common activity among the CSIR's experienced researchers. The low metadata usage of emerging researchers also clashes with several international study findings, but not with all RDM studies. As such, the results of comparing the metadata findings of this study with global trends remain inconclusive.

- **Data confidentiality:** A higher percentage of CSIR experienced researchers have indicated that their data are confidential than was the case with the current study.
- **Data sharing methods:** Worldwide, researchers show a higher tendency to make use of a data repository than is the case with the CSIR's emerging, as well as experienced researchers.

Implications of findings relating to similarities and differences are the following:

- Although RDM practices shown by the CSIR's emerging researchers are in general not seen to deviate widely from global trends, RDM at the CSIR is still in its infancy. It could also be stated that perhaps RDM is not as established globally as was expected by this researcher, and that CSIR RDM, when compared with global RDM trends, is not as far behind as was suspected.
- CSIR RDM similarities with global practices should not be seen as a sign that RDM at the CSIR is of a high standard, or adhering to best practices.
- In general, it can be said that the lack of RDM training undergone by the CSIR's emerging researchers as well as researchers worldwide is expressed via sub-standard RDM practices.
- Several RDM practices, worldwide as well as at the CSIR, were found to portray sub-par standards. These practices include the non-use of DMPs, backing up of data on an ad-hoc basis, using at-risk devices when backing up data, and researchers not having received any RDM training.
- A high degree of RDM overlap was shown to exist between the CSIR's emerging researchers and researchers internationally. This can be seen as proof of the novelty of RDM as an international research practice as well.
- The similarity in current RDM practices paves the way for requesting advice from, or following the examples of other institutions implementing RDM services.
- Areas shown to deviate strongly from international trends could be an indication that these are areas in need of extra concern. Deviating practices were shown to be frequency of data backups, low metadata usage, and the non-use of a curated digital data repository.

Implications of the RDM practices of the CSIR's emerging researchers are numerous, and will be discussed in more detail in section 5.2.1.3 (Sub-question 3: What data are collected and held by emerging researchers in the CSIR), section 5.2.1.4 (Sub-question 4: What are the current RDM practices and trends among emerging researchers in the CSIR?) and section 5.2.1.5 (Sub-question 5: What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?).

5.2.1.3 Sub-question 3: What data are collected and held by emerging researchers in the CSIR?

The online questionnaire used in this survey contained questions about the data formats held, the size of the dataset, as well as software tools used to collect or analyse the data. These findings were discussed in section 4.5 (Types of research data), 4.6 (Volume of research data) and 4.7 (Software applications used for analysis/manipulation of data).

Data formats used by the CSIR's emerging researchers are numerous and diverse. Survey results indicate that 15 different data formats are used, and that, on average, an emerging CSIR makes use of nearly six different data formats. The most common datatypes were found to be text documents, spreadsheets, images, and raw data files. Datatypes not commonly used include audio, questionnaires/transcripts, and digital objects generated/acquired during data creation.

Findings relating to dataset size show that the CSIR's emerging researchers hold data volumes belonging in all dataset size ranges. Data held by emerging researchers were most often found to be in the 1 TB to 50 TB range. It is also worth noting that 15% of respondents had no idea how much data they held. Data volumes least prevalent are seen to be the very large collections.

In total, 31 different software applications were mentioned by respondents. Software applications most commonly used by emerging researchers were found to be Microsoft Excel, Microsoft Word, and Matlab.

Implications of findings relating to data held by emerging researchers are the following:

- The use of a variety of data types and formats by emerging researchers is a reality that is global, expected, and cannot be eradicated. It is part and parcel of scientific research.
- The use of a variety of data types and formats by emerging researchers indicate the need for this phenomenon's acceptance by all CSIR RDM stakeholders.
- RDM services, especially those pertaining to data storage, data access, data preservation, as well as RDM training topics, will need to make provision for all data formats when implemented.
- It is suspected that due to the diverse range of formats used, some of the formats might fall victim to data obsolescence, as well as data access and retrieval problems in future.

- RDM services, especially those dealing with long-term preservation and issues surrounding format obsolescence, as well future data access and retrieval, are vital.
- A one-size-fits-all approach when creating data format guidelines cannot be used; different researchers use different data formats. Furthermore, every researcher tends to make use of nearly six different data types.
- A one-size-fits-all approach with regards to data storage and data backups cannot be used; different researchers use different data formats. In addition, differing dataset size indicates the need for different storage options.

Recommendations based on these findings will be made in section 5.4 (Recommendations), in particular sections 5.4.5 (RDM training), 5.4.7 (Preservation) and 5.4.9 (Data storage).

5.2.1.4 Sub-question 4: What are the current RDM practices and trends among emerging researchers in the CSIR?

The biggest part of the online questionnaire used in this study contained questions about the RDM practices and trends of the CSIR's emerging researchers. The results of these findings were documented in Chapter 4: Results and discussion. A short summary of findings is shown below and the implications of findings are added to each section.

Findings revealed the following:

- **DMPs:** The majority of CSIR's emerging researchers do not create and submit DMPs. Implications of DMP findings are numerous: funders are likely not requesting a DMP, and as a result of this, data are not being managed well. Adding on to this: the major benefits of a DMP are not available to emerging CSIR researchers. It could be argued that as a result of this, research data management could be sub-standard, data security might be suspect, data access might be compromised, and data quality and integrity might be affected. Recommendations pertaining to DMPs are put forward in section 5.4.4 (Data management plans).
- **Funder requirements:** The majority of CSIR's emerging researchers are not aware of funders' RDM requirements. Thus: even though receiving a research grant, bursary or funding from an authority, there is lack of awareness by the recipient regarding management of the research data. The conclusion can be made that communications between funders and researchers are in need of improvement. The implication of not being aware of funder requirements is that researchers run the risk of failing to meet funder demands and requirements, are failing to adhere to RDM best practices, and could compromise future grant awards. Recommendations to

address this issue are put forward in section 5.4.3 (Marketing and awareness) and section 5.4.12 (Funders).

- **Data storage:** CSIR emerging researchers most commonly store their data on the hard-drive of the office computer, on an external hard-drive, or on a USB stick. Curated discipline-specific repositories are rarely used by emerging CSIR researchers when storing data. Implications of findings are numerous, but in brief it can be said that the use of many locations is an RDM reality, and one that needs to be taken note of when creating an RDM procedure or drawing up best practice guidelines. Furthermore, when looking at the range of storage locations being used, it could be argued that not all options are equally secure, protected against theft, data loss, data corruption, or being in a shareable format. Format obsolescence is another storage issue that needs to be considered. Recommendations will be put forward in section 5.4.7 (Data storage), section 5.4.9 (Preservation), and to a lesser extent in section 5.4.5 (RDM training).
- **Data backup:** All emerging CSIR researchers back up their data. Data are most commonly backed up on an ad-hoc basis. A subset of emerging researchers has indicated not being aware of how often their data are backed up. The implications of data being backed up on an ad-hoc basis is that emerging researchers are making use of a backup strategy that is possibly lacking in justification, or are doing so without consideration of wider application, or the ramifications of such a decision. Data loss is a major risk factor when backing up without justification or forethought. Recommendations in this regard will be made in section 5.4.5 (RDM training).
- **Data backup location:** Data created by emerging CSIR researchers are most commonly backed up to an external hard drive, a CSIR drive, a cloud service, or to a USB device. These findings have implications in that backup locations might not meet security standards, possibly resulting in data loss, data theft, or data corruption. It could then be stated that the same dangers mentioned in a previous paragraph relating to data storage implications, could be applicable when discussing data backup locations. The use of external hard drives as preferred backup location, while in full agreement with the findings of studies conducted elsewhere, is a worrisome RDM aspect. Recommendations in this regard will be put forward in section 5.4.5 (RDM training).
- **Metadata:** A minority of emerging researchers (15%) always add metadata to their datasets. A third of emerging researchers never add metadata. A tiny minority of emerging researchers (2%) always make use of a metadata standard. Findings with regards to low metadata usage have serious implications, in that the advantages and

benefits of using metadata are not available to emerging CSIR researchers. It would mean that researchers would often be unable to trace and reuse their own data, trace and share data, or be able to understand secondary data created by other CSIR emerging researchers. In addition, not adding metadata would impact on being able to deposit data sets in archives or data repositories. In short: not adding metadata could have serious consequences for emerging researchers, and could also impact records management and research at the CSIR. Similarly, not abiding by a metadata standard could have similar serious research implications. Advantages of abiding to a metadata standard, including interoperability, quality of the data collection, quality of the data documentation, and enhancing of data sharing, would be benefits lost when refraining from standard adherence. Often, adhering to metadata standards is a prerequisite for collaboration and participation. Recommendations with regards to metadata usage will be put forward in section 5.4.5 (RDM training).

- **Data ownership:** The majority of emerging CSIR researchers (71%) have indicated that their data are owned by the CSIR. 6% of emerging researchers stated not knowing who owns their data. Data ownership impacts and influences data sharing, and knowledge of copyright, intellectual property rights, and accompanying legal issues, is vital. Findings also show that most of emerging researcher data are CSIR property, and should be viewed as an intellectual asset and valuable research output, often of national importance. Recommendations in this regard will be put forward in section 5.4.5 (RDM training).
- **Data confidentiality:** With regards to data confidentiality, it is seen that half of emerging researchers do not deal with confidential data. At the same time, just more than a third of research data are identified as being confidential, while 12.5% of respondents are not sure of the confidentiality of their data. Implications of findings include the need for proper treatment of confidential data, signalling the importance of training in the areas of data anonymization, data de-identification, gaining informed consent, and restricting access to data. Recommendations in this regard will be put forward in section 5.4.5 (RDM training).
- **Data sensitivity/Privacy concerns:** Obtaining informed consent was shown to be the most prevalent method of dealing with sensitive data and privacy issues. Recommendations for dealing with sensitive data will be put forward in section 5.4.5 (RDM training).
- **Data sharing:** Emerging researchers most often share their data with researchers who helped create the data, others in their group/unit, or their study supervisor/s.

Data are not shared with the general public, and hardly shared with other non-CSIR or non-collaborating researchers in the same discipline. Lack of wide sharing has the implication for researchers as sharing benefits are lost; such benefits include the incentive to produce and ensure high quality data, the promotion of research within a field, encouragement of collaboration and increased data acquisition, as well as new findings, and a reduction in the redundancy or replication of data production, saving time and money. Findings could also imply that emerging researchers are concerned about sharing their data widely, as reluctance to share outside of the CSIR environment, was observed. Recommendations relating to data sharing, especially dealing with fears and concerns when sharing widely, as well as the implementation of tools making data sharing easier, will form part of the RDM training recommendations, to be stated in section 5.4.5.

- **Data requests:** 54% of researchers had not received any requests for their data during the last five years. Responses from the remainder of the group reveal that the frequency of sharing requests is varied among emerging researchers. By implication this would mean that emerging researchers are, in general, not very experienced when it comes to data sharing. It is also implied and expected that requests for sharing would surely increase once data are archived in curated searchable data repositories.
- **Data access when sharing:** When asked for their data, emerging researchers were generally able to provide access. Question wording makes it impossible to state whether not providing access is a function of not being asked for access, or not being able to provide access. If not being able to provide access is a result of problems with data access and data sharing tools/mechanism, it stands to reason that recommendations should be made in this regard.
- **Data sharing method:** Data are most commonly shared via email, a USB device, or FTP. Making use of a cloud service was also seen to be a sharing method used by emerging researchers. Curated digital data repositories are rarely used. These findings imply that many sharing tools are used, that sharing tools are varied, and that emerging researchers seem to be keeping track of online options and trends. The under-utilization of certain sharing platforms, such as curated data repositories, indicate the need for implementation of such a platform, marketing the use of such a platform, as well as marketing and highlighting existing curated digital data repositories during training sessions or in best practice guidelines. Recommendations relating to data sharing methods will be stated in section 5.4.5 (RDM training), section 5.4.7 (Data storage) and section 5.4.9 (Data preservation).

- **Secondary data:** The majority of emerging researchers have requested data from others. The frequency of requests shows some variance. These findings imply that secondary data usage is a reality and a viable option for many emerging researchers, and that there is some awareness of the benefits of requesting data that have already been generated. While the percentage of emerging researchers not making use of secondary data is not overwhelming, the possibility exists that not being aware of the benefits of secondary data usage, or not being aware of how to search for or obtain secondary data, could be a challenge faced by some emerging researchers. Recommendations in this regard will be put forward in section 5.4.5 (RDM training).
- **Data post-publication:** After publishing their findings, emerging researchers most commonly store their data on the office computer. Digital repositories are hardly ever used. The findings show emerging researchers' data as not being preserved and curated for the long term. As a result of this, data are prone to loss, theft, corruption or format obsolescence. Recommendations in this regard, as well as steps to be taken to make the use of a curated digital repository part of good RDM practice, will be listed in section 5.4.7 (Data storage), as well as section 5.4.9 (Preservation services).
- **Other RDM tasks:** Apart from RDM tasks discussed earlier (adding metadata, backup of data), other RDM tasks frequently performed by the CSIR's emerging researchers include labelling data files, and storing data in a standard format for the field. Making an inventory of data versions, and making an inventory of location of data versions, are rarely performed. It is not clear what the reasons for non-prevalence of certain RDM tasks are; possible reasons could include not having time, not having the resources, not being aware of the importance of tasks, or not having the knowledge to perform the tasks. Recommendations around this issue will be put forward in section 5.4.3 (RDM awareness), section 5.4.5 (RDM training), as well as section 5.4.13 (RDM funding).
- **RDM training:** The overwhelming majority of emerging CSIR researchers had not undergone any RDM training. Implications are quite straightforward: emerging researchers are often not familiar with the concept of RDM, are not managing their data, and current data practices would not be adhering to best practices. Recommendations are put forward in section 5.4.5 (RDM training).
- **RDM training requirements:** Emerging researchers have indicated an interest in receiving training in the following fields:
    - developing a DMP,

- data storage,
- data documentation,
- copyright and IPR issues, and
- creating metadata for data.

Findings implicate a need for the above RDM activities, and show that there is a lack of confidence in their current understanding of these vital RDM practices. Furthermore, as could be seen in a previous paragraph ('other RDM tasks'), some RDM tasks are rarely performed. What is worrying is that results seem to indicate that not all researchers are aware of RDM benefits, or of the existence of certain RDM activities. There is not a single RDM activity included in this question that is required by all researchers; in fact, the highest required percentage indicated for a single area was found to be 58% (Developing a DMP), with the rest of RDM areas included in the question scoring even lower.

A subset of emerging researchers was shown to not be interested in receiving RDM training. While the possibility exists that the emerging researchers not interested in data training might actually be adhering to best practices, it is more likely that disinterest is a result of RDM regarded as unimportant activity, or too demanding of valuable research time to be invested in fully. Recommendations as to possible steps in overcoming emerging researcher RDM negativity, and making emerging researchers more aware of the benefits of RDM, will form part of section 5.4.5 (RDM training). Recommendations with regards to these all stated training requirements can be found in section 5.4.5 (RDM training), section 5.4.4 (Data management plans), as well as section 5.4.7 (Data storage).

- **Regard for data services:**

Emerging researchers regard the following RDM services as being very important:

- having access to data in their discipline,
- having the infrastructure at the CSIR to allow CSIR data to be preserved,
- having teaching and learning materials so that researchers can work with data,
- having CSIR data repositories where researchers can deposit data,
- having user-friendly tools to support RDM, and
- having a data citation impact factor, or recognition, when data are cited.

Emerging researchers regard the following RDM services as less important than others:

- Having services necessary to assign a permanent DOI to their data

A subset of emerging researchers has stated not being familiar with the following RDM services and practices:

- data curation expert,
- a data citation impact factor/recognition,
- funders' timelines for depositing data, and
- a CSIR data ethics policy.

These findings mainly indicate the need for RDM training, as well as the importance of RDM awareness and marketing. It is clear from the findings that although there is some level of awareness as to what services are required to assist researchers in managing their data, more awareness is needed. In particular, it is hoped that awareness training would provide emerging researchers with a more informed perspective regarding RDM services, and that they will be informed of the nature and benefits of RDM services seen to be not understood by them, or not regarded as important. These recommendations will be put forward in section 5.4.5 (RDM training), and section 5.4.3 (RDM marketing and awareness).

5.2.1.5 Sub-question 5: What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?

RDM-related challenges and concerns were investigated in this survey, with information gathered via an open-ended survey question requesting respondents to mention any RDM concerns. Although this question was only completed by a tiny minority of respondents, findings revealed that lack of sufficient data storage space, lack of RDM funding, and network inefficiency were areas of concern.

It could also be argued that findings from several of the survey's other questions, in particular the findings relating to RDM training requirements, as well as findings relating to perceptions of importance of RDM services and practices/policies, are proof of additional areas of concern. For example, when an RDM service such as 'having the infrastructure at the CSIR to allow CSIR data to be preserved' is found to be regarded as an RDM service seen as very important, and it is currently not an RDM service provided, the deduction can be made that the lack of such an RDM service currently, is a challenge facing emerging researchers. Thus, given the fact that none of the RDM services or RDM policies/practices rated by the respondents are currently available at the CSIR, and all of these RDM aspects rated as 'important' or 'very important', these could be flagged as being RDM issues of concern to emerging researchers.

This researcher has taken into account the following information when trying to establish RDM concerns and challenges: the concerns explicitly mentioned in an earlier paragraph, lack of RDM training received, and perceived high levels of importance of certain RDM

services as well as policies/practices. Based on obtained information, the following issues are put forward as RDM concerns and challenges:

- insufficient research data storing space,
- lack of RDM funding,
- network inefficiency,
- lack of RDM training,
- lack of access to data in their discipline, as well as access to cross-disciplinary data,
- lack of guidelines, services and tools relating to RDM, data deposit, and metadata,
- lack of DOI services,
- lack of a data preservation infrastructure, and
- lack of a CSIR data repository.

These findings show that researchers experience several barriers to performing research data management tasks. The concerns listed here cannot be ignored by either this researcher or the CSIR research support environment: without adequate funding, storage space and network capability, CSIR RDM will battle to take off and will fail to prosper. Recommendations addressing listed issues will be put forward in several sections, namely 5.4.5 (RDM training), 5.4.7 (Data storage), 5.4.9 (Data preservation), 5.4.10 (DOI), as well as 5.4.13 (Other RDM stakeholders).

5.2.2 Research question, findings and implications

As stated in Chapter 1, the main research question of this study is the following:

How can an organisation like the CSIR ensure that emerging researchers apply best practices when managing the CSIR's research data?

Section 5.4 attempts at answering this main research question by putting forward several recommendations. In summary, the majority of the RDM deviations observed between the CSIR's emerging researchers and worldwide trends/best practices, could be addressed with the following:

- RDM becoming a mandatory part of the CSIR research process and viewed as a vital research activity: policies and procedures used by researchers should address RDM, or be RDM-specific,
- creation and implementation of a user-friendly online DMP tool,
- RDM training/guidelines: researchers are in need of training, areas such as data ethics, completing a DMP, how to secure data, reasons for data sharing, how to

share data, and using metadata should be included in training sessions or in guidance documents,

- CSIR RDM infrastructure to be established; data storage options, data preservation infrastructure, sharing tools, and indexing architecture to be amended to assist and enhance RDM practices,
- RDM to be marketed and RDM awareness to be boosted: an RDM blog, regular intranet articles, CSIRIS frontline involvement and unit-specific RDM teams are needed, and
- the appointment of a data librarian, tasked with realistic, specific and reachable RDM targets. The formation and assistance of an RDM team in each research unit, RDM-trained information specialists, as well as a CSIR RDM working group, would enhance the quality and speed of implementing RDM services at the CSIR.

While some of these suggested steps have already been implemented or are in progress, others need to be investigated and implemented from scratch. It needs to be stressed that adherence to RDM best practices is, as has been shown in previous studies, dependent on institutional support. The hope is expressed that RDM infrastructures, tools and services be put in place with a view to improved RDM practices, and as a result of this, more efficient research.

5.3 Conclusions reached

This study's objective was to advise how to improve the current RDM practices at the CSIR in order to be aligned with international trends and practices. Furthermore, this researcher was also interested in sub-objectives comprising the establishing of international RDM trends as well as CSIR RDM practices, and comparing emerging CSIR RDM behaviours with RDM behaviours elsewhere. By answering several research questions, this study was able to show findings enabling the study objectives to be achieved.

Knowledge of international RDM trends as well as CSIR RDM practices have been obtained, making possible the comparison between the relevant groups. It could be stated that in general, RDM practices performed by CSIR emerging researchers, CSIR experienced researchers as well as researchers elsewhere, could be improved upon. Although it was found that some RDM differences do exist between CSIR emerging researchers and CSIR experienced researchers, and between CSIR emerging researchers and researchers elsewhere on the globe, the deviations are neither excessive nor cause for alarm.

While RDM practices among CSIR emerging researchers, in general, are in need of amendment, current shortcomings are not entirely unexpected. RDM is a fairly new

discipline in South Africa, and the CSIR currently not aggressively marketing a RDM policy, not having an RDM procedure, not providing RDM training, and having a virtually non-existent RDM infrastructure, is not an unanticipated state of affairs. As a result of these shortcomings, CSIR researchers are not likely to exhibit or portray RDM best practices.

In an effort to address current deficiencies existing in CSIR RDM practices, a set of recommendations has been put forward. These recommendations are aimed at improving RDM infrastructure, RDM awareness, and RDM skills. Recommendations will be discussed fully in the next section.

5.4 Recommendations

Based on the results of this study (Chapter 4: Results and discussion), the summaries in section 5.2, and the conclusions reached in section 5.3, several recommendations can be made. Recommendations are made in order to address RDM practices of emerging CSIR researchers showing deviance from international trends, as well as RDM practices not deviating, but still in need of improvement. These recommendations will be categorised, and will address a wide range of RDM activities, services and tools. It is the intention of this researcher to make recommendations pertaining to the following CSIR RDM-related issues:

- assist with the implementation of an RDM-regime at the CSIR, including the adherence to the data management section of the CSIR Records Management procedure, and adherence to CSIR RDM procedures,
- indicate the need for RDM marketing, and an increase in RDM awareness,
- provide CSIR researchers with guidelines, tools and services with which to manage their data, in particular guidance relating to the creation of a DMP, data storage, data preservation and DOI services. Indexing of datasets will also be included,
- indicate measures by which adherence to ethical data treatment is ensured,
- stipulate the role of the data librarian and said position's contribution and responsibilities,
- indicate the role and required contributions of all CSIR RDM stakeholders, and
- indicate which aspects of this study need to be elaborated on, or studied in more detail, in further studies.

Bearing these intentions in mind, and coupled with her knowledge of and experience with current CSIR systems and services, the following recommendations are made:

5.4.1 CSIR management and RDM

The first recommendation entails gaining CSIR managerial approval for, and support of an RDM 'regime' at the CSIR. Without this being achieved, RDM as a part of good research practice will not realise and most recommendations put forward below will fail to be achieved or even implemented.

It is therefore suggested that a concerted effort be made to involve and convince CSIR research management of the importance of RDM in a research institution such as the CSIR. The need for RDM services at the CSIR, the benefits of RDM, and a plan of envisaged steps that need to be followed to establish RDM at the CSIR, need to be conveyed to CSIR research management and relevant parties involved at institutional policy level. Steps needed to get RDM established at the CSIR show great overlap with the recommendations put forward in the remainder of this section.

5.4.2 RDM policy and RDM procedure

A CSIR Records Management Policy is already in existence; this policy, having a concise section on data management, will suffice for RDM purposes too. The current Records Management Policy (CSIR, 2011), which views research data as a research record, addresses RDM by stating that managing, storing and retaining data for specified periods of time forms a CSIR responsibility. While it might be stated that the RDM coverage of the policy is rather brief and vague, a further recommendation – the development of an RDM procedure – is a next step. This RDM procedure would be a detailed document and would address specific aspects such as the RDM framework in the CSIR, RDM workflow and steps to be followed by all stakeholders when managing data, meeting the requirements of different funders, the role of ethics, legal issues, and guidance on issues mentioned. A further characteristic of the RDM procedure would be its inclusiveness: discipline-specific RDM aspects, as well as funder-specific specific requirements, would need to be included. So while the CSIR policy is succinct, the CSIR RDM procedure would be comprehensive document taking into account the roles of all stakeholders including researchers, their managers, supports staff, and funders.

This researcher, in her capacity as CSIR data librarian, and based on conclusions reached via this study, has compiled a draft-procedure that has yet to be ratified. After approval and ratification, the formal procedure would need to be brought to the attention of relevant parties, and demonstrated to be part of the research process. Parties would include all researchers, research managers, the CSIR Research Ethics Committee, and parties indirectly involved with research outputs, such as indexers and information specialists.

It is foreseen that the approved RDM procedure would be available electronically on the CSIR intraweb.

5.4.3 Marketing and awareness

In principle it needs to be acknowledged that RDM as an important part of the research process which needs to be promoted and marketed CSIR-wide.

- A first step to be taken in this regard could be an RDM awareness session.
    - This event should be open to all interested CSIR staff and should address issues such as the mandate and drivers of RDM, benefits of RDM, and what RDM entails.
    - Furthermore, representatives from the CSIR library and information services could report on the current RDM status within the organisation.
    - A representative of the Data Intensive Research Initiative for South Africa (DIRISA), as responsible authority for storing datasets of national importance, could explain their role and how it benefits and relates to CSIR research data storage. In addition, the DIRISA spokesperson also should provide detail regarding the progress made with the creation and implementation of an online DMP. The current state of affairs relating to the development of an online DOI assignment facility could also be explained.
    - It is recommended that a representative of the CSIR Research Ethics Committee (REC) is invited to participate in the session.
- A second important marketing/awareness recommendation entails the CSIR RDM team embarking on a series of informal round-table discussions with each of the CSIR research units.
    - At the time of writing (August 2016), this RDM team consisted of the portfolio manager of CSIR Library and Information Services, the CSIR Research and Archives Records Specialist, and the CSIR data librarian.
    - Unit representatives for recommended unit discussions would ideally include the strategic research manager, research group leaders, and other senior staff involved in the unit's research decisions.
    - Issues to be addressed during these sessions would include unit feedback with regards the policy draft mentioned earlier. Suggested changes or adaptations to the document, taking into account unit- or discipline-specific data requirements, would be debated.

- In addition to discussing the RDM policy draft document, input from units pertaining to the online DMP tool mentioned at the awareness session, as well as any RDM requirements, would be vital.

- It is recommended that additional future awareness steps include the following events and activities:
  - an up-to-date RDM blog on the CSIR intraweb,
  - contents of the RDM blog should include the following:
    - a section explaining the drivers, need and advantages of RDM; aimed especially at those who are not familiar with the practice,
    - links to the CSIR records management policy, and RDM procedure (procedure to be finalised in due course),
    - a research data glossary,
    - links to the online DMP, completed DMP examples, DMP guidelines,
    - step-by-step explanation on dataset indexing,
    - best practice guidelines: data formats, data storage, adding metadata, backups, data security, obtaining a DOI, data ownership issues,
    - data sharing: benefits, making data shareable, sharing tools, and obtaining and evaluating secondary data,
    - a link to online training tools,
    - reports on both CSIR RDM studies conducted,
    - links to RDM-related sites of importance; NeDICC as well as the DCC spring to mind, and
    - funder requirements.
  - regular intraweb articles informing CSIR researchers of RDM training opportunities and webinars, RDM developments and new RDM tools, and
  - information specialists to assist in marketing RDM-related events.

5.4.4 Data management plan (DMP)

A DMP, as revealed in section 2.4.9 (Data management plans), forms an important part of the RDM process**.** The findings of this study (see section 4.8: Development of a research data management plan) revealed the use of DMPs in the CSIR to be uncommon, non-mandatory and subject to individual whim. To address these RDM shortcomings, the following recommendations were put forward:

- The completion and submission of a DMP is identified and recognised as being part of the research process. Ideally, it should be submitted at the start of every research project, and form part of the research proposal.
- Completion of a DMP when the project proposal is submitted, should be mandatory.
- An online DMP tool is to be developed, tested, refined and the output of the tool added to the list of all documents to be submitted when a research proposal is made. Important developments in this regard have already begun: DIRISA has made available an online DMP facility featuring several preliminary DMP templates which need to be stress-tested. In addition, data librarians or data workers are also invited to submit institution-specific templates to be added to the website. This project is still in its infancy and future planned steps include institutional branding on DMPs, alerts to notify institutions of support needs, and introduction of a plan lifecycle. It is therefore recommended that the CSIR should fast track the process by collaborating with DIRISA
- It is recommended that the DIRISA-hosted online DMP tool be customizable as to allow for the requirements and preferences of different funders and/or research units/research projects.
- It is recommended that a DMP tool be user-friendly, and a completed DMP limited to a maximum length of two A4 pages. An intricate DMP, or a DMP template resulting in the plan being very long, cumbersome and time-consuming, is not advisable.
- It is recommended that completed DMPs be submitted to the relevant research project parties, as well as to funders, the CSIR Research Ethics Committee, and indexing staff at CSIR Library and Information Services. Where there are data storage concerns (i.e. dataset described in DMP would require additional storage space), it is recommended that CSIR ICT also be involved in the DMP submission process.
- The complete DMP should be added to the project file.
- Accompanying DMP activities, such as marketing and training, need to form part of DMP implementation too. These activities will be discussed in the applicable segment of this recommendations section.

5.4.5 RDM training and guidance

As portrayed and discussed in section 4.26 (Research data management training), lack of RDM training was a reality experienced by the majority of the CSIR's emerging researchers. Furthermore, as seen in in section 4.27 (Research data management training: areas of interest), emerging researchers indicated an interest in receiving training in many RDM

activities. Based on these findings, the following RDM-related training recommendations were put forward:

- RDM online training materials to be developed, marketed and added to the CSIR's RDM blog.
- Available online training tools to be marketed and advertised, and links added to the RDM blog. MANTRA, the RDM tool developed by the University of Edinburgh, is a good example (http://datalib.edina.ac.uk/mantra/).
- It is recommended that a session be held whereby relevant parties discuss training formats other than the online variety. Individual training, group training, training on demand, and the possibility of training forming part of the on-boarding sessions, need to be discussed.
- It needs to be established which parties would be responsible for providing RDM training. Options are manifold, and include the senior research staff/project leaders, information specialists, or a designated CSIR RDM trainer. Combinations of training providers, or training being only provided in online format, should be considered too. Workshop-based training in collaborative format (e.g. a NeDICC effort), is another possibility.
- It is recommended that once RDM training tools are set up and ready for release, training sessions or online training tools be attended/used by all CSIR research staff. In addition, it is recommended that RDM training also be undergone by CSIR information specialists, as dealing with indexing-related queries forms part of their duties.
- It is recommended that RDM practices included in RDM training include:
    - benefits of RDM,
    - what RDM entails,
    - role of the DMP, DMP guidelines, DMP examples, DMP workflow in the CSIR,
    - best practices with regards to formats, storage, backups, metadata, data sharing, and preservation, to name but a few RDM practices that would be covered, and
    - a reiteration of the RDM activities and contents of the RDM blog.
- The recommendation is made that the CSIR's Good Research Guide, published in 2003, be updated and that research data practice be included in the guide. An

updated guide should be available electronically and be marketed during the CSIR's on-boarding[5] session.

5.4.6 RDM and CSIR indexing

- It is recommended that RDM form part of the CSIR indexing process; datasets linked to CSIR publications should be indexed. This would mean that metadata pertaining to datasets, as is the case with published scientific output, need to be added to the CSIR's Technical Outputs Database (TOdB). It is also envisaged that metadata pertaining to datasets be added to ResearchSpace, the CSIR's institutional repository. Should certain criteria pertaining to confidentiality, sensitivity, as well as dataset size be met, the dataset itself would also be uploaded to ResearchSpace.
- Implementation of this recommendation would entail making changes to the current indexing workflow procedure. Steps to index datasets need to be detailed in RDM training guidelines, and fully grasped by CSIR information specialists. Also, Oracle Workflow, the workflow management system used in the CSIR when submitting publications for indexing, need to be customised by CSIR ICT in order to cater for datasets indexing. A first step in this regard has been made with an email request sent to CSIR ICT, requesting a few basic changes be made to the current online indexing submission form. These requested changes, once implemented, would enable the researcher to indicate that a dataset is related to an indexed publication. It would also provide the researcher with an online form when, once completed and electronically submitted by him/her, would contain metadata related to the dataset. This metadata is required by the indexer when indexing onto the CSIR's outputs database as well as to the institutional repository.
- It is recommended that dataset indexing include all datasets currently being created. Datasets could also be indexed retrospectively, especially with datasets identified to be of national significance.
- It is recommended that relevant parties (CSIR management, CSIR library management, CSIR Research Ethics Committee) incorporate into RDM procedural guidelines the conditions making datasets eligible for publication onto the institutional repository.
- A decision also needs to be made regarding the level of indexing, and which metadata standards would need to be adhered to when indexing datasets.

[5] A day-long session for new employees, where institutional support services such as Library and Information Services, Finance, and Human Resources inform new staff members of services available to them

5.4.7 Data storage

As was shown in section 4.10 (Data storage location), 4.12 (Research data backup: location), 4.23 (Data storage after publication), as well as 4.28 (Importance of RDM-related standards, policies, principles and practices), data storage concerns shared by CSIR emerging researchers are issues which cannot be ignored. As such, the following storage-related recommendations are made:

- ResearchSpace, the institutional repository of the CSIR, to be used for small datasets (smaller than 2 megabytes in size), meeting confidentiality/sensitivity requirements.
- CSIR ICT to be consulted and involved with regards to data storage provision should a DMP indicate that data storage needs cannot be met. The precise nature of involvement will be stipulated in due course, during the RDM data procedure finalization stages.
- The CSIR will need to take data storage into consideration for all approved projects. DIRISA, as a repository of last resort, will only consider data sets that are of national importance.
- Adding on to the previous point: it is recommended that the CSIR assesses what proportion of datasets are of national importance, and on what criteria.
- The role of DIRISA as it pertains to data storage provision for CSIR researchers, will form part of the CSIR RDM awareness session (discussed earlier). Information relating to DIRISA as storage provider will need to form a permanent part of CSIR RDM training, blog guidelines, and marketing strategy, with updates made as the need arises.
- Another recommendation relating to data storage involves the creation of an RDM blog document explaining data storage guidelines/best practices (discussed earlier).
- Storage-related RDM training and guidelines need to consider that emerging researchers make use of a wide range of data types and data volumes. A one-size-fits-all approach when creating best practices guidelines will not suffice. An example of discipline-specific practices would be data preservation, requiring researchers in different disciplines to convert their working data to formats with a longevity record.

5.4.8 RDM librarian and RDM working groups

In an ideal world with unlimited resources, each of the CSIR research units would have access to a data manager in possession of an advanced post-graduate qualification suited to the research unit. This candidate would be able to clean, anonymise and curate the unit's

data, and provide RDM training. In addition, the candidate would have a thorough understanding of the unit's RDM requirements as well as of the data created, software tools used, and metadata standards most applicable to the unit's data. However, in a resource-constrained environment such as the CSIR, the following recommendations are made:

- A suitable candidate be identified, and appointed as CSIR data librarian/data curator/data administrator.
- A candidate would ideally be in possession of a master's degree in an SET-related field. However, taking into account that some research units at the CSIR also deal with qualitative data more related to the humanities (i.e. interview data, data generated via focus groups), the holding of a master's degree in the humanities could be seen to suffice.
- It is foreseen that apart from academic qualifications, such an individual also portray the necessary qualities to create training guidelines, assist in creating an RDM procedure, and provide training. Furthermore, it is expected that this candidate display traits required for dataset indexing, as this activity would be a big part of the candidate's daily activities.
- At the time of writing, a CSIR employee, in possession of a master's degree in humanities, as well as a library qualification, had been appointed as a data librarian. The data librarian formed part of CSIR Library and Information services (CSIRIS), and had previously been involved in indexing onto the CSIR's Technical Outputs Database, as well as to the CSIR's institutional repository.
- With RDM training for researchers not common in South Africa, and training in becoming a data curator/data manager/data librarian equally uncommon, it is recommended that suitable training opportunities be investigated and actively pursued. At the time of writing, the appointed data librarian had already attended various RDM workshops, joined an RDM community of practice, and embarked on an informal training programme involving available online RDM training courses, RDM-related reading matter, RDM webinars and subscribing to RDM alerts and email listservs. While it recommended that a data librarian be in possession of a recognised data management qualification, such a SET-based qualification is not yet obtainable in South Africa.
- It is recommended that the data librarian's RDM training be an ongoing activity. The hope is expressed that a formal RDM qualification be available in the future and that the incumbent, or her successor, obtain stated qualification.
- It is recommended that due to workload and the novelty of data manager/data librarian as a professional position, the data librarian be assisted by a CSIR RDM

working group. At the time of writing (June 2016), the data librarian was assisted in her duties by the portfolio manager of CSIRIS, as well as the CSIRIS Research and Archives Record Specialist.

- It is recommended that due to workload and the novelty of data manager/data librarian as a professional position, the data librarian further be supported in her duties by the CSIR's information specialists. Such support would be vital should the CSIR data librarian not be in possession of a SET-based qualification, and not familiar with discipline-specific research data formats, software, and metadata standards, to name but a few RDM aspects. It is envisaged that after a training period, information specialists responsible for providing subject specific leading-edge information services to unit researchers, be able to provide unit researchers with RDM assistance. Assistance would involve helping with DMP completion, helping with submitting metadata required for dataset indexing, helping with DOI assignment requests, and answering basic questions around RDM activities.
- Ideally, it would be advisable for each of the CSIR's research units to appoint two or three persons to a discipline-specific RDM team or working group. It is recommended that members of the working group be familiar with characteristics of data created in their unit, and and be This team would be able to provide assistance and insight when RDM procedures are being drawn up, and also be the first point of contact when future RDM decisions, affecting each research discipline, are being made. In addition, RDM representatives from each of these disciplines could also be called upon to take part in CSIR-wide RDM meetings, discussions, or brain-storming sessions.
- It is foreseen that as soon as possible after the establishment of the working group, a briefing session would be held, outlining the tasks of group members to be performed. Expectation will be stated, and concerns will be addressed.
- Following the briefing session, the data librarian in consultation with immediate managers to create a 6-month work schedule for RDM working group members. It is suspected that the schedule will consist of several single assignments, each to be completed within the next three weeks. Examples of assignments would be:
    - Critiquing the draft RDM procedure and providing feedback
    - Supplying information on data formats and software tools used within their unit/research group/discipline
    - Supplying information on RDM requirements as stipulated by their funders; providing funder DMP templates, providing links to funder RDM requirements

    - Investigating and supplying feedback on metadata standards used in their discipline
    - Investigating and supplying feedback on accredited repositories used in their discipline
- As stated earlier, tasks would each be given a due date for completion.
- Feedback would be evaluated and consolidated into the RDM draft procedure.
- Of prime importance are the facts that membership of the working group is voluntary, and that CSIR researchers are all full-time employees with heavy research loads. As such, meetings would be held to a minimum, virtual interaction would be preferred as far as possible, and non-submission of feedback would be seen as 'par for the course'.
- It is recommended that an RDM working group become a permanent feature in the CSIR RDM environment. It is expected that the group's composition might change due to workload, resignations, transfers and other circumstances. It is however recommended that even after formalisation of the RDM procedure, and its implementation into the CSIR research environment, the RDM working group be kept in existence. It is foreseen that procedural updates and revisions, changes in funder requirements and other changes in the CSIR research sphere would necessitate and benefit from continual input from working group members.

5.4.9 Preservation services

Preservation-related findings show that data are being stored in various locations (4.10, Data storage location), many software applications are being used to access data (4.7, Software application used), curated digital data repositories are not being used (4.21, Data sharing methods), and data preservation activities not being performed (4.23, Data storage after publication). Furthermore, this researcher could not find proof that the CSIR currently has either a data preservation infrastructure or data preservation procedures. As a result of this, the following recommendations with regards to data preservation are made:

- data preservation to be included in the planned RDM policy and procedure,
- procedural information to include details on infrastructures to be used for long term archival purposes,
- adding to the above, RDM procedure to detail the conditions around the preferred use of ResearchSpace, DIRISA storage and other digital data repositories,
- influencing DIRISA to clarify and draft a policy and procedures around the topic of 'datasets of national importance', and of their role and involvement in such cases, and

- a list of curated digital data repositories to be drafted, placed on RDM blog, and made part of awareness sessions.

5.4.10 Data Citation/DOI

A Digital Object Identifier (DOI) is not issued to a dataset on an ad-hoc or unmanaged basis; it is assigned, by authorised agencies or institutions, to datasets that are well described and managed for long-term access. Assigning a Digital Object Identifier (DOI) to a dataset therefore indicates that the dataset will be well managed and accessible for long-term use. Benefits of an assigned DOI include the dataset being easier accessible on the internet, and the increase in acceptance of the dataset as a citable contribution. It is also expected that CSIR researchers would share data more readily if data would be properly cited; a scenario more likely should a dataset possess a DOI. Similarly, sharing citable datasets with DOIs brings the data-creating researcher a step closer to being rewarded for data citation, in that DOIs form part of a scholarly structure that in future could recognise and reward data producers.

With these benefits in mind, it is recommended that the CSIR RDM procedure incorporates assigning DOIs for datasets meeting sharing requirements.

As mentioned earlier under the heading ‘Awareness and Marketing’, DIRISA has already begun with a DOI trial project, in which the CSIR is one of the project participants. This trial project is expected to run for the duration of the 2016/2017 financial year, after which the online DOI assignation tool will be implemented and made available to all dataset creators, dataset curators, and similar.

5.4.11 Research data and ethics

Currently, all CSIR researchers are required to apply for ethical clearance when embarking on a research project. In light of this fact, and the difficulty this researcher had in obtaining ethical clearance for this study, the following recommendations are made:

- CSIR Research Ethics Committee to stipulate and explain the steps, activities and RDM requirements needed for obtaining ethical approval,
- these steps need to be discussed at the CSIR RDM awareness session (discussed earlier) as well as stipulated in a guiding document to be accessible either on the CSIR RDM blog, or via a link to the CSIR Research Ethics webpage, and
- ethics procedure to include details on how applicants should proceed when ethical approval from more than one body (for the same research project) needs to be

obtained. As was experienced by this researcher, obtaining ethical approval from both the CSIR Research Ethics Committee as well as the university she was registered at, proved to be a difficult task. In this instance, both authorities requested proof of ethical approval from the other before giving ethical approval. It is suggested that a procedure describing the required steps for dual ethical approval be devised, thereby eliminating time-consuming deliberation between the applicant and all ethics bodies involved.

5.4.12 Funder requirements

Funder-related findings have shown that the majority of emerging CSIR researchers are not aware of funder requirements with regard to RDM (4.9, Awareness of policy/funder requirements regarding RDM), and that data is rarely shared with funders (4.18, Data sharing: sharing parties). In addition, a quarter of respondents have indicated that they require training in understanding funder requirements (4.26, RDM training: areas of interest). Furthermore, the majority of respondents have shown that stated timelines by funders for data is an important RDM practice (4.28, Importance of RDM-related standards, policies, principles and practices).

As a result of these shortcoming and perceptions, the following recommendations are made:

- funder requirements to be given its fair share of prominence when creating RDM best practice guidelines or training material,
- funder requirements should feature on the RDM blog,
- funder requirements should feature on the online DMP tool,
- once the DIRISA online DMP tool has been implemented, it is vital that researchers are made aware of its presence, and the benefits that its effective use provide, and
- the online DMP should be customisable and able to cater for all funders of CSIR research.

5.4.13 Other stakeholders: ICT and RDM funding

RDM concerns and challenges discussed earlier in 4.29 (Additional RDM concerns), as well as section 5.2.1.5. (Sub-question 5: What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?), have revealed that emerging researchers experience several barriers when it comes to applying RDM best practices. Recommendations addressing these issues, and not fitting into any of the previous recommendations sections, are the following:

- CSIR ICT Services: this CSIR division should be informed of extreme data storage requirements as soon as such a requirement has been indicated on submitted RDM plans. Although DIRISA would also be able to provide storage, it is at this stage not clear to this researcher what the relationship between DIRISA and CSIR ICT is, and who would best assist with storage of large datasets. This relationship should be clarified.
- CSIR ICT services: It is recommended that mentioned slow network speed as well slow data transfer rates be investigated further, either by this or another researcher.
- RDM funding: it is recommended that the concerns about RDM funding be addressed in at least the following manner:
    - RDM best practices be viewed and accepted as an activity requiring funding,
    - funding proposal as well as DMP to include a section on expected RDM-related expenditure, and
    - a document be drafted giving clarifying information on RDM costs, storage costs, and other RDM expenditures.

5.4.14 Further studies

When looking at possible further research emanating from this study, this researcher identified two types of studies, namely small studies, and advanced academic research.

Smaller studies would investigate aspects such as discipline-specific RDM trends, and would attempt at establishing inter-disciplinary RDM differences. Such a study, for example, would look at disciplinary differences in RDM data held, RDM practices executed, RDM services required and RDM concerns expressed. Furthermore, such a study need not be, as is the case with this study, formal in nature, but could form part of discussions mentioned earlier in in 5.4.3 (Marketing and awareness). Unit-specific roundtable RDM discussions, and in essence discipline-specific RDM sessions, would provide insight into the unique or non-standard/non-generic RDM practices and requirements expressed by different researchers, disciplines or units. Taking cognisance of these differences is vital when finalizing the data management part of the CSIR Records Management Policy, CSIR RDM procedures, the online DMP tool, and storage requirements, to name but a few aspects.

A second type of research stemming from this study would be more formal in nature, and be said to constitute ‘real research’. This researcher foresees the establishing of RDM services at a South African research institution, and implementing its accompanying policies, procedures, tools and infrastructure as an obvious next step, and a theme worthy of a topic for doctoral level studies.

5.5 Study conclusion

This study aimed at establishing the current RDM practices of the CSIR's emerging researchers, in an attempt to put forward suggestions and recommendations on how to address and improve the CSIR RDM status quo. In order to ascertain the current RDM practices of emerging researchers, and identify areas lacking, it was necessary to study not only the RDM practices of the target population, but also to delve into their RDM expectations, perceptions and concerns. Furthermore, in order to get a view of the global RDM situation, a glimpse into the RDM practices of researchers elsewhere, as well as experienced CSIR researchers, was necessary.

Findings revealed that although RDM practices performed by CSIR emerging researchers are in need of improvement, not much deviation from global practices could be found. In order to address deficiencies in emerging researchers' RDM practices, this study has put forward a set of recommendations aimed at assisting CSIR researchers in applying best practices when managing their data.

To recapitulate: the main objective of this study was to improve the current RDM practice at the CSIR in order to be aligned with international trends and practices. This objective, as well as several sub-objectives, can be seen as attainments that were reached via the investigation of current RDM practices at the CSIR, analysing RDM practices globally, and using these findings as a stepping stone for putting forward recommendations aimed at eventually improving CSIR RDM practices.

Several research questions needed to be answered before the objectives could be attained. This study reported on:

- the RDM practices of all researchers globally as well as at the CSIR,
- the current international RDM trends, and
- the data held, RDM practices, and RDM concerns of CSIR emerging researchers.

By doing so, this researcher was able to put forward several recommendations on how to ensure that future researchers at an organisation such as the CSIR, apply best practices when managing research data.

In general, it was found that stated RDM practices of the CSIR's emerging researchers do not deviate from the global state of affairs, in that the global situation also revealed traces of less-than-ideal RDM practices. Recommendations were put forward by this researcher in an effort to address current RDM deficiencies at the CSIR. The creation of and adherence to an

RDM policy and RDM procedures, RDM awareness and marketing, RDM training, an online DMP tool, and including datasets in the CSIR indexing workflow are some of the suggestions put forward as mechanisms to improve the quality of RDM at the CSIR.

This study was successful in demonstrating that although in need of improvement in terms of portraying ideal RDM habits, the current situation at the CSIR is far removed from a doom-and-gloom scenario. Some RDM practices, for example the percentage of research data backed up, reveal good practice adherence even at this moment, prior to the implementation of RDM procedures. It is hoped that the implementation of recommended services, policies, practices and tools will provide the impetus needed to put into effect RDM best practices for the majority of RDM activities to be performed by researchers.

Study limitations could be identified, and centre mainly around the absence of discipline-specific requirements, the absence of inter-disciplinary investigations, and the possible ambiguity of certain survey questions. Adding on to this, it can be seen that several possibilities for future studies are revealed through this study's findings. Specific RDM requirements of different study disciplines, addressing the RDM requirements of multi-disciplinary research, and establishing RDM services at institutions dealing with widely-differing RDM requirements, are a few topics which come to mind. This researcher also foresees a follow-up study emanating from the recommendations put forward in 5.4; a study revolving around the establishing of an RDM service at a South African research institution, and implementing its accompanying services, policies, tools and infrastructures is an obvious next step.

The current study fills a gap in current knowledge, in that it is the first of its kind in South Africa, aiming to establish what emerging researchers are doing with their research data, identifying deficiencies in practices, and based on this knowledge, coming forward with ways to improve research data management practices. Its contribution therefore lies in the fact that study findings, as well as recommendations, provide insight into this previously non-investigated aspect of research behaviour. Following on from that, this study also shares information regarding the manner in which a leading science institute will address inadequacies in RDM habits, or deal with RDM concerns and challenges expressed. It is hoped that the findings of this study will not only serve as justification and motivation for CSIR RDM improvement and enhancement, but also encourage the undertaking of similar studies at comparable research institutions. Future collaborations with investigators or research institutions, pertaining to RDM studies, RDM infrastructure or RDM services (e.g. workshops and training sessions), are additional envisaged outcomes.

REFERENCES

AKERS. K.G. & DOTY, J. 2012. Differences among faculty ranks in views on research data management. *IASSIST Quarterly*, 36(2):16-20. [Online]. Available from: http://www.iassistdata.org/sites/default/files/iq/iqvol36_2_doty.pdf [Accessed 15 July 2014]

AKERS, K.G. & DOTY, J. 2013. Disciplinary differences in faculty research data management practices and perspectives. *The International Journal of Data Curation*, 8(2):5-26. [Online]. Available from: http://www.ijdc.net/index.php/ijdc/article/viewFile/8.2.5/332 [Accessed 18 July 2014]

ALEXOGIANNOPOULOS, E., MCKENNYE, S. & PICKTON, M. 2010. *Research data management project: a DAF investigation of research data management practices at the University of Northampton*. [Online]. Available from: http://nectar.northampton.ac.uk/2736/ [Accessed 9 July 2014]

AUSTRALIAN BUREAU OF STATISTICS. 2014. *Statistical language – Census and Sample*. [Online]. Available from: http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+census+and+sample Accessed 29 January 2015

AUSTRALIAN NATIONAL DATA SERVICE. 2014*. Funders guidelines.* [Online]. Available: http://ands.org.au/working-with-data/data-management/funders-guidelines [Accessed 30 March 2016]

AVERKAMP, S., GU, X. & ROGERS, B. 2014. *Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs*. 34pp. http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1246&context=lib_pubs

BALL, A. 2010. *Review of the state of the art of the digital curation of research data*. [Online]. Available from: http://opus.bath.ac.uk/18774/2/erim1rep091103ab11.pdf [Accessed 8 July 2014]

BARDYN, T.P., RESNICK, T. & CAMINA, S.K. 2012. Translational researchers' perceptions of data management practices and data curation needs: findings from a focus group in an academic health sciences library. *Journal of Web Librarianship*, 6(4):274-287. [Online]. Available from: http://www.tandfonline.com/doi/pdf/10.1080/19322909.2012.730375 [Accessed 24 October 2014]

BEILE, P. 2014. *The UCF Research Data Management Survey: A report of faculty practices and needs*. [Online]. Available from: http://www.ist.ucf.edu/hpc/rcd/Beile_rcd2014.pdf [Accessed 8 July 2014]

BEZUIDENHOUT, R. 2014. *Unisa Data Curation*. PowerPoint presentation, Unisa. Pretoria, South Africa. Unpublished presentation.

BEZUIDENHOUT, R. & MACANDA, M. 2014. *Research Data Management.* [Online]. Available from: http://eprints.rclis.org/22717/ [Accessed 20 October 2014]

BORGMAN. C. WALLIS, J.C. & ENYEDY, N. 2006. Building digital libraries for scientific data: an exploratory study of data practices in habitat ecology. *Lecture Notes in Computer Science*, 4172:170-183. [Online]. Available from: http://link.springer.com/chapter/10.1007%2F11863878_15#page-1 [Accessed 8 July 2014]

BLUMENTHAL, D., CAMPBELL, E.G., GOKHALE, M., YUCEL, R., CLARRIDGE, B., HILGARTNER, S. & HOLTZMAN, N.A. 2006. Data withholding in genetics and other life sciences: prevalences and predictors*. Academic Medicine*, 81(2):137-145. [Online]. Available from: http://journals.lww.com/academicmedicine/Fulltext/2006/02000/Data_Withholding_in_Genetics_and_the_Other_Life.6.aspx [Accessed 8 July 2014)

BRADBURY, S.J. & BORCHERT, M. 2010. *Survey of eResearch practices and skills at QUT, Australia*. [Online]. Available from: http://eprints.qut.edu.au/33111/3/c33111.pdf. [Accessed 8 July 2014]

BRAUN, V. & CLARKE, V. 2013. *Successful qualitative research: a practical guide for beginners*. London: Sage. 382p.

BRINKMANN, S. & KVALE, S. 2006. Confronting the ethics of qualitative research. *Journal of Constructivist Psychology*, vol. 18(2): 157-181. [Online]. Available from: http://www.tandfonline.com/doi/abs/10.1080/10720530590914789 [Accessed 28 January 2015]

BUYS, C.M. & SHAW, P.L. 2015. Data Management Practices Across an Institution: Survey and Report. *Journal of Librarianship and Scholarly Communication*, 3(2), eP1225. http://jlsc-pub.org/articles/abstract/10.7710/2162-3309.1225/

CAMPBELL, E.G., CLARRIDGE, B.R., GOKHALE, M., BIRENBAUM, L., HILGARTNER, S., HOLTZMAN. N.A. & BLUMENTHAL, D. 2002. Data withholding in academic genetics: evidence from a national survey. *The Journal of the American Medical Association*, 287(4):473-480. [Online]. Available from: http://jama.jamanetwork.com/article.aspx?articleid=194592 [Accessed 8 July 2014]

CARLSON. J.R. 2011. *Demystifying the data interview: developing a foundation for reference librarians to talk with researchers about their data*. [Online]. Available from: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1186&context=lib_research [Accessed 23 October 2014]

CARLSON, J.R., FOSMIRE, M., MILLER, C. & NELSON, M.S. 2011. *Determining Data Information Literacy Needs: A Study of Students and Research Faculty*. [Online]. Available from: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1031&context=lib_fsdocs [Accessed 21 October 2014]

CARVALHO, E.C.A.D., BATILANA, A.P., SIMKINS, J., MARTINS, H., SHAH, J., RAJGOR, D., SHAH, A., ROCKART, S. & PIETROBON, R. 2010. Application description and policy model in collaborative environment for sharing of information on epidemiological and clinical research data sets. *PLoS ONE* 5(2): e9314. [Online] Available from: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0009314 [Accessed 8 July 2014]

CHERRY, K. 2014. *What is a representative sample?* [Online]. Available from: https://www.verywell.com/what-is-a-representative-sample-2795798 Accessed 29 January 2015

CONCORDIA UNIVERSITY. 2014. *How to write a literature review*. [Online]. Available from: http://library.concordia.ca/help/howto/litreview.php [Accessed 22 October 2014]

CRAGIN, M.H., PALMER, C.L., CARLSON, J.R. & WITT, M. 2010. Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A*, 368:4023-4038. [Online]. Available from: http://rsta.royalsocietypublishing.org/content/368/1926/4023.full.pdf+html [Accessed 24 October 2014]

CRAGIN, M.H., PALMER, C.L., KOGAN, M., CARLSON, J.R. & WITT, M. 2009. *Data sharing, small science, and institutional repositories*. [Online]. Available from: http://www.allhands.org.uk/2009/09/Cragin_UK_All-Hands_2009_final2.pdf [Accessed 24 January 2014]

CSIR. 2011. Records Management. 12pp. Unpublished document.

CSIR. 2015. *2014/2015 Annual Report*. [Online]. Available from: http://www.csir.co.za/publications/pdfs/CSIR%20Annual%20Report%202014-15_low%20res.pdf [Accessed 28 March 2016]

CURTIN LIBRARY. 2015. *Managing research data*. [Online]. Available: http://libguides.library.curtin.edu.au/research-data-management [Accessed 30 March 2016]

DCC. 2014. *Data Management Plans*. [Online]. Available from: http://www.dcc.ac.uk/resources/data-management-plans [Accessed 24 October 2014]

DELASALLE, J. 2013. *Research Data Management at the University of Warwick: recent steps towards a joined-up approach at a UK university*. [Online]. Available from: http://libreas.eu/ausgabe23/10delasalle/ [Accessed 21 October 2014]

DE VOS, A.S. & FOUCHE, C.B. 2000. General introduction to research design, data collection methods and data analysis. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

DE VOS, A.S., SCHURINK, E.M. & STRYDOM, H. 2000. The nature of research in the caring professions. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

DIEKMANN, F. 2012. Data practices of agricultural scientists: results from an exploratory study. *Journal of Agricultural & Food Information*, 13(1):14-34. [Online]. Available from: http://www.tandfonline.com/doi/abs/10.1080/10496505.2012.636005#.U7vgg7FlPF9 [Accessed 8 July 2014]

DIETRICH, D., ADAMUS, T., MINER, A. & Steinhart, G. 2012. De-mystifying the data management requirements of research funders*. Issues in Science and Technology Librarianship*, 70 (22p). [Online]. Available from: http://www.istl.org/12-summer/refereed1.html [Accessed 3 March 2014]

DIGITAL CURATION CENTRE. 2016. *Overview of funders' data policies*. [Online]. Available: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies [Accessed 30 March 2016]

DIGITAL CURATION CENTRE. 2016. *Using Metadata Standards*. [Online]. Available from: http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/using-metadata-standards [Accessed 1 February 2016]

DILLMAN. D.A. 2007. *Mail and Internet Surveys: The Tailored Design Method*. Hoboken: John Wiley & Sons

DILLO, I. & DOORN, P. 2011. *The Dutch data landscape in 32 interviews and a survey*. [Online]. Available from: https://pure.knaw.nl/portal/files/478516/The_Dutch_Datalandscape_DEF.pdf [Accessed 8 July 2014]

DONNELLY, M. 2013. *Research data management: definitions, drivers and resources*. [Online]. Available: http://www.slideshare.net/martindonnelly/donnelly-eifl-open-research-datafinal#btnLast [Accessed 30 March 2016]

DOTY, J., AKERS, K.G., BODNAR, J. & JAHNKE, L. 2013. *Assessing research data management needs at Emory University*. [Online]. Available from: http://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1015&context=southeasternlac [Accessed 27 January 2014]

DOUCETTE, L. & FYFE, B. 2013. *Drowning in research data: addressing data management literacy of graduate students*. Paper presented at ACRL2013, Indianapolis, 10-13 April 2013:165-171. [Online]. Available from: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/2013/papers/DoucetteFyfe_Drowning.pdf [Accessed 8 July 2014]

DUBLIN CORE METADATA INITIATIVE. 2014. *Metadata basics*. [Online]. Available from: http://dublincore.org/metadata-basics/ [Accessed 27 October 2014]

ECAR. 2009. *Research Data Management*. ECAR Research Study: 119-140. [Online]. Available from: http://net.educause.edu/ir/library/pdf/ers0908/rs/ers09087.pdf [Accessed 20 October 2014]

EKMEKCIOGLU, C. & RICE, R. 2009. *Edinburgh Data Audit Implementation Project: Final Report.*[Online]. Available from: http://repository.jisc.ac.uk/283/ [Accessed 8 July 2014]

ELIOT, S. 2011. *Information Rich Sampling*. [Online]. Available from: http://www.qualitative-researcher.com/qualitative-research-2/information-rich-sampling/ [Accessed 11 March 2015]

EMORY LIBRARIES & INFORMATION TECHNOLOGY. 2016. *Benefits of Research Data Management.* [Online]. Available: http://guides.main.library.emory.edu/datamgmt [Accessed 30 March 2016]

ENKE, N., THESSEN, A., BACH, K., BENDIX, J., SEEGER, B., GEMEINHOILZER, B. 2012. The user's view on biodiversity data sharing – Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, 11: 25-33. Available from: http://www.sciencedirect.com/science/article/pii/S1574954112000222 [Accessed 8 July 2014]

EUROPEAN COMMISSION. 2016. *Guidelines on Data Management in Horison 2020*. [Online]. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [Accessed 31 March 2016]

EXECUTIVE OFFICE OF THE PRESIDENT. OFFICE OF SCIENCE AND TECHNOLOGY POLICY. 2013. *Memorandum for the Heads of Executive Departments and Agencies, 22 February 2013.* [Online]. Available from: https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [Accessed 28 June 2016]

FARRIMOND, H. 2013. *Doing ethical research*. New York: Palgrave Macmillan.

FEARON, D.J., GUNIA, B., LAKE, S., PRALLE, B.E. & SALLANS, A.L. 2013. *SPEC Kit 334: Research Data Management Services* (July 2013). [Online]. Available from: http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/ [Accessed 21 October 2014]

FOUCHE, C.B. 2000. Single-system designs. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

FOUCHE, C.B. & DE VOS, A.S. 2000. Selection of a research design. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

FREIMAN, L., WARD, C., JONES, S., MOLLOY, L. & SNOW, K. 2010. *Scoping study and implementation plan*. [Online]. Available from: http://www.lib.cam.ac.uk/preservation/incremental/documents/Incremental_Scoping_Report_170910.pdf [Accessed 28 August 2014]

GARRETT, L., GRAMSTADT, M-T., BURGESS, R., MURTAGH, J., NADIM, T. & SPALDING, A. 2012. *Kaptur environmental assessment report*. [Online]. Available from: http://www.research.ucreative.ac.uk/1054/7/Kaptur_environmental_assessment_v1_7.pdf [Accessed 21 October 2014]

GIBBS, H. 2009. *Southampton data survey: our experience and lessons learned*. [Online]. Available from: http://www.disc-uk.org/docs/SouthamptonDAF.pdf [Accessed 9 July 2014]

GIBSON, D. & GROSS, J. 2013. Research data management in a collaborative network. *Research Global*, 33:12-15. [Online]. Available from: http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1585&context=ecuworks2011 [Accessed 28 August 2014]

GLAZIER, J.D. & POWELL, R.R. 1992. *Qualitative research in information management*. Englewood: Libraries Unlimited. 238p.

GRIFFITHS, A. 2009. The publication of research data: researcher attitudes and behaviour. *International Journal of Digital Curation*, 4(1):46-46. [Online]. Available from: http://www.ijdc.net/index.php/ijdc/article/view/101/76 [Accessed 20 October 2014]

GU, X. & AVERKAMP, S. 2012. *Report on the University of Iowa libraries' data management needs survey*. [Online]. Available from: http://blog.lib.umn.edu/lmcguire/hslm/Data_Management_at_UIowa_SurveyReport_20121121.pdf [Accessed 24 January 2014]

HALBERT, M. 2013. The problematic future of research data management: challenges, opportunities and emerging patterns identified by the DataRes Project. *International Journal of Digital Curation,* 8(2):111-122. [Online]. Available from: http://www.ijdc.net/index.php/ijdc/article/view/8.2.111/321 [Accessed 20 October 2014]

HALL, N. 2013. *Environmental Studies Faculty Attitudes Towards Sharing of Research Data*. Paper presented at the 13th ACM/IEEE-CS joint conference on Digital libraries, Indianapolis, USA, 22-26 July 2013:383-384. [Online]. Available from: http://delivery.acm.org/ [Accessed 21 October 2014]

HANK, C., JORDAN, M.W. & WILDEMUTH, B.M. 2009. Survey research. In: Wildemuth, B.M. *Applications of social research methods to questions in information and library science*. Westport: Libraries Unlimited

HENTY, M., WEAVER, B., BRADBURY, S. & PORTER, S. 2008. *Investigating data management practices in Australian universities*. [Online]. Available from: http://apsr.anu.edu.au/orca/investigating_data_management.pdf [Accessed 28 August 2014]

HUMPHREY, C. 2012. *Canada's Long Tale of Data: A Body of Evidence*. [Online]. Available from: http://preservingresearchdataincanada.net/tag/national-data-archive-consultation/ [Accessed 28 October 2014]

JAHNKE, L.M. & ASHER, A. 2012. *The problem of data: data management and curation practices among university researchers*. [Online]. Available from: http://www.clir.org/pubs/reports/pub154/problem-of-data [Accessed 8 July 2014]

JISC. 2014. *An Analysis of Training Needs from JISC MRD Project Surveys*. [Online]. Available from: https://figshare.com/articles/An_Analysis_of_Training_Needs_from_JISC_MRD_Project_Surveys/1092560

JOHARE, S. 2014. *InterPARES 3: TEAM Malaysia Preliminary Findings*. [Online]. Available from: http://www.interpares.org/display_file.cfm?doc=ip3_isym05_presentation_03-03--yunus.pdf [Accessed 8 July 2014]

JOHNSTON, L. & JEFFRYES, J. 2014. Data management skills needed by structural engineering students: case study at the University of Minnesota. *Journal of Professional Issues in Engineering Education and Practice*, 140(2)05013002 (29p). [Online]. Available from: http://conservancy.umn.edu/bitstream/handle/11299/156947/JohnstonEngEdu.pdf?sequence=1&isAllowed=y [Accessed 4 October 2014]

JONES, K. 2011. *Assessing institutional data storage and management using the Data Asset Framework (DAF) methodology at the University of Bath*. [Online]. Available from: http://opus.bath.ac.uk/24960/1/DAF_report_May_2011.pdf [Accessed 20 October 2014]

JONES, S. 2011. *How to develop a Data Management and Sharing Plan*. [Online]. Available from: http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How%20to%20Develop.pdf [Accessed 30 March 2016]

JONES, S., ROSS, S. & RUUSALEPP, R. 2009. [Online]. *Data Audit Framework methodology*. Available from: http://www.data-audit.eu/DAF_Methodology.pdf [Accessed 15 August 2014]

KENNAN, M.A. & MARKAUSKAITE, L. 2015. Research Data Management Practices: A Snapshot in Time. International Journal of Digital Curation, vol. 10(2): 69-95 http://www.ijdc.net/index.php/ijdc/article/view/10.2.69

KERALIS, S.D.C., STARK, S., HALBERT, M. & MOEN, W.E. 2012. *DataRes Project Primary Survey*. [Online]. Available from: http://digital.library.unt.edu/ark:/67531/metadc228265/ [Accessed 20 October 2014]

KEY PERSPECTIVES LTD. 2010. *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability*. [Online]. Available from: https://www.era.lib.ed.ac.uk/bitstream/1842/3364/1/SCARP%20SYNTHESIS.pdf [Accessed 21 October 2014]

KNIGHT, G. 2013. *Research Data Management at LSHTM: Web Survey Report*. [Online]. Available from: http://blogs.lshtm.ac.uk/rdmss/files/2013/04/LSHTM-RDM-Web-Survey-Report.pdf [Accessed 28 August 2014]

KOUPER, I., AKERS, K.G., NICHOLLS, N.H. & SFERDEAN, F.C. 2013. *A roadmap for data services*. Paper presented at the 2013 Joint Conference on Digital Libraries, Indianapolis, USA:375-376. [Online]. Available from: http://dl.acm.org/ft_gateway.cfm?id=2467763&ftid=1384624&dwn=1&CFID=444998685&CFTOKEN=20001173 [Accessed 21 October 2014]

KUIPERS, T. & VAN DER HOEVEN, J. 2009. *Insight into digital preservation of research output in Europe: survey report*. [Online]. Available from: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf [Accessed 8 July 2014]

LAERD RESEARCH LIMITED. 2012. *Total population sampling*. [Online]. Available from: http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+census+and+sample Accessed 28 January 2015

LICHTMAN, M. 2014. *Qualitative research for the social sciences*. London: Sage. 418p.

LORD, P. & MACDONALD, A. 2003. *e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*. [Online]. Available from: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf [Accessed 20 October 2014]

LÖTTER, L. 2014. *HSRC Research Data Management*, Powerpoint presentation, HSRC, Pretoria, South Africa. Unpublished presentation.

LYON, L., RUSBRIDGE, C., NEILSON, C., & WHYTE, A. 2010. *Disciplinary approaches to sharing, curation, reuse and preservation*. [Online]. Available from: http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf [Accessed 8 July 2014]

MARCHIONINI, G. 2012. *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership*. University of North Carolina, Chapel Hill, NC. 72pp.

[Online]. Available from: http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf

MARCUS, C., BALL, S., DELSERONE, L., HRIBAR, A. & LOFTUS, W. 2007. *Understanding research behaviors, information resources, and service needs of scientists and graduate students: a study by the University Of Minnesota Libraries*. [Online]. Available from: http://conservancy.umn.edu/bitstream/handle/11299/5546/1/Sciences_Assessment_Report_Final.pdf [Accessed 20 October 2014]

MARTINEZ-URIBE, L. 2008. *Findings of the scoping study interviews and the research data management workshop*. [Online]. Available from: http://ora.ox.ac.uk/objects/uuid%3A4e2b7e64-d941-4237-a17f-659fe8a12eb5/datastreams/ATTACHMENT02 [Accessed 9 July 2014]

MATHABE, C. & DLAMINI, T. 2010. *Policies of the CSIR. Strengthening the SET Base: Research Ethics Policy*. (Unpublished)

MCCLURE, C.R. & HERNON, P. (eds). 1991. *Library and information science research: perspectives and strategies for improvement.* Norwood: Ablex.

MIT LIBRARIES. (n.d.). *Data management and publishing*. [Online]. Available: http://libraries.mit.edu/data-management/plan/write/#Resources [Accessed 30 March 2016]

MOCH, S.D. & GATES, M.F. (eds). *The researcher experience in qualitative research*. London: Sage. 193p.

MORSE, J.M. 2000. Determining sample size. *Qualitative health research*, vol. 10(1): 3-5. [Online]. Available from: http://qhr.sagepub.com/content/10/1/3.full.pdf [Accessed 29 June 2016]

MOSSINK, W. & BIJSTERBOSCH, M. 2013. *European Landscape Study of Research Data Management*. [Online]. Available from: http://www.sim4rdm.eu/sites/default/files/uploads/documents/SIM4RDM%20landscape%20report%20final%2025.01.12.pdf [Accessed 8 July 2014]

MOWERS, S., HUMPHREY, C. & PERRY. C.M. 2013. *Summary Report: Survey of Researchers' Needs and Practices regarding Research Data Management in Canada*. [Online]. Available from: http://gsg.uottawa.ca/data/open/aa-interim-survey-report/20130801-en.pdf [Accessed 21 October 2014]

MUGERA, W. 2013. *Non-probability sampling techniques*. [Online]. Available from: https://learning.uonbi.ac.ke/courses/LDP603/work/assig_2/Non-probability_sampling_techniques_assignment_1.pdf Accessed 28 January 2015

NASSIRI, S. & WORTHINGTON, B. 2012. *Digital Asset Framework (DAF) Analysis*. [Online]. Available from: http://research-data-toolkit.herts.ac.uk/2012/08/data-asset-survey-results/ [Accessed 20 October 2014]

NATIONAL RESEARCH FOUNDATION. 2015. *Statement on Open Access to Research Publications from the National Research Foundation (NRF)-Funded Research*. [Online]. Available from: http://ir.nrf.ac.za/bitstream/handle/10907/103/oastatement_2015.pdf?sequence=1 [Accessed 28 March 2016]

NCSU LIBRARIES. 2014. *Data Management Planning*. [Online]. Available: http://www.lib.ncsu.edu/guides/datamanagement [Accessed 30 March 2016]

ONDRACEK, C. 2013. *What's the data about Research Data*? [Online]. Available: http://www.slideshare.net/ctondracek/data-about-data-management [Accessed 30 March 2016]

OPEN EXETER PROJECT TEAM. 2012. *Summary findings of the Open Exeter Data Asset Framework Survey*. [Online]. Available from: https://ore.exeter.ac.uk/repository/bitstream/handle/10036/3689/daf_report_public.pdf?sequence=1 [Accessed 28 August 2014]

OSSWALD, A. & STRATHMANN, S. 2012. *The role of libraries in curation and preservation of research data in Germany: findings of a survey*. Paper presented at the 78th IFLA General Conference and Assembly, Helsinki, Finland, 11-17 August 2012(10p). [Online]. Available from: http://conference.ifla.org/past-wlic/2012/116-osswald-en.pdf [Accessed 8 July 2014]

PARHAM, S.W., BODNAR, J. & FUCHS, S. 2012. Supporting tomorrow's research: assessing faculty data curation needs at Georgia Tech. *College & Research Libraries News*, 73(1):10-13. [Online]. Available from: http://crln.acrl.org/content/73/1/10.full [Accessed 20 October 2014]

PARSONS, T., GRIMSHAW, S. & WILLIAMSON, L. 2013. *Research data management survey*. [Online]. Available from: http://admire.jiscinvolve.org/wp/files/2013/02/ADMIRe-Survey-Results-and-Analysis-2013.pdf [Accessed 14 July 2014]

PATRICK, M. 2012. *DaMaRO Survey Results – Research Data Management Training for the Sciences*. [Online]. Available from: http://blogs.it.ox.ac.uk/damaro/2012/11/21/damaro-survey-results-research-data-management-training-for-the-sciences/ [Accessed 20 October 2014]

PATRICK, M. 2014. *What researchers want out of RDM training.* [Online]. Available from: http://www.ses.ac.uk/wp-content/uploads/sites/79/2015/08/Dr-Meriel-Patrick-What-researchers-want-out-of-RDM-training-MJP-revised.pdf

PATTERTON, L.H. 2014a. *Research data management at the CSIR: an exploratory survey.* Unpublished research report

PATTERTON, L.H. 2014b. *Survey of CSIR research data management practices*. Unpublished presentation

PESET, F., FERNÁNDEZ, A., FERRER-SAPENA, A., GARCÍA-GARCÍA, A., ALEIXANDRE, R., VIDAL, A. & GARCÍA, C. 2015. *Research data management in Spain: results from a DATASEA project survey.*, 2015 . In 10th International Digital Curation Conference, London, 9-12 February 2015. (Unpublished) [Conference poster] http://eprints.rclis.org/24601/

PETERS, C. & DRYDEN, A.R. 2011. Assessing the academic library's role in campus-wide research data management: a first step at the University of Houston. *Science & Technology Libraries*, 30(4):387-403. [Online]. Available from: http://www.tandfonline.com/doi/abs/10.1080/0194262X.2011.626340#tabModule [Accessed 20 October 2014]

PICKARD, A.J. 2013. *Research methods in information*. London: Facet

PIENAAR, H. n.d. *Findings of a survey of research data management practices over the period October 2009 - March 2010 at the University of Pretoria*. Internal communication.

PIENAAR, H. 2010. *Survey of research data management practices at the University of Pretoria, South Africa: October 2009-March 2010.* [Online]. Available from: http://repository.up.ac.za/handle/2263/15154 [Accessed 22 October 2014]

PIENAAR, H. 2011. *An analysis of data management practices at a large South African University*. [Online]. Available from: http://www.slideshare.net/heila1/survey-of-research-data-management-practices-up2010digschol2011?from_search=7 [Accessed 8 July 2014]

PIENAAR, H. Heila.Pienaar@up.ac.za. 2013. *Re: Data versoek*. [email] Message to L.H. Patterton (lpatterton@csir.co.za). Sent 15 October 2013: 18:16

PIENAAR, H. 2015. *Dr heila pienaar cvOct2015*. [Online]. Available from: http://www.slideshare.net/heila1/dr-heila-pienaar-cvoct2015 [Accessed 3 June 2016]

PILLAY, T., thashni.maistry@nrf.ac.za. 2016. *Data management requirements for NRF grant applicants.* [email] Message to L.H. Patterton (lpatterton@csir.co.za). Sent 27 January 2016: 15:07.

PINK, C.J., COPE, J., JORDAN, K.M. & JONES, K. 2013. *Research360: Faculty-Industry Data Requirements Report*. [Online]. Available from: http://opus.bath.ac.uk/36361/2/R360_RequirementsReport_FINAL.pdf [Accessed 20 October 2014]

PIWOWAR, H.A., BECICH, M.J., BILOFSKY, H. & CROWLEY, R.S. 2008. Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. *PLoS Medicine,* vol. 5 (9): e183. [Online]. Available from: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050183 [Accessed 26 January 2017]

PIWOWAR, H.A. 2011. Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*, 6(7):e18657 (13p). [Online]. Available from: http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0018657&representation=PDF [Accessed 8 July 2014]

POLYDORATOU, P. 2009. *JISC Final Report*. [Online]. Available from: http://discovery.ucl.ac.uk/15053/1/15053.pdf [Accessed 20 October 2014]

PRYOR, G. 2009. Multi-scale data sharing in the life sciences: some lessons for policy makers. *The International Journal of Digital Curation,* 4(3):71-82. [Online]. Available from: http://www.ijdc.net/index.php/ijdc/article/view/135/178 [Accessed 24 October 2014]

RAGGETT, M. 2012a. *Analysis of results from the research data preservation survey – full report.* [Online]. Available from: http://lsedice.files.wordpress.com/2012/04/survey-report-full.pdf [Accessed 9 July 2014]

RAGGETT, M. 2012b. *DICE final report*. [Online]. Available from: http://eprints.lse.ac.uk/47261/1/DICE%20final%20report(lsero).pdf [Accessed 24 October 2014]

RANKIN, P., BUTTENFIELD, B., DUERR, R., HAUSER, T., JOHNSON, A., MANESS, J., PARSONS, M., RAJARM, H., SHOEMAKER, R., STACEY, K., VIGGIO, A. & WAKIMOTO, J.C. 2012. *Research Data Management at the University Of Colorado Boulder:*

Recommendations in support of fostering 21st Century research excellence. [Online]. Available from: http://digitool.library.colostate.edu/exlibris/dtl/d3_1/apache_media/L2V4bGlicmlzL2R0bC9kM18xL2FwYWNoZV9tZWRpYS8xNzQ3MzM=.pdf [Accessed 20 October 2014]

REIDPATH, D.D. & ALLOTEY, P.A. 2001. Data sharing in medical research: an empirical investigation. *Bioethics*, 15(2):125-134. [Online]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/1467-8519.00220/pdf. [Accessed 8 July 2014]

RESEARCH INFORMATION NETWORK. 2011. *Reinventing research? Information practices in the humanities*. [Online]. Available from: http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities [Accessed 8 July 2014]

RESEARCH INFORMATION NETWORK AND THE BRITISH LIBRARY. 2009. *Patterns of information use and exchange: case studies of researchers in the life sciences.* [Online]. Available from: http://www.bl.uk/reshelp/experthelp/science/science@blevents/previousevents/rincasestudy/Patterns_information_use-REPORT_Nov09.pdf [Accessed 21 October 2014]

RESOURCES FOR RESEARCH ETHICS EDUCATION. 2013. *What is Research Ethics?* [Online]. Available from: http://research-ethics.net/introduction/what/ Accessed 13 April 2015

RIBEIRO, C. & FERNANDES, M.E.M. 2011. *Data curation at U.Porto*. [Online]. Available from: http://repositorio-aberto.up.pt/bitstream/10216/62536/2/1862.pdf [Accessed 8 July 2014]

RIRMIT University. Not dated. *Literature review*. [Online]. Available from: http://www.rmit.edu.au/library/literaturereview [Accessed 22 October 2014]

SALLANS, A. 2013. *Addressing research data management needs as the Scientific Data Consulting Group.* [Online]. Available from: http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1063&context=escience_symposium [Accessed 27 January 2014]

SAYOGO, D.S. & PARDO, T.A. 2013. Exploring the determinants of scientific data sharing: understanding the motivation to publish research data. *Government Information Quarterly*, 30(supplement 1):S19-S31. [Online]. Available from: http://www.sciencedirect.com/science/article/pii/S0740624X12001529. [Accessed Downloaded 8 July 2014]

SCARAMOZZINO, J.M., RAMIREZ, M.L. & McGAUGHEY, K.J. 2012. A study of faculty data curation behaviors and attitudes at a teaching-centered university. *College & Research Libraries*, 73(4):349-365. [Online]. Available from: http://crl.acrl.org/content/73/4/349.full.pdf+html [Accessed 20 October 2014]

SCHOLES, R.J. (ed.) 2003. *Good research guide*. Pretoria: CSIR.

SCHURINK, E.M. 2000. Deciding to use a qualitative approach. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

SEWERIN, C., DEARBORN, D., HENSHILWOOD, A., SPENCE, M., & ZAHRADNIK, T. 2015. *Research Data Management Faculty Practices: a Canadian Perspective*. IATUL 2015 'Strategic Partnerships for Access and Discovery' Conference, Hannover, Germany, July 5-9, 2015. https://tspace.library.utoronto.ca/handle/1807/69145 [Accessed 8 June 2016]

SHAI, N. shaigs@unisa.ac.za. 2014. *Re: surveymonkey*. [email] Message to L.H. Patterton (lpatterton@csir.co.za). Sent 28 February 2014: 12:24

SIMUKOVIC, E., KINDLING, M., & SCHIRMBACHER, P. (2014). *Unveiling Research Data Stocks: A Case of Humboldt Universität zu Berlin*. In iConference 2014 Proceedings (p. 742–748). doi:10.9776/14351 https://www.ideals.illinois.edu/bitstream/handle/2142/47259/351_ready.pdf?sequence=2 [Accessed 8 June 2016]

SINGLETON, R.A. & STRAITS, B.C. 2005. *Approaches to social research*. New York: Oxford University Press

SOCIAL SCIENCES AND HUMANITIES RESEARCH COUNCIL OF CANADA & NATIONAL ARCHIVES OF CANADA. 2001. *National Research Data Archive Consultation. Phase One: Needs Assessment Report*. [Online]. Available from: http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/da_phase1_e.pdf [Accessed 28 October 2014]

SOEHNER, C., STEEVES, C. & WARD, J. 2010. *E-science and data support services: a study of ARL member institutions*. [Online] Available from: http://www.arl.org/storage/documents/publications/escience-report-2010.pdf [Accessed 20 October 2014]

STEINHART, G., SAYLOR, J., ALBERT, P., ALPI, K., BAXTER, P., BROWN, E., CHIANG, K., CORSON-RIKERT, J., HIRTLE, P., JENKINS, K., LOWE, B., McCUE, J., RUDDY, D., SILTERRA, R., SOLLA, L., STEWART-MARSHALL, Z. & WESTBROOKS, E.L. 2008. *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the*

Cornell University Library. [Online]. Available from: http://ecommons.library.cornell.edu/handle/1813/10903 [Accessed 22 October 2014]

STRYDOM, H. 2000. Ethical aspects of research in the caring professions. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

STRYDOM, H. & DE VOS, A.S. 2000. Sampling and sampling methods. In: De Vos, A.S. (ed.) *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik

TASK FORCE, VICE CHANCELLOR FOR RESEARCH'S DATA MANAGEMENT. 2012. Research Data Management at the University of Colorado Boulder: Recommendation in Support of Fostering 21st Century Research Excellence. Office of the Vice Chancellor for Research, Paper 1. [Online]. Available from: http://scholar.colorado.edu/ovcr/1/ [Accessed on 26 January 2017]

TAKEDA, K., BROWN, M., COLES, S., CARR, L., EARL, G., FREY, J., HANCOCK, P., WHITE, W., NICHOLS, F., WHITTON, M., GIBBS, H., FOWLER, C., WAKE, P. & PATTERSON, S. 2010. *Digital Management for All – The Institutional Data Management Blueprint project*. [Online]. Available from: http://eprints.soton.ac.uk/169533/1/6th_international_digital_curation_conference__idmb_final_paper_revised.pdf [Accessed 20 October 2014]

TAM, W., FRY, J. & PROBETS, S. 2014. *The disciplinary shaping of research data management practices*. [Online]. Available from: https://ideals.illinois.edu/handle/2142/47276 [Accessed 8 July 2014]

TENOPIR, C., ALLARD, S, DOUGLASS, K, AYDINOGLU, A.U., WU, L., READ, E., MANOFF, M. & FRAME, M. 2011. Data sharing by scientists: practices and perceptions. *PLoS ONE*, 6(6):e21101 (21p). [Online]. Available from: http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0021101&representation=PDF [Accessed 8 July 2014]

UNIVERSITY OF BRISTOL. 2014. *Qualitative methodologies*. [Online]. Available from: http://www.bris.ac.uk/Depts/DeafStudiesTeaching/dissert/Qualitative%20Methodologies.htm [Accessed 21 March 2014]

UNIVERSITY OF CAMBRIDGE. 2012. *Preservation*. [Online]. Available from: http://www.lib.cam.ac.uk/dataman/pages/preservation.html [Accessed 27 October 2014].

UNIVERSITY OF EDINBURGH, INSTITUTE FOR THE STUDY OF SCIENCE, TECHNOLOGY AND INNOVATION AND THE DIGITAL CURATION CENTRE. 2009.

Patterns of information use and exchange: case studies of researchers in the life sciences. Annex to the Research Information Network and British Library Report. [Online]. Available from: http://www.rin.ac.uk/system/files/attachments/Patterns_information_use_Annex_Nov09.pdf [Accessed 8 June 2016]

UNIVERSITY OF LEICESTER. n.d. *Writing a dissertation*. [Online]. Available from: http://www2.le.ac.uk/offices/ld/resources/writing/writing-resources/writing-dissertation [Accessed 22 October 2014]

UNIVERSITY OF MELBOURNE. 2014. *Literature reviews*. [Online]. Available from: http://unimelb.libguides.com/content.php?pid=87165&sid=651752 [Accessed 22 October 2014]

UNIVERSITY OF OXFORD. 2016. *What is RDM?* [Online]. Available: http://researchdata.ox.ac.uk/home/introduction-to-rdm/ [Accessed 30 March 2016]

UNIVERSITY OF SHEFFIELD. 2014. *Organising research data*. [Online]. Available from: http://www.sheffield.ac.uk/library/rdm/organising [Accessed 27 October 2014]

UNIVERSITY OF WESTERN AUSTRALIA. 2015. *Research Data Management Planning*. [Online]. Available: http://guides.is.uwa.edu.au/RDMtoolkit [Accessed 30 March 2016]

VAN TUYL, S. & MICHALEK, G. 2015. Assessing Research Data Management Practices of Faculty at Carnegie Mellon University. *Journal of Librarianship and Scholarly Communication*, 3(3), eP1258. http://jlsc-pub.org/articles/abstract/10.7710/2162-3309.1258/

WALLACE, D.P. & VAN FLEET, C. 2012. *Knowledge into action: research and evaluation in library and information science*. Oxford: Libraries Unlimited. 388p.

WARD, C., FREIMAN, L., JONES, S., MOLLOY, L. & SNOW, K. 2011. Making sense: talking data management with researchers. *The International Journal of Digital Curation*, 2(6): 265-273. [Online]. Available from: http://www.ijdc.net/index.php/ijdc/article/view/197/262

WESTRA, B. 2010. Data services for the sciences: a needs assessment. *Ariadne*, 64: (11p). [Online]. Available from: http://www.ariadne.ac.uk/issue64/westra [Accessed 1 April 2014]

WHITMIRE, A.L., BOOCK, M., & SUTTON, S.C. 2015. *Variability in academic research data management practices: implications for data services development from a faculty survey.* Program, 49(4). https://ir.library.oregonstate.edu/xmlui/handle/1957/57240

WHITMIRE, A., BRINEY, K., NUMBERGER, A., HENDERSON, M., ATWOOD, T., JANZ, M., KOZLOWSKI, W., LAKE, S., VANDEGRIFT, M. & ZILINSKY, L. 2015. *A table summarizing the Federal public access policies resulting from the US Office of Science and Technology Policy memorandum of February 2013*. [Online]. Available from: https://docs.google.com/spreadsheets/d/1PYOhBh6bglh6BkQFlpvNLOwlpzvQyguWAG8AkQMtU0s/edit#gid=0 [Accessed 31 March 2016]

WILDEMUTH, B.M. 2009a. Correlation. In: Wildemuth, B.M. *Applications of social research methods to questions in information and library science*. Westport: Libraries Unlimited

WILDEMUTH, B.M. 2009b. Descriptive statistics. In: Wildemuth, B.M. *Applications of social research methods to questions in information and library science*. Westport: Libraries Unlimited

WILDEMUTH, B.M. 2009c. Frequencies, cross-tabulation, and the chi-square statistic. In: Wildemuth, B.M. *Applications of social research methods to questions in information and library science*. Westport: Libraries Unlimited

WILEY. 2014. *Researcher Data Sharing Insights.* [Online]. Available from: http://exchanges.wiley.com/blog/wp-content/uploads/2014/11/Researcher-Data-Insights-Infographic-FINAL-REVISED-2.jpg

WILLIAMS, S.C. 2012. Data practices in the crop sciences: a review of selected faculty publications. *Journal of Agricultural and Food Chemistry*, 13(4): 308-325. [Online]. Available from: http://www.tandfonline.com/doi/pdf/10.1080/10496505.2012.717846 [Accessed 21 October 2014]

WILSON, A.J. & PATRICK, M. 2010. *Sudamih Research Requirements Report*. [Online]. Available from: http://sudamih.oucs.ox.ac.uk/docs/Sudamih%20Researcher%20Requirements%20Report.pdf [Accessed 20 October 2014]

WILSON. J.A., FRASER, M.A., MARTINEZ-URIBE, L., PATRICK, M., AKRAM, A., & MANSOORI, T. 2010. Developing infrastructure for research data management at the University of Oxford. *Ariadne*, 65. (7p). [Online]. Available from: http://www.ariadne.ac.uk/issue65/wilson-et-al [Accessed 23 October 2014]

WILSON, J.A.J., MARTINEZ-URIBE, L., FRASER, M.A. & JEFFREYS, P. 2011. An institutional approach to developing research data management infrastructure. *International*

Journal of Digital Curation, 6(2):274-287. [Online]. Available from: http://www.ijdc.net/index.php/ijdc/article/view/198 [Accessed 20 October 2014]

WILSON, J. 2013. *University of Oxford Research Data Management Survey 2012: The Results*. [Online]. Available from: http://blogs.it.ox.ac.uk/damaro/2013/01/03/university-of-oxford-research-data-management-survey-2012-the-results/ [Accessed 21 October 2014]

WITT, M. 2009. *Eliciting faculty requirements for research data repositories*. [Online]. Available from: https://smartech.gatech.edu/bitstream/handle/1853/28509/92-446-1-PB.pdf?sequence=4 [Accessed 13 July 2014]

WYNHOLDS, L., FEARON, D.S., BORGMAN, C.L. & TRAWEEK, S. 2011. *When use cases are not useful: data practices, astronomy, and digital libraries*. Paper presented at the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottowa, Canada, 13-17 June 2011:383-386. [Online]. Available from: 10.1145/1998076.1998146 [Accessed 20 October 2014]

YEUMO, E.D. 2014. *Survey results: Data standards in the Wheat research community*. [Online]. Available from: https://rd-alliance.org/sites/default/files/RDA%20survey-overview.pdf [Accessed 8 June 2016]

APPENDICES

Appendix 1: Data management plan

Data Management Categories	
Administrative Data	
DMP number (system-generated)	
Funder	CSIR
Grant Reference Number	
Project Name	MIS studies: Research data management practices of emerging CSIR researchers
Project Description	A master's degree study at the University of Pretoria comprising a survey investigating the research data management practices of the CSIR's emerging researchers. This study is also registered as a CSIR project.
Principal investigator / Researcher	Louise Patterton, data librarian, CSIR
Principal investigator / Researcher ID	0000-0003-5000-0001 (ORCID)
Project Data Contact	See PI/Researcher
Date of First Version	15/09/2015
Date of Last Update	28/06/2016
Related Policies	No funder data management policies
Data Collection	
What data will you collect or create?	• Data will be qualitative in nature. • Project will not be using existing data.
How will the data be collected or created?	• This study is a survey. • Data will be collected via an online questionnaire. E-surv is used as survey platform. • The online questionnaire consists of 31 questions. Three of the questions are demographic in nature while the rest are concerned with research data management practices, challenges and perceptions. • Target population and sample will be identified from a list compiled by CSIR Human Resources. • Responses will be captured by the E-surv platform and captured in an Excel spreadsheet. • File formats: MS Excel spreadsheet (.xls; and comma-delimited .csv) MS Word for text documents (.doc) • These formats are in widespread use, and are accepted standards in this research discipline. • Data volume: about 50 KB of data, in digital format, will be collected. Storage space is not an anticipated problem.
Documentation and Metadata	
What documentation and metadata will accompany the data?	• The data captured is not self-explanatory. • A project information sheet, methodology description including description of research setting and data collection instruments, need to accompany the data. These documents will be captured in MS Word. • Metadata will be created. Dublin Core to be used as metadata standard.
Ethics and Legal Compliance	
How will you manage any ethical issues?	• Informed consent will be gained from participants prior to data collection. • This project has been given CSIR as well as University of Pretoria ethical clearance.

	• Confidentiality: The project will be clearly explained to each participant. Personal details will not be collected from participants. • Personal details or information accidently revealed via responses to open-ended questions will be anonymised and de-identified. • Participants will be asked to sign a consent form which specifies what data will be collected and how it will be managed and used
How will you manage copyright and Intellectual Property Rights (IPR) issues?	• The research data is owned by the University of Pretoria. • Anonymised data will be made freely available after completion of the studies and project.
Storage and Backup	
How will the data be stored and backed up during the research?	• Storage space is not an anticipated problem. • Data will be stored at the CSIR, on the hard drive of the researcher's office computer. Data are also stored on the E-surv platform. All data are password-protected. • The CSIR-based data is automatically backed up by CSIR IT-services. Data is also backed up to the researcher's personal external hard drive, and password-protected. • Responsibility for data storage and back up is with the principal researcher.
How will you manage access and security?	• User ID/password required to access personal CSIR computer drive. • User ID/password required to access backed up data in all instances • User ID/password required to access data stored on E-surv platform. • Security of sensitive data: this is not an anticipated issue. However, data will be anonymised and de-identified, should it be necessary. Any mention of personal grievance issues or identifiable research practices will not be included in the anonymised data.
Selection and Preservation	
Which data should be retained, shared, and/or preserved?	• Data will be kept for ten years. • Data will be kept on a CSIR personal drive, with a backup on a CSIR drive as well as an external device. • Metadata related to the dataset will be added to the CSIR's Technical Outputs Database (TOdB) as well as the CSIR's Institutional Repository (ResearchSpace). • Data will be added to an accredited data archive as soon as a suitable platform has been identified. • Data will be of interest to researchers in the field of Information Science.
Are any restrictions on data sharing required?	• Data will be shared, freely and publicly. • Email as well as an online platform (the CSIR's institutional repository: ResearchSpace) will be the data sharing media.
Responsibilities and Resources	
Who will be responsible for data management?	• Principal investigator
What resources will you require to deliver your plan?	• No additional resources required

Appendix 2: Questionnaire outline and informed consent form

Informed consent form

- I hereby voluntarily grant my permission for participation in the project as explained to me by Louise Patterton
- The nature, objective, possible safety and health implications have been explained to me and I understand them.
- I understand my right to choose whether to participate in the project and that the information furnished will be handled confidentially. I am aware that the results of the investigation may be used for the purposes of publication.

Multiple choice, one answer only, 'yes' and 'no' are the options

Questions

1. CSIR unit *textbox*

2. Please indicate the academic discipline your Phd, or intended PhD, is part of:
   - Humanities (e.g. History, Linguistics, Arts, Philosophy, Religion)
   - Social Sciences (e.g. Anthropology, Archaeology, Economics, Geology, Psychology)
   - Natural Sciences (e.g. Biology, Chemistry, Physics, Space Sciences)
   - Formal Sciences (e.g. Mathematics, Computer Sciences, Statistics, Logic)
   - Other

   Multiple choice, one answer only, text box for 'other'

3. What types of research data do you create or work with as part of your research?
   Select all that apply:
   - Documents (text, PDF, Microsoft Word, etc.)
   - Spreadsheets (e.g. Excel)
   - Images (pictures, photos, etc.)
   - Audio
   - Video
   - Databases (e.g. Access, MySQL, Oracle)
   - Geospatial data
   - Slides, artefacts, specimens, samples
   - Websites
   - Raw data files generated by software, sensors, or instrument files
   - Models, algorithms and scripts
   - Questionnaires, transcripts, codebooks
   - Notebooks, diaries
   - Contents of an application (input, output, log files for analysis software, simulation software, schemas)
   - Collection of digital objects acquired and generated during the process of research
   - Other

   Multiple choice, select all applicable answers, text box for 'other'

4. Please estimate the volume of research data across all of your CSIR work:
   - less than 1 GB
   - 1-50 GB
   - 51-100 GB
   - 101-500 GB
   - 501 GB – 999 GB
   - 1-50 TB
   - 51-100 TB
   - more than 100 TB
   - I don't know

   Multiple choice, one answer only

5. Which applications are used for analysis or manipulation of your data?
   (Please select all that apply)
   - Access
   - Adobe Photoshop
   - ArcGIS
   - Eviews, ImageJ
   - Excel
   - FileMakerPro
   - Galaxy
   - Gaussian
   - Labview
   - Matlab
   - Minitab
   - MS Word
   - NVIVO
   - R
   - SAS
   - SigmaPlot
   - SPSS
   - Stata
   - Other

   Multiple choice, select all applicable answers, text box for 'other'

6. Have you developed or submitted a research data management plan for any of your projects?
   - Yes
   - No
   - I don't know

   Multiple choice, one answer only

7. Are you aware of any policy or requirements from your funder/s regarding research data management?
   - Yes, I am aware and there are requirements
   - No, I am not aware, there might be requirements
   - Not applicable....my funders have no RDM requirements

   Multiple choice, one answer only

8. Where is your research data stored?
   Select all that apply:
   - Hard disk drive of office computer/laptop
   - Hard disk drive of home computer/laptop/tablet
   - Hard disk drive of instrument which generates data
   - Shared drive
   - Server in unit

- External hard drive/USB/Flash drive
- Web-based service, e.g. Dropbox, Flickr, Google Docs
- CD/DVD
- On paper
- NAS (Network-attached storage)
- Discipline-specific repository
- Email client, e.g. Groupwise
- Other (please specify below)

Multiple choice, select all applicable answers, text box for 'other'

9. How frequently is your research data backed up?
   - Daily
   - Weekly
   - Monthly
   - Annually
   - Ad-hoc
   - Don't know
   - Never

   Multiple choice, one answer only

10. If data is backed up, where is it backed up?
    (please select all that apply)
    - External hard drive
    - CSIR drive
    - NAS (Network-attached storage)
    - Server in unit
    - CD/DVD
    - Server managed by CSIR ICT
    - USB/memory stick
    - External/cloud e.g. Dropbox, Google Docs (please specify in 'Other')
    - Don't know
    - Other
    - Not applicable, data not backed up

    Multiple choice, select all applicable answers, text box for 'other'

11. Do you document or record any metadata about your data? (Metadata is data about research data, making it more meaningful or easier to search for)
    - Yes
    - No
    - Sometimes
    - I don't know

    Multiple choice, one answer only

12. If metadata is added, do you use any standards or guidelines?
    - Not applicable
    - Yes (please specify below)
    - No
    - Sometimes (please specify below)
    - I don't know
    - Not applicable: metadata is not added

    Multiple choice, one answer only. Text box for 'yes' and 'sometimes'

13. Who owns the Intellectual Property Right for your research data?
    - CSIR

- University
- Funder
- Don't know
- Other

Multiple choice, one answer only, text box for 'other'

14. Is your data subject to confidentiality rules or sensitive data restrictions?
    - Yes
    - No
    - Unsure

    Multiple choice, one answer only

15. What steps have you taken, or will you be taking, to ensure the privacy of your data?
    - I obtained informed consent from the participants to share my data
    - I anonymized/de-identified sensitive or confidential data
    - I destroyed the data within a short time of publishing research results
    - I destroy all data within a specified time period
    - No steps are taken to ensure privacy
    - Other
    - Not applicable; data is not sensitive or confidential

    Multiple choice, select all applicable answers, text box for 'other'

16. Data sharing entails sharing informally with researchers, as well as more formal sharing such as data repositories, data banks and data centres, or submission to a journal to support publication.

    With whom do you share your data? Who can typically access the research data you are creating? *(select all that apply)*
    - Researchers who help create the data
    - Others in the research unit/research group
    - Others within the CSIR
    - Study supervisor/s
    - Others in the discipline/field
    - Funders
    - Journal publishers
    - General public
    - No-one
    - Other

    Multiple choice, select all applicable answers, text box for 'other'

17. In the last 5 years have you received requests from other researchers for access to your data?
    - No, never
    - Yes, once
    - Yes, 2-5 times
    - Yes, 6-10 times
    - Frequently (more than 10 times)

    Multiple choice, one answer only

18. How often have you been able to provide access to your data?
    - Never
    - Less than half the time
    - About half the time or more
    - Always

Multiple choice, one answer only

19. If the answer to the previous question (question 18) was 'never', please explain why.

Text box provided

20. Which methods/infrastructures have you used to share your data? *(select all that apply)*
    - A web portal for download access
    - Provided online contact information for data requests
    - File transfer protocol (FTP)
    - CD/DVD
    - E-mail
    - A curated digital data repository
    - Flashdrive/USB stick
    - Other

Multiple choice, select all applicable answers, text box for 'other'

21. If making use of a curated digital data repository, please provide more details on its location, curator, characteristics, users and usage

Text box

22. In the last 5 years, have you asked other researchers to gain access to their research data?
    - No, never
    - Yes, once
    - Yes, 2-5 times
    - Yes, 6-10 times
    - Often: I make use of such data regularly

Multiple choice, one answer only

23. If the answer to the previous question (question 22) was 'no, never'. Please explain why.

Text box provided

24. After publishing your research results, where do you preserve or store most of the data you produce?
    - I do not do anything special to store my data after publishing
    - I store only a part of my data because of storage limitations
    - I store my data locally for my access, e.g. on my office computer
    - I store my data on a server, repository or database at the CSIR
    - I store my data on a digital repository or database outside the CSIR
    - Other

Multiple choice, one answer only, text box for 'other'

25. Which of the following data management tasks do you generally perform? *(select all that apply)*
    - Organise my data into structured data files with clear labelling
    - Ensure that my data is stored in a format that makes them easy to share
    - Ensure that my data is stored in a format that is standard in my field/ discipline
    - Prepare information material (documentation or metadata) describing my methods of data collection, sampling, testing and processing
    - Document and keep a copy of the computer code used to process and analyse my data
    - Create metadata files in an accepted standard
    - Ensure storage and backup of my original raw data files (the master version)
    - Ensure storage and backup of my most recent data files
    - Maintain an inventory of the various versions of my data and documentation files
    - Maintain an inventory of the locations of the various versions of my data and documentation files

- Other

Multiple choice, select all applicable answers, text box for 'other'

26. Have you ever received any research data management training?
    - Yes (please specify below)
    - No
    - I cannot remember

    Multiple choice, one answer only, with text box for explanatory information if 'yes' was chosen

27. Please indicate the areas you would be interested receiving data management training in:
    - Developing a research data management plan
    - Documenting your data
    - Formatting your data
    - Storing your data
    - Sharing your data
    - Creating metadata for data
    - Ethics and consent
    - Funders requirements and research data management
    - Copyright and Intellectual Property Right
    - Data Repositories and Open Access
    - Data citation
    - Not interested in receiving training
    - Other (please specify):

    Multiple choice, select all applicable answers, text box for 'other'

28. Please rate the importance of the following data-related services to help improve research in your discipline:
    - Having increased access to research data in my discipline
    - Having increased access to cross-disciplinary data, or data from a variety of sources
    - Guidelines and services supporting researchers in 'managing' their research data
    - Guidelines and services supporting researchers in 'depositing' their research data
    - Guidelines and services supporting researchers in creating metadata (for describing data)
    - Having teaching and learning materials so that researchers can work with data
    - Having the services necessary to assign a permanent digital object identifier (DOI) to my data to help others to find and cite my data
    - Having the infrastructure at the CSIR to allow CSIR data to be preserved, and made available to others

    All services listed have to be rated as either 'not important', 'somewhat important', 'important', 'very important', or 'unfamiliar with this service'

29. Please rate the importance of the following standards, policies, principles and practices in helping CSIR researchers to manage their data.
    - Stated timelines by funding agencies for depositing data
    - Available time to implement research data management plans, and manage data
    - Available funds to implement research data management plans, and manage data
    - Discipline-specific metadata standards and easy-to-use metadata tools
    - Compulsory submittal of a research data management plan
    - User-friendly tools/assistance to support research data management
    - CSIR ethics policies pertaining to research data sharing and confidentiality
    - CSIR data repositories where researchers can deposit data
    - One-stop data access/discovery across CSIR disciplines
    - Steps and assistance with data securing and data anonymization
    - Data citation impact factor or recognition (e.g. promotion, grants) when my data is cited
    - An accepted national standard for data citation and related support services
    - Having dedicated data curation experts to curate research data

All concepts listed have to be rated as either 'not important', 'somewhat important', 'important', 'very important', or 'unfamiliar with this service'

30. Please state any data-related service/policy/practice, not mentioned in questions 28 and 29, deemed important to you when managing research data.

Open-ended question with text block

31. Please state any other data-related concern, issue or problem, not already covered in this survey:

Open-ended question with text block

Appendix 3: Ethics approval as stated by the University of Pretoria

Reference Number: EBIT/57/2015 29-Jul-2015

Louise LH Patterton
Information Science
UNIVERSITY OF PRETORIA

Dear Patterton,

FACULTY COMMITTEE FOR RESEARCH ETHICS AND INTEGRITY

Your recent application to the EBIT Ethics Committee refers.

1. I hereby wish to inform you that the research project titled *"Research data management practices of emerging CSIR researchers"* has been approved by the Committee.

This approval does not imply that the researcher, student or lecturer is relieved of any accountability in terms of the Codes of Research Ethics of the University of Pretoria, if action is taken beyond the approved proposal.

2. According to the regulations, any relevant problem arising from the study or research methodology as well as any amendments or changes, must be brought to the attention of any member of the Faculty Committee who will deal with the matter.

3. The Committee must be notified on completion of the project.

The Committee wishes you every success with the research project.

Prof. J.J. Hanekom
Chair: Faculty Committee for Research Ethics and Integrity

FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY

Appendix 4: Ethics approval as stated by the CSIR

CSIR Research Ethics Committee

PO Box 395 Pretoria 0001 South Africa
Tel: +27 12 841 4060
Fax: +27 12 841 2476
Email:R&DEthics@csir.co.za

Permission Certificate

18 September 2015

Dear: **Ms Louise Patterton**

Title: **Survey of Research Data Management Practices of Emerging CSIR Researchers (CSIR REC Ref: 141/2015)**

Thank you for submitting your application to the CSIR Research Ethics Committee (REC). The CSIR REC notes the University of Pretoria (UP) ethical clearance and grants permission for the study to proceed as per the UP clearance.

We wish you all of the best with your research project.

Kind regards

Dr Clemence Tarirai
(CSIR REC Vice-Chair)

Appendix 5: Memorandum to all CSIR Research Unit Directors

To: All Research Unit Directors
From: Martie van Deventer
Unit name: CSIR Library and Information Services (CSIRIS)
Subject: Survey into the Research Data Management (RDM) practices of emerging CSIR researchers
Date: 6 April 2015

With all NRF-funded research data now required to be deposited in an accredited Open Access repository, and responsible parties at the CSIR currently drafting a data management policy, procedures and guidelines, CSIR Library and Information Services (CSIRIS) are planning to conduct a **survey into the Research Data Management (RDM) practices of emerging CSIR researchers.** This survey will also be the focus of a master's degree study in Information Science at the University of Pretoria, the student being Louise Patterton, a permanent CSIR employee.

Information gathered from our emerging researchers regarding their RDM practices is a crucial piece of information as it is anticipated that a better understanding of the RDM habits, trends, expectations and needs of the CSIR's emerging researchers could assist in ensuring that all research staff would benefit from the formal procedures that should result from the study.

Data will be collected using an online questionnaire circulated to all emerging researchers in the CSIR. For the purposes of this study, an emerging researcher is defined as:

- a researcher employed full-time by the CSIR,
- currently aged 35 or younger, and
- in possession of a PhD degree, or currently registered for a PhD degree.

The online questionnaire consists of 31 questions, will take about 20 minutes to complete, and will be distributed during May 2015. A document detailing the questionnaire structure and format of questions, is attached. Anonymity will be guaranteed, and respondents are under no obligation to supply personal details. Emerging researchers are free to refuse to take part in the survey. Should they decide to take part, they will be free to opt out at any time.

Five emerging researchers from your research unit will be involved.

Your permission as Director of (unit's name) is required in order to proceed with this survey.

Please note that if no response is received by 10 May 2015, it will be regarded as permission to proceed.

Kind regards

Martie van Deventer
Portfolio Manager: CSIRIS
Tel: 012 841 3278
Cell: 082 924 6650

Appendix 6: First contact with emerging researchers

Dear emerging researcher

A few days from now you will receive, via email, a request to complete an online questionnaire for an important research project being conducted by CSIR Information Services.

This project is focussed on the research data management (RDM) behaviours of emerging CSIR researchers. For the purposes of this project, a CSIR emerging researcher is identified as a permanent employee of the CSIR, aged 35 years and younger, and currently busy with, or in possession of a PhD.

I am writing in advance because we have found that many people like to know ahead of time that they will be contacted. The study is an important one that will help the CSIR understand what young researchers are currently doing with their research data, what their RDM-related training requirements are, and what services they would like to see implemented to enable quality management of research data. This study, investigating the RDM practices of emerging researchers at a research institute, is also the first of its kind in South Africa, and will act as a benchmark for future similar studies.

Thank you for your time and consideration. It is only with the generous help of people like you that this type of behaviour-related research can be successful.

Sincerely

Louise Patterton
Research Data Librarian
CSIR Information Services
Pretoria

South Africa

Tel: 012-8413767/073 169-0935

Appendix 7: Second contact with emerging researchers

Dear emerging researcher

I am writing to ask your help in a study investigating the research data management (RDM) behaviours of emerging CSIR researchers. This study is part of an effort to learn more about the data management practices of CSIR scientists, the data-related training requirements indicated by them, as well as the research data services deemed important by the CSIR's emerging researchers.

With all NRF-funded research data now required to be deposited in an accredited Open Access repository, and responsible parties at the CSIR currently drafting a data management policy, procedures and guidelines, CSIR Library and Information Services (CSIRIS) felt the need to conduct a **survey into the Research Data Management (RDM) practices of emerging CSIR researchers.** This survey will also be the focus of a master's degree study in Information Science at the University of Pretoria, the student being Louise Patterton, a permanent CSIR employee.

Information gathered from our emerging researchers regarding their RDM practices is a crucial piece of information as it is anticipated that a better understanding of the RDM habits, trends, expectations and needs of the CSIR's emerging researchers could assist in ensuring that all research staff would benefit from the formal procedures that should result from the study.

For the purposes of this study, an emerging CSIR research is identified as a permanent CSIR employee, age 35 years and younger, and in possession of, or currently busy with a PhD degree. It is my understanding, after studying a researcher list supplied by CSIR Human Resources, that you meet all requirements to be considered an emerging CSIR researcher. I am contacting all CSIR emerging researchers, and asking them to complete an online questionnaire consisting of 31 questions related to research data management practices. The survey should take about 20 minutes to complete and can be located at this address: http://esurv.org/surveyEditor.php?survey_ID=LHIIMF_e2e40c29 .

Your answers are completely confidential and the data released, as well as results made public, will not be identifying individual respondents or an individual's answers. In addition, the online survey software, eSurv, guarantees that online responses are visible to the study researcher, only. Completion of the questionnaire, as well as the completion of individual questions, is voluntary. However, you can help my very much by taking a few minutes to share your data management habits and requirements. If at all possible, please complete the questionnaire by 30 June 2015.

Ethics approval has been obtained. The director of your research unit has also given the go-ahead for this study.

Results of this study will be made available to all emerging CSIR researchers.

If you have any questions or comments about this study, I would be happy to talk with you. My email address is lpatterton@csir.co.za; alternatively I can also be contacted at 012-8413767 or 0731690935.

Thank you very much for helping with this important study.

Sincerely,

Louise Patterton
(Research Data Librarian)
CSIR Information Services
P.O. Box 395
Pretoria
0001

Tel: 012-8413767/073 1690935

Appendix 8: Third contact with emerging researchers (Reminder/thank you letter)

Dear emerging researcher

Last week a questionnaire seeking your responses to questions pertaining to your research data management practices was sent to you. Your name was selected as you were identified as an emerging CSIR researcher; the target population being investigated in this study.

If you have already completed and returned the questionnaire, please accept my sincere thanks. If not, please do so today. I am especially grateful for your help because it is only by asking CSIR researchers like you to share your practices, training requirements and service needs that CSIR management, CSIR ICT and CSIRIS can understand what needs to be done to support researchers in managing their research data.

It might be possible that you did not receive my previous letter with a link to the online questionnaire; the online questionnaire can be accessed at: https://eSurv.org?s=LHIIMF_e2e40c29

Sincerely

Louise Patterton
Research Data Librarian
CSIR Information Services
CSIR
Pretoria
South Africa
tel: 012-8413767/0731690935

Appendix 9: Fourth contact (final email)

Dear emerging researcher

During the last month I have sent several mailings about an important research project being conducted by the CSIR Library and Information Services.

Its purpose was to help understand the research data management practices of emerging CSIR researchers, and identify the services they require to enable quality management of their research data.

The study is drawing to a close, and this is the last contact that will be made with the sample of people identified as being emerging researchers employed by the CSIR.

I am sending this final contact as I am concerned that researchers who have not responded, might have different data management experiences that those who did. Hearing from everyone in this small CSIR sample helps assure that the survey results are as accurate as possible. Access to the online survey can be found here:
https://eSurv.org?s=LHIIMF_e2e40c29

I also want to assure you that your response to this study is voluntary, and if you prefer not to respond, that's fine. If you are not an emerging CSIR researcher (not 35 years and younger, and not in possession of or busy with a PhD), and you feel that I have made a mistake including you in the study, please let me know. That would be very helpful.

Finally, I appreciate your willingness to consider this request as CSIRIS concludes this effort to better understand data management practices, concerns and needs facing emerging researchers.

Thank you very much

Sincerely

Louise Patterton
Research Data Librarian
CSIR Information Services
CSIR
Pretoria
South Africa
tel: 012-8413767/0731690935

Appendix 10: Dataset

dataset.xlsx

Appendix 11: Data documentation

DATA DOCUMENTATION

Project background

Management of research data is globally being seen as part of good research practice. As a result of this, funders are increasingly insisting on proof of good research data management (RDM) practices when funding proposals are submitted. This study aimed at establishing the data management practices of emerging researchers at the Council for Scientific and Industrial Research (CSIR), South Africa. With no official RDM procedures currently being implemented at the CSIR, it was hoped that by gaining information about the RDM practices of emerging CSIR researchers, as well as insight into the RDM challenges experienced by them, this researcher would be able to put forward recommendations enabling the establishing of an RDM regime at the CSIR.

This study was the topic of a master's dissertation in the Department of Information Science at the University of Pretoria, South Africa.

Aims and objectives

The study aimed at answering several research questions. The main research question was:

How can an organisation like the CSIR ensure that future researchers apply best practices when managing the CSIR's research data?

Five research sub-questions were identified:

1. What are the international RDM requirements, standards, best practices and expectations that are being developed?
2. What are the current research data management practices among all researchers: CSIR, nationally, internationally?
3. What data are collected and held by emerging researchers in the CSIR?
4. What are the current RDM practices and trends among emerging researchers in the CSIR?
5. What are the RDM-related challenges, issues and concerns facing emerging researchers at the CSIR?

Investigator

Louise Patterton (position: CSIR data librarian)
Contact: lpatterton@csir.co.za

Population studied

This study aimed at identifying the RDM practices of emerging researchers at the CSIR. For the purposes of this study, an emerging research was identified as a permanent employee of the CSIR, 35 years of age or younger, and in possession of a PhD-degree or currently busy with doctoral studies.

Data collecting

This study was a survey, and data was collected via an online questionnaire.

Sampling design

Total population sampling was used. This researcher obtained a name list of emerging researchers at the CSIR from the CSIR Human Resources Division. A decision was made to use total population sampling, meaning the all CSIR emerging researchers would be contacted via email, informed about the study and survey tool, and asked to complete the online questionnaire.

Instruments used

The study made use of Esurv as online survey tool. The survey tool was created by the study investigator. The survey consisted of 31 questions, related to RDM practices, RDM needs and RDM challenges. An introductory letter and electronic letter of consent also formed part of the online survey.
Respondents accessed the online survey via a link sent to them in an email.

Software used

Esurv (http://esurv.org/) was used as online survey tool. To view the online survey:


survey questions PDF.docx

Calculations were performed by the online software tool, Esurv.

Results were exported to a spreadsheet. Microsoft Office is required when viewing the dataset; the dataset is saved in Excel.

Data files

Data for this study comprises one data file, an Excel spreadsheet. The set is 32.3 KB in size.

Data validation and cleaning

Data was checked and cleaned. Email addresses were removed from the dataset to ensure data anonymity.

Data confidentiality, access, use

Data will be placed in the institutional repository of the CSIR (ResearchSpace) after publication. It will be freely available to the public for reuse and sharing. All data have been anonymized.

Dataset details

The dataset is in spreadsheet format, with file type being Microsoft Excel. To view the dataset:


dataset.xlsx

The dataset contains the responses of 48 emerging researchers. Respondents are listed in column A, and responses are listed in columns to the right of column A.

The dataset variables are explained in the table below:

Column ID: the column as found in the spreadsheet

Question number: the corresponding survey question (see 'survey questions' inserted earlier under 'Software used'

Description of variable: clarifying information on the survey variable being investigated, as well as the responses given

Table 1: Explanation of dataset variables and values

COLUMN ID	SURVEY QUESTION	DESCRIPTION OF VARIABLE
the column as found in spreadsheet	corresponding survey question no.	clarifying information on the survey variable being investigated, as well as the responses given
A	**n/a**	**Respondent ID as generated by online software (an 8-digit number)**

B	**n/a**	**The number 1 is shown as indication of 'yes' to informed consent**
C	**1**	**CSIR unit**
D	**2**	**Indication of academic discipline the respondent's PhD is part of:** 1. Humanities 2. Social sciences 3. Natural sciences 4. Formal sciences 5. Multidisciplinary/other
E		Clarifying details if previous answer was 'multidiscipinary/other'
F - V	**3**	**Types of research data created or worked with as part of research**
F		Documents (text, PDF, Microsoft Word)
G		Spreadsheets
H		Images
I		Audio
J		Video
K		Databases
L		Geospatial data
M		Slides, artefacts, specimens
N		Websites
O		Raw data files generated by software/sensors
P		Models. algorithms
Q		Questionnaires, codebooks
R		Notebooks, diaries
S		Contents of an application
T		Collection of digital objects acquired
U		Other
V		Clarifying information
W	**4**	**Volume of research data across all of respondent's CSIR**

		work 1. less than 1 GB 2. 1-50 GB 3. 51-100 GB 4. 100-500 GB 5. 501-999 GB 6. 1-50 TB 7. 51-100 TB 8. more than 100 TB 9. I don't know
X-AQ	**5**	**Software applications used**
X		Access
Y		Adobe Photoshop
Z		ArcGIS
AA		Eviews, ImageJ
AB		Excel
AC		Filemaker Pro
AD		Galaxy
AE		Gaussian
AF		Labview
AG		Matlab
AH		Minitab
AI		Microsoft Word
AJ		NVIVO
AK		R
AL		SAS
AM		SigmaPlot
AN		SPSS
AO		Stata
AP		Other
AQ		Clarifying information
AR	**6**	**Submission of RDM plan** 1. Yes 2. No

		3. I don't know
AS	**7**	**Awareness of funder RDM policy** 1. Yes, I am aware and there are requirements 2. No, I am unaware 3. Not applicable, there are no requirements
AT-BH	**8**	**Where is your research data stored?**
AT		Hard disk drive of office PC/laptop
AU		Hard disk of home PC/laptop/tablet
AV		Hard disk drive of instrument which generates data
AW		Shared drive
AX		External hard drive/USB/Flash drive
AY		Web-based service
AZ		CD/DVD
BA		Paper
BB		Network-attached storage
BC		Server in unit
BD		Discipline-specific repository
BE		Email client
BF		Vibe
BG		Other
BH		Clarifying information
BI	**9**	**Frequency of data backup** 1. Daily 2. Weekly 3. Monthly 4. Annually 5. Ad-hoc 6. Don't know 7. Never
BJ-BU	**10**	**Data backup location**
BJ		External hard drive
BK		CSIR drive

BL		Network-attached storage
BM		Server in unit
BN		CD/DVD
BO		Server managed by ICT
BP		USB/memory stick
BQ		External (dropbox, google docs)
BR		Don't know
BS		Other
BT		Data not backed up
BU		Specify other or external (clarify BQ answer)
BV	**11**	**Metadata creation** 1. Yes 2. No 3. Sometimes 4. I don't know
BW-BX	**12**	**Metadata standard adherence**
BW		1. Yes 2. No 3. Sometimes 4. I don't know 5. Not applicable, metadata not added
BX		Clarifying information
BY-BZ	**13**	**Intellectual Property Rights Owner**
BY		1. CSIR 2. University 3. Funder 4. Don't know 5. Other
BZ		Clarifying information
CA	**14**	**Data confidentiality** 1. Yes 2. No

		3. Unsure
CB-CJ	**15**	**Data privacy steps**
CB		Informed consent
CC		Data anonymization/de-identification
CD		Data in secure setting
CE		Data destruction after publishing
CF		Data destruction within specified time
CG		No steps taken
CH		Other
CI		Not applicable, data is not sensitive
CJ		Clarifying information
CK-CU	**16**	**Data sharing**
CK		Researchers who helped create data
CL		Others in research group/unit
CM		Others in CSIR
CN		Study supervisors
CO		Others in discipline/field
CP		Funders
CQ		Journal publishers
CR		General public
CS		No-one
CT		Other
CU		Clarifying information
CV	**17**	**Data sharing requests received in last 5 years** 1. No, never 2. Yes, once 3. Yes, 2-5 times 4. Yes, 6-10 times 5. More than 10 times
CW	**18**	**Providing access to data** 1. Never

		2. Less than half the time 3. Half the time or more 4. Always
CX	**19**	**Clarifying 'never' in previous question response**
CY-DG	**20**	**Data sharing methods/infrastructure**
CY		Web portal for download access
CZ		Online contact information provided
DA		FTP
DB		CD/DVD
DC		E-mail
DD		Curated digital data repository
DE		USB stick
DF		Other
DG		Clarifying information
DH	**21**	**More details on 'curated digital data repository' in previous response**
DI	**22**	**Requesting secondary data during last 5 years** 1. No, never 2. Yes, once 3. Yes, 2-5 times 4. Yes, 6-10 times 5. Often, I use such data regularly
DJ	**23**	**Explanation of 'no, never' in previous question**
DK-DQ	**24**	**Where is data preserved after publishing?**
DK		Nothing done with data after publishing
DL		I store part of data (storage limitations)
DM		I store my data locally for own access
DN		I store my data on a server, repository or database at CSIR
DO		I store my data in a repository outside CSIR
DP		Other

DQ		Clarifying information
DR-EC	**25**	**RDM tasks performed**
DR		Organise data into structured files with labelling
DS		Data stored in sharable format
DT		Data stored in format standard to discipline
DU		Prepare metadata and documentation
DV		Document computer code used
DW		Create metadata in accepted standard
DX		Ensure storage and backup of raw data
DY		Ensure storage and data of most recent files
DZ		Maintain inventory of various data versions
EA		Maintain inventory of various data locations
EB		Other
EC		Clarifying information
ED-EE	**26**	**Data management training received** 1. Yes 2. No 3. Cannot remember
EF-ES	**27**	**RDM areas interested receiving training in**
EF		Developing a data management plan
EG		Documenting data
EH		Formatting data
EI		Storing data
EJ		Sharing data
EK		Creating metadata for data
EL		Ethics and consent
EM		Funder requirements
EN		Copyright, IPR
EO		Data repositories, Open Access
EP		Data citation
EQ		Not interested
ER		Other areas

ES		Clarifying information
ET-GG	**28**	**Rating the importance of data-related services**
ET		Not important: Increased access to data in my discipline
EU		Somewhat important: Increased access to data in my discipline
EV		Important: Increased access to data in my discipline
EW		Very important: Increased access to data in my discipline
EX		Unfamiliar: Increased access to data in my discipline
EY		Not important: Increased access to cross-disciplinary data
EZ		Somewhat important: Increased access to cross-disciplinary data
FA		Important: Increased access to cross-disciplinary data
FB		Very important: Increased access to cross-disciplinary data
FC		Unfamiliar: Increased access to cross-disciplinary data
FD		Not important: RDM guidelines and supporting services
FE		Somewhat important: RDM guidelines and supporting services
FF		Important: RDM guidelines and supporting services
FG		Very important: RDM guidelines and supporting services
FH		Unfamiliar: RDM guidelines and supporting services
FI		Not important: Data depositing guidelines
FJ		Somewhat important: Data depositing guidelines
FK		Important: Data depositing guidelines
FL		Very important: Data depositing guidelines
FM		Unfamiliar: Data depositing guidelines
FN		Not important: Metadata guidelines and services
FO		Somewhat important: Metadata guidelines and services
FP		Important: Metadata guidelines and services
FQ		Very important: Metadata guidelines and services
FR		Unfamiliar: Metadata guidelines and services
FS		Not important: Teaching/learning materials to work with data
FT		Somewhat important: Teaching/learning materials to work with data
FU		Important: Teaching/learning materials to work with data
FV		Very important: Teaching/learning materials to work with data
FW		Unfamiliar: Teaching/learning materials to work with data
FX		Not important: DOI services

FY		Somewhat important: DOI services
FZ		Important: DOI services
GA		Very important: DOI services
GB		Unfamiliar: DOI services
GC		Not important: Preservation services
GD		Somewhat important: Preservation services
GE		Important: Preservation services
GF		Very important: Preservation services
GG		Unfamiliar: Preservation services
GH-IT	**29**	**Rating the importance of RDM standards, policies and practices**
GH		Not important: Timelines by funders for data deposit
GI		Somewhat important: Timelines by funders for data deposit
GJ		Important: Timelines by funders for data deposit
GK		Very important: Timelines by funders for data deposit
GL		Unfamiliar: Timelines by funders for data deposit
GM		Not important: Time for data management plans, and RDM
GN		Somewhat important: Time for data management plans, and RDM
GO		Important: Time for data management plans, and RDM
GP		Very important: Time for data management plans, and RDM
GQ		Unfamiliar: Time for data management plans, and RDM
GR		Not important: Funds for data management plans, and RDM
GS		Somewhat important: Funds for data management plans, and RDM
GT		Important: Funds for data management plans, and RDM
GU		Very important: Funds for data management plans, and RDM
GV		Unfamiliar: Funds for data management plans, and RDM
GW		Not important: Discipline-specific metadata standards, metadata tools
GX		Somewhat important: Discipline-specific metadata standards, tools
GY		Important: Discipline-specific metadata standards, metadata tools
GZ		Very important: Discipline-specific metadata standards, metadata tools

HA		Unfamiliar: Discipline-specific metadata standards, metadata tools
HB		Not important: Compulsory submittal of RDM plan
HC		Somewhat important: Compulsory submittal of RDM plan
HD		Important: Compulsory submittal of RDM plan
HE		Very important: Compulsory submittal of RDM plan
HF		Unfamiliar: Compulsory submittal of RDM plan
HG		Not important: User-friendly RDM tools
HH		Somewhat important: User-friendly RDM tools
HI		Important: User-friendly RDM tools
HJ		Very important: User-friendly RDM tools
HK		Unfamiliar: User-friendly RDM tools
HL		Not important: CSIR ethics policy pertaining to RDM
HM		Somewhat important: CSIR ethics policy pertaining to RDM
HN		Important: CSIR ethics policy pertaining to RDM
HO		Very important: CSIR ethics policy pertaining to RDM
HP		Unfamiliar: CSIR ethics policy pertaining to RDM
HQ		Not important: CSIR data repositories
HR		Somewhat important: CSIR data repositories
HS		Important: CSIR data repositories
HT		Very important: CSIR data repositories
HU		Unfamiliar: CSIR data repositories
HV		Not important: One-stop data access across CSIR disciplines
HW		Somewhat important: One-stop data access across CSIR disciplines
HX		Important: One-stop data access across CSIR disciplines
HY		Very important: One-stop data access across CSIR disciplines
HZ		Unfamiliar: One-stop data access across CSIR disciplines
IA		Not important: Assistance with data securing/anonymization
IB		Somewhat important: Assistance with data securing/anonymization
IC		Important: Assistance with data securing/anonymization
ID		Very important: Assistance with data securing/anonymization
IE		Unfamiliar: Assistance with data securing/anonymization
IF		Not important: Data citation impact factor
IG		Somewhat important: Data citation impact factor

IH		Important: Data citation impact factor
II		Very important: Data citation impact factor
IJ		Unfamiliar: Data citation impact factor
IK		Not important: Accepted standard for data citation
IL		Somewhat important: Accepted standard for data citation
IM		Important: Accepted standard for data citation
IN		Very important: Accepted standard for data citation
IO		Unfamiliar: Accepted standard for data citation
IP		Not important: Data curation experts
IQ		Somewhat important: Data curation experts
IR		Important: Data curation experts
IS		Very important: Data curation experts
IT		Unfamiliar: Data curation experts
IU	**30**	**Other RDM service/policy/practice deemed important**
IV	**31**	**Any other data-related concern not already covered**

METADATA

Abstract

48 Emerging researchers from the Council for Scientific and Industrial Research (CSIR), South Africa completed a survey investigating their research data management (RDM) practices. RDM practices investigated included the use of data management plans, data storage and backup locations, creation of metadata, metadata standard adherence, and data sharing practices. Challenges faced when managing research data, as well as RDM needs and requirements, also formed part of the survey. Results of the online questionnaire revealed that the RDM practices of the group studied do not show to differ significantly from experienced CSIR researchers, or from researchers surveyed elsewhere on the globe. Findings enabled this researcher to put forward several recommendations which would assist in the implementing of a formalised RDM structure at the CSIR. Recommendations addressed, but were not limited to: formalization of RDM procedures, RDM marketing, and RDM training.

Keywords

Research data

Research data management

RDM

Researcher behaviour

Data management plan

Collection date

Data was collected over a 6 week period, starting September 2015 and finishing October 2015.

Location

CSIR, South Africa