Readiness for Research Data Management in the Life Sciences at the University of the Witwatersrand

Salomé Potgieter: PTGSAL001

A minor dissertation submitted in *partial fulfillment* of the requirements for the award of the degree of Master of Library and Information Studies

Faculty of the Humanities

University of Cape Town

2021

COMPULSORY DECLARATION

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signature: Signed by candidate Date: 17 December 2021

DEDICATION

To Suzy, my dearest sister who is no longer with us. Thank you for believing in me – this is for you.

ACKNOWLEDGEMENTS

I am grateful to my Heavenly Father, who gave me the opportunity, ability and strength to undertake this study and to persevere despite so many challenges along the way.

I would further like to extend my gratitude to the following people, without their help I would not have been able to finish this dissertation:

My supervisor, Michelle Kahn, for her excellent guidance and unwavering support throughout this study.

University of the Witwatersrand for financial support, study leave and for allowing me to conduct my research amongst researchers at the University.

Researchers in the Schools of Animal, Plant and Environmental Sciences and Molecular and Cell Biology that took part in this study, without your participation I would not have been able to do my research.

My colleagues, Mr Paiki Muswazi and Maryna van den Heever, who had the faith in me to complete my studies, as well as Nina Lewin for all the conversations we could have around the topic of Research Data Management.

Vanessa-Lynn Neophytou, for editing my thesis in such a thorough and patient way.

John Kaye, from Jisc, for the insight he shared with me regarding the Jisc Research Data Lifecycle.

My mom and dad, although no longer with us, for all their love and support through many years.

All my family and friends, there are so many of you, who kept believing in me and encouraged me to see this through.

ACRONYMS

ACRL	Association of College and Research Libraries
APES	Animal, Plant & Environmental Sciences
ARDC	Australian Research Data Commons
DataONE	Data Observation Network for Earth
DCC	Digital Curation Centre
DMP	Data management plan
DOI	Digital object identifier
ELN	Electronic laboratory notebook
EMBL-EBI	European Bioinformatics Institute
EML	Ecological Metadata Language
EnvO	Environment Ontology
EPSRC	Engineering and Physical Sciences Research Council
FAIR	Findable, Accessible, Interoperable & Reusable
GCD	Global Charcoal Database
GEO	Gene Expression Omnibus
GIS	Geographical information system
HEI	Higher Education Institution
HPC	High-performance computing
ICT	Information and communications technology
IP	Intellectual property
IR	Institutional repository
IT	Information technology

MCB	Molecular & Cell Biology
MRCSA	Medical Research Council of South Africa
NCBI	National Centre for Biotechnology Information
NCBO	National Centre for Biomedical Ontology
NRF	National Research Foundation
NSF	National Science Foundation
NYU	New York University
OA	Open access
OECD	Organisation for Economic Co-operation and Development
OSF	Open Science Framework
PDB	Protein Data Bank
RDM	Research Data Management
RDS	Research data services
REDcap	Research Electronic Data Capture
RO	Research Office
SAEON	South African Environment Observation Network
Wits	University of the Witwatersrand

ABSTRACT

Because of the importance of Research Data Management (RDM) in the life sciences, where vast amounts of research data in different complex formats are being produced, this study aimed to assess the state of RDM readiness in the life sciences at Wits to ascertain what support is needed with regards to RDM. In order to achieve the aim, the current RDM practices and needs of researchers, as well as the challenges they face, were investigated.

The Jisc Research Data Lifecycle (Jisc, 2021a) was used to guide the literature review, frame data collection, analyse data and advise on some of the main findings and recommendations.

A mixed methods approach and an explanatory sequential design were used to achieve the research objectives. For the quantitative phase of research, an online questionnaire was used to collect data. As the total target population (282) was not big, a census was conducted. The questionnaire was administered using SurveyMonkey software. During the qualitative part of the research, semi-structured interviews were used to explain the quantitative results. Five participants were purposively sampled to take part in interviews. The statistical package, MS Excel, was used to analyse quantitative data whilst qualitative data were analysed by thematic analysis.

The study showed that life sciences researchers at Wits have adopted many RDM practices, and researchers are increasingly becoming aware of the importance of the openness of data. However, they are dealing with similar RDM issues as their peers worldwide. Results highlighted challenges of, amongst others, the lack of an RDM policy as well as the lack of, or unawareness of, appropriate RDM training and support at Wits. As formal implementation of RDM still needs to take place at Wits, it is recommended that Wits puts an RDM policy in place, followed by suitable RDM infrastructure and awareness making of current services.

Key words: Research data management, RDM, data management plans, DMPs, policy, academic libraries, life sciences, researchers

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction and background to the study

Data has been described as the primary building block of information in any field of knowledge (Pryor, 2012: 2). Research data, to be more specific, is “any information that has been collected, observed, generated or created to validate original research findings” (Jisc, 2021b: para.3).

The creation of the internet and the World Wide Web last century transformed the world (Benioff, 2015), including research. Developments in computational simulation and modelling, automated data acquisition and communication technologies resulted in vast amounts of research data being collected and analysed (Tenopir et al., 2011). Research data has increased exponentially in the 21st century (Pryor, 2012: 2; Ray, 2014: 3). The data deluge has been described as the most challenging aspect of 21st-century research administration (Pryor, Jones & Whyte, 2014: vii). The massive increase in research data created a need to develop policies, infrastructure and services within institutions to manage data over the long term (Pinfield, Cox & Smith, 2014: 2; Ray, 2014: 3).

As research became more data intensive and collaborative over the past two decades, data sharing has also become important (Tenopir et al., 2011), leading to the Open Science Movement. Open Science supports the open access (OA) publishing of research articles, but also extends to the open publishing of datasets, workflows and details of research processes (Grand, Bultitude & Winfield, 2010). There is a growing recognition of the benefits of sharing data, such as verifying research outcomes and facilitating data reuse (Pinfield, Cox & Smith, 2014: 2; Tenopir et al., 2011). Data sharing can also increase the return on large investments made in research and should ultimately result in the advancement of human knowledge being in the service of the public good (Ray, 2014: 1).

The acknowledgement of the importance of the openness and sharing of data has seen a change in the attitude of major funders towards data and research output – they are now viewing data as an important asset that needs attention and proper management (Pryor, 2012: 4). Funders started instituting data policies, especially in cases where research is publicly funded (Jones, 2012: 47). In 2007, the Organisation for Economic Co-operation and Development (OECD) issued *OECD Principles and Guidelines for Access to Research Data from Public Funding* (OECD, 2007). The main theme of these principles is “the notion that publicly funded research is a public good, produced in the public interest, and should be openly

available to the maximum extent possible" (Jones, 2012: 47). In South Africa, the National Research Foundation (NRF) published a statement on OA in March 2015. Part of the statement encouraged authors of research papers either partially or fully funded by the NRF to deposit the data supporting their publications in an accredited OA repository (NRF, 2015). This statement was an important event locally regarding Research Data Management (RDM) as it required researchers and research organisations to re-evaluate and develop their roles in terms of data curation and management (Patterton, Bothma & van Deventer, 2018: 14).

According to McLure et al. (2018a: 8), 'data curation' ensures data "are retrievable for future research or reuse". Research data management can be defined as the "organization, storage, preservation, and sharing of data collected and used in a research project" (Bordelon, 2021: para.1) – the activities that result in curated data.

The effective management of data is crucial for many reasons, one being that research data is the evidence base of research findings. Research data are not only the result of the time invested in conducting research but often that of a significant amount of funding.

Institutions can benefit from good RDM as it can ensure compliance with funders' data policies, ensure that research data are safe and secure, and increase research efficiency and improve research integrity. Other important reasons for good RDM are ensuring research reproducibility and facilitating online sharing (Jisc, 2021b). In order for research data to be of current and future value, they need to be managed so that they are discoverable, accessible and reusable (Procter, Halfpenny & Voss, 2012: 135). Providing support services towards effective RDM requires defining role-players and roles and relevant organisational processes and technological infrastructure (Pryor, Jones & Whyte, 2014: 41).

1.1.1 RDM in the life sciences

Although an increase in research data can be found in all disciplines, the most dramatic increase is occurring in the sciences (Pryor, 2012: 2). Researchers are increasingly dealing with large volumes of information about humans, animals or microbes and sometimes need to handle massive datasets (Li & Chen, 2014: 187). Biologists, for example, are being described as part of the "big-data club" (Marx, 2013), with many scientists working with large volumes of data as a result of the rapid advance in high-throughput technologies (Altaf-Ul-Amin et al., 2014; Li & Chen, 2014: 187). In the field of genomics, it has been predicted that by 2025, around 2–40 exabytes of annual storage capacity for human genetic data would be needed (Stephens et al., 2015: 5). The challenge with big data for these scientists is not only the size of the data but also the increasing complexity of it (Altaf-Ul-Amin et al., 2014; Griffin et al.,

2018: 15). According to Griffin (2018: 15), best practice data management, including associated challenges, is highly domain-specific even within the life sciences, which is the subject area being investigated for this study.

The life sciences are confined to the scientific investigation of living organisms as well as living processes and include subject fields such as biology, botany, zoology, microbiology, physiology and biochemistry ("Life Sciences", 1993; "Life Sciences", 2010). Within the life sciences, the data deluge has led to an urgent need to understand complex and global data phenomena and improve data management (Thessen & Patterson, 2011: 15). Best practice data management is critical to ensure the quality and interoperability of shared life science data (Griffin et al., 2018: 15).

Requirements for data storage vary among sub-disciplines in the life sciences, and cause different challenges. Some disciplines, such as research that involves taxa with large genomes, require large storage space (Griffin et al., 2018: 14). In research with human data, data privacy may have specific requirements, for example, the requirement of local storage and access controls (Griffin et al., 2018: 14). Data types in the life sciences also pose data management challenges as they range from highly structured data to complex images and textual data.

Major funders of research in the life sciences, the focus area of this study, in South Africa, include the NRF and the Medical Research Council of South Africa (MRCSA). International funders include the Royal Society (UK), the Andrew Mellon Foundation and the Carnegie Corporation of New York (Davids, personal communication 2019, March 18). Each of these funders has a data policy to which grant holders must adhere.

1.1.2 Institutional context: University of the Witwatersrand

The University of the Witwatersrand (Wits) is a research-intensive university and one of the leading research institutions in Africa (Wits, 2010: 7, Wits, 2018b: 4). Over a period of time, Wits earned international recognition for its academic excellence and exceptional research (University of the Witwatersrand, 2020). During the period 2015–2019, research output at the university increased by 50% (Wits, 2019a: 5). The university has 5 faculties and 36 schools. In 2019 the student population consisted of approximately 40 800 students, and more than a third of the student body were postgraduate students (Wits, 2020a).

Within Wits, the broad subject area of the life sciences falls largely under two faculties, namely the Faculty of Science and the Faculty of Health Sciences. The focus for this study was on the

Faculty of Science both because the researcher serves this faculty as a librarian and because she wanted to ensure a manageable target population. The Faculty of Science consists of nine schools covering the broad areas of mathematical, physical, biological and earth sciences. Research and tuition in the life sciences mainly fall under two schools in the faculty, namely the School of Animal, Plant & Environmental Sciences (APES) and the School of Molecular & Cell Biology (MCB) (Wits, 2020b).

The School of APES had its centenary celebrations in 2017. The school conducts a wide variety of research, with some of the most critical research areas being African ecology, herpetology, higher plant systematics research, insect physiology, plant biotechnology, and the biological control of alien weeds and dung (Wits, 2020c). The school houses the SARChI chair for Global Change & Systems Analysis as well as the African Ecology and Conservation Biology Research Group (Wits, 2018c). The research output of the school, as measured by the Scopus database, included 122 publications in 2019 (Elsevier, 2019). Postgraduates in the school included the following enrolments in 2019: 73 MSc. (coursework and research report), 80 MSc. (research) and 54 PhD students (van Tonder, personal communication 2021, December 13).

The School of MCB is home to the SARChI Chair in Protein Biochemistry and Structural Biology (Wits, 2018c). Other research focus areas include cell biology research that looks into cellular events that trigger gene expression in relation to cell growth, differentiation and control; research on mechanisms that regulate the expression of genes that play a role in cancer and innate immunity; collaborative research in cassava looking into its genetic engineering, food quality, pest and disease resistance traits (Wits, 2020d). The research output of the school, as measured by the Scopus database, included 29 publications in 2019 (Elsevier, 2019). Postgraduates in the school included the following enrolments in 2019: 44 M.Sc. (research) and 46 PhD students (Botes, personal communication 2021, December 1).

1.1.3 RDM at Wits

The Wits RDM journey started in 2012 when the Wits library appointed a Data Services Librarian. Over time, generic RDM services began being introduced in the library, including metadata creation services, consultative services on developing Data management plans (DMPs), and advisory services on choosing and depositing data in suitable repositories. RDM Libguides were created to inform users about data curation good practices.

Besides the library, RDM stakeholders at Wits include the Research Office (RO) and the eResearch Office (including IT entities involved with big data and high-performance computing [HPC]).

The RO is involved with the management of grants, and the university's chief data protection officer resides under the RO. The RO plays a key role in linking researchers, university management and funders (Lewin, personal communication 2021, September 28).

The eResearch Office falls under Wits Information and Communications Technology (ICT). The main purpose of the eResearch unit is to support the five faculties as well as intra-faculty research institutes in their attempts to use large and complex data sets. eResearch services include IT guidance, data management planning, centralised infrastructure sharing, archiving and providence, and data reuse (Wits, 2021).

The library, eResearch Office and the RO collaborate in providing guidance for data management planning in the grant writing process (Lewin, personal communication 2019, March 18).

Despite the fact that RDM services have been introduced at Wits and RDM stakeholders have been identified, no formal RDM needs assessment has been conducted amongst researchers and RDM has not been formalised at the university.

Wits does not have a policy that explicitly addresses the management of research data though the university has indicated in different ways that it is important to have a policy that informs RDM. There are however RDM-related policies at the university which include an OA Policy (Wits, 2018d), an Institutional Repository (IR) Policy (Wits, 2008), an Intellectual Property (IP) Policy (Wits, 2012), as well as an Information Classification and Handling Policy (Wits, 2019b). In terms of research integrity, which underpins the ethical management of research data, the university subscribes to the research integrity standards as set out in the Singapore Statement on Research Integrity (Wits, 2020e; World Conferences on Research Integrity, 2010). RDM stakeholders at the university are considering RDM policy in line with privacy legislation (Lewin, personal communication 2019, March 18), such as the Protection of Personal Information Act (POPIA) that came into effect on the 1st July 2020 (The Presidency, n.d.).

The library's 2017–2022 Strategic Plan (Wits, 2018a: 6) mentions its participation in developing and implementing RDM policy. The mission of the university's *Strategic Plan for supporting eResearch Information Systems* is to "provide the infrastructure, policy framework,

technical support and data handling/manipulation training to manage digital data for the benefit of most Wits researchers in an unequal world" (Wits, 2017: 11).

1.2 Problem statement

Though there has been extensive discussion around RDM at Wits, RDM has not been formally adopted at the institution. While, over the past few years, stakeholders have developed data support services, these services are general, catering broadly to the Wits research community. The institution has not investigated the RDM practices and requirements of researchers in different fields that would allow them to tailor support services where necessary. Because of the special importance of RDM in the life sciences subject area, where vast amounts of research data in different complex formats are being produced, this study will address the gap in the institution's knowledge about the readiness of life sciences researchers to manage their data, which would assist in identifying their RDM needs and ultimately result in appropriate support being offered.

1.3 Study aim and research questions

1.3.1 Study aim

The aim of the research was to assess the state of RDM readiness in the life sciences at Wits to ascertain what support is needed regarding RDM.

1.3.2 Research questions

The study attempted to answer the following research questions in order to achieve its aim:

- What are the current RDM practices and needs of researchers in the life sciences at Wits?
- What are the RDM challenges researchers in the life sciences at Wits face?

1.4 Rationale/motivation of the study

According to Patterton, Bothma and van Deventer (2018: 16), information about the RDM habits of South African researchers in Science, Engineering and Technology (SET) is scarce. A study of the RDM practices and requirements of researchers in the life sciences will therefore contribute to the body of literature in this subject area and fill a gap in RDM literature pertaining to South Africa that includes a wider spectrum of life sciences sub-disciplines.

1.5 Overview of the research methodology

A mixed methods approach employing an explanatory sequential design was chosen for this study. A survey method in the form of a questionnaire was used for the quantitative part of the study, where a census was conducted amongst the 282 researchers within the Schools of APES and MCB at Wits. Semi-structured interviews then took place as part of the qualitative component of the study for which five researchers were purposively sampled.

1.6 Study delimitations

The study was delimited to Wits, and within Wits, the study population was further confined to researchers in the Schools of APES and MCB.

1.7 Research report structure

The following five chapters cover the study:

Chapter 1: Introduction – This chapter comprises the introduction and background to the study, the problem statement, aim and research questions, motivation for the study, as well as study delimitations and an overview of the research methodology.

Chapter 2: Literature review – This chapter provides an overview of important studies; model guiding the study; RDM practices, needs and challenges of life sciences researchers across the research data lifecycle and RDM in Higher Education Institutions (HEIs), role-players, infrastructure and services.

Chapter 3: Research methodology – This chapter comprises the chosen worldview, research approach and design of the study. This is followed by research methods used as well as methods used for population and sampling. Ethical considerations and a discussion about data collection, data analysis and interpretation follow. The chapter concludes with a discussion on validity and reliability in mixed methods design.

Chapter 4: Data analysis and presentation – Quantitative results are first presented, followed by a presentation of qualitative results that attempt to explain quantitative results.

Chapter 5: Data interpretation, recommendations and conclusion – The final chapter comprises mixing the results to summarise findings, followed by a section on the RDM readiness of researchers in the life science at Wits. Recommendations, study limitations and suggestions for further studies are also provided.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The topic of RDM has been written about for longer than a decade. As the life sciences cover many sub-disciplines, not all material on RDM in the life sciences is included in this literature review. Rather, the chapter focuses on the different aspects of the research data lifecycle as practised by researchers within the life sciences, with a focus on the sub-disciplines residing under the Schools of APES and MCB at Wits. The review mainly covers material published over the past decade. However, the main focus is on literature published within the past five years as data practices, needs, and challenges might have changed due to rapid changes in the scientific research environment. Because of this study's focus on support services, this chapter also reviews some of the literature on RDM stakeholders.

2.2 Overview of important studies

Some studies were particularly significant within the literature because of their focus on RDM practices, needs and challenges in the life sciences. These studies were by Johnson and Steeves (2019), Saaed and Ali (2019), RIN & British Library (2009), Kvale (2012), Griffin et al.(2018) and Thessen and Patterson (2011).

Studies consulted also included those that focus on specific sub-disciplines of the life sciences. While several studies on data practices and needs of researchers have been conducted in the many sub-disciplines of the life sciences, significant to this study were those in the areas of environmental sciences and biology (Alves et al., 2018; Kim, 2021; Renaut et al., 2018).

Two local RDM studies were particularly significant in terms of data practices of researchers. Koopman and de Jager's (2016) paper, from Koopman's (2015) dissertation, investigated the data management and archiving habits of researchers in the biological sciences, whilst a study by Patterton, Bothma and van Deventer (2018) – also emanating from dissertation research (Patterton, 2017), investigated the data management habits of experienced researchers and emerging researchers (including those in the life sciences) at a large research council. A survey method was used for both of the above studies.

Numerous studies looked into RDM practices, needs and challenges of researchers where the population consisted of researchers from different subject fields. Some of these studies were consulted, but the researcher attempted to focus on studies where the life sciences component of respondents was substantive, for example, studies that were conducted across

multiple institutions or worldwide, such as those by Tenopir et al. (2011, 2015, 2020) and Zhu (2019).

2.3 Model guiding the study

Studies conducted to investigate the RDM practices of researchers often use data lifecycle approaches as frameworks. The use of a lifecycle to approach the management of digital data ensures continuity of service throughout the lifecycle, despite technological and organisational changes that may influence data provenance (Pennock, 2007: 2).

Examples of data lifecycle models include the Research Data Lifecycle from the UK Data Archive Service (2019), Data Life Cycle from the Data Observation Network for Earth (DataONE) (2020), the Digital Curation Centre (DCC) Curation Lifecycle Model (2019) and the Research Data Lifecycle (Jisc, 2021a). While intended for life science researchers, the Data Life Cycle framework for bioscience, biomedical and bioinformatics data (Griffin et al., 2018: 4) focuses on bioinformatics. This lifecycle was described in Griffin et al.'s (2018) review paper of best practice life cycle approaches in the life sciences.

In light of the above this study, the Research Data Lifecycle (Jisc, 2021a) (Figure 2.1) was used to guide the literature review, data collection instruments, data analysis and some of the main findings and recommendations.

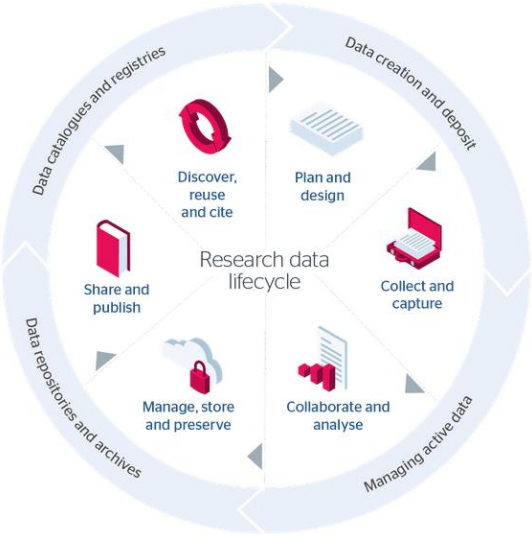

Figure 2. 1: Research Data Lifecycle

Source: Jisc (2021a)

The main sequential steps of the Jisc Research Data Lifecycle are summarised as follows (John Kaye, Jisc, personal communication 2021, October 13):

- 'Data creation and deposit' involves the collection of new or existing data according to a DMP and depositing it into a place where one can manage active data.
- 'Managing active data' involves managing data in such a way that it allows data analysis and collaboration.
- 'Data repositories and archives' firstly involves the selection of data that would be needed to reproduce findings and secondly adherence to FAIR (Findable, Accessible, Interoperable & Reusable) criteria for storing or sharing of data, such as in institutional and other repositories, for example Zenodo and figshare. Important processes in this stage include assigning of metadata, preservation activities and curation decisions.
- 'Data catalogues and registries' involves moving data from the repository space into a data discovery space; this is not always a physical move as some repositories are also registries. Catalogues and registries are aggregations of datasets, for example, Google Dataset Search, Australian Research Data Commons (ARDC), re3data and DataCite Search. The step includes publishing and making datasets openly available as part of the scholarly publications ecosystem and includes data citation.

Each step has related stages (plan and design; collect and capture; collaborate and analyse; manage, store and preserve; share and publish; and discover, reuse and cite). These six stages will be unpacked in relation to the life sciences in the sections that follow.

2.4 RDM in the life sciences: practices, needs and challenges across the research data lifecycle

2.4.1 Plan and design

Planning for RDM forms an important foundation for a research project. It determines how data will be organised, stored and preserved, all of which should be done according to FAIR principles (Jisc, 2021c). The planning phase is the first stage of the research data lifecycle, and according to Michener (2015), it might be considered an ongoing activity throughout all stages of the lifecycle. A plan can be seen as a tool to communicate requirements, changes and restrictions to role-players throughout the data life cycle (DataONE, 2021a).

At the 'Plan and design' stage, it is important to consider what requirements funders have in terms of RDM, for example, the requirement for a DMP (Jisc, 2021c; Renaut et al., 2018: 404). The creation of a DMP forms an important component of this stage as it describes how the

researcher will manage and document data throughout the project to ensure long-term preservation (Jisc, 2021d).

According to the literature, the requirement for a DMP among life sciences researchers is becoming more common. Tenopir et al. (2020: 19), for example, found that the primary funder of 52% of researchers in ecology/environmental science and of 41% of researchers in biology require submission of a DMP.

Johnson and Steeves (2019) conducted a quantitative study that investigated the actual and perceived RDM needs of life sciences researchers at New York University (NYU) to see how the library could best contribute towards RDM services. The study showed that for 57% of researchers, a DMP is a funder requirement, whilst 66% of researchers said that funders or journal publishers required that they share their data when published (Johnson & Steeves, 2019: 8). Saaed and Ali's (2019) study that investigated the perceptions of scholars in the life sciences and social scientists towards research data management and sharing, found that 86% of life sciences researchers confirmed having to write DMPs and the reasons given by most for not using a DMP, was "not knowing how to make it" followed by "not knowing what it is", "thinking that it is not necessary" and "not having time for it" (Saeed & Ali, 2019: 294). Patterton's (2017: 223) study, which focused on the data management habits of emerging researchers at a South African research council, is in contrast with the above studies as it found that researchers were not aware of RDM requirements from funders and for that reason did not create and submit DMPs. Patterton's study took place shortly after the NRF's OA statement (NRF, 2015), and therefore, researchers were not yet that aware of the need for a DMP.

Institutional policies relevant to RDM should also be considered at the RDM planning phase (Jisc, 2021c). Renaut et al. (2018: 407) stated that it is important that institutions play a leadership role in mandating RDM. A lack of institutional mandates for RDM may result in data not being managed and, therefore, not being preserved and made accessible for future research (Halbert, 2013a: 7). The DataRes Project survey (Halbert, 2013b: 117) in the United States found that most researchers were unaware of their institutions having an RDM policy, though a large majority agreed that an institutional policy for RDM would be valuable.

2.4.2 Collect and capture

During the data collection and capture stage, it is important to organise and document data as it is collected and decide where the collected data should be kept during analysis (Jisc, 2021e). When documenting data, it is crucial to add metadata in order for data to be found,

contextualised and reused (Jisc, 2021e). When collecting and capturing data, it is also necessary to consider file formats to ensure future compatibility and data reuse, as well as file naming and folder structure, as this can assist in locating information (Jisc, 2021e).

Diepenbroek et al. (2014: 1713) describe data collection as a crucial step in the data life cycle as errors that occur at the point of data collection, such as missing data or metadata, can be difficult and expensive to overcome later on in the data life cycle.

The nature, type and volume of research data vary in life sciences disciplines and can influence the way in which data need to be managed, including data collection and capturing. It is therefore important to know how these data characteristics feature in the life sciences.

2.4.2.1 Generation of research data

The RIN & British Library's (2009) study, that looked at patterns of information use and exchange among researchers in life sciences, found that novel data collection/generation in the life sciences include many different methods such as experimental, remote sensing and sequenced data collection, as well as the collection of field and observational data. For the study, seven case studies were conducted across a diverse range of researchers and laboratories, and methodologies used included self-administered questionnaires, semi-structured interviews and focus groups (RIN & British Library, 2009).

According to Zozus (2017) methodology impacts the way in which data are collected and managed and ultimately influences data quality and reproducibility. Data collection in ecology and biodiversity, for example, use a variety of protocols to address many diverse topics such as marine ecosystems or genetics (Alves et al., 2018: 87).

2.4.2.2 Data capture

A review article by Thessen and Patterson (2011) focused on the technical and sociological issues facing researchers in the life sciences. The article highlighted RDM practices, challenges and needs of researchers in the life sciences at the time. It further described the lack of comprehensive standards for life sciences data as a major issue in the transition of life sciences to a more data-focused discipline (Thessen & Patterson, 2011: 15).

Two types of standards essential for the management of data are metadata and ontologies (Thessen & Patterson, 2011: 30). Griffin et al. (2018: 5) stated it is important to use controlled vocabularies that are built on ontologies. An ontology can be described as a formal statement of relationships between concepts that are represented by metadata (Thessen & Patterson,

2011: 31). Scientific metadata standards are formulated to document details of who collected the data, how it was collected and what the data content is about (Qin, 2013: 218). This is important for data management as it enhances data discovery and reuse.

Several ontological structures or standards are available for use in the life sciences. They include the NCBO (National Centre for Biomedical Ontology) BioPortal, EML (Ecological Metadata Language), Darwin Core, EnvO (the Environment Ontology), the Gene Ontology Resource and the Plant Ontology (EMBL, 2021; "Environment Ontology", n.d.; "Gene Ontology Resource", 2020; "Metadata Standards Directory - Life Sciences", n.d.; NCBO, 2021; Thessen & Patterson, 2011: 32;).

While the addition of metadata can make data more findable and usable, not all researchers assign metadata to their datasets. Patterton's (2017: 178) study found that about half of respondents (52%) assigned metadata. The study by Kvale (2019: 10) conducted at the Norwegian University of Life Sciences that investigated researchers' attitudes towards data sharing, as well as researchers' data storage and sharing routines, also found that about half of respondents assigned metadata in their research. The Tenopir et al. (2015: 9) study found the same - that about half of respondents used metadata to describe their data.

In terms of standards, Tenopir et al. (2020: 16) found that more than a third of respondents used metadata standards to describe their data, with almost an equal amount of respondents not using metadata standards and nearly a quarter saying they used "metadata standardized within my institution/lab". Metadata standards used by researchers included Dublin Core, EML and Darwin Core (Tenopir et al., 2020: 18).

A challenge reported by researchers in ecology was the lack of comprehensive metadata standards to describe the variety of data types in the ecology domain (Alves et al., 2018: 93). They also experienced the complexity of metadata standards as a challenge (Alves et al., 2018: 93).

2.4.2.3 Volume, nature and types of data

Handling of large datasets is common in the life sciences and has been described as a challenge in several life sciences sub-disciplines (Li & Chen, 2014: 187; RIN & British Library, 2009: 44). Data volumes and the complexity of data differ between sub-disciplines. For example, within botanical curation, data volumes are low compared with systems biology or genomics research (RIN & British Library, 2009: 34). Although research in proteomics and genomics are characterised by high volumes of data, data are largely standardised (RIN &

British Library, 2009: 34), making the management of these datasets simpler than, for example, in systems biology where there are large volumes of data which are also heterogeneous (RIN & British Library, 2009: 34). Biodiversity research, which brings together many facets of environmental research, is also characterised by large data volumes and vast data heterogeneity (Diepenbroek et al., 2014: 1711). Local studies however, found that very large data sets (more than 100 TB) were held by the smallest percentage of respondents (Koopman, 2015: 67; Patterton, 2017: 163).

The nature of biological data has been described as being hierarchical (data are generated at different levels ranging from molecules and cells to systems), heterogenous (data are generated using different methods), complex (data can be simultaneously recorded in the forms of multi-level information) and dynamical (biological processes or states change with conditions and over time) (Li & Chen, 2014: 188). Managing the variety and complexity of biological data has been described as a major challenge of 21st-century biology (Lin &. Wooley, 2005: 35).

A wide variety of data types are also noticeable in the life sciences, and researchers must be cognisant of different types of data collected as this will determine its management. Data types include quantitative data, images, clinical, laboratory-derived/experimental data, remote sensing, observational and field data (Diepenbroek et al., 2014: 1713; RIN & British Library, 2009: 11). Varieties of data include genome data, flora and fauna data, protein data and nucleotide sequences (Qin, 2013: 217). Life science studies have shown that the most-used data types are spreadsheets/tabular data, images and documents (Koopman, 2015: 68; Kvale, 2012: 49; Patterton, 2017: 160; Saeed & Ali, 2019: 160). Image files are the most common file type in the DataONE database, which houses earth and environmental data, making up 34% of all file types (DataONE, 2021b). Figure 2.2 shows the different data types found in the DataONE database (2021b).

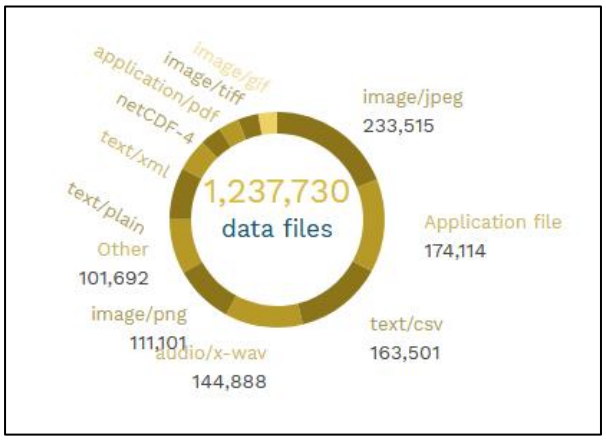

Figure 2. 2: DataONE data types

Source: (DataONE, 2021b)

2.4.3 Collaborate and analyse

Research is often conducted in groups that may be local, national or international. Collaborative research requires creating and documenting relevant data management processes before the start of a project (Jisc, 2021f).

With collaborations, the sharing of data files is important and dependent on how active data storage is managed (Jisc, 2021f). Good storage practices can prevent data loss and enable successful collaboration (Jisc, 2021g). Finkel et al. (2020: 641) stated that the effective management of research data in long-term and large-scale collaborative research is crucial for the success of such projects.

In data management, it is also essential to document the data analysis methods and protocols to ensure that data reuse amongst collaborators occurs smoothly (Jisc, 2021f). Data analysis requires considering important issues such as the handling of raw versus analysed data, the performance of quality control, spreadsheet best practices and the management of research codes (Briney, 2015: 12).

The reporting of the processing and analysis of data, including the software used, is crucial for the reproducibility of results (Griffin et al., 2018: 5). Not many studies have focused on software used by life science researchers to analyse and process data. Patterton (2017: 165) found that software used mainly by researchers were Microsoft Excel and Microsoft Word, whilst other software used included MATLAB, SigmaPlot, ArcGIS, R and ImageJ (Patterton, 2017: 165). R is free software used for statistical computing and graphics (R Foundation, 2021). A researcher in ecology described it as having “completely transformed the landscape for data management and analysis” (British Ecological Society, 2014: 24). ImageJ is a free, open-source image analysis program widely used in the biological sciences (Rueden et al., 2017: 1). ArcGIS is a geographical information system (GIS) used to visualise and analyse geographic data via contextual tools. It is maintained by Esri [Environmental Systems Research Institute] (Esri, 2021).

Researchers in the life sciences have expressed concern about analysing high quantities of varied data (RIN & British Library, 2009: 44), and thus this can be seen as a challenge.

2.4.4 Manage, store and preserve

Although managing and storing data are a priority during active stages of the research process, such as during data collection and analysis, they must also be considered beyond the research process, along with preservation (Jisc, 2021h).

2.4.4.1 Data storage (short term)

Tenopir et al. (2020: 9) described short-term data storage as storage of data "during the life of the project". Some of the aspects of data storage during the active part of a research project that should be considered for data management are the location of storage, the frequency of data backups as well as the location of backups (Jisc, 2021g).

Koopman & de Jager (2016: 5) found that most researchers (83%) used an external hard drive for storage of their data, followed by personal computers/laptops (55%). Cloud storage was also becoming more popular, with 39% of researchers making use of it. In Patterton's (2017: 172) study, the overwhelming majority of respondents (96%) used an office laptop/desktop as a storage method compared to 67% who used a hard drive/USB/flash drive and a third of respondents using cloud storage.

Tenopir (2020: 10) found that most researchers (61.3%) still store their data on a personal computer, followed by storage on an institution's server (42.9%), USB/external drive (29.8%) and cloud storage (12%). Storage in paper format decreased from 7.1% (Tenopir et al., 2015: 25) to 5.1% (Tenopir, 2020:10). Tenopir (2015: 14) found that researchers based in Africa are more likely to store their data in paper format than North American researchers, who were less likely to store data in paper format.

Storage is the most significant RDM need researchers have, including the amount of available space and the selection of storage space (Johnson & Steeves, 2019: 4).

When it comes to keeping data backups, the NYU study (Johnson & Steeves, 2019: 11) showed that 82% of researchers kept backups of their data other than through a data repository. Patterton (2017: 176) found that 92% of emerging researchers do keep backups. The most popular location for backups was an external hard drive (71%), followed by an organisational drive (46%) and cloud storage (29%). In Koopman (2015: 61), 94% of respondents (all biologists) confirmed keeping backups.

2.4.4.2 Data storage (long term)

Long-term data storage can be defined as storage “beyond the life of the project”, and good data practices can be described as practices that facilitate long-term preservation (Tenopir et al., 2020: 9). Storing data for the long term, for example, in a repository, would be partly for the intention of preservation. The terms ‘digital archiving’ and ‘digital preservation’ are often used interchangeably (Digital Preservation Coalition, 2015). Long-term preservation of data can be described as the actions and procedures needed to ensure the long-term sustainability of and access to data (Library of Congress, 2021; USGS, 2021). Data archiving can be described as the process whereby data that is no longer actively used is stored in such a way that it can be easily discovered and retrieved for future use (Renaut et al., 2018: 400).

Although it is important to store data for the long-term, it is impractical to keep all data, and thus a process of data selection should be followed (Jones, 2012: 100).

A drive towards preserving data for the long-term has been seen amongst some biology sub-disciplines, for example, the DataONE project and the National Centre for Biotechnology Information (NCBI) GenBank (Renaut et al., 2018: 400).

Tenopir et al. (2020: 11) found that storage in a repository (publisher, discipline-based, national data archive or institutional) accounted for 43.3% of researchers’ storage practices. The study also found that researchers in only four disciplines reported good long-term data storage practices, and of these, two were life sciences sub-disciplines, namely marine sciences (46.5%) and environmental science (28.3%) (Tenopir et al., 2020). However, only 11.4% of researchers in biology were found to have good data storage practices (Tenopir et al., 2020: 11). Johnson and Steeves (2019: 10) found that only 28% of life sciences researchers stored their data in a repository.

When investigating long-term storage/preservation practices of researchers, Patterton, Bothma and van Deventer (2018: 22) found that researchers did not use curated digital data repositories, and preservation activities did not take place. Koopman (2015: 45) found that 42% of researchers in biology did not archive their data for long-term purposes. Researchers who archived their data (58%) used a range of repositories, with GenBank and SAEON (South African Environment Observation Network) being the most used (Koopman & de Jager, 2016: 5).

It has been found that researchers archive their data in repositories for different reasons, such as funder and publisher requirements. It seems that unless data archiving is mandatory, researchers will not easily archive their data (Koopman & de Jager, 2016: 2).

Some studies tried to establish what researchers thought the purpose for the long-term preservation of data was. Responses by researchers in the life sciences included making sure data are available for future use, stimulating the advancement of science and allowing the re-analysis of existing data (Kvale, 2012).

Various repositories are available in the life sciences. In the area of molecular biology, bioinformatics organisations such as the NCBI and the European Bioinformatics Institute (EMBL-EBI) host public data repositories (Griffin et al., 2018: 5). Some of these repositories include Gene Ontology, GenBank, Gene Expression Omnibus (GEO) and Protein Data Bank (PDB). DataONE is a network of repositories in the earth and environmental sciences; Dryad Digital Repository is amongst numerous repositories in the network (Renaut et al., 2018: 402–403).

In some disciplines, such as molecular genetics and genomics, researchers are used to routine archiving in repositories such as EMBL-EBI (Laloë, 2017; Renaut et al., 2018: 407). Renaut et al. (2018: 400) stated that the following are barriers limiting data preservation: the lack of training and technological resources for data management and archiving, governing regulations, and individual views on data sharing. Locating a suitable repository is also problematic in some life sciences sub-disciplines; for example, scientists in agriculture and natural resources and in biology find it more onerous to locate a repository than other sub-disciplines (Tenopir et al., 2020: 14).

2.4.5 Share and publish

Data sharing can be described as the process that takes place when “scientists intentionally make their own data available to other people for their use in research or other related scientific endeavours” (Tenopir et al., 2015: 3). Sharing research data is recommended regardless of discipline and can be achieved in many different ways, such as the informal sharing of data during the duration of the research project or after the end of the project via repositories, data journals or as supplementary material to publications (Jisc, 2021i).

Data sharing in the life sciences goes back as far as 1996 when the principles for open sharing of data among the genomics community were agreed upon in Bermuda (Cook-Deegan & McGuire, 2017: 897). Representatives from DNA sequencing centres in five nations met and

agreed to release DNA sequence information daily, making data available to laboratories worldwide (Cook-Deegan & McGuire, 2017: 897). However, in practice, it was difficult to convince researchers to follow the principles of the Bermuda agreement (Zhu, 2019: 2). Indeed, Scaramozzino, Ramírez and McGaughey (2012) found that, in the life sciences, whilst most researchers acknowledged the importance of sharing data, far fewer researchers in practice share their data openly. A study conducted amongst UK academics (Zhu, 2019: 1) found that although academics recognised the importance of sharing data, most had never shared or reused data. Thirty-five per cent of respondents in this study were from the medical and life sciences disciplines, and 21% of respondents from these disciplines confirmed depositing primary research data in an online repository that allows reuse of data (Zhu, 2019: 8). In contrast, the NYU study (Johnson & Steeves, 2019: 10) found that 89% of researchers indicated that they either have shared, would share or are required to share data with other researchers.

In the area of environmental research, the Belmont Forum's survey on Open Data found that 82% of respondents agreed that open data were "very important" (Schmidt, Gemeinholzer & Treloar, 2016: 10). This correlates with the Tenopir et al. (2020: 15) study that showed that 96.1% of scientists in environmental science and ecology are willing to share their data with a broad group of researchers. The willingness to share data amongst other life sciences sub-disciplines in the Tenopir et al. (2020: 15) study included marine and ocean science (89.7% of respondents willing to share their data), biology (84.9%) and agriculture and natural resources (80.2%). A study by Herhold (2015: 1) found that ecology and evolution scientists shared their data at the highest rate (70% of their articles) in contrast with fisheries, wildlife and conservation biologists (18%).

A cross-sectional study that examined changes in the sharing and withholding practices among geneticists between 2000 and 2013 found that researchers showed a major shift in data sharing habits, away from a peer-to-peer sharing model towards sharing in central repositories (Zinner, Pham-Kanter & Campbell, 2016: 433).

Although the willingness to share is part of the ethos of life science, researchers choose what to share, with whom and when (RIN & British Library, 2009: 7). When investigating with whom researchers share their data, local studies (Bangani & Moyo, 2019: 11; Patterton, 2017: 187) found that the majority of scientists share their data either with researchers who helped to create their data, who were members of their research group or who were their supervisors. Koopman and de Jager (2016: 4) found that researchers in biology share their data in different ways, for example, on request via e-mail, within published papers (38%) and through

collaboration (15%). Only 12% of researchers shared their data through a repository. On the international front a study by Saaed and Ali (2019: 297) found that 29% of life sciences researchers shared their data via academic social networks, followed by sharing data with peers on request (24%) and publishing in a research journal (23%). Only 4% of researchers deposited data in open data repositories.

Another aspect of data sharing practices investigated in previous studies is the conditions under which researchers are willing to share their data for future use. The precondition for sharing data only post-publication was confirmed by Koopman and de Jager (2016) and Tenopir et al. (2015: 16, 2020: 16) and Schmidt, Gemeinholzer and Treloar (2016: 25), particularly in the area of environmental research. Koopman and de Jager (2016: 4) found that 62% of researchers in biology had this precondition. Tenopir et al. (2015: 16) found that the requirement first to publish was significantly more in disciplines such as biology and physical science. The same study also found that more researchers from Africa and Asia required permission from others to access their data than researchers from North America (Tenopir et al., 2015: 12).

Other pre-conditions for sharing were being offered co-authorship where one's data were used, citing or acknowledgement of a dataset, sharing of data only 'on request at my discretion' and having the opportunity to collaborate on a project using the data (Koopman & de Jager, 2016: 4; Tenopir et al., 2015: 11).

The literature has shown many barriers researchers encounter with the sharing of data. Zhu (2019: 1) stated that barriers in the full-scale adoption of data sharing are not only technical but also include social and cultural barriers. The RIN and British Library (2009: 7) study found that barriers to sharing data included concerns over potential misuse of data, and ethical restrictions and IP issues in terms of data. Researchers view data as a critical part of their 'intellectual capital', and they have reservations about the way and timing in which they share information (RIN & British Library, 2009: 7). Researchers are wary of sharing their data as someone else might analyse it and get the credit (RIN & British Library, 2009: 38).

Zhu (2019: 7) stated that one of the major barriers for academics not sharing their data was 'academic competition' as they rely mainly on their primary data to publish and for promotion. The most noted barriers to sharing data in Tenopir et al.'s study (2020: 16, 18) were: the need to publish first, inadequate time to make the data available, insufficient rights to make the data open, as well as the lack of funding. Zhu (2019: 6) identified additional challenges researchers had with data sharing – ethical issues and the time and effort required to produce data for

reuse. Furthermore, a lack of incentives and standards could also be barriers to data sharing (Zhu, 2019: 10). Apart from the need to publish first, the Belmont Forum's survey on Open Data also found concerns about legal issues and potential data misuse related to sharing data in the area of environmental research (Schmidt, Gemeinholzer & Treloar, 2016: 25).

Policies often drive data sharing, and data policies are more established in some disciplines than others, such as the biomedical sciences, where many journals require authors to share their primary datasets (Zhu, 2019: 2). Well-known journals that have mandated public data archiving include *Nature* and *PloS* (Koopman & de Jager, 2016: 2; Renaut et al., 2018: 406; Roche et al., 2015: 1).

Johnson and Steeves (2019: 8) found that 66% of life sciences researchers said that funders or journal publishers required that they share their data when published. In the area of environmental research, it was found that amongst different policies that influence the sharing of data, funder policies were seen as the most important motivator for sharing data (Schmidt, Gemeinholzer & Treloar, 2016: 1).

2.4.6 Discover, reuse and cite

This stage of the data lifecycle focuses on the researcher as a data user and can assist in understanding why the management of research data is important (Jisc, 2021j). In order to discover and reuse data, good RDM practices are necessary, such as the assigning of metadata and digital object identifiers (DOIs), earlier in the data life cycle (Jisc, 2021j). Thessen and Patterson (2011: 16) refer to data-driven discovery as the "discovery of scientific insights through the novel management and analysis of pre-existing data", which relies on the access to and reuse of data.

The literature on the reuse of data revealed that most academics have never reused or shared data, despite recognising the importance of data sharing (Zhu, 2019: 1). The reuse of data differs among life science sub-disciplines. Tenopir et al. (2020: 16), for example, found that 32.2% of respondents from environmental science/ecology and 27.7% of respondents from agriculture and natural resources used data generated by others. Only 16.7% of respondents in the field of biology were regular users of data collected by others. In the field of molecular biology, however, data-driven discovery has become an integral part of research (Thessen & Patterson, 2011: 16).

Researchers' opinions on the importance of access to data generated by others also differ, with researchers in the environmental sciences expressing a significantly higher agreement

with the statement that the lack of access to data generated by other researchers is a major impediment to progress in science (Tenopir et al., 2015: 15).

Different factors affect life sciences researchers' willingness to reuse data. According to Tenopir et al. (2020: 21), the most important criteria influencing scientists' confidence in reusing data are the use of metadata standards and the availability of complete provenance data. It has also been found that researchers are not willing to reuse data collected by other researchers, as there are differences in experimental design and data collection practices in the life sciences (RIN & British Library, 2009: 39).

2.5 RDM in HEIs: role-players, infrastructure and services

As universities are trying to deal with the challenges being posed by a dramatic increase in research data, they need not only to provide for technical infrastructure but also need the human capacity to support researchers in the management of data (Procter, Halfpenny & Voss, 2012: 136). Determining the level of RDM support needed is part of the aim of this study. Therefore, literature on RDM role-players, infrastructure and service is incorporated into this literature review.

2.5.1 Role-players

In order to support the management of a Higher Education Institution's (HEI) research data, a range of stakeholders is needed from both within and outside the institution (Pryor, Jones & Whyte, 2014: 43). In the majority of cases, it has been seen that support teams for managing research data at HEIs include the library, information technology (IT) as well as research administration (Pinfield, Cox & Smith, 2014: 1). While each has its own area of strength, in reality, RDM support that one provides overlaps with what the other provides. Stakeholders in RDM outside an institution may include external data centres and archives that facilitate the sharing of data (Pryor, Jones & Whyte, 2014: 57).

2.5.1.1 The library

Libraries have traditionally been seen as the providers of information literacy and have played a role in curation. Thus it is a natural step for the library to be involved with RDM-related activities such as training of researchers in managing their data, creating metadata and managing digital repositories (Latham, 2017: 263). 'Research data services' (RDS) was listed as one of the top trends in academic libraries by the Association of College and Research Libraries' (ACRL) Research and Planning Committee in both 2016 and 2020 (ACRL Research Planning and Review Committee, 2016, 2020). In the 2016 top trends report, it was stated that

libraries that are providing RDS have taken a traditional approach that aligns more with current liaison and outreach roles, with fewer libraries offering technical RDM support (ACRL Research Planning and Review Committee, 2016).

2.5.1.2 Information technology (IT)

The technological infrastructure and services needed for RDM require a wide range of technology for the collection, storing, processing, organising, transmitting and preservation of data. These can include networks, databases, authentication systems as well as software applications that are equipped to handle scientific data from different sources and data that comes in a variety of types and formats (Qin, 2013: 216). According to Yu (2017: 787), technical RDM services mainly include providing repository access, discovery systems, the preparation of data or datasets to be deposited into a repository as well as the creation or transformation of metadata. Patterton, Bothma and van Deventer (2018: 23) describe preservation assistance and data storage as important RDM activities in which IT can take the lead. Many of these technical activities would require working hand-in-hand with the library.

2.5.1.3 Research administration

Research administration can include divisions at HEIs that handle grant administration as well as units involved with commercialisation and innovation. These functions are often performed by the 'research office' (Pryor, Jones & Whyte, 2014: 48). The research office must ensure compliance with funder policies, including their RDM requirements.

2.5.1.4 Researchers

Kennan and Markauskite (2015: 70) expressed that the awareness of researchers' needs is essential for developing RDM policy and infrastructure. It is important to be aware that the responsibility of RDM lies with both the researcher and the institution (Singh, Monu & Dhingra, 2018: 113). Pryor, Jones & Whyte (2014: 49) stated that the active involvement of researchers in the development of RDM services is crucial as they would be the users of the service with specific motivations and priorities in terms of RDM.

2.5.2 RDM infrastructure and services

RDM stakeholders provide a range of RDM services in response to supporting researchers with data management during different points in the research cycle (Matusiak & Sposito, 2017: 754).

A study (Yu, 2017: 792) that reviewed RDS studies conducted since 2009 showed an increase in the scope and level of RDS offerings by academic libraries. The study revealed that services provided by libraries covered the data life cycle from research data planning up to data discovery. RDS offerings included consultation, training on data management planning, data guidance during research, research documentation and metadata, as well as data sharing and curation (Yu, 2017: 792).

Jones (2014: 89) suggests a broad governance framework for RDM. As can be seen in Figure 2.3, support services are categorised into: data management planning, managing active data, data selection and handover, and sharing and preserving data (Jones, 2014: 90), mirroring to some extent the Jisc Research Data Lifecycle discussed earlier. Guidance, training and support are needed for the uptake and use of these services. RDM policy, strategy and business case drive RDM infrastructure and services.

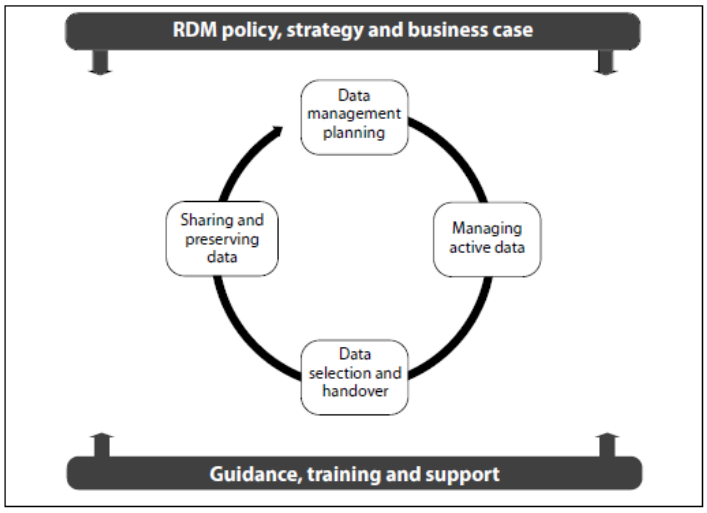

Figure 2. 3: Components of RDM support services

Source: (Jones, 2014: 90)

The following section discusses RDM support services according to the above components.

2.5.2.1 RDM or related policies

Qin (2013: 216) describes policy as one of three dimensions of data infrastructure as policies govern the management, use and sharing of data. The development of policies, strategies and procedures with regards to RDM can be seen as the first step towards delivering RDM support services (Pryor, Jones & Whyte, 2014: 91), and the lack of a well-defined RDM policy framework has been identified as a barrier for effective RDM (Chiware & Becker, 2018: 11).

Policies that influence and govern data management can be reviewed from a funder (international and national) and institutional or organisational perspective.

Data policies at the national and institutional level provide a framework for researchers to establish a routine of RDM practices (Qin, 2013: 18). The National Science Foundation (NSF), for example, requires a DMP that complies with the NSF policy on dissemination and sharing of research results (Radboud University, 2021), and the Engineering and Physical Sciences Research Council (EPSRC) has a policy framework on research data (EPSRC, 2021).

Institutional RDM policy development requires an awareness of the roles of all stakeholders as well as their needs and issues (Jones, Pryor & Whyte, 2013: 6).

2.5.2.2 Data management planning

Institutions can play a role in offering assistance with conceptualising data management and thus assist researchers with creating DMPs. Many libraries have taken on this role by providing consultancy services for DMPs (Latham, 2017: 263; Matusiak & Sposito, 2017: 755; Yu, 2017: 787). Libraries can give support for data management planning by providing templates of DMPs. They can also give advice on suitable data management tools such as DMPonline (DCC, 2020) from the DCC in the UK, the DMPTool (California Digital Library, 2020) used in the USA (Jones, 2014: 96) or the SA-DMP Tool (DIRISA, n.d.). A local study (Patterton, 2017: 201) found that training on creating DMPs was rated as being the most important RDM training requirement amongst scientists.

2.5.2.3 Managing active data

According to Jones, Pryor and White (2013: 13), the two main things to consider when delivering RDS during the active phase of research are to provide for adequate volumes of data storage and for suitable applications that can assist researchers to store, access and share their research when collaborating. Institutions need to advise researchers on recommended data storage and backup approaches (Jones, 2014: 96). Services for managing active data may also include HPC facilities and the facilitation of cloud storage services (Jones, Pryor & Whyte, 2013: 13). Those traditionally involved in ICT are usually recognised as the most suitable for the aforementioned support services (Pryor, Jones & Whyte, 2014: 48). RDS in this stage may also include metadata services giving advice to researchers during the active phase of research to ensure that data are accessible after the life of the project (Ray, 2014: 65). Matusiak and Sposito (2017: 755) stated that consultation on metadata is one of the RDS focus areas of libraries.

2.5.2.4 Data selection and handover

A selection process is necessary in order to retain data that needs to be preserved and remain accessible for future research (Jones, 2014: 100). Several data selection guidelines exist, such as the *How to Appraise and Select Research Data for Curation* publication (Whyte & Wilson, 2010) and the data value checklist from the Natural Environment Research Council (Jones, 2014: 101; NERC, 2019). Data handover typically occurs when data is being transferred to a repository for the long-term curation of data (Pryor, Jones & Whyte, 2014: 102). An RDS may include establishing guidelines, processes and best practices for data selection and handover (Jones, Pryor & Whyte, 2013: 14).

2.5.2.5 Sharing and preserving data

Once the decision has been made as to what data should be kept, a decision needs to be made as to how the data would be preserved and shared. There are several RDS that can support preservation and sharing.

Many data repository services exist, and this number is growing. Researchers often use a mix of institutional and external repository services (Pryor, Jones & Whyte, 2014: 103). Choosing where to deposit research data is critical for researchers as the correct choice can lead to increased reuse and citations (Jisc, 2021k) because the data have been preserved. Several websites assist with identifying general and discipline-specific repository options such as re3data and FAIRsharing (“FAIRsharing”, 2019; Jisc, 2021k). Institutional RDS may include developing and maintaining a data IR and guiding researchers on choosing suitable repositories for the long-term access to data (Jones, Pryor & Whyte, 2013: 18). Giving advice on the choice of a repository, as well as consultancy on data citation, archiving and sharing, are some of the RDS libraries have undertaken (Matusiak & Sposito, 2017: 755).

In order for data to be discovered and reused, it is important that the necessary metadata and persistent identifiers are assigned to datasets. RDS can include advising on assigning permanent identifiers to datasets (Ray, 2014: 65).

2.5.2.6 Guidance, training and support

As shown, there are many RDS that an institution can provide. Different groups might lead the services within the institution that provide different levels of guidance, training and support. However, it is important that there is a coherent vision across services delivered by different departments (Pryor, Jones & Whyte, 2014: 106). In a local study, the lack of RDM training was

identified as a major challenge that hinders effective RDM (Patterton, Bothma & van Deventer, 2018: 21).

2.6 Summary

A great deal of literature is available on RDM in the broad area of the life sciences. However, a review of the literature showed that RDM studies conducted amongst life sciences researchers in South Africa are limited. This chapter provided an overview of the RDM practices, needs and challenges of life sciences researchers based on stages of the Jisc Research Data Life Cycle. As the study also aimed to identify the support needed for researchers at Wits, the literature review provided an overview of RDM in HEIs and the infrastructure and services needed to support researchers throughout the data lifecycle.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research methodology used for this study. The research approach, the worldview underpinning it, and the research design are outlined. Research methods are discussed as well as the population and sampling methods used. Ethical considerations and a discussion about data collection, data analysis and interpretation follow and the chapter concludes with a section on ensuring the validity and reliability of the study.

3.2 Worldview, research approach and design

The choices of worldview, research approach and design for this study were chosen in order to achieve the study aim and answer the two research questions.

The study is approached from a pragmatic worldview. In pragmatism, researchers base their approach to research on the intended consequence of the research rather than antecedent conditions (Creswell, 2009: 11). The pragmatic paradigm seeks solutions that would “work” rather than what might be seen as the absolute “truth” or “reality” (Frey, 2018).

Pragmatism is not bound to one specific system of philosophy as researchers draw from both quantitative and qualitative methods (Creswell, 2009: 10). In order to answer the research questions, the researcher for this study chose to include both quantitative and qualitative research elements, and therefore a mixed methods approach was followed.

A mixed methods approach was preferred over using a single method approach as results from one method can either expand or complement results from the other method; for example, qualitative methods can be used to investigate unexpected findings from quantitative data (Maree, 2012: 129). The core assumption of a mixed methods approach is that the collective strength of combining both quantitative data (statistical trends) as well as qualitative data (for example, personal views or experience collected via interviews or focus groups) can provide a better understanding of the research problem than using only one method (Creswell, 2015: 2).

The mixed methods approach that was employed in this study is explanatory sequential. An explanatory sequential design uses quantitative methods first, and then qualitative methods are used to explain the quantitative results (Creswell, 2015: 6). The initial quantitative data are used to identify the qualitative data to be collected (DeCuir-Gunby & Schutz, 2017: 5). Mixing

of the data occurs when the initial quantitative data informs the qualitative data (Creswell, 2009: 211).

This study used a survey design for both the quantitative component and the qualitative component of the study. Pickard (2013: 111) describes survey research as the process of collecting and analysing data via the questioning of individuals. Data can either be collected from a sample of individuals representative of the population or from the entire population, in which case it is a 'census' (Pickard, 2013: 111). Survey research is particularly helpful in describing characteristics of a large population (Babbie, 2010: 287), as is the case with the quantitative strand of this study.

3.3 Research methods

In survey research, information can be gathered either via self-administered surveys or interviewer-administered surveys.

With a self-administered survey, individuals complete the survey without the assistance of the researcher. In most cases, this format allows respondents to complete the survey at their own pace, which allows them to be reflective and thoughtful (Andres, 2012: 47). An example of a self-administered survey is an online/web-based questionnaire (Andres, 2012: 50).

In an interviewer-administered survey, an interviewer guides an individual through questions in the survey instrument. It allows the interviewer to establish a rapport with the respondent, clarify ambiguities and ask probing questions in the case of open-ended questions (Andres, 2012: 53).

3.3.1 Questionnaire

As an explanatory sequential design was used, quantitative data collection took place during the first phase of the research, and the instrument used was a questionnaire. Using a questionnaire as a means of data collection is an efficient way of reaching many participants in a relatively quick and cheap way (Maree, 2012: 93). An advantage of an online survey is that it can be programmed to allow sequencing of questions; for example, a 'Yes' answer can automatically guide respondents to a follow-up question (Andres, 2012: 51–52).

The questionnaire (See Appendix E) was designed to cover the following main themes: RDM practices (guided by stages in the Jisc Research Data Life Cycle), RDM challenges and RDM training and support needs.

3.3.2 Semi-structured interviews

During the second, qualitative phase of the research, results from the quantitative analysis were used to determine what data needed further exploration using interviews (Creswell, 2015: 38).

Interviewing is a data collection method where the researcher asks participants open-ended questions to get to know and understand their ideas, beliefs, views and opinions. Interviews can provide rich and descriptive information that will help the researcher understand the participant's social reality and can also lead to saturation of the data (Maree, 2012: 89).

Semi-structured interviews were conducted as they allow for flexibility (Miller & Brewer, 2003). With a semi-structured interview, the interviewer decides in advance what broad topics need to be covered and what main questions need to be answered (Miller & Brewer, 2003) but can ask additional questions as the interview progresses. In an explanatory sequential design, it can be challenging for the researcher to determine what quantitative results need further explanation. Interviews may include expanding the investigation into important variables or variables that might not have given the expected results gained from the quantitative data collection (Creswell, 2015: 38). See Appendix D for the interview schedule.

3.4 Population and sampling

Once the research methods were determined, the target population and sample needed to be established.

3.4.1 Population

A population can be described as the entire set of individuals important to a researcher (Gravetter & Forzano, 2012: 138). One can refer to a target population as the group that is defined by the researcher's specific interest (Gravetter & Forzano, 2012: 138). Individuals that form part of a target population typically have one common characteristic (Creswell, 2012: 142; Gravetter & Forzano, 2012: 138). In the case of this study, the characteristic that all the individuals in the population share is that they are researchers in the life sciences in the Schools of APES and MCB at Wits. Although many researchers in the life sciences at Wits reside under the Faculty of Health Sciences, the focus for this study was the Faculty of Science. For the purpose of the study, 'researchers' were classified as all academic staff and registered masters, doctoral and postdoctoral students in the two schools for 2019 (the year in which the study began). The total population of this study was 282 researchers. Table 3.1 shows an overview of the population compilation.

Table 3. 1: Overview of population compilation

Researcher group	Population
School of MCB	
M.Sc. students	46
PhD students	33
Postdoctoral fellows	7
Academic staff	21
Total MCB population	**107**
School of APES	
M.Sc. students	46
PhD students	33
Postdoctoral fellows	7
Academic staff	21
Total APES population	**175**
TOTAL POPULATION	**282**

3.4.2 Sampling

Two major groups of sampling procedures exist, namely probability sampling and non-probability sampling (De Vos et al., 2011: 228).

Probability sampling typically involves randomly selecting participants (Babbie, 2010: G9; De Vos et al., 2011: 228). With probability sampling, each person has the same known probability to be selected (De Vos et al., 2011: 228). Quantitative research often relies on probability sampling techniques, although non-probability sampling can also be used (De Vos et al., 2011: 228).

For the quantitative phase of this study, the total population was used instead of using a sample, as the total target population is not that big (282 researchers). A census was therefore conducted. Using the total population could compensate for a potential low survey response rate. The respective schools provided the researcher with the exact numbers of all researchers in each of the researcher categories.

Non-probability sampling is used when the probability of selecting a specific person in a population is not the same, unlike probability sampling, where everyone has the same chance of being selected (Maree, 2012: 70). Non-probability sampling is widely used in social research and particularly in qualitative research where the focus is on the in-depth description of a phenomenon rather than the need to generalise findings across a population, as may be the case with quantitative research (Merriam & Tisdell, 2015: 39).

For this study, non-probability sampling was used for the qualitative part of the study. Purposive or judgemental sampling was used. With purposive sampling, researchers intentionally seek participants that will assist them in understanding the central phenomenon (Creswell, 2012: 206). Patton (2014: 264) describes purposive or purposeful sampling as the strategic selection of “information-rich cases that by their nature and substance will illuminate the inquiry question being investigated”.

Five participants were purposively sampled to take part in interviews. Creswell & Clark (2011: 174) suggest that for qualitative sampling, rather than selecting a large number of participants, the researcher should select a small number of participants that will provide in-depth information on the central phenomenon. The number of participants typically used in qualitative research when cases are studied ranges from 4 to 10 (Creswell, 2008: 209; Creswell & Clark, 2011: 174).

Participants were chosen in order to get input from a variety of life sciences sub-disciplines. Research areas covered by the participants included the following: biochemistry, bioinformatics, biotechnology, cell biology, developmental biology, genetics, botany, climate change, conservation, ecology, evolutionary biology and zoology. Choosing participants from different subject areas was of interest to the researcher as more RDM literature is available from certain fields, especially molecular biology and ecology. It was anticipated that questionnaire results could then be compared with previous research in these sub-disciplines. All interviewees were established researchers and had insight and experience regarding RDM and, therefore, would be able to explain issues that were brought to light in the quantitative phase of research and which needed further exploration.

3.5 Ethical considerations

As this research included the participation of humans, it was important to consider ethics applicable to the study of human beings. The following ethical aspects were taken into consideration:

3.5.1 Informed consent

Getting informed consent ensures that participants understand what they agree to when taking part in a study; that they are fully aware of the aim of the research and the intended use of the data (Pickard, 2013: 89). Participants should be made aware that their participation is entirely voluntary, and they need to be made aware of any possible risks that may be involved in taking part in the research (Babbie, 2010: 66). The researcher was given informed consent for their participation in the study from both questionnaire respondents and interviewees via a formal informed consent letter.(See Appendix E: Online questionnaire p. 1 and Appendix C: Letter of introduction).

3.5.2 Confidentiality and anonymity

Ensuring the confidentiality of data provided in the research process means that a participant's identity would not be revealed when using data provided by that participant (Pickard, 2013: 93). Researchers were assured of confidentiality and anonymity in the informed consent letters. Personal details of questionnaire respondents were not collected in the survey, and therefore data collected in this way was kept confidential. However, where specific researchers could be identified via the information supplied in interviews and the open-ended question at the end of the questionnaire, such information was anonymised.

3.5.3 Ethical clearance

For this study, ethical clearance was obtained from the Department of Knowledge & Information Stewardship, Faculty of Humanities (UCT), as this is the institution where the researcher is registered as a master's student (See Appendix A: Ethical clearance). Once clearance was granted, the office of the deputy registrar at Wits issued a letter of permission that allowed the researcher to conduct the study among students and staff at Wits (See Appendix B: Clearance from gatekeepers). After this permission was granted, an e-mail with a link to the questionnaire was sent to all researchers via contacts in the two schools. Interview participants were personally invited via e-mail.

3.6 Data collection

Data collection comprised quantitative data collection during the first phase of the research, followed by qualitative data collection.

For quantitative data collection, the SurveyMonkey software application (SurveyMonkey, 2019) was used to create the online questionnaire. The survey ran from the 16th of October till the 5th of November 2019. A first e-mail reminder was sent a week after the initial invitation, and a final reminder was sent to all researchers two weeks after the survey opened.

Interviews, used for qualitative data collection, took place between February and September 2020. As a result of the COVID-19 pandemic, two interviews were conducted online via the MS Teams platform; three were conducted in person. All interviews were recorded with permission. The duration of the interviews ranged between 25 and 40 minutes.

3.7 Data analysis and interpretation

As an explanatory sequential design was used for this study, the analysis of quantitative data took place before qualitative data collection and analysis. Data analysis in an explanatory sequential design occurs in three phases: analysis of the initial quantitative data, analysis of the follow-up qualitative data, and finally, an analysis of the mixed methods question to indicate how qualitative data inform quantitative data (Creswell & Clark, 2011: 221). After analysis of both quantitative and qualitative data, mixed methods interpretation takes place, which involves reviewing both quantitative and qualitative results as well as assessing how the information addressed the mixed methods research question (Creswell & Clark, 2011: 212).

In this study, analysis of quantitative data took place immediately after data collection in the first phase of the research. Because this analysis may reveal extreme cases or deviation from the expected results, or the majority of results, these cases are followed up using qualitative interviews to gain insight about why these cases diverged from the other quantitative results (Creswell, 2009: 218). Analysis of the quantitative results can be used to identify results that need further explanation (Creswell & Clark, 2011: 218), as was the case with this study.

The statistical package, MS Excel, was used to analyse quantitative data obtained from the questionnaire.

Thematic analysis was used to organise data collected via interviews which were transcribed via the Otter transcription software (Koopman & de Jager, 2016: 5). Thematic analysis is a

process whereby a researcher looks for “recognisable reoccurring topics, ideas, or patterns (themes) within the data that provide insight into communication” and can be used to investigate a phenomenon that needs better understanding (“Thematic analysis”, 2017).

The point where the quantitative and qualitative phases of the research intersect in mixed methods research is known as integration or mixing (Creswell, 2015: 82). In explanatory sequential mixed methods research, this point occurs when qualitative data are used to explain the results of the quantitative data (Creswell, 2015: 83). However, according to Creswell (2015: 87), integration can also be included in the data collection, data analysis as well as the discussion or conclusion parts of the research. For this study, integration took place in the final chapter.

3.8 Validity and reliability in mixed methods design

Within mixed methods research, it is important to consider steps that need to be taken to check the validity of the quantitative data as well as the accuracy of the qualitative data (Creswell, 2009: 219). The following validity procedures for both the quantitative and qualitative strands of the research were considered.

In quantitative research, internal validity refers to the extent to which changes in a dependant variable are indeed caused by changes in an independent variable (Maree, 2012: 137; Pickard, 2013: 22). External validity refers to the extent to which findings from the research can be generalised to a wider context (Maree, 2012: 138; Pickard, 2013: 22). Reliability refers to the extent to which the same results can be produced over a time period and across locations (Pickard, 2013: 22).

In qualitative research, validity is also referred to as trustworthiness. Trustworthiness can be described in terms of credibility, confirmability, transferability, dependability, and authenticity (Maree, 2012: 38; Pickard, 2013: 20).

Credibility or authenticity can be described as the extent to which the subject of research has been accurately identified and described (De Vos et al., 2011: 420).

Confirmability refers to the objectivity of data as well as the absence of errors in research (Maree, 2012: 141).

Transferability in qualitative research refers to the extent to which the findings from research can be transferred to other contexts or situations (De Vos et al., 2011: 420; Maree, 2012: 140).

With dependability, research should be logical and well documented (De Vos et al., 2011: 420).

The following steps were taken to ensure the validity, reliability and trustworthiness of the study:

- Pre-testing of the questionnaire as well as the interview questions on an academic in the School of MCB.
- Scrutiny of the questionnaire and interview schedule by the Department of Knowledge & Information Stewardship Research Ethics Committee at UCT.
- The same questionnaire and semi-structured interview schedule were maintained for all respondents.
- Questions asked in the questionnaire pertained to research questions and were informed by extensive consultation of the literature and using a well-known framework for data management.
- The interview schedule also resulted from a thorough review of the literature and from input from a researcher familiar with research data concepts in the life sciences.
- Care was taken to document research carefully and report on it accurately.

3.9 Summary

This chapter explored the worldview, research approach, and design used to conduct the research. This was followed by a description of research methods used for data collection, sampling, data analysis and interpretation, and a discussion around ethical considerations. The chapter concluded with steps to ensure the validity and reliability of a mixed methods design.

CHAPTER 4: DATA ANALYSIS AND PRESENTATION

4.1 Introduction

This chapter presents the data collected from the questionnaire and interviews. Quantitative data collected from the questionnaire are mainly presented through descriptive statistical tools such as tables and figures. Data collected from interviews are presented thematically.

4.2 Quantitative data analysis

Of the total population of 282 invited to participate in the quantitative phase of the research, a total of 58 researchers responded to the questionnaire, which is a response rate of 20.6%. The low response rate is commented on in Chapter 5.

This study did not distinguish between the two schools in results obtained from the questionnaire as there is an overlap between disciplines across the two schools (for example, the topic of 'evolution' is studied in both the School of APES and the School of MCB). Instead, the study wanted to distinguish between different life sciences sub-disciplines (for example, ecology and molecular biology).

While 58 researchers in total responded to the questionnaire, not everyone answered all questions. The 'n' alongside tables and figures denotes the number of respondents to a particular question and therefore may differ from question to question.

4.2.1 Researcher characteristics

The first four questions in the online questionnaire were used to collect researcher characteristics of respondents. (See Appendix E).

Table 4. 1: Researcher categories (n = 58)

Researcher Category	Masters	PhD	Academic staff
Number of responses	25	12	21
% of total responses	43	21	36

From Table 4.1, it can be seen that master's students in the life sciences were the largest researcher group (25) that took part in the survey, followed by academic staff (21) and PhD students (12). Although respondents were requested to select all options that apply to them, for example, a PhD student that is also a member of the academic staff, no overlap was found between different categories. No postdoctoral fellows took part in the survey.

Table 4. 2: Researcher qualifications (n = 57)

Researcher Category	Honours	Masters	PhD
Number of responses	23	14	20
% of total responses	40	25	35

The largest number of respondents (23) possessed an honours degree, followed by respondents with PhDs (20) and master's degrees (14), as can be seen in Table 4.2.

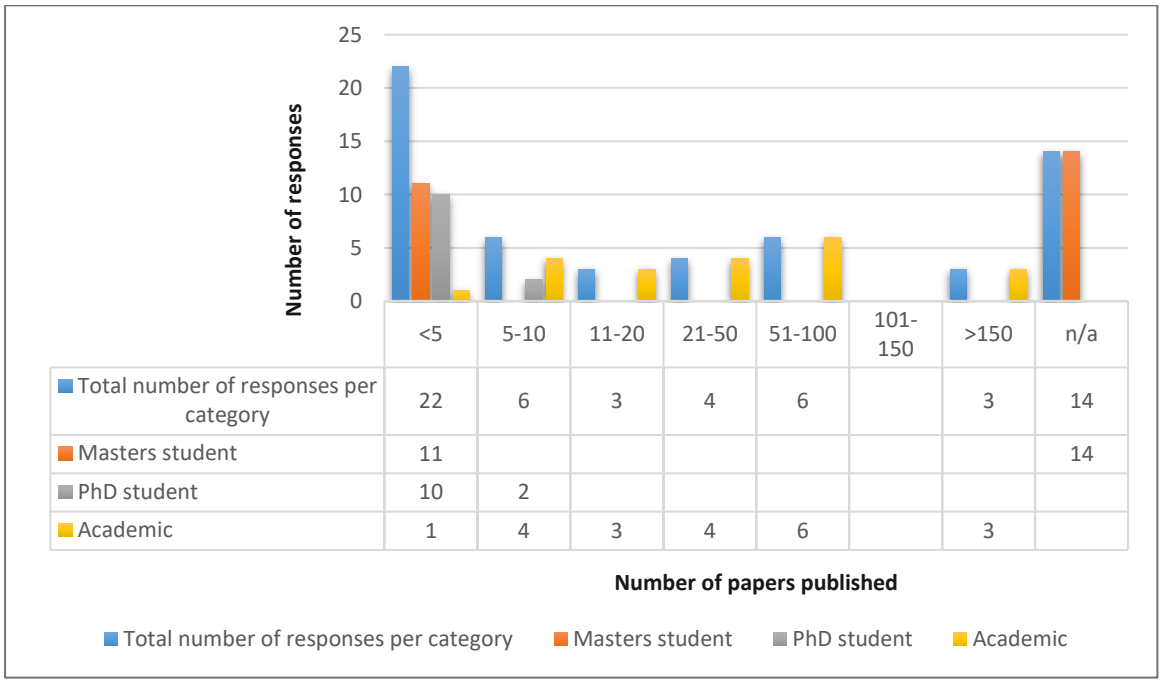

Figure 4. 1: Scientific papers published (n = 58)

Figure 4.1 shows that most researchers (22; 38%) fell in the category of having published fewer than 5 papers, whilst 14 (24%) respondents indicated that the questions do not apply to them as they have not been published. The remainder of respondents (22; 38%) published

papers from 5 up to more than 150 papers (3; 5% of respondents). No respondents published papers in the 101–150 range.

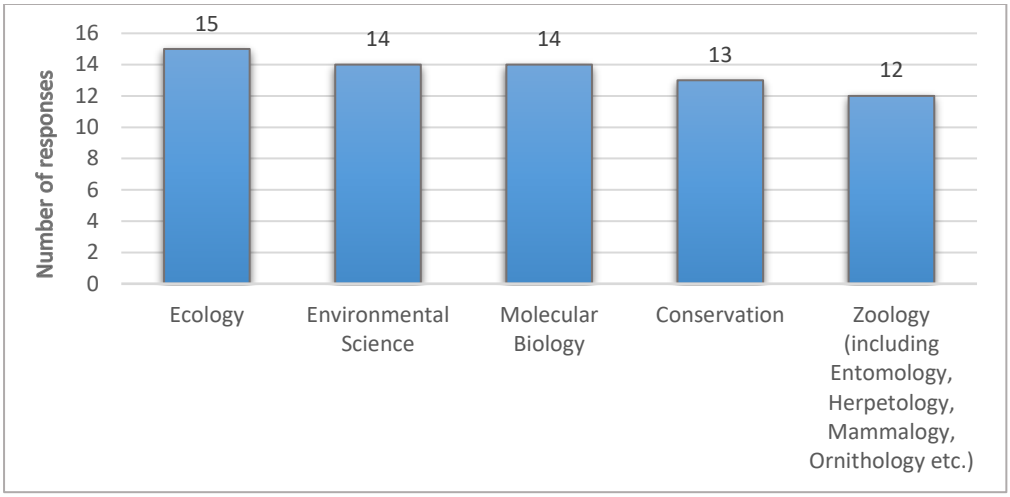

Figure 4. 2: Most represented areas of research in life sciences sub-disciplines (n = 58)

Respondents were requested to select all the sub-disciplines that apply to them. Figure 4.2 shows the sub-disciplines that were represented by most researchers. The sub-discipline with the largest occurrence amongst respondents was ecology (26%), followed by environmental science (24%) and molecular biology (24%), conservation (22%), and zoology (21%). The remainder of responses were scattered amongst the other sub-disciplines respondents could select (See Appendix E).

4.2.2 RDM practices

In order to establish the RDM practices of researchers, questions were posed to cover different stages of the Jisc Research Data Lifecycle. Data are presented in the order that it was collected via the questionnaire.

4.2.2.1 Funding

The requirements of funding agencies play an important part in the planning phase of the data lifecycle. Therefore this study included questions related to funders and funding requirements.

Publicly funded research

As funders of research increasingly require researchers to adhere to some aspects of RDM, this study wanted to determine how many respondents are publicly funded. As Figure 4.3 shows, the majority of respondents (57%) said that they are publicly funded, whilst the remaining respondents either advised they do not receive public funding (22%) or they do not know if they are publicly funded or not (21%).

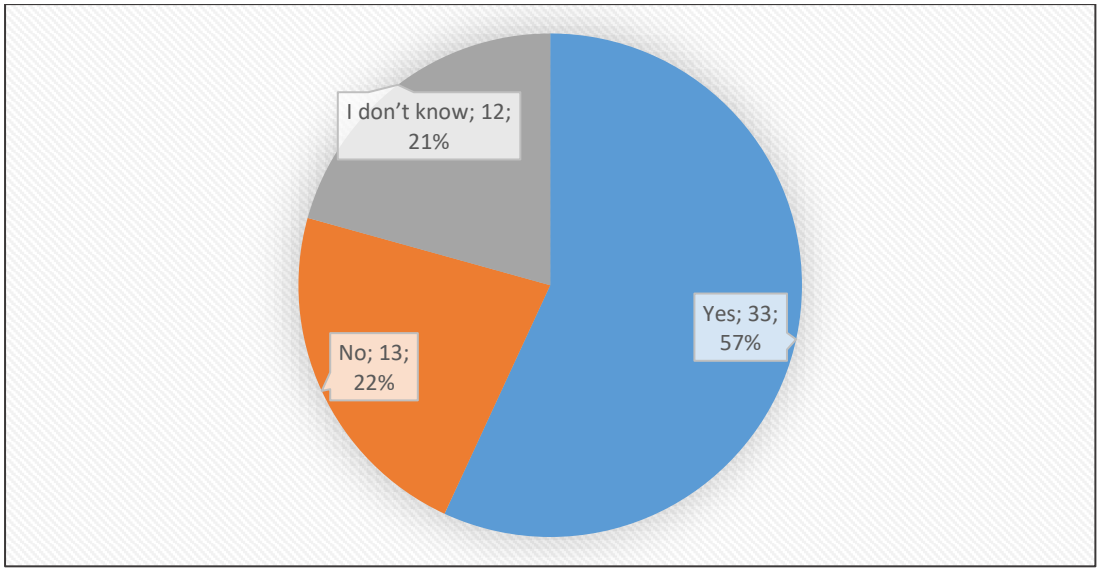

Figure 4. 3: Public funding of research or research unit (n = 58)

Funder identities

The study further wanted to establish who the funders involved are. Respondents were given a listing of funders from which to select as many as were applicable to them. An 'other' option was added that allowed respondents to specify any funders not listed. Table 4.3 lists all the funders of researchers that responded to this question. A total of 13 funders were identified as public funders of respondents. However, the overwhelming majority of funded respondents (93%) are funded by the NRF.

Table 4. 3: Research funders (n = 29)

Funder	Total responses
National Research Foundation (NRF)	27
Andrew Mellon Foundation	1
CABI	1
Carnegie Corporation of New York	1
Centre for Invasion Biology	1
Dept Environmental Affairs	1
Dept Water Affairs Forestry	1
Gauteng Dept Agriculture Rural Development	1
Medical Research Council of South Africa	1
National Science Foundation	1
NPO via foreign donor money	1
Royal Society	1
US Military	1

Funder requirements for data management

A question was asked to determine to what extent funders have policies in terms of data management. In response to whether their funders require a DMP or whether they require the sharing of data in a repository, and as shown in Figure 4.4, 10 researchers (34%) responded positively to each of these two requirements. Nine respondents (31%) said that their funders require them to make their research data completely open. None of the respondents said that their funders required them to add metadata to their research data. Only 8 respondents (28%) said that their funder does not have any requirements for data management. Respondents could select as many options as applicable.

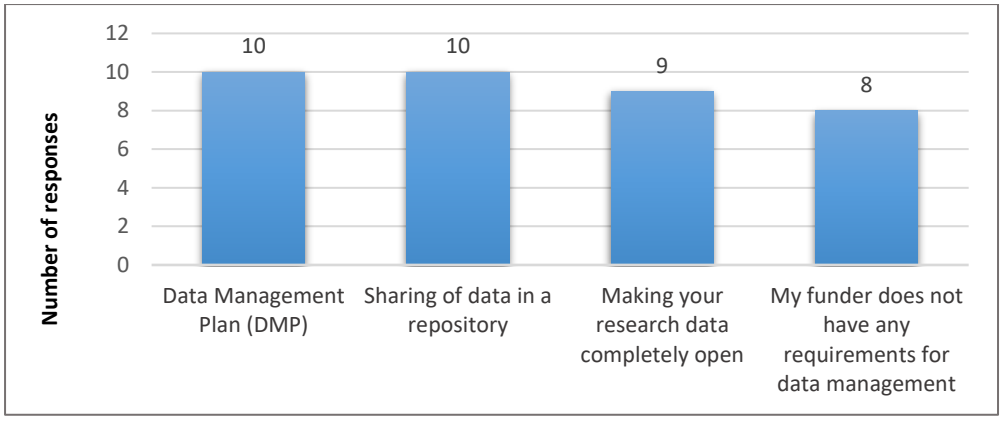

Figure 4. 4: Funder requirements for data management (n = 29)

4.2.2.2 Data collection and capture

As data volumes and types can have an influence on how data are collected and managed, questions in this regard were included in the questionnaire. The use of metadata is an important element of the data capturing process, and thus questions on this were also posed.

Volumes of research data

As shown in Figure 4.5, researchers hold data in a range of volumes, from less than 1 GB (8; 17%) to more than 100 TB (1). As no researchers responded to the categories of 51–100 GB and 51–100 TB, these categories were not included in the figure. Eight respondents did not know how much data they held. Data held by respondents were most often found in the range of 1–50 GB (20). Half of researchers (50%) held data volumes in the gigabyte range (1–50GB, 100–500GB, and 501–999GB). Very large data sets (more than 100 TB) were held by only one respondent.

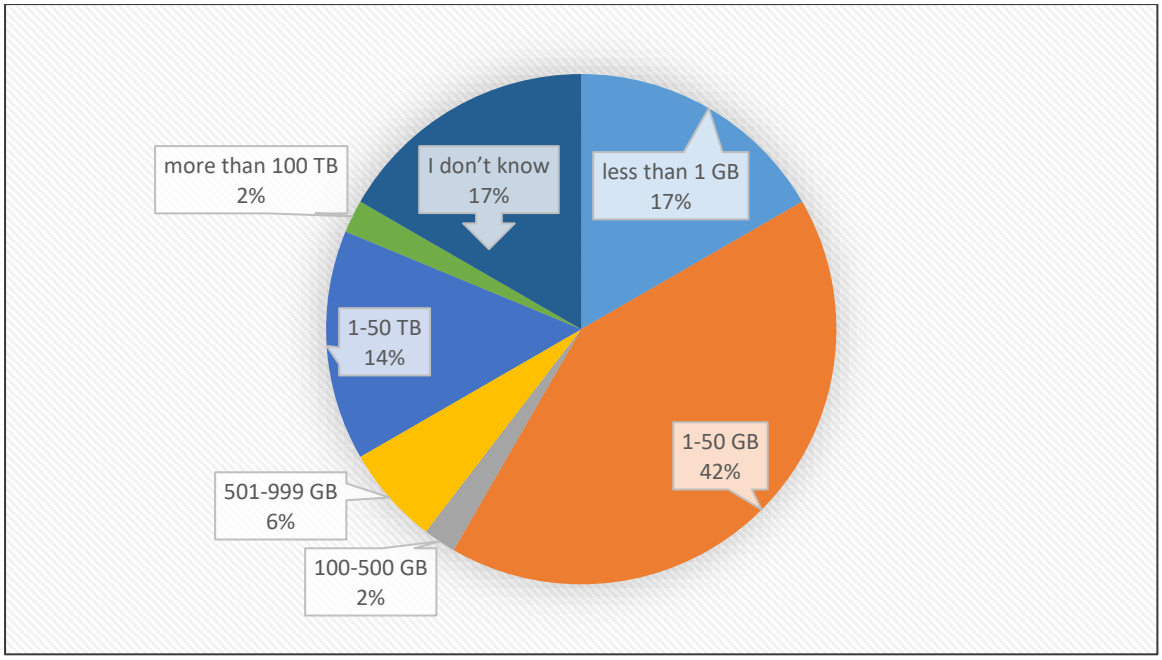

Figure 4. 5: Research data volume (n = 48)

Types of research data

Respondents were asked what types of data they create or collect for their research and could select all relevant data types listed. As evident in Figure 4.6, all 15 data types listed are used by respondents, with the most common types being spreadsheets/tabular data and images used by 40 respondents each (83%), followed by documents used by 39 respondents (81%) and raw data used by 26 respondents (54%). The least used data types include electronic laboratory notebooks (ELNs) and structured text used by 3 respondents each (6%), and structured graphics used by 1 respondent.

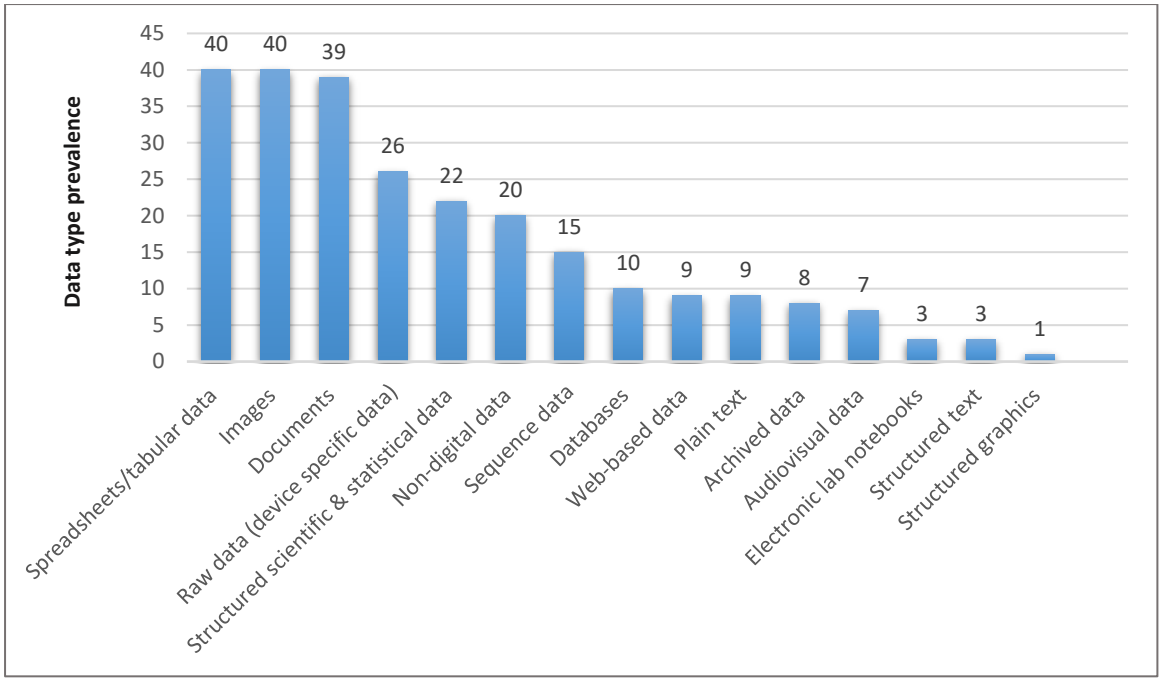

Figure 4. 6: Types of research data collected (n = 48)

Generation of research data

This study further wanted to establish to what extent researchers generate their own data or use existing data as this could contribute to understanding the data-sharing practices of researchers. Figure 4.7 shows that most respondents (47; 98%) create/collect/produce new data. However, 65% of these researchers also used existing data from other sources such as data from their research group (10; 21%), data from an open archive/repository (19; 40%) or other sources that included using survey data from government agencies (1) and the use of data from a weather station in the study area (1).

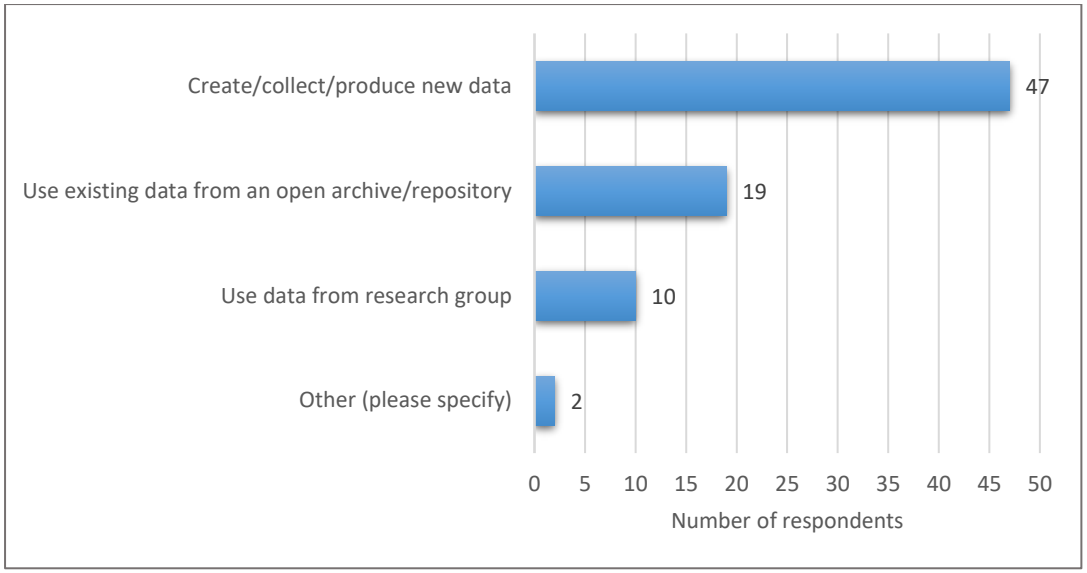

Figure 4. 7: Methods of generating research data (n = 48)

Metadata usage

As the assigning of metadata improves discovery and reuse of data, respondents were asked if they assign metadata to their data. Figure 4.8 shows that 21% of respondents always assign metadata to their research data, whilst 48% only sometimes assign metadata to their research data. Respondents that indicated that they do assign metadata, therefore, totalled 69% (33). Eight per cent of respondents did not know if they assigned metadata or not, which might indicate that they are not aware of metadata or the importance thereof.

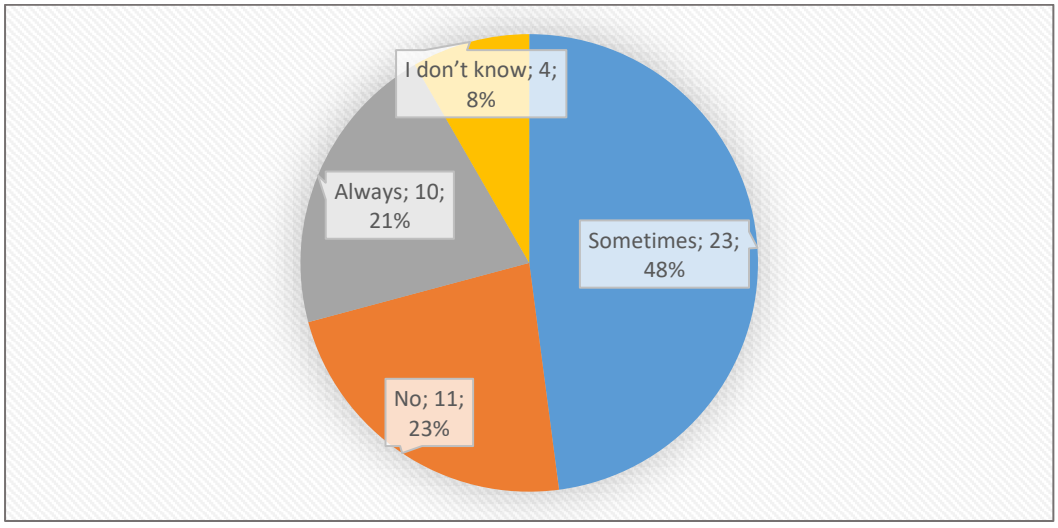

Figure 4. 8: Metadata usage (n = 48)

Metadata standards usage

Figure 4.9 shows to what extent those respondents who assigned metadata to their datasets adhere to metadata standards/guidelines when assigning metadata.

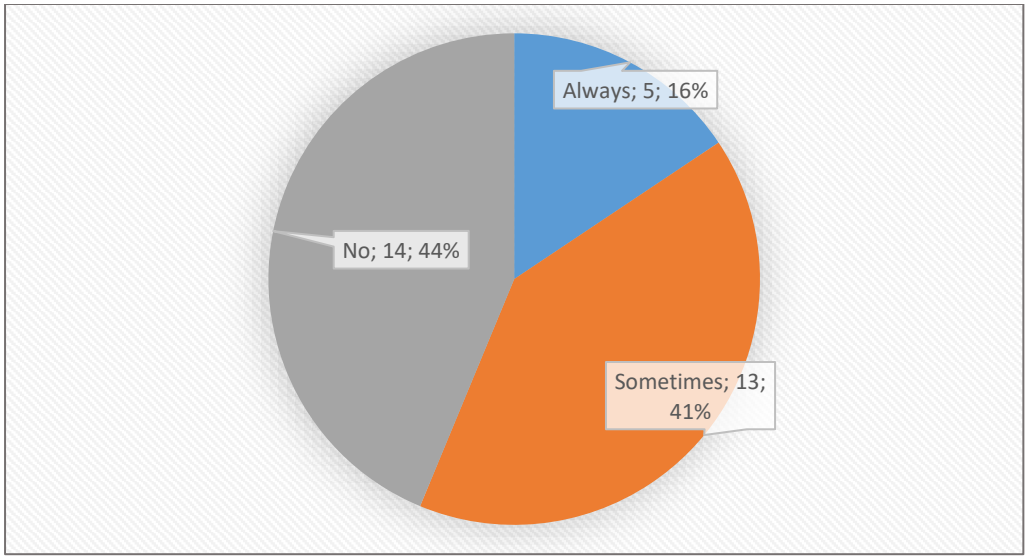

Figure 4. 9: Usage of metadata standards/guidelines/ontological structures (n = 32)

Five respondents (16%) always use metadata standards, whilst 13 respondents (41%) only sometimes use metadata standards. A total of 18 respondents (57%) who assign metadata, therefore, make use of metadata standards. Not using metadata standards may be the result of either not being aware of metadata standards or not being satisfied with available metadata standards.

Types of metadata standards being used

To determine which metadata standards are used, respondents were provided with a list of standards/guidelines used in different sub-disciplines of the life sciences. Figure 4.10 presents the standards used by respondents in the study:

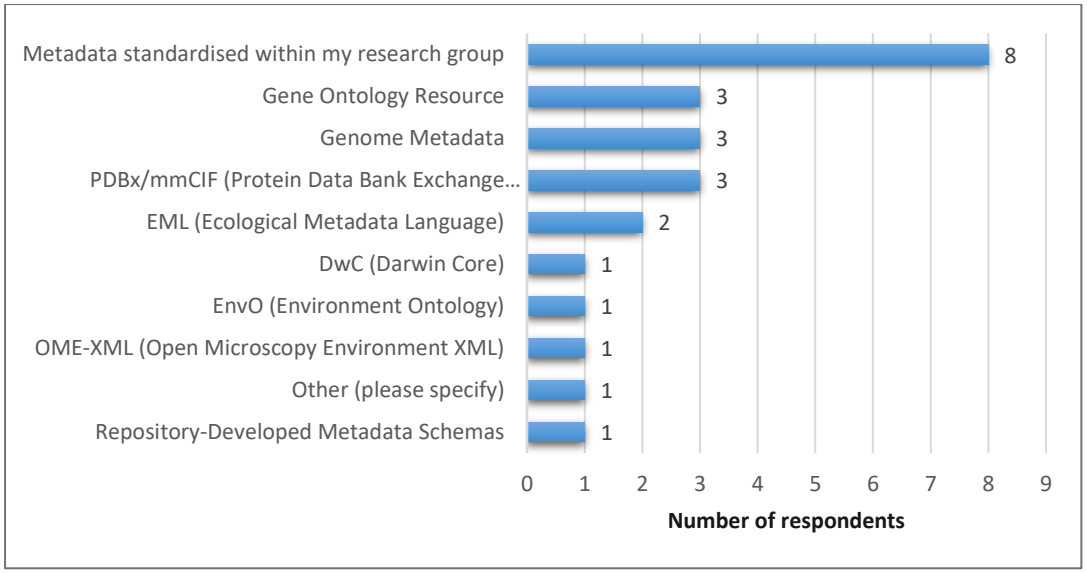

Figure 4. 10: Types of metadata standards/guidelines/ontological structure used (n = 12)

The majority of respondents to this question (8; 67%) stated that they used metadata standardised within their research group. The second most used standards (three respondents each) were Gene Ontology Resource, Genome Metadata, and PDBx/mmCIF (Protein Data Bank Exchange Dictionary & Macromolecular Crystallographic Information Framework). Two respondents use EML while Darwin Core, EnvO, OME-XML (Open Microscopy Environment XML), and Repository-Developed Metadata Schemas had one respondent each. One respondent indicated under the 'other' option that he has his own way of 'keeping track'. As this does not refer to a standard, it was not included with the above results.

4.2.2.3 Data analysis

Software applications

Figure 4.11 shows all responses received when respondents were requested to select the software they used to analyse and manipulate data. It includes a total of 27 applications. Respondents could choose all options that apply to them, and an 'other' category also allowed them to list any applications that might not have been listed in the questionnaire.

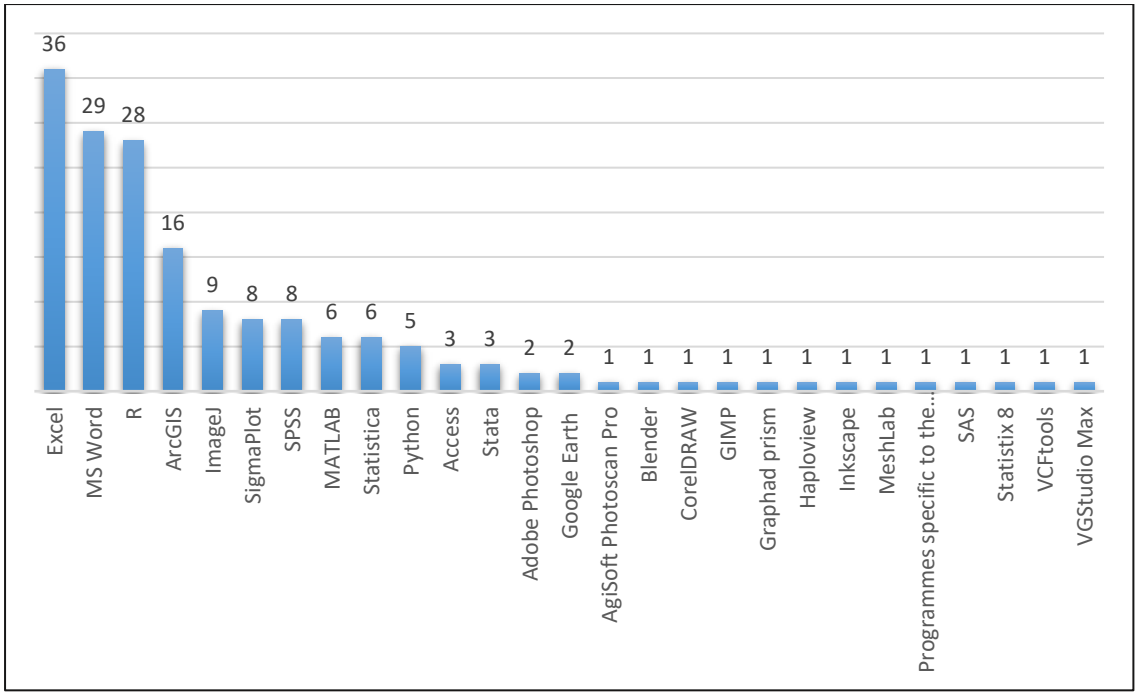

Figure 4. 11: Software applications used for data analysis and number of respondents per application (n = 42)

The software applications most commonly used by respondents were Excel (36 respondents; 86%), MSWord (29; 69%), R (28; 67%) and ArcGIS (16; 38%). Within the 'other' category, 16 different software applications were listed by respondents. All applications in 'other' were only mentioned once, except Statistica, with six mentions. All are included in Figure 4.11.

4.2.2.4 Data storage (short term)

Location

The study wanted to determine where researchers store their data during the active phase of research. Respondents were given the option to choose all locations they use. Figure 4.12 shows that the storage location that most researchers (36; 86%) used for the short-term storage of data are an external hard drive/USB/flash drive. This was followed by storage on a hard disk drive of an office desktop/laptop and cloud storage for example, Dropbox, Google Drive, Microsoft OneDrive, and Google Docs (35 respondents; 83% each). It was further found that 16 respondents (38%) still used paper or paper laboratory notebooks to store data whilst only 2 respondents (5%) used an ELN for the short-term storage of data (also see Figure 4.6).

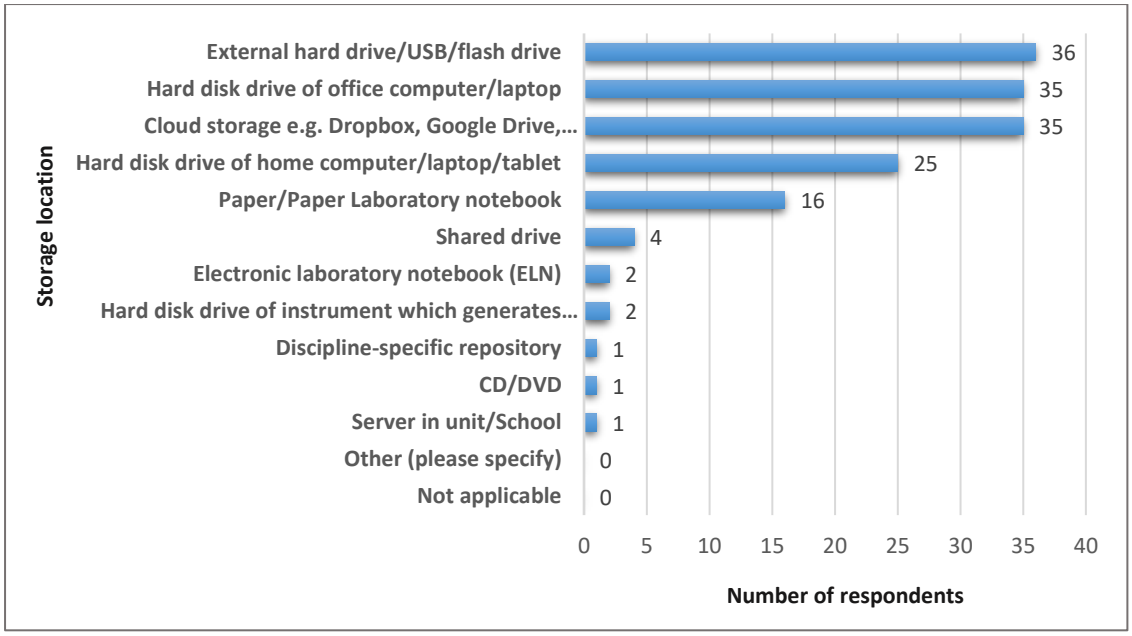

Figure 4. 12: Storage of research data during the active research phase (short-term storage) (n = 42)

Frequency of research data backups

Respondents were questioned about how frequently they back up data during data collection and analysis. They had to choose one option from a list. Figure 4.13 shows that the majority of respondents (15; 36%) back up their data daily.

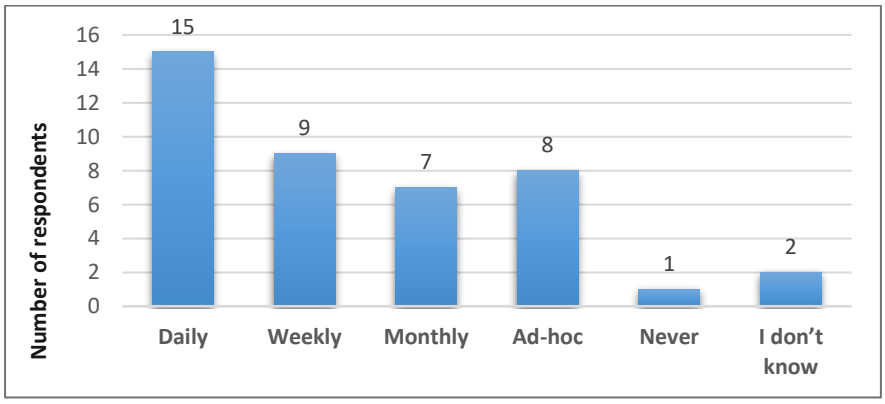

Figure 4. 13: Frequency of research data backups whilst collecting and analysing data (short-term storage) (n = 42)

Nine respondents (21%) said that they backed up their data weekly, whilst 7 respondents (17%) ran backups monthly. In total, 74% of respondents backed up their data either daily, weekly or monthly, and when adding responses from the ad-hoc category (8; 19%), 93% of

researchers indicated that they keep backups. This can be an indication of the high premium respondents place on backing up their data. No researchers indicated that they back up their data only on a 6-monthly or annual basis. Two respondents said that they do not know how frequently they keep backups, and 1 respondent said they never make backups.

Location of data backups

The study was not only interested in the frequency of backups but was also interested in where researchers kept them. Respondents were given the choice of selecting all options that apply to them from a list provided.

Figure 4.14 shows that the two locations researchers used most frequently to store their backups are cloud storage and storage on an external hard drive/USB/flash drive, with 33 (79%) respondents each. This was followed by storage on the hard disk drive of an office desktop/laptop (27; 64%) and the hard disk drive of a home desktop/laptop/tablet (22; 52%).

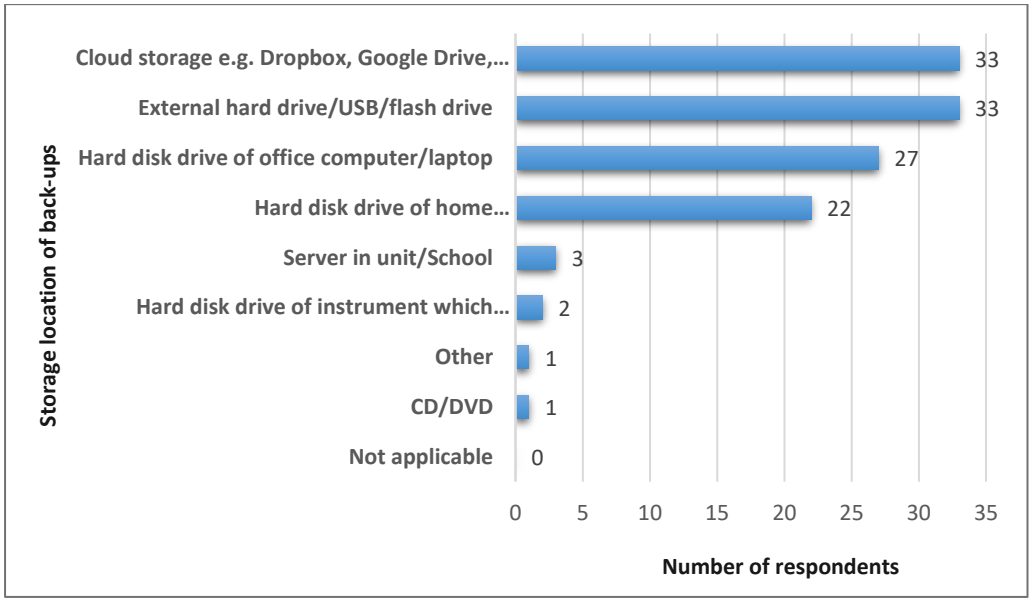

Figure 4. 14: Location of data backups during the life of the project (short-term storage) (n = 42)

The fact that the study showed that 79% of researchers use the cloud for backup storage shows that there is a move by researchers to using the cloud for both short term and backup storage whilst using CD/DVDs is falling out of fashion.

4.2.2.5 Data storage (long term)

Long-term storage/preservation behaviour of researchers

Figure 4.15 shows the responses received from researchers on the question of whether respondents store their data in the long-term for preservation:

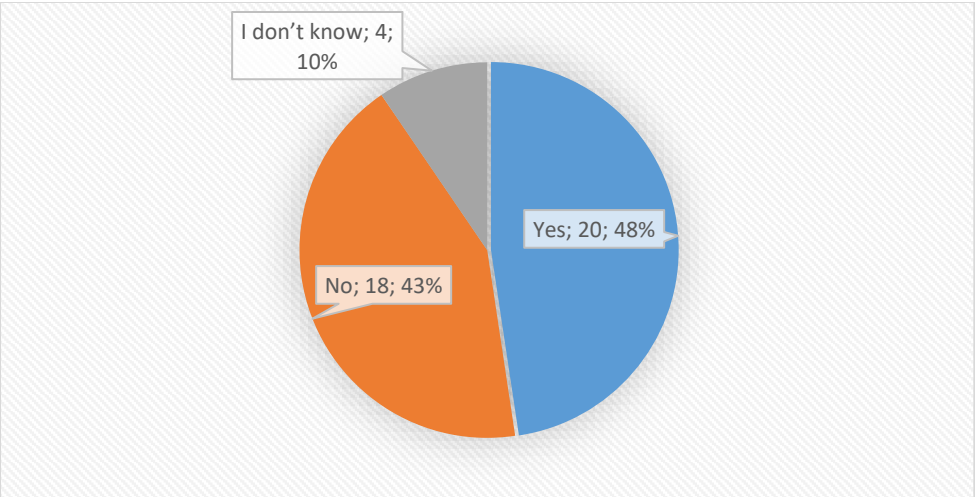

Figure 4. 15: Research data storage for long-term preservation (n = 42)

Almost half of the researchers (20; 48%) indicated that they store data for long-term preservation. Eighteen researchers (43%) said that they do not preserve their data long-term. Four respondents (10%) said that they do not know if their data are stored for long-term preservation or not.

Reasons for storing data for long-term preservation

Following the previous question, the study wanted to determine the 20 researchers' motives for storing their data for long-term preservation. Figure 4.16 shows that most respondents (9; 45%) said they did so as it is a journal publication requirement, followed by 8 respondents (40%) who said it is a research group requirement.

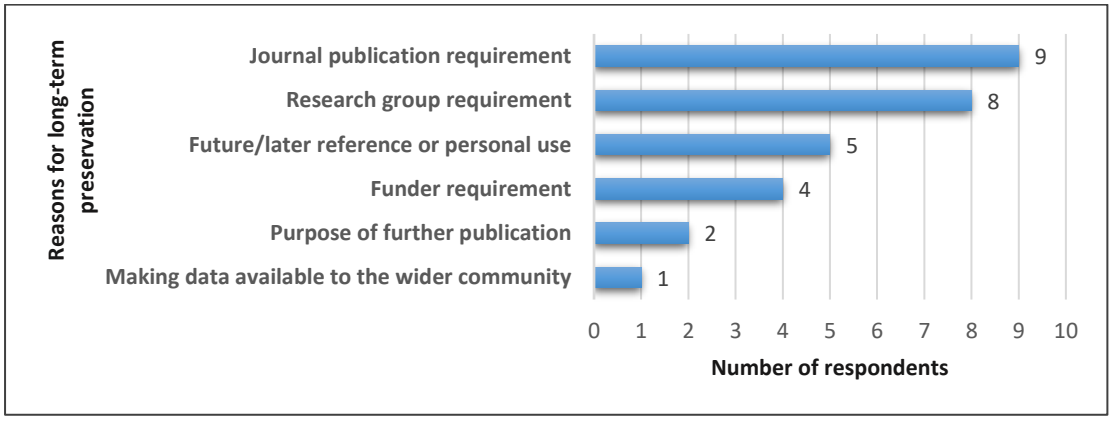

Figure 4. 16: Reasons for archiving data for long-term preservation (n = 20)

Four respondents (20%) indicated that they store data as it is a funder requirement. Most respondents stored their data because of a mandate from either a journal publication, a research group, or a funder. Other responses included storing data for future use (5 respondents; 25%) or the purpose of further publication (2 respondents). Only one respondent said they store data for long-term preservation to make it available to the broader community.

Repositories used for the archiving of research data

Researchers were asked which repositories they have used for the archiving of research data. Figure 4.17 shows all responses selected from the list provided (including repositories common in different life sciences sub-disciplines) and the repositories specified when the 'other' field was selected.

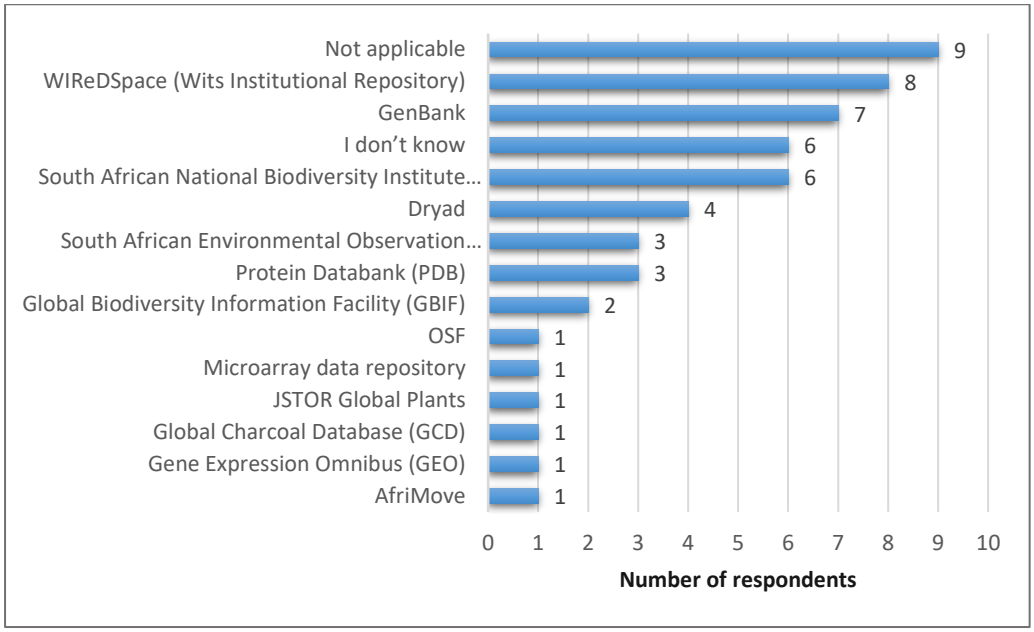

Figure 4. 17: Repositories used for the archiving of research data (n = 40)

Repositories used by researchers include the IR, WIReDSpace (8 respondents; 20%), followed by GenBank (7;18%), the South African National Biodiversity Institute (SANBI) (6;15%), and Dryad (4;10%). Although eight respondents said they use WIReDSpace for archiving research data, this researcher concluded that respondents might have misunderstood the question as currently only one dataset from the Schools that took part in the study has been stored in this data repository, and respondents might have been referring to published output. Nine respondents (23%) chose the 'Not applicable' option, indicating that they do not use repositories to archive research data. Seven respondents used the 'other' option to list repositories they use, for example, AfriMove, Global Charcoal Database (GCD),

Open Science Framework (OSF), and 'Microarray data repository'. Some of the repositories listed were not selected and therefore not included in Figure 4.17. In total, only 12 repositories were used by respondents. As there are many repositories available for the archiving of research data, the fact that only 12 repositories are being used can either be an indication that researchers are not aware of available repositories, funders/journal publishers do not require the long-term preservation of data or for other reasons further being investigated as part of this study.

4.2.2.6 Data sharing

Data sharing practices

In order to determine with whom researchers share their data, they were presented with a choice of both informal and more formal ways of sharing data. Respondents were asked to choose all applicable options. Figure 4.18 shows that the majority of researchers do not have a problem with informal data sharing.

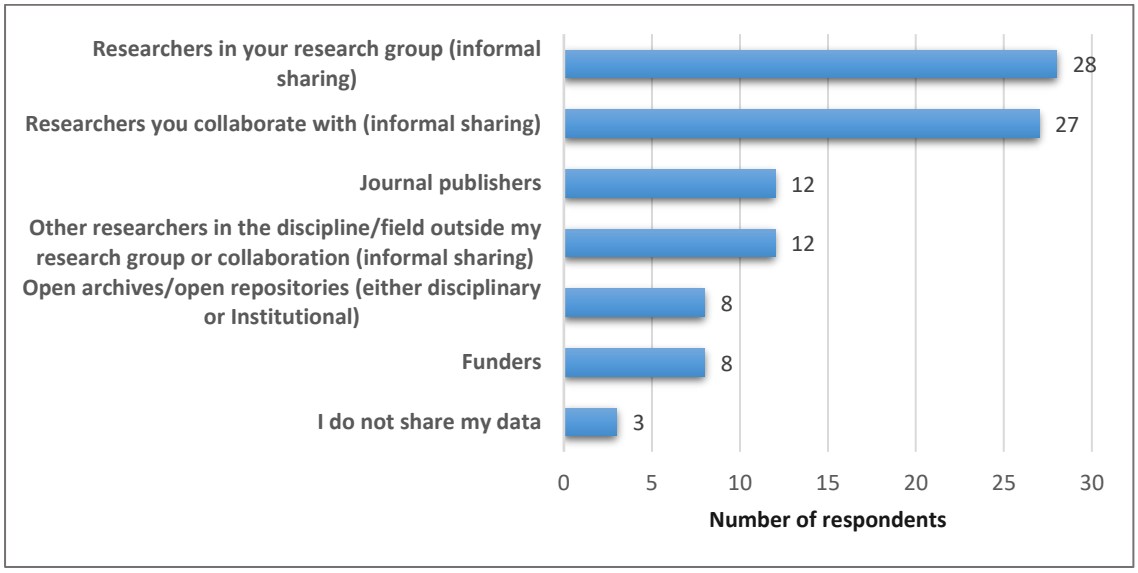

Figure 4. 18: Data sharing practices (n = 41)

Twenty-eight respondents (68%) said that they share data with researchers in their research group and 27 (66%) said they shared data with researchers with whom they collaborate. Only 12 respondents (29%) indicated that they share data with researchers outside their research group.

More formally, 12 respondents (29%) shared their data with journal publishers, and 8 respondents (20%) shared their data with funders. Only 8 respondents (20%) shared their

data in open archives/open repositories (either disciplinary or institutional). Three respondents (7%) do not share their data.

Conditions for sharing data for future research

This study further wanted to establish under which conditions researchers were willing to make their data available for future research. Respondents had the opportunity to choose various options. Figure 4.19 shows all the conditions under which researchers were willing to make their data open.

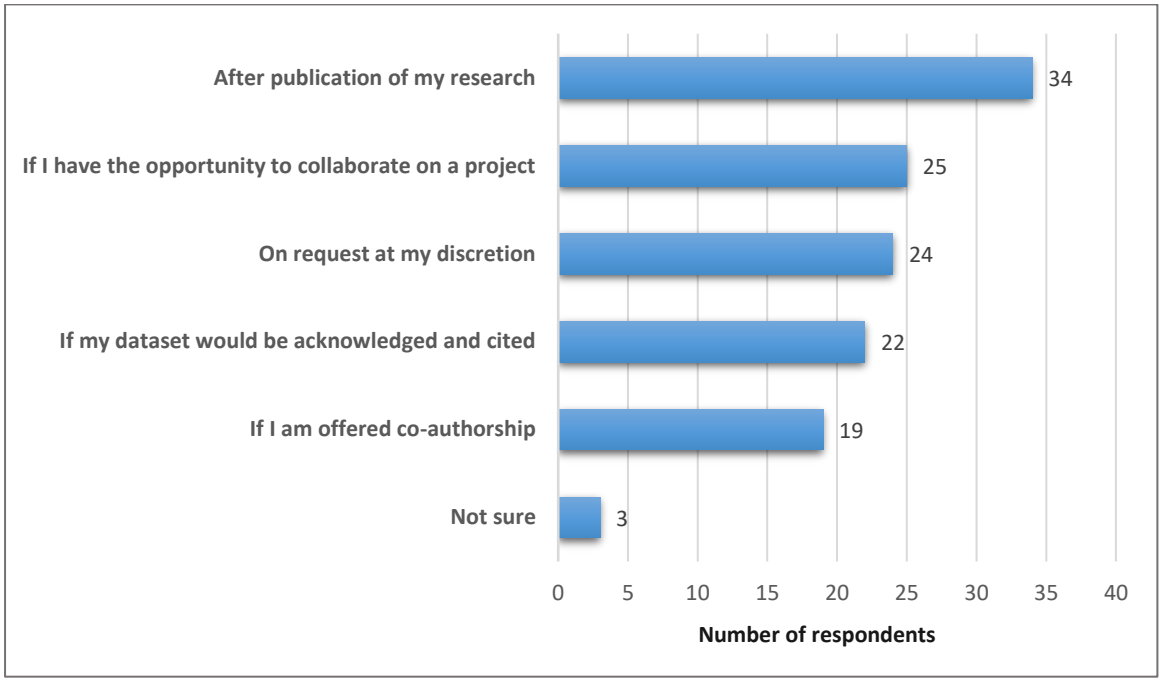

Figure 4. 19: Conditions for sharing data (n = 41)

It is evident that researchers are willing to share data, but only under certain conditions. Most researchers (34; 83% of respondents) have a precondition that they will only make their research data available after publication. Other pre-conditions for sharing data (collaboration, on request, acknowledgement, co-authorship) received a significant number of responses: between 46% and 61%. Only 3 respondents (7%) indicated that they were not sure if they were willing to share their data for future research. None of the respondents was unwilling to make their data available for future research. Therefore, question 23 of the questionnaire, which asked for reasons for not making data available for future research, was not posed to anyone.

Linking of datasets to published papers

The study wanted to establish to what extent researchers in the life sciences at Wits publish their datasets as supplementary information to published research. Figure 4.20 shows that only 10 respondents (25%) indicated that they had published datasets as supplementary information to published papers contrasting with 26 respondents (65%) who have never linked a dataset to a published paper.

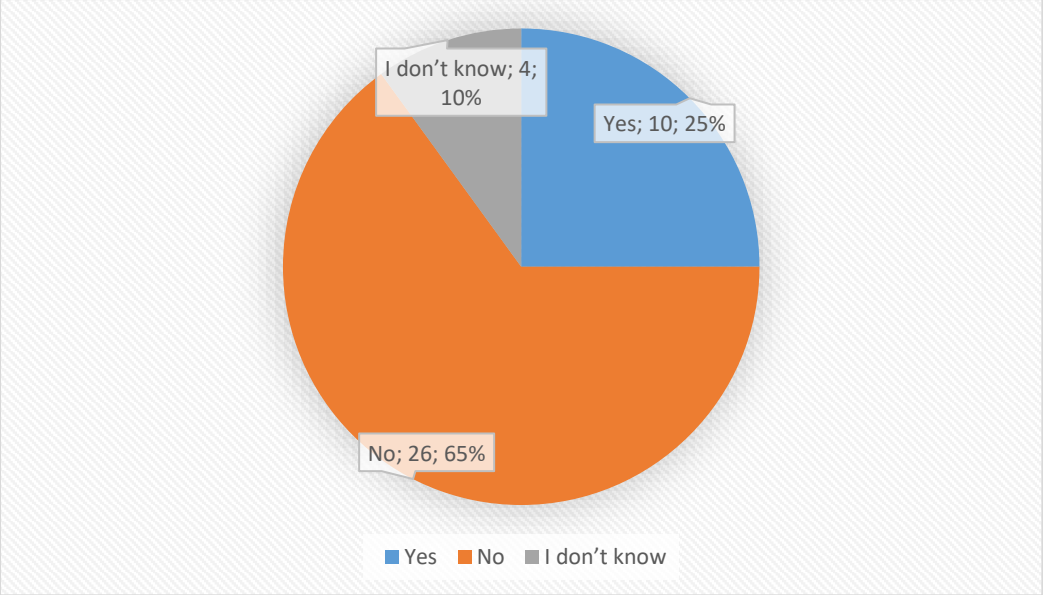

Figure 4. 20: Linking of datasets to published papers (n = 40)

4.2.3 RDM challenges and barriers

The study was not only interested in the research data practices of respondents but was also interested in what researchers perceived as challenges and barriers that hinder effective RDM. Potential challenges were listed, drawing from previous studies, and respondents had to indicate to what extent they agreed or disagreed with the statements about them according to a Likert scale. Figure 4.21 displays the results.

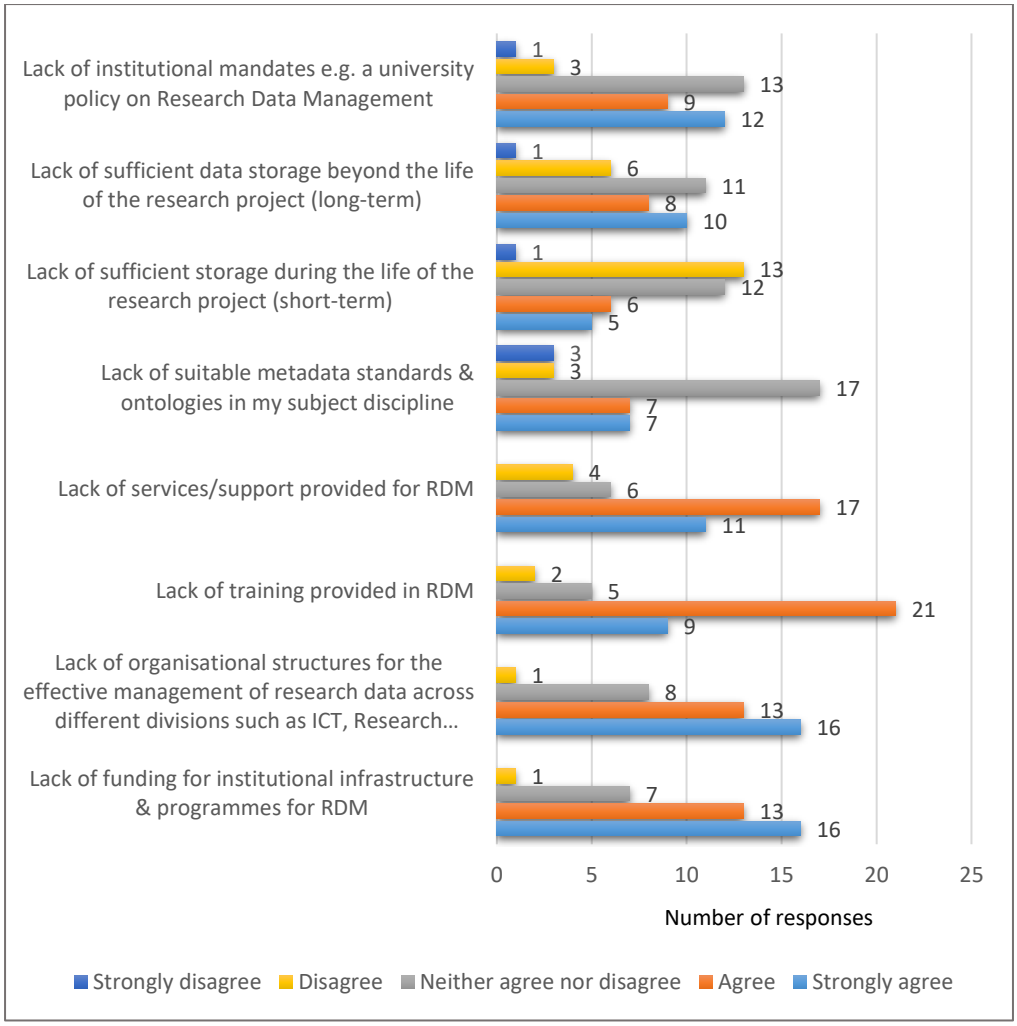

Figure 4. 21: RDM challenges and barriers (n = 38)

Significantly, respondents perceived the following as most challenging when it comes to effective data management:

- lack of RDM training – 30 respondents (81%) either strongly agreed/agreed, with the majority of respondents agreeing rather than strongly agreeing that this is a challenge
- lack of funding for infrastructure and programmes for RDM – 29 respondents (78%) either strongly agreed or agreed with this statement
- lack of organisational structures for the effective management of research data across different divisions such as ICT, Research Office and the library – 29 respondents (76%) strongly agreed or agreed
- lack of services/support provided for RDM – 28 respondents (74%) either strongly agreed or agreed, with the majority of respondents agreeing rather than strongly agreeing.

Therefore, the above factors can all be seen as major challenges for respondents in managing data effectively.

The factors that were perceived as least challenging were: the lack of sufficient storage during the life of the research project – only 11 respondents (30%) agreed/strongly agreed, followed by the lack of suitable metadata standards & ontologies in a subject discipline – only 14 respondents (38%) agreed/strongly agreed with this statement.

For both lack of sufficient data storage beyond the life of the research project (long-term) and lack of institutional mandates for RDM, about half of respondents (18 and 21, respectively) either agreed/strongly agreed that these were a challenge. The remainder selected the 'neither agree nor disagree' option (11 and 13 respectively), indicating that they were not sure that these factors caused a challenge. Only 4 respondents (11%) either disagreed/strongly disagreed with the statement that the lack of institutional mandates for RDM is a barrier to effective data management. Only 21 respondents (55%) saw the lack of institutional mandates for RDM as a challenge. A further 13 respondents (34%) were indecisive about this being a challenge. This indicates that respondents are either unaware of the importance of institutional mandates for effective RDM or are guided by funder and journal data management requirements.

4.2.4 RDM training and support needs

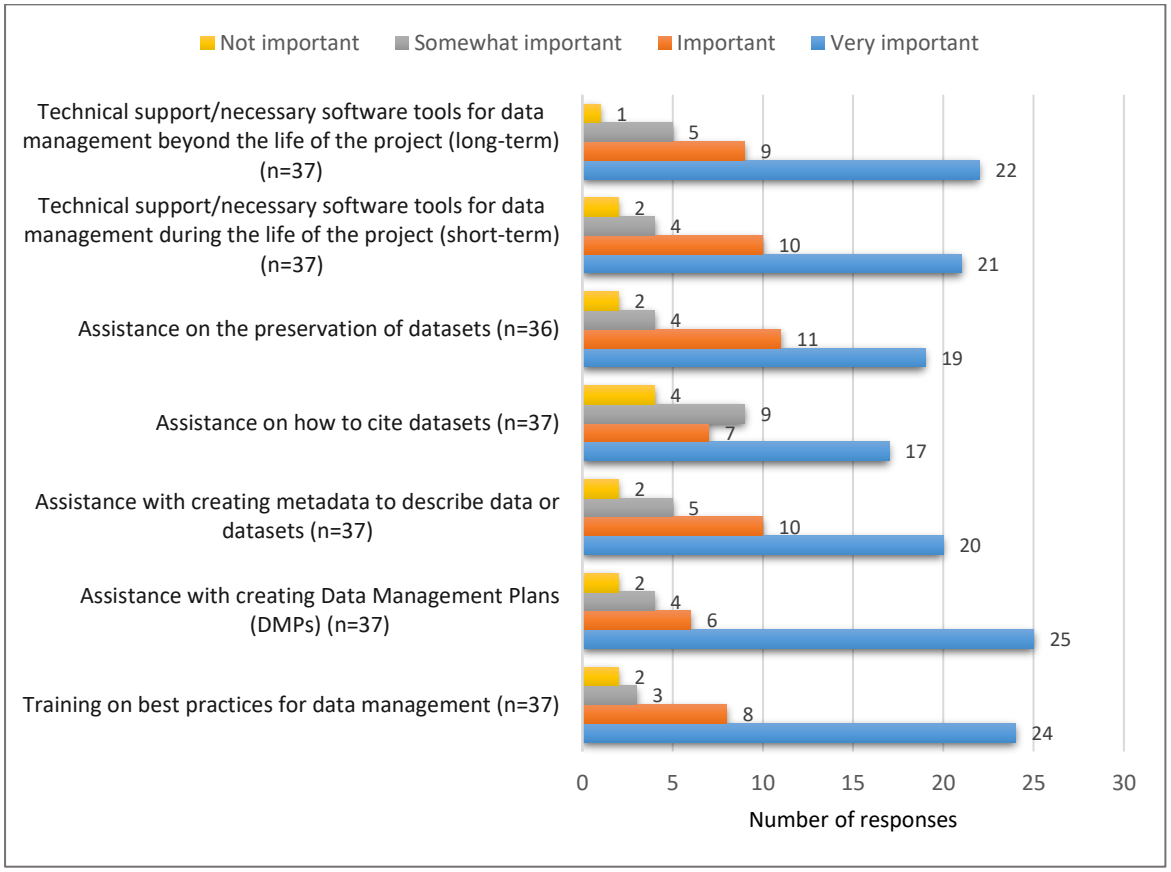

Figure 4. 22: RDM training and support needs

Respondents were asked to indicate the importance of certain types of training and support via a Likert scale. Figure 4.22 shows that respondents require different training and support services in the area of RDM. For each of the trainings/services listed, more than 80% of respondents indicated that it is either very important or important. The only exception was assistance needed with citing datasets, where 24 respondents (65%) indicated that this service is either very important or important. The majority of respondents indicated that a training/service is very important rather than only important, demonstrating how strongly they feel about training/services for RDM. The training/service chosen by most respondents (32; 87%) as being very important/important was training on best practices for data management, followed by assistance needed with creating DMPs (31 respondents; 84%).

4.2.5 Additional RDM-related comments, concerns, or issues

The last question in the questionnaire provided an opportunity for respondents to add comments or raise issues not already covered in the questionnaire. Seven researchers replied to the question, and below are the responses received:

- The nature of my particular data is such that I can easily manage it myself. It is simple and straightforward. Do not need any assistance.
- I mostly consider data storage as a personal endeavour, not the university's responsibility.
- Data management is a key component in the potential for future work to generate new information from existing data. Insufficient RDM structures and support severely reduce the potential of data quality and quantity of future publications.
- Journals and funding agencies are requesting that data be available or stored. Wits need to assist with this data storage, but to do this, a reliable ICT network is required. People with training and who are willing to listen to researchers are also needed.
- It is imperative that the university subscribe to sufficient cloud-based and local storage to ensure that every researcher's work is completely safe and secure. I have approached them … and they seem completely unwilling to address the issue. Other initiatives … are more supportive, such as WIReDSpace, but they have limited data capacity. We desperately need an efficient, secure, and safe place to store [data] both in the short term and in the long term.
- It would be great to get some overview of where and how to manage, store, and name data.
- It is unclear what options are available to staff. By putting info on websites makes it challenging for academics to figure out the university's system – no one has time to read a tonne of text or take lots of training sessions. If there is something already (freely) available to staff, I don't know about it.

From the above, it can be seen that researchers have concerns about insufficient RDM structures and support. This corresponds with findings from the previous question that identified challenges/barriers that hinder the effective management of data. Respondents also expressed the need for sufficient cloud-based and local storage as well as an efficient, safe and secure place to store data both in the short term and the long term. In addition to this, they need trained staff that are “willing to listen to researchers”. Training needs include

learning how to manage, store and name data. Another researcher expressed the concern that what Wits offers in terms of RDM is not clear.

4.3 Qualitative data analysis

Five researchers from a range of life sciences sub-disciplines were interviewed to gain in-depth detail of the central phenomenon and inform results obtained from the quantitative part of the study.

The semi-structured interviews consisted of sixteen questions.

Responses are divided into the following broad themes: RDM practices; RDM training and support needs; and RDM challenges and barriers.

4.3.1 RDM practices

4.3.1.1 RDM planning and design

Institutional policies and procedures

One of the questions posed in the questionnaire was to what extent respondents agreed with the statement that a lack of an institutional mandate, for example, a university policy on RDM, was a barrier to effective RDM. The researcher wanted to find out about awareness of policies/procedures for RDM at Wits or within schools or research groups. None of the interview respondents was aware of any policies/procedures, although some mentioned that there had been talks about them.

Considering the answers received, the researcher wanted to know whether respondents think formal processes or policies would make a difference to the management of research data in their research group or the university. Two respondents agreed that formal policies/processes would make a difference, whilst three agreed to an extent but had some reservations, as shown in their statements below.

> Formal processes … would make it easier. If they have something that can say – here's a way to do it and streamline things and make it easy. (R1)

> I think a lot of our data is collected by students who write theses and then disappear with that data, or they leave data on a computer … and you never know what is the actual data. (R3)

> I think a formal process would be helpful, but I think it would also be awful with [more] administration and bureaucracy. (R2)
>
> I don't think formal processes will make a difference … I would never put data that has not been published on any repository, whether it is internal or external. (R4)
>
> I think if Wits has some sort of policy or plan to tell people that you have to keep your data safe, to first file IP and [then] publish, and where to deposit, that would be helpful … especially when responding to funders. (R4)
>
> All we really need is proper IT support. Having policies in place is great, but if there's no adequate, strong, transparent, effective IT support, we are not going to manage this ... the policy on its own is not going to do anything. (R5)
>
> It should make a difference … if you know that's available … and if we make sure that our data is organised and stored properly, then there is no excuse for people losing data. (R5).

Funder requirements for data management

As most researchers in this study that are publicly funded need to adhere to funder requirements, a question about the extent to which funder requirements are a challenge was posed to respondents. Three of them expressed finding it difficult to deal with DMPs, especially the lack of guidance offered:

> They just want a section on data management, and that’s it. You have to do something without anything saying what you need to do. (R1)
>
> They don't really say what they want included, they just want to know what your data management plan is. (R2)
>
> No, trust me, we seriously don't know what they want, but they ask us for a DMP, and then it becomes super confusing as what do you write about it? (R3)

The two other respondents did not have any challenges with funder requirements. Both of them used data from and stored data in OA repositories and were familiar with funder requirements on depositing data in these repositories as well as with DMPs. Other funder requirements, such as data sharing, did not seem to pose a challenge to respondents.

4.3.1.2 Data collection and capture

Volumes and types of research data

Quantitative results indicated that half of researchers held data volumes in the gigabyte range with 16% in the terabyte range. The researcher wanted to determine if data volumes and the many different kinds of data types used posed a challenge to researchers in terms of data management. Two respondents indicated that increasing data volumes are becoming a challenge for them as sufficient storage space is needed, which can be costly. They said the following.

> Data that is GPS specific increasingly involves a lot more data. Handling data volumes is therefore a challenge. (R1)

> The volume of data that we produce now is quite large. We ended up purchasing terabyte hard drives, but those are expensive. (R2)

The other interviewees did not have a challenge with the volume or types of data they had to deal with. They either used raw/experimental data that were manageable or existing data stored somewhere else.

Metadata and metadata standards usage

Two-thirds of questionnaire respondents indicated that they assign metadata to their data, whilst 44% of those who assign metadata also used metadata standards. Following on from this, the researcher wanted to establish if respondents view the use of metadata and metadata standards as important or not and possible reasons for their views.

All respondents agreed that assigning metadata to your research data is important, but they highlighted other issues, as follows:

> It is important, but again we don't have any training on exactly what to do or how to do it, and I think that's a problem. (R1)

> Metadata is incredibly important, but it's something that we are very lazy to use. (R3)

In response to why it is important to assign metadata, most respondents said that it is necessary as it can provide information on the data, for example, how and when it was collected and by whom. It can also provide information on how the data were manipulated and how the analysis was done.

> When we download the data sets, then we need information on how it was generated and what kind of technology was used. We are personally not generating anything big, so at this stage, we don't personally worry about it too much, but when we are presenting data to the public, they really need to know the description of the data. (R4)

A further reason respondents gave why metadata needs to be assigned is that it enhances discoverability and thus reusability.

> You generate data for basically everything that happens in a tissue [group of cells] at that time, and you'll be looking at a tiny aspect of what you're analysing and reporting on that. The rest of the data is already generated … we are not interested in it for now, but someone else may be. They will need to have access to all of the data so that they don't have to repeat the experiment. (R5)

When interviewed about metadata standards, one of the respondents admitted not knowing what metadata standards are. Others indicated that usage of standards could be useful, though they admitted that they did not use them:

> I've never used it, but it would make sense to use it if such a standard is available in one's field (R1)

> I think it could be useful if there's several points in the standard that could be included but are not necessarily required. With data being somewhat personal, sometimes it's challenging to meet all of the requirements, but just because they may not be relevant. (R2)

> Having any sort of standard, I think, would be awesome … I think a lot of the data that is collected can be archived in very similar ways, if adhering to standards, which would assist with things like doing meta-analysis. (R3)

Following from questionnaire results (Figure 4.11), interviews were used to establish why researchers used certain metadata standards. The one respondent who used standards remarked that some publishers require deposit in certain repositories, which require adhering to metadata standards.

4.3.1.3 Data storage (long term)

Almost half of the questionnaire respondents indicated that they have stored data for long-term preservation. The interviews explored what researchers think about long-term preservation and the reasons for storing data for long-term preservation.

Although not all respondents indicated having stored data for the long term, all agreed that it is important for different reasons such as making data accessible for future research, saying:

> It will always be useful to future scientists. (R2)

> We are revisiting some questions 10 years down the line. It's great to be able to go back to the historical record. (R3)

> I think none of us think through it properly, and then at one point, you realise you need to dig something up from 10 years ago, and you can't open it … (R5)

One of the respondents admitted not knowing how to store data for the long term.

> I've certainly not done that, but I have thought about it, but kind of become stuck in where do I start? I do feel like we're limited in terms of knowing what our options are to better keep data safe for other people to use. (R1)

One respondent both agreed and disagreed that data should be stored for the long term saying:

> You need to keep a proof of whatever you have done … but in terms of long-term storage, [ensuring that data] is available even maybe after 30 years … I don't know if it is feasible to actually revisit that data set every few months, and then make file format changes and things like that. (R4)

The above respondent (R4) felt that that long-term preservation is not practical.

> I am thinking that it would be good to have old data somewhere where it could be used for the betterment of science … that you can use it for more than just what may have been a narrow focus at the time, and there could be other things that people would do with it better. (R1)

> Side projects are just sitting there. You never got around to publishing, there's always someone who is interested in that or will be at some point, and having the data handy, or being able to share, is quite important. (R5)

As many researchers are mandated by publishers and funders to store data for the long-term, this study wanted to determine whether respondents would consider storing data for the long-term even if it was not mandated and what the reason for this could be, but unlike the questionnaire respondents, none of the researchers said that they would only store data for the long-term if it was mandated.

Questionnaire data showed that researchers used a variety of repositories (Figure 4.18). Respondents were questioned about their usage of repositories for the long-term preservation of data to better understand why they have used certain repositories and highlight challenges they may have with certain repositories. Responses were as follows.

> I'm aware of Dryad, which is fairly common in our field. I have not used it because of the costs associated with it. I also use OSF … it's a free repository. I found it to be very useful because I can keep my data to myself or share with my collaborators, and then once it's published, I can make it readily available. (R2)

> One of the researchers used microarray data repositories such as MIAME for long-term preservation, as it is a requirement by the journals that any high-throughput data that we produce should be stored in an international repository at the time of submission of the paper … So that is what we use, and that is where it is safe. (R4)

When questioning the researcher about whether it was a challenge in finding and using these repositories, the following response was received:

> We know the repositories, and even journals ask you – have you submitted your dataset in here? (R4)

R5 confirmed the experiences of R4 saying:

> They usually do require that all sequencing data is in the public domain by the time you submit it. You get clear guidelines when you're submitting articles. (R5)

4.3.1.4 Data sharing

Following a question posed in the questionnaire about under which conditions researchers were willing to share data for future research (Figure 4.20), respondents were asked to elaborate on the topic by commenting on how they feel about sharing data. In general, these respondents did not have a problem with sharing data on the condition that it is already published:

> I'm more than happy to share any published data. I am more wary of sharing unpublished data if I am not a collaborator or included in a project. (R2)

Two of the respondents, both from MCB/genetics, advised that they have to share data as it is a prerequisite for publishing in some journals.

> It has to be shared publicly, that is why we have so many repositories available. There is no hiccup in sharing data. We don't like to share data before we have actually made use of it and published it. (R4)

> For certain quite prestigious journals, it's a prerequisite that whatever material you used in publication of the article must be made freely accessible to whoever asks you for it. You don't really have a choice about it, and generally, we don't have issues sharing. It's nice when someone actually asks you for something … especially if it's someone well established in the field … though, we don't necessarily like if it's a large chunk of work that's already been completed. (R5)

A respondent who researches in the area of conservation ecology mentioned that, as researchers in Africa have limited resources to analyse data, international researchers with more resources could be in a better position to do something with that data after it is made open.

> I need to publish a paper. [After publishing] I've still got three other papers I'd like from that data, but the journal forces you to make that data open. Other researchers can have access to that data, they may be able to, with much more support and sophisticated systems, beat you to ideas you might have had … after you have put so much effort into getting funding and collecting data. That's one of the risks of making data open. Happy to share, concerned about other things in terms of what that sharing means, and would ideally like to share so that you're still benefiting from the sharing. (R1)

One of the respondents working in the field of environmental science and ecology had no problem with sharing data as the researcher did not produce raw data, and their specific research field required constant use of existing data and data sharing across different OA repositories. The researcher stressed the importance of openness:

> By having a collaborative consciousness, we are going to ask better questions. People are naive in some ways to think that they've collected the world's best data, and only they can work on it, and only they can answer questions. (R3)

Quantitative results showed that only 25% of respondents had published data sets as supplementary information to published papers (Figure 4.21). The interview probed whether respondents who have linked a dataset to a paper published have noted an increase in their

citation count when compared to papers without datasets. The reason for this question was to see if the presence of datasets published with a paper may be seen in a more positive light.

Only one of the respondents advised that they've seen a substantial increase in citations from a paper where the data were deposited in an open repository:

> I think that my masters' paper gets more citations than it would normally get because those data are part of that repository. (R3)

The respondents who did not notice increased citations from linked datasets commented that citations mainly depend on the type of publication, whether one has international collaborators, and at what scale work was done.

4.3.1.5 Discover and reuse

Results from the quantitative part of the study showed that although 98% of researchers generated new data, 65% used existing data. Interviews were used to determine why some researchers are not using existing data. For those respondents who did use existing data, the question was asked if they experienced any challenge in finding and reusing data and, if so, what the reasons for these are.

R1 and R2 mostly generated their own research data and rarely experienced challenges in cases where they tried to source existing data.

Respondents who used existing data experienced limited challenges with finding and using this data:

> I almost exclusively use existing data that is openly available. As it's citizen science, there's errors in the data, and we need to put quite rigorous protocols in place to clean the data … You need to be very aware of how data were collected and what you can and can't say by using them. (R3)

> I did not really have any challenge in finding the data and reusing it. The data in the repository is submitted according to certain standards. The only challenge can be the experimental side of that data set, but searching for it is not an issue. (R4)

> We always rely on genetics databases such as NCBI's sequencing databases for existing data…. and usually do not experience any challenge. It requires a basic skill level … we're managing fine. (R5)

4.3.2 RDM training and support needs

In the questionnaire, assistance with creating DMPs was identified as one of the most important training needs for researchers. In the interview, the researcher wanted to establish why respondents thought there was a need for such training. Their responses echoed their earlier responses about funder requirements.

> Because it is so unknown and came about recently. You know there is a need for it but just don't know where to start. We don't know what's good practice or what's not. (R1)

> There isn't many guidance, and my experience with students and even staff is that people don't really know how to manage data. (R3)

> I think we need to know exactly what is required of us when we write data management plans for funders. If there is some assistance … and we are aware of resources available within the Institute, or a standard format that is created by Wits, I would be most happy to use that. (R4)

> I think younger researchers or new academics struggle with creating DMPs. From my experience, [a DMP] could make a difference between your grant getting funded versus being thrashed. (R5)

Two of the researchers in the field of environmental science and ecology said that students should be made aware of the importance of data management from an undergraduate level. All other respondents agreed that solid basic training in creating DMPs should take place from a postgraduate level, and some also said postdocs and young academics/supervisors should be trained.

> Anybody collecting data [such as] postgraduates, but maybe also the supervisors so that they can make sure the postgrads are implementing [the training]. (R1)

> Our students are collecting data constantly … we should start talking about data management in second year and upwards … so they're aware of it and so it contributes to being organised [when it comes to] data. (R2)

> Even in first and second year we should be talking to them about data management … but honours would be a more realistic starting point. (R3)

> I think that all the researchers, including our postgraduates, master's and PhD students, should be given a basic training [in data management]. (R4)

> PhDs, postdocs and young academics because they do have to put data management plan into their proposals. (R5)

The interview also explored by whom training should be conducted. R1 said a data specialist should train supervisors in the university and who can then pass that information onto their students. Other respondents said that training in DMPs should be conducted by a mixture of people with IT/data management/data science backgrounds and researchers from the discipline in question.

R3 suggested that training should be conducted by someone “used to working with data and [who] understands ecological data”… someone that “knows the type of data we're likely to encounter and how best to store it”. The respondent suggested that, unless the university can have a designated position for a data scientist, an external person should be consulted.

> Someone trained in data management …and specialised in the discipline. (R4)

> … an IT person with an idea of what's available in terms of IT support and researchers who can advise on the discipline’s specifics. (R5)

4.3.3 RDM challenges and barriers

Towards the end of the interview, respondents were asked about their major RDM challenges, and to expand on why they see them as challenges. Responses included the following:

> Not knowing enough about managing research data … We need the very basics ... (R1)

> Not knowing, first of all, what is available at Wits as it is not really publicised … and who is the person who can help us. There is no standard [data management] plan available at Wits or by any of the funders within South Africa – that is definitely a major hiccup. (R4)

> For me, it's primarily storage space … It would be ideal if Wits reduced our admin so that we can focus more on our science and data management. If Wits isn't willing to set up a repository, not having us write motivation letters for every hard drive, or computer or server that we need to purchase to store the data that we're generating [would be ideal]. (R2)

Two of the respondents did not have major challenges in terms of RDM. R3 mostly used existing data, and R5 stated that RDM is not a major challenge as their work is not very data intensive.

> I don't [have challenges], and it is because so much of [my data] is stored online already or is accessible via repositories. (R3)

> [RDM] has not been as big a problem for me as trying to get funding for my work … it's something that we get done, and we are managing … If my work was very data intensive, that would be one of the top three priorities. (R5)

4.3.4 Suggestions on how the university can assist towards efficient RDM

The last question was posed to respondents to get their thoughts on how the university can assist towards best practice and efficient RDM. The question was asked to allow respondents to freely express their views to further inform any of the questions posed in both the questionnaire and interviews.

Four respondents stated that the university could assist with more storage space, such as repositories that can handle different kinds of data. Below are some comments received regarding data storage:

> Having a straightforward system where researchers can have their data and then move it over, create a DOI to go with the manuscript, then it's readily available, but the unpublished data would not be available. (R2)

> There needs to be enough space, consistent backup and make sure [data] integrity is not compromised and that it’s protected. (R5)

Four respondents also indicated that the university should assist with more guidance and training on different aspects of RDM, such as creating a DMP, RDM best practices and implementing different aspects of RDM. Their answers confirmed questionnaire results where more than 80% of respondents rated the need for training in DMPs and best practice and technical and software support and tools during the research lifecycle as either very important or important.

Some respondents also mentioned that users need to know what is available from the university in terms of RDM, for example, a standardised DMP and certain repositories, IT support and RDM services.

> They need to create something like a pamphlet or a small, not large policy. Something informative saying … these are the repositories where you should submit your data, and this should be your standard data management plan. (R4)

> There needs to be clarity on who to deal with and clarity on the proper level of support in terms of IT etc. (R5)

Some further suggestions with regards to training also corresponded with questionnaire results that highlighted the need for training in best practice, using metadata and choosing suitable repositories:

> Something like a practical workshop to say that these are the kinds of data; these are different repositories; these are your options; this is metadata … (R1)

> Some designated individuals should work with a data scientist to learn about data management and different types of storage facilities, and then those people can be running courses maybe on a yearly basis. (R5)

In addition to the above, respondents made some further suggestions on how the university can assist towards good and efficient RDM:

> Reducing administration in other areas where it seems unnecessary. (R2)

> First thing would be to identify really key people whose work is data intensive … that really require intense data support, one would focus on them. There will be general trends emerging, and then something can be built to accommodate them, or existing facilities repurposed. (R5)

4.4 Summary

In this chapter, data collected from the questionnaire were presented by means of tables and figures. Qualitative data, collected by interviews during the second phase of research, were then presented thematically.

CHAPTER 5: DATA INTERPRETATION, RECOMMENDATIONS AND CONCLUSION

5.1 Introduction

The final chapter of this dissertation mixes and interprets qualitative and quantitative results to address the research questions and inform the study problem.

Main findings from the data presented in Chapter 4 will be discussed and practical recommendations provided in order to achieve the aim of the study, that is, to assess RDM readiness in the life sciences at Wits in order to ascertain what support is needed with regards to RDM. The chapter will also provide recommendations for further studies.

5.2 Study findings

The findings of the study are presented according to the two research questions.

5.2.1 Current RDM practices and needs of researchers in the life sciences at Wits

The main findings related to the first research question are discussed below according to the different stages in the Jisc Research Data Lifecycle, which was the framework that guided the literature review and framed the data collection instruments and data analysis.

5.2.1.1 Plan and design

At the time of this study, Wits University had not created or implemented a formal institutional RDM policy. This study found that respondents were not aware of any other RDM policies/procedures at Wits. Although some respondents thought that formal policies/procedures would be very helpful, some had reservations and expressed the fear of increased administration and bureaucracy resulting from a policy. The opinion was also raised that having policies in place is appreciable, but on their own and without proper IT support, would be worthless.

Most respondents in this study were publicly funded, the overwhelming majority by the NRF with requirements from funders including the need for a DMP, sharing data in a repository and making research data completely open. These requirements were not surprising to discover considering the NRF's OA statement (NRF, 2015). The findings from this study are in sharp contrast to Patterton's (2017) study, which showed little awareness of funder RDM requirements among those participants. The NRF's OA statement has thus, over time, had a visible effect on local RDM practices where researchers are now more aware of funder requirements than they were a few years ago.

5.2.1.2 Collect and capture

Dataset sizes held by researchers showed great variance, with half of researchers using datasets in the gigabyte range and 16% using datasets in the terabyte range, confirming findings from other studies that the handling of large datasets are common in the life sciences (Marx, 2013: 255; RIN & British Library, 2009: 44). The finding that very large data sets (more than 100 TB) were held by the smallest percentage of respondents corresponds with some local studies (Koopman, 2015: 67; Patterton, 2017: 163).

Researchers used several data types, with the most common being spreadsheets/tabular data, images and documents. This finding corresponds with data types used mainly by life science researchers internationally (DataONE, 2021b; Kvale, 2012: 49; Saeed & Ali, 2019: 293), as well as locally (Koopman, 2015: 68; Patterton, 2017: 160).

Most researchers (70%) indicated that they assign metadata to their research. Although these researchers are aware of metadata and understand the importance of assigning metadata, some expressed that they did not know how to use metadata. Only about half of questionnaire respondents who used metadata indicated that they also use metadata standards; not all researchers knew what metadata standards are. The relatively low use of metadata standards has also been found in other studies that include life sciences researchers (Tenopir et al., 2020: 22) and is a concern as the use of metadata standards is vital for data discovery and reuse. Researchers who had not used standards before expressed that using metadata standards could be useful. The study found that the metadata standards mostly used are those standardised within research groups (this finding corresponds with Tenopir et al.'s [2020: 17] study) followed by standards in molecular and cell biology. A reason for the latter could be that publications in these fields often require data deposits in repositories that use metadata standards.

5.2.1.3 Collaborate and analyse

Software applications for data analysis mostly used by researchers were Excel, MSWord, R, ArcGIS, ImageJ, SigmaPlot and SPSS (Figure 4.11). The popularity of ImageJ and R indicate the move to open-source programs in data analysis (Rueden et al., 2017; R Foundation, 2021).

In terms of collaboration, researchers in this study preferred to share their data with researchers with whom they work, rather than with other researchers. This sharing practice is

common, having been seen in other studies amongst life sciences researchers (Kvale, 2012: 50; Tenopir et al., 2011: 12).

5.2.1.4 Manage, store and preserve

During the active research phase, almost an equal number of researchers opted for storage on an external hard drive (86%), a hard disk drive of an office desktop/laptop and cloud storage (83% each). Compared to other local studies that included life sciences researchers, the amount of cloud storage is significantly higher. For instance, only 39% of Koopman and de Jager's (2016: 5) study and 33% of Patterton's (2017: 172) study indicated that they used the cloud for storage. This could indicate that cloud storage is on the increase and that it has become more popular and reliable in recent years.

Many researchers (38%) still used paper or paper laboratory notebooks for data storage, whilst only 5% used ELNs for short-term storage. Tenopir et al. (2015: 14) found that researchers based in Africa are more likely to store their data on paper than North American researchers. Tenopir et al. (2015: 17) also found that researchers in biology are more likely to store data in their offices on paper than in other fields.

Those researchers who indicated that they store data for long-term preservation did so because they are mandated to do so, much like those in the Koopman and de Jager (2016: 5) and Renaut et al. (2018: 404) studies. Some, however, were of the opinion that they would consider data preservation even if it was not mandated. The 50% of researchers that did not store their data for the long-term is a concern as this data can be lost for future use if not preserved.

Some life sciences sub-disciplines are more used to depositing in repositories than others. The fact that researchers in molecular biology are used to routinely archiving in repositories (Laloë, 2017) was confirmed by respondents from this field. As molecular biologists in this study practised data preservation regularly, they did not express any difficulty using repositories for long-term preservation.

5.2.1.5 Share and publish

The majority of respondents were willing to share data to some extent, but with conditions. Mostly respondents did not mind sharing their data informally, such as within research groups. Similar results were found in local studies (Bangani & Moyo, 2019: 11; Patterton, 2017: 187).

Researchers in the fields of molecular biology and genetics especially noted that they had little choice about sharing data in repositories, which is not surprising as using repositories for the sharing of data has been part of these disciplines for some time (Laloë, 2017; Renaut et al., 2018: 407). An optimistic attitude towards the sharing of data was especially noted in the field of environmental science and ecology, which correlates with a recent study that indicated that the attitude of these researchers towards sharing was the most positive of the group investigated (Tenopir et al., 2020: 15). Researchers valued the openness of data sharing, with one interviewee referring to the need for researchers to have a "collaborative consciousness", suggesting that data openness could ensure that research questions are better answered.

Nevertheless, most researchers only want to share their data after publication. The same was noted in other studies, locally (Koopman & de Jager, 2016: 4) as well as internationally (RIN & British Library, 2009: 7; Tenopir et al., 2015: 16, 2020: 16). The study by Tenopir et al. (2015: 16) also found that the desire first to publish was significantly higher in disciplines such as biology.

5.2.1.6 Discover and reuse

As two-thirds of respondents used existing data in their research, it is evident that there is a move to open access in the life sciences disciplines and that data are being reused. Data-driven discovery is an integral part of the molecular field (Thessen & Patterson, 2011: 16), and researchers from the molecular disciplines that took part in this study confirmed that they rely on repositories to a large extent. As repositories in this field are mostly standardised, researchers had limited challenges finding and using existing data.

5.2.2 RDM challenges of researchers in the life sciences at Wits

Researchers were specifically asked about the challenges they face in terms of RDM in both the quantitative and qualitative phases of research. However, challenges may also have emanated from needs that were highlighted in the study.

Lack of training stood out as the major RDM challenge experienced by researchers. The lack of RDM training was noted by another local study (Patterton, Bothma & van Deventer, 2018: 21). Most of the listed RDM training and support needs in this study were regarded as either very important/important, supporting the significance of this challenge. A need for the basics, such as a practical workshop in data management planning, was expressed.

Although researchers were aware of RDM requirements from funders, they expressed having a specific challenge with the requirement of a DMP. South African researchers are fairly new

to funder requirements such as the need for a DMP (Patterton, Bothma & van Deventer, 2018; van Deventer & Pienaar, 2015), and that some are still experiencing challenges with creating a DMP is therefore not surprising. Researchers expressed frustration with “not knowing” what should be included in a DMP. As far as researchers were aware there is no standard DMP available at Wits or from any of the funders within South Africa. This was highlighted as a major challenge.

Other major challenges hindering effective RDM were the lack of funding for infrastructure, the lack of organisational structures for the effective management of data across different divisions, and the lack of services/support provided for RDM (Figure 4.21). The lack of awareness of available services and tools was raised as a major challenge in the qualitative part of the study.

Another opinion raised in this study was that insufficient RDM structures and support severely reduced data quality and subsequent publication quality. Institutions having organisational structure issues in terms of RDM was noted as the top challenge in Chiware and Becker’s (2018: 11) study that investigated RDS in Southern African libraries. The study noted the lack of IT infrastructure as a significant challenge. Some researchers in this study believed that a reliable ICT network is truly essential for RDM infrastructure.

Although only about half of respondents to the questionnaire agreed with the statement that the lack of long-term storage hindered effective RDM, the qualitative part of the study highlighted it as a major issue, with respondents strongly expressing challenges of storage/repository space.

Most researchers did not see the lack of an institutional mandate for RDM as a major barrier to effective RDM, in contrast to Chiware and Becker (2018: 11). Qin (2013: 218) considers mandating RDM to be an important driver for adopting RDM practices. In addition, Pryor, Jones and Whyte (2014: 91) say that it is needed for RDM support services to be created and used.

5.3 RDM readiness in the life sciences at Wits

The study showed that life sciences researchers at Wits had adopted many RDM practices, and researchers are increasingly becoming aware of the importance of the openness of data. In the area of molecular and cell biology, it was noticeable that researchers are accustomed to data practices such as long-term data storage and data sharing, mainly as these practices are mandated in the discipline. These researchers also did not experience challenges with

data discovery and reuse, as repositories in the field are mostly standardised. A positive attitude towards the sharing of data was also noted amongst researchers in ecology and environmental sciences. However, it was found that researchers generally struggle with similar data management issues as their peers both locally and internationally, such as long-term data storage, insufficient knowledge of DMPs and a lack of awareness of available services and tools as well as training. It seems that the RDM readiness of researchers would be enhanced by an institutional RDM policy. Data policy provides a framework for researchers and projects to establish a routine of RDM practices (Qin, 2013: 18). Policy is also seen as an important step towards delivering RDM support services (Pryor, Jones & Whyte, 2014: 91).

5.4 Recommendations for RDM support

This study sought to discover the RDM practices and needs of life sciences researchers at Wits and identify related challenges. By doing so, it has established the RDM readiness of this group of researchers. The purpose of the investigation was ultimately to inform the level of support that should be offered to them. Based on the main findings from this study, recommendations for services and infrastructure are now provided. Although the recommendations are based on results from the study conducted among life sciences researchers at Wits, the recommendations could be of value to the whole Wits university research community.

5.4.1 Institutional RDM policy

While many respondents in this study did not see the need for formal policy for effective RDM (possibly as this may lead to more administration for them), the literature supports the creation and implementation of RDM policy. An institutional RDM policy should be prioritised as not only funder policies, but institutional policies are important driving forces for the adoption of RDM practices (Pryor, Jones & Whyte, 2014: 91; Qin, 2013: 218). Renaut et al. (2018: 409–410) stated that “Research Institutes granting degrees should enforce explicit data management, archiving and sharing policies”. Policy would also ensure better uptake of RDM services offered by the library, IT and the RO.

5.4.2 Data management plans

Although funders usually have their own DMP templates, the need for a generic template for Wits researchers was expressed during the study, especially for researchers that conduct research without funding. Ideally, an RDM policy should mandate the completion of a DMP for all researchers, including masters and PhD students. Master’s level studies usually generate

a significant amount of research data that might be used in future if preserved; this should be planned for. Examples of online DMP tools that can be used or adapted by the university are DMPonline (DCC, 2020), DMPTool (California Digital Library, 2020) as well as the SA-DMP Tool (DIRISA, n.d.). Currently, examples of DMPs and links to some available tools such as the SA-DMP Tool are available at Wits on the LibGuides and the intranet. The Data Services Librarian also created an NRF DMP template pilot. Not all researchers are aware that these are available and how they can be accessed.

5.4.3 Long-term data storage

The study highlighted the need for long-term data storage facilities with the following requirements: sufficient data storage for large data volumes and the ability to store both published and unpublished data safely and securely.

Recommendations for long-term data storage would be based on funder and journal requirements or specific researcher needs for long-term data storage. If long-term storage in an OA repository is required, researchers can be referred to general-purpose OA repositories such as Zenodo, OSF or Dryad. The Wits IR, WIReDSpace, does not sufficiently cater for the storage of open data. The library is investigating the integration of CKAN (“Ckan”, n.d.) with the current DSpace module in order to cater for the storage of open datasets on WIReDSpace (Lewin, personal communication 2021, May 20). Using the local IR to host data of Wits researchers will not only make data openly available, but it can contribute to the showcasing of university research output.

For long-term data that needs to be secured, iRODS (the Integrated Rule-Oriented Data System), a type of middleware that offers highly secured, controlled access to both published and unpublished data (“IRODS”, 2021), is recommended. Implementation of iRODS at Wits is in its early stages, and thus awareness making is needed at this stage.

In addition to the above, researchers can be referred to re3data.org, a comprehensive global registry of research data repositories that can assist in identifying a suitable repository for specific needs.

As different stakeholders are currently involved with development of long-term storage options, it is recommended that researchers are referred to one central platform, such as the RDM LibGuides, where information and contact details of long-term storage options could be found.

5.4.4 Funding for infrastructure and programmes for RDM

Effective RDM is crucial not only for current research but also for future research. Therefore, funding for RDM should become a standard budget item within the university, and provisions should be made to cover RDM items not always covered by funders, such as university infrastructure development and support programmes for RDM. It is further recommended that DMPs should include a section on expected RDM-related costs so that researchers can plan for costing of data management activities such as equipment, software and staff ("UK Data Archive", 2015). Researchers can also be referred to a helpful data management costing tool such as one from the UK Data Service ("UK Data Archive", 2015).

5.4.5 Organisational structures for RDM

Lack of organisational structures for the effective management of research data across different institutional divisions was identified as one of the major challenges hindering effective RDM. Evidently, researchers at Wits are unaware of the different RDM role-players such as the eResearch unit, Library and RO or the role that each one plays. Clarity is needed on how these units are structured and are collaborating to support RDM. Successful implementation of RDM will require increased collaboration (Patterton, Bothma & van Deventer, 2018: 19) and evaluation of the roles of the different stakeholders such as the Library, ICT and RO.

5.4.6 RDM training and support

Wits already offers some RDM training and support services, but the study highlighted that researchers are either not aware of them or they do not meet their needs. This study, therefore, recommends either an increase or redesigning of the following training and support services for researchers, including those at postgraduate level of study:

- Assistance with creating DMPs, including the implementation of a DMP tool
- Assistance and training on the creation of metadata and the use of metadata standards
- Technical support for RDM during the life of the research project, including the availability of necessary software tools for data management, for example, increased cloud storage, Research Electronic Data Capture (RedCAP) and iRODs
- Technical support for RDM beyond the life of the project (long-term), including the availability of necessary software tools for data management, for example, OSF, Dryad or iRODS
- Assistance with preservation activities, such as ensuring data integrity and considering data formats for preservation

- Training on general best practices for data management, such as sharing data in repositories/archives that comply with FAIR principles

In addition, assistance and training should be provided by a mixture of people with an IT/data science background and subject specialists/researchers to address subject-specific data needs for different life sciences disciplines. Griffin et al. (2018: 15) highlighted the value of domain-specific RDM training in the life sciences.

Libraries are increasingly involved with RDM services such as offering training and support for creating DMPs, best practices for data management, preservation of datasets and using metadata to describe datasets (Latham, 2017: 263). These services are all offered by the Wits Library, but need tailoring for specific disciplines. Awareness is also needed, and this is discussed below.

5.4.7 Marketing and awareness

The study found that many researchers are not aware of RDM services already available from the university. Other studies found a limited awareness amongst researchers in the life sciences of available resources that can assist with handling data throughout the data lifecycle (Griffin et al., 2018: 15). RDM services at Wits include the availability of a Data Services Librarian, LibGuides for RDM, as well as services to assist with DMPs available from both the Data Services Librarian and the eResearch office. As information on RDM seems to be spread over too many platforms, services need to be streamlined, and clarity is needed on the role of each of the stakeholders. The role of stakeholders should also be made clear and marketed on already-established platforms such as the LibGuides. A suggestion is that the Research Support LibGuide, which has very high usage, has a link to the RDM LibGuides. These guides should then be marketed amongst all postgraduates and researchers. Tenopir et al. (2020: 13,21–22) stressed the importance to “better publicise support”, whilst Patterton, Bothma and van Deventer (2018: 20) recommended that marketing and awareness form an integral part of an RDM implementation plan.

5.5 Further studies

The findings from this study highlighted the need for mandating RDM at Wits. Further studies may include the development of an implementation program for formal RDM at the university. As the study focused on the practices, needs and challenges of researchers in the life sciences, future studies could investigate the RDM practices and needs of researchers from other subject fields at the university to better inform RDM training and implementation.

5.6 Study limitations

The time during which the online questionnaire was distributed was seen as a potential limitation of this study. The questionnaire ran towards the end of the academic year, and a large portion of the study population was masters' students who had to submit their research projects. As a poor response was anticipated, the researcher tried to manage this limitation by sending two follow-up reminders to the study population. This improved the response rate, but a higher response rate would have yielded more significant results.

The response rate to the questionnaire was not high enough to test for significance which decreased the generalisability of the results. In addition, the small number of respondents did not allow in-depth comparison between sub-disciplines – no radical differences in practices or challenges were noted – or much comparison with the literature. Also, no significant results showed up in terms of different researcher qualifications and research output.

Due to time constraints, only five participants were sampled to be interviewed. Although five is an acceptable number for a small study, this limited the range of qualitative results. One or two more interviews could have strengthened the study as it could have given more insights into explaining quantitative results.

5.7 Conclusion

The study concluded that life sciences researchers in the Schools of APES and MCB at Wits adopted many RDM practices, and researchers are increasingly becoming aware of the importance of the openness of data. They seem to be in support of the Open Science Movement underpinned by the move to OA, including the open publishing of datasets. However, they are trying to deal with similar RDM issues as their peers worldwide, some of them attributable to lack of awareness of available services and tools.

The development of RDM policy as well as development and marketing of RDM training and support are important steps to be taken by an institution wanting to implement RDM services effectively. As formal implementation of RDM still needs to occur at Wits, it is crucial to have an RDM policy, followed by suitable RDM infrastructure and awareness making of current services. The results from this study are in support of the university's strategic plan to provide RDM policy framework and infrastructure. This should then be accompanied by having an online DMP in place as well as increased RDM guidance and training, some of it discipline-specific, for all levels of research. As the proper management of research data is of increasing

importance, the results from this study can be of note for Wits as one of the top research institutes in Africa.

REFERENCES

ACRL Research Planning and Review Committee. 2016. 2016 top trends in academic libraries: A review of the trends and issues affecting academic libraries in higher education. *College and Research Libraries News*. 77(6). DOI: 10.5860/crln.77.6.9505.

ACRL Research Planning and Review Committee. 2020. 2020 top trends in academic libraries: A review of the trends and issues affecting academic libraries in higher education. *College and Research Libraries News*. 81(6). DOI: 10.5860/crln.81.6.270.

Altaf-Ul-Amin, M., Afendi, F.M., Kiboi, S.K. & Kanaya, S. 2014. Systems biology in the context of big data and networks. *BioMed Research International*. 2014. DOI: 10.1155/2014/428570.

Alves, C., Castro, J.A., Ribeiro, C., Honrado, J.P. & Lomba, A. 2018. Research data management in the field of ecology: An overview. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Porto, Portugal. 87–94.

Andres, L. 2012. *Designing and doing survey research*. London: SAGE.

Babbie, E.R. 2010. *The practice of social research*. 12th ed. Australia ; United Kingdom: Wadsworth Cengage Learning.

Bangani, S. & Moyo, M. 2019. Data sharing practices among researchers at South African universities. *Data Science Journal*. 18(1):28. DOI: 10.5334/dsj-2019-028.

Benioff, M. 2015. *The digital revolution needs a transparency revolution*. Available: https://www.weforum.org/agenda/2015/01/the-digital-revolution-needs-a-transparency-revolution/ [2021, October 30].

Bordelon, D. 2021. *What is research data management?* Available: https://pitt.libguides.com/managedata/understanding [2021, September 28].

Briney, K. 2015. *Data management for researchers: Organize, maintain and share your data for research success*. (Research skills series). Exeter, UK: Pelagic Publishing.

British Ecological Society. 2014. *A guide to data management in ecology and evolution*. London: British Ecological Society.

California Digital Library. 2020. *DMPTool*. Available: https://dmptool.org/ [2020, December 17].

Chiware, E.R.T. & Becker, D.A. 2018. Research data management services in Southern Africa: A readiness survey of academic and research libraries. *African Journal of Library, Archives & Information Science*. 28(1):1–16.

ckan. n.d. Available: https://ckan.org/ [2020, December 17].

Cook-Deegan, R. & McGuire, A.L. 2017. Moving beyond Bermuda: Sharing data to build a medical information commons. *Genome Research*. 27(6):897–901. DOI: 10.1101/gr.216911.116.

Creswell, J.W. 2008. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, N.J.: Pearson/Merrill Prentice Hall.

Creswell, J.W. 2009. *Research design: Qualitative, quantitative, and mixed methods approaches*. 3rd ed. Los Angeles: SAGE.

Creswell, J.W. 2012. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. 4th ed. Boston: Pearson.

Creswell, J.W. 2015. *A concise introduction to mixed methods research*. (SAGE mixed methods research series). Thousand Oaks, California: SAGE.

Creswell, J.W. & Clark, V.L.P. 2011. *Designing and conducting mixed methods research*. 2nd ed. Los Angeles: SAGE.

DataONE. 2020. *Data Life Cycle*. Available: https://old.dataone.org/data-life-cycle [2020, November 26].

DataONE. 2021a. *Best practice: Revisit data management plan throughout the project life cycle*. Available: https://dataoneorg.github.io/Education/bestpractices/revisit-data-management [2021, September 19].

DataONE. 2021b. *DataONE Data Catalog*. Available: https://search.dataone.org/profile [2021, May 01].

DCC. 2019. *DCC Curation Lifecycle Model*. Available: https://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf [2021, October 24].

DCC. 2020. *DMPonline*. Available: http://www.dcc.ac.uk/dmponline [2020, December 17].

De Vos, A.S., Strydom, H., Fouché, C. B., & Delport, C.S.L. 2011. *Research at grass roots: For the social sciences and human services professions*. 4th ed. Pretoria: Van Schaik.

DeCuir-Gunby, J.T. & Schutz, P.A. 2017. Mixed methods designs: Frameworks for organizing your research methods. In *Developing a mixed methods proposal: A practical guide for beginning researchers*. Thousand Oaks, CA: SAGE Publications, Inc.

Diepenbroek, M., Glöckner, F.O., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., Kostadinov, I., Nieschulze, J., et al. 2014. Towards an integrated biodiversity and ecological research data management and archiving platform: The German federation for the curation of biological data (GFBio). Bonn: Gesellschaft für Informatik e.V. Available: http://dl.gi.de/handle/20.500.12116/2782 [2021, May 01].

Digital Preservation Coalition. 2015. *Glossary: Digital Preservation Handbook*. Available: https://www.dpconline.org/handbook/glossary#D [2021, June 30].

DIRISA. n.d. *SA-DMP Tool*. Available: https://secure.dirisa.ac.za/SADMPTool/about/ [2021, August 09].

Elsevier. 2019. *Scopus preview*. Available: https://www.scopus.com/ [2019, September 07].

EMBL. 2021. *Plant Ontology*. Available: https://www.ebi.ac.uk/ols/ontologies/po [2021, May 08].

Environment Ontology. n.d. Available: http://www.environmentontology.org/home [2021, May 08].

EPSRC. 2021. *EPSRC policy framework on research data*. Available: https://epsrc.ukri.org/about/standards/researchdata/ [2021, August 29].

Esri. 2021. *About ArcGIS*. Available: https://www.esri.com/en-us/arcgis/about-arcgis/overview [2021, July 07].

FAIRsharing. 2019. Available: https://fairsharing.org/ [2019, May 16].

Finkel, M., Baur, A., Weber, T.K.D., Osenbrück, K., Rügner, H., Leven, C., Schwientek, M., Schlögl, J., et al. 2020. Managing collaborative research data for integrated, interdisciplinary environmental research. *Earth Science Informatics*. 13(3):641–654. DOI: 10.1007/s12145-020-00441-0.

Frey, B. 2018. Pragmatic paradigm. In *The Sage encyclopedia of educational research, measurement, and evaluation*. V. 4. Thousand Oaks, CA: SAGE Publications, Inc.

Gene Ontology Resource. 2020. Available: http://geneontology.org/ [2019, May 08].

Grand, A., Bultitude, K. & Winfield, A.F.T. 2010. Muddying the waters or clearing the stream? Open Science as a communication medium. New Delhi, India. Available: http://eprints.uwe.ac.uk/13540/2/Muddying_the_waters_or_clearing_the_stream_Open_Science_as_a_communication_medium.pdf [2019, March 20].

Gravetter, F.J. & Forzano, L.B. 2012. *Research methods for the behavioral sciences*. 4th ed. Australia ; Belmont, CA: Wadsworth.

Griffin, P.C., Khadake, J., LeMay, K.S., Lewis, S.E., Orchard, S., Pask, A., Pope, B., Roessner, U., et al. 2018. Best practice data life cycle approaches for the life sciences [version 2; peer review: 2 approved]. *F1000Research*. 6:1618. DOI: 10.12688/f1000research.12344.2.

Halbert, M. 2013a. Prospects for research data management. In *Research data management: Principles, practices, and prospects*. Washington, D.C.: Council on Library and Information Resources. Available: https://libres.uncg.edu/ir/uncg/f/M_Halbert_Prospects_2013.pdf [2019, July 16].

Halbert, M. 2013b. The problematic future of research data management: Challenges, opportunities and emerging patterns identified by the DataRes project. *International Journal of Digital Curation*. 8(2). DOI: 10.2218/ijdc.v8i2.276.

Herold, P. 2015. Data sharing among ecology, evolution, and natural resources scientists: An analysis of selected publications. *Journal of Librarianship and Scholarly Communication*. 3(2):eP1244. DOI: 10.7710/2162-3309.1244.

iRODS. 2021. Available: https://irods.org/ [2021, August 30].

Jisc. 2021a. *Research Data Lifecycle (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit [2021, September 19].

Jisc. 2021b. *How and why you should manage your research data: A guide for researchers*. Available: https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data [2021, September 19].

Jisc. 2021c. *Plan & Design (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/plan-and-design [2021, September 19].

Jisc. 2021d. *Data management planning (RDM toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/data-management-planning [2021, September 19].

Jisc. 2021e. *Collect & capture (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/collect-and-capture [2021, September 19].

Jisc. 2021f. *Collaborate & analyse (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/collaborate-and-analyse [2021, September 19].

Jisc. 2021g. *Active data storage and backup (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/active-data-storage-and-backup [2021, September 19].

Jisc. 2021h. *Manage, store & preserve (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/manage-store-and-preserve [2021, September 19].

Jisc. 2021i. *Share & Publish (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/share-and-publish [2021, September 19].

Jisc. 2021j. *Discover reuse and cite (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/discover-reuse-and-cite [2021, September 19].

Jisc. 2021k. *Where should I deposit my data? (RDM Toolkit)*. Available: https://www.jisc.ac.uk/guides/rdm-toolkit/where-should-i-deposit-my-data [2021, September 19].

Johnson, K.A. & Steeves, V. 2019. Research data management among life sciences faculty: Implications for library service. *Journal of eScience Librarianship*. 8(1):7. DOI: 10.7191/jeslib.2019.1159.

Jones, S. 2012. Research data policies: Principles, requirements and trends. In *Managing Research Data*. London: Facet Publishing.

Jones, S. 2014. The range and components of RDM infrastructure and services. In *Delivering research data management services: Fundamentals of good practice*. London: Facet Publishing.

Jones, S., Pryor, G. & Whyte, A. 2013. *How to develop research data management services: A guide for HEIs.* Edinburgh: DCC. Available: https://www.dcc.ac.uk/guidance/how-guides/how-develop-rdm-services [2021, August 20].

Kennan, M.A. & Markauskaite, L. 2015. Research data management practices: A snapshot in time. *International Journal of Digital Curation*. 10(2). DOI: 10.2218/ijdc.v10i2.329.

Kim, Y. 2021. A study of the roles of metadata standard and data repository in science, technology, engineering and mathematics researchers' data reuse. *Online Information Review*. DOI: 10.1108/OIR-09-2020-0431.

Koopman, M.M. 2015. Data archiving, management initiatives and expertise in the Biological Sciences. Masters Mini Dissertation. University of Cape Town. Available: http://open.uct.ac.za/handle/11427/13656. [2019, March 11].

Koopman, M.M. & de Jager, K. 2016. Archiving South African digital research data: How ready are we? *South African Journal of Science*. 112(7–8):1–7. DOI: 10.17159/sajs.2016/20150316.

Kvale, L.H. 2012. Data sharing in the life sciences: A study of researchers at the Norwegian University of Life Sciences. Masters Thesis. Norwegian University of Life Sciences (UMB). Available: http://hdl.handle.net/10642/1269 [2019, July 14].

Laloë, A. 2017. Archives of and for science. *EMBO reports*. 18(8):1273–1278. DOI: 10.15252/embr.201744733.

Latham, B. 2017. Research data management: Defining roles, prioritizing services, and enumerating challenges. *Journal of Academic Librarianship*. 43(3):263–265. DOI: 10.1016/j.acalib.2017.04.004.

Li, Y. & Chen, L. 2014. Big biological data: Challenges & opportunities. *Genomics Proteomics Bioinformatics*. 12(5). DOI: 10.1016/j.gpb.2014.10.001.

Library of Congress. 2021. *Life cycle management & digital preservation*. Available: https://www.loc.gov/preservation/about/prd/presdig/preslifecycle.html [2021, June 30].

"Life Sciences". 1993. In *Bloomsbury guide to human thought*. Available: https://0-search-credoreference-com.innopac.wits.ac.za [2021, January 20].

"Life Sciences". 2010. In *Oxford Dictionary of English*. 3rd ed. Oxford University Press. Available: http://0-www.oxfordreference.com.innopac.wits.ac.za/ [2021, January 20].

Lin, H.S. & Wooley, J.C. Eds. 2005. *Catalyzing inquiry at the interface of computing and biology*. Washington, D.C.: National Academies Press.

Maree, J.G. Ed. 2012. *Complete your thesis or dissertation successfully: Practical guidelines*. Claremont [Cape Town, South Africa]: Juta.

Marx, V. 2013. Biology: The big challenges of big data. *Nature*. 498:255–260. DOI: 10.1038/498255a.

Matusiak, K.K. & Sposito, F.A. 2017. Types of research data management services: An international perspective. *Proceedings of the Association for Information Science and Technology*. 54(1):754–756. DOI: 10.1002/pra2.2017.14505401144.

Merriam, S.B. & Tisdell, E.J. 2015. *Qualitative research: A guide to design and implementation*. 4th ed. San Francisco, CA: John Wiley & Sons.

Metadata Standards Directory: Life Sciences. n.d. Available: https://rd-alliance.github.io/metadata-directory/subjects/life-sciences.html [2021, May 08].

Michener, W.K. 2015. Ten simple rules for creating a good data management plan. *PLOS Computational Biology*. 11(10):e1004525. DOI: 10.1371/journal.pcbi.1004525.

Miller, R.L. & Brewer, J.D. 2003. Semi-structured interviews. In *The A-Z of social research: a dictionary of key social science research concepts*. London, UNITED KINGDOM: SAGE Publications. ProQuest Ebook Central [2019, March 23].

NCBO. 2021. *Home: BioPortal*. Available: https://www.bioontology.org/ [2021, May 08].

NERC. 2019. *NERC: Data policy*. Available: https://nerc.ukri.org/research/sites/data/policy/ [2019, May 16].

NRF. 2015. *Statement on Open Access to research publications from the National Research Foundation (NRF)-funded research*. Available: https://www.nrf.ac.za/media-room/news/statement-open-access-research-publications-national-research-foundation-nrf-funded [2019, March 03].

OECD. 2007. *OECD Principles and guidelines for access to research data from public funding*. Available: https://www.oecd.org/sti/inno/38500813.pdf [2021, July 27].

Patterton, L.H. 2017. Research data management practices of emerging researchers at a South African research council. Masters (Research) Dissertation. University of Pretoria. Available: http://hdl.handle.net/2263/59502 [2019, February 25].

Patterton, L., Bothma, T. & van Deventer, M. 2018. From planning to practice: An action plan for the implementation of research data management services in resource-constrained institutions. *South African Journal of Libraries and Information Science*. 84(2). DOI: 10.7553/84-2-1761.

Patton, M.Q. 2014. *Qualitative research & evaluation methods: Integrating theory and practice*. Thousand Oaks, California: SAGE Publications.

Pennock, M. 2007. Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library & Archives*. 1(1):1–3.

Pickard, A.J. 2013. *Research methods in information*. 2nd ed. London: Facet.

Pinfield, S., Cox, A.M. & Smith, J. 2014. Research data management and libraries: Relationships, activities, drivers and influences. *PLOS ONE*. 9(12):e114734. DOI: 10.1371/journal.pone.0114734.

Procter, R., Halfpenny, P. & Voss, A. 2012. Research data management: Opportunities and challenges for HEIs. In *Managing Research Data*. G. Pryor ed. London: Facet Publishing.

Pryor, G., Jones, S. & Whyte, A. Eds. 2014. *Delivering research data management services: Fundamentals of good practice*. London: Facet Publishing.

Pryor, G. 2012. *Managing research data*. Facet Publishing.

Qin, J. 2013. Infrastructure, standards, and policies for research data management. In *Sharing of scientific and technical resources in the era of big data: The Proceedings of COINFO 2013*. Beijing: Science Press. 214–219. Available: https://surface.syr.edu/istpub/164/ [2019, May 01].

R Foundation. 2021. *R: The R Project for Statistical Computing*. Available: https://www.r-project.org/ [2021, May 22].

Radboud University. 2021. *RDM-NSF*. Available: https://www.ru.nl/rdm/vm/nsf/ [2021, August 29].

Ray, J.M. 2014. *Research data management: Practical strategies for information professionals*. Purdue University Press.

Renaut, S., Budden, A.E., Gravel, D., Poisot, T. & Peres-Neto, P. 2018. Data management, archiving, and sharing for biologists and the role of research institutions in the technology-oriented age. *BioScience*. 68(6):400–411. DOI: 10.1093/biosci/biy038.

RIN & British Library. 2009. *Patterns of information use and exchange: Case studies of researchers in the life sciences*. London: Research Information Network and British Library.

Roche, D.G., Kruuk, L.E.B., Lanfear, R. & Binning, S.A. 2015. Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*. 13(11):e1002295. DOI: 10.1371/journal.pbio.1002295.

Rueden, C.T., Schindelin, J., Hiner, M.C., DeZonia, B.E., Walter, A.E., Arena, E.T. & Eliceiri, K.W. 2017. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*. 18(1):529. DOI: 10.1186/s12859-017-1934-z.

Saeed, S. & Ali, P. 2019. Research data management and data sharing among research scholars of life sciences and social sciences. *DESIDOC Journal of Library & Information Technology*. 39(6). DOI: 10.14429/djlit.39.06.14997.

Scaramozzino, J.M., Ramírez, 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. DOI: 10.5860/crl-255.

Schmidt, B., Gemeinholzer, B. & Treloar, A. 2016. Open data in global environmental research: The Belmont Forum's Open Data Survey. *PLOS ONE*. 11(1):e0146695. DOI: 10.1371/journal.pone.0146695.

Singh, N.K., Monu, H. & Dhingra, N. 2018. Research data management policy and institutional framework. In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*. 111–115. DOI: 10.1109/ETTLIS.2018.8485259.

Stephens, Z.D., Skylar, Y.L., Faraz, F., Campbell, R.H. & Zhai, C. 2015. Big data: Astronomical or genomical? *PLoS Biology*. 13:e1002195. DOI: 10.1371/journal.pbio.1002195.

SurveyMonkey. 2019. *SurveyMonkey: The world's most popular free online survey tool*. Available: https://www.surveymonkey.com/ [2019, June 16].

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. DOI: 10.1371/journal.pone.0021101.

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. & Dorsett, K. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*. 10(8). DOI: 10.1371/journal.pone.0134826.

Tenopir, C., Rice, N.M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., et al. 2020. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS ONE*. 15(3):e0229003. DOI: 10.1371/journal.pone.0229003.

The Presidency. n.d. *Commencement of certain sections of the Protection of Personal Information Act, 2013*. Available: http://www.thepresidency.gov.za/press-statements/commencement-certain-sections-protection-personal-information-act%2C-2013 [2021, November 14].

"Thematic analysis". 2017. In *The SAGE encyclopedia of communication research methods*. V. 4. Thousand Oaks, CA: SAGE Publications, Inc.

Thessen, A.E. & Patterson, D.J. 2011. Data issues in the life sciences. *ZooKeys*. (150). DOI: 10.3897/zookeys.150.1766.

UK Data Archive. 2015. Available: https://ukdataservice.ac.uk/media/622368/costingtool.pdf [2021, August 10].

UK Data Service. 2019. *Research data lifecycle*. Available: https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx [2019, May 04].

University of the Witwatersrand. 2020. *About-Wits*. Available: https://www.wits.ac.za/about-wits/history-and-heritage/ [2021, January 19].

USGS. 2021. *Data Lifecycle*. Available: https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle [2021, June 30].

van Deventer, M. & Pienaar, H. 2015. Research data management in a developing country: A personal journey. *International Journal of Digital Curation*. 10(2). DOI: 10.2218/ijdc.v10i2.380.

Whyte, A. & Wilson, A. 2010. *How to appraise and select research data for curation*. (DCC How-to Guides). Edinburgh: Digital Curation Centre. Available: https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data [2021, October 24].

Wits. 2008. *WIReDSpace (Wits Institutional Repository environment on DSpace) Policy*. (Unpublished).

Wits. 2010. *Wits VISION 2022: Strategic Framework*. Available: https://www.wits.ac.za/media/wits-university/footer/about-wits/governance/documents/Wits%20Vision%202022%20Strategic%20Framework.pdf [2021, January 19].

Wits. 2012. *Intellectual Property Policy*. (Unpublished).

Wits. 2017. *Strategic Plan for supporting eResearch Information Systems.* (Unpublished).

Wits. 2018a. *University of the Witwatersrand (Wits) Libraries*|*Strategic Plan, 2017-2022.* (Unpublished).

Wits. 2018b. *Research Report 2018*. Available: https://www.wits.ac.za/media/wits-university/research/documents/Wits-Report-Web-2018.pdf [2021, January 21].

Wits. 2018c. *Faculty of Science Research Report 2017*. (Unpublished).

Wits. 2018d. *Open Access (OA) Policy*. Available: https://libguides.wits.ac.za/openaccess_a2k_scholarly_communication/Wits_OA_Policy [2021, January 19].

Wits. 2019a. *Research Report 2019*. Available: https://www.wits.ac.za/media/wits-university/research/documents/Wits-Report-2019-compressed.pdf [2021, July 08].

Wits. 2019b. *Information Classification & Handling Policy*. (Unpublished).

Wits. 2020a. *Facts & figures 2019/2020*. Available: https://www.wits.ac.za/media/wits-university/about-wits/documents/WITS%20Facts%20%20Figures%202020%20Hi-Res.pdf [2020, December 20].

Wits. 2020b. *Faculty of Science*. Available: https://www.wits.ac.za/science/ [2021, January 19].

Wits. 2020c. *Animal, Plant & Environmental Sciences*. Available: https://www.wits.ac.za/apes/ [2021, January 19].

Wits. 2020d. *Molecular & Cell Biology*. Available: https://www.wits.ac.za/mcb/ [2021, January 19].

Wits. 2020e. *Research ethics*. Available: https://www.wits.ac.za/research/researcher-support/research-ethics/ [2021, January 19].

Wits. 2021. *eResearch office*. (Unpublished).

World conferences on research integrity. 2010. *Singapore Statement on Research Integrity*. Available: https://wcrif.org/statement [2021, January 19].

Yu, H.H. 2017. The role of academic libraries in research data service (RDS) provision: Opportunities and challenges. *The Electronic Library*. 35(4):783–797. DOI: 10.1108/EL-10-2016-0233.

Zhu, Y. 2019. Open-access policy and data-sharing practice in UK academia. *Journal of Information Science*. 0165551518823174. DOI: 10.1177/0165551518823174.

Zinner, D.E., Pham-Kanter, G. & Campbell, E.G. 2016. The changing nature of scientific sharing and withholding in academic life sciences research: Trends from national surveys in 2000 and 2013. *Academic Medicine : Journal of the Association of American Medical Colleges*. 91(3):433–440. DOI: 10.1097/ACM.0000000000001028.

Zozus, M. 2017. *The Data Book: Collection and management of research data*. Boca Raton: CRC Press.

APPENDICES

Appendix A: Ethical clearance

Department of Knowledge & Information Stewardship
University of Cape Town
Upper Campus

Private Bag X3, RONDEBOSCH, 7701 South Africa
Level 6 Hiliegaard, The Chancellor Oppenheimer Library
Tel: +27 (0) 21 650 4546 Fax: +27 (0) 21 650 2529
E-mail: lisc@uct.ac.za
Internet: www.lhuct.ac.za/lisc

Ref No.: UCTDKIS201910-07 10 September 2019

Ms Salome Potgieter
Department of Knowledge & Information Stewardship
Chancellor Oppenheimer Library
University of Cape Town

Ethics approval for Master's research

Dear Ms Potgieter

I am pleased to inform you that ethics clearance has been granted by an Ethics Review Committee of the Department of Information & Knowledge Stewardship, Faculty of Humanities, for you to proceed with collecting data for your Master's study on '**Readiness for Research Data Management (RDM) in the Life Sciences at the University of the Witwatersrand**'.

As a next step, please ensure that you obtain approval from the relevant ethics committees to collect data at your data collection site(s), as necessary.

We wish you well with your data collection and the completion of your research.

Yours faithfully,

Mr Richard Higgs
Chair: Department (DKIS) Research Ethics Committee

Appendix B: Clearance from gatekeepers

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

OFFICE OF THE DEPUTY REGISTRAR

11 October 2019

Salome Potgieter
Master of Library and Information Studies
University of Cape Town

TO WHOM IT MAY CONCERN

"Readiness for Research Data management (RDM) in the Life Sciences at the University of the Witwatersrand"

This letter serves to confirm that the above project has received permission to be conducted on University premises, and/or involving staff and/or students of the University as research participants. In undertaking this research, you agree to abide by all University regulations for conducting research on campus and to respect participants' rights to withdraw from participation at any time.

If you are conducting research on certain student cohorts, year groups or courses within specific Schools and within the teaching term, permission must be sought from Heads of School or individual academics.

Ethical clearance has been obtained. (Protocol number: UCTDKIS201910-07)

Nicoleen Potgieter
University Deputy Registrar

Appendix C: Letter of introduction

Interview Schedule

Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand.

LETTER OF INTRODUCTION TO PARTICIPANTS

The purpose of this interview is to explore the Research Data Management (RDM) practices of researchers in the Life Sciences at Wits University in order to get an in-depth understanding of certain needs and challenges they may have with regards to Research Data Management. The results from interviews will be used for my mini-dissertation towards the Master of Library and Information Studies for which I am registered at the University of Cape Town.

The interview should take no longer than 30-45 minutes. The research is strictly for academic purposes and all collected data would be treated confidentially. Collected data would also be anonymised and de-identified.

Your answers will enable me to identify areas where infrastructure might be addressed or training & services could be offered in response to identified needs in the area of RDM.

This study has been approved by the UCT Humanities Faculty Research Ethics Committee - Reference: UCTDKIS201910-07, as well as the University of the Witwatersrand (Office of the Deputy Registrar).

Thank you for your time and cooperation.

Salomé Potgieter — Researcher (PTGSAL001@myuct.ac.za)

Michelle Kahn — Supervisor (michelle.kahn@uct.ac.za)

INFORMED CONSENT FORM

I hereby grant my permission to take part in this interview as explained in the above Introduction

I understand that participation in this interview is completely voluntary and that I may refuse to answer certain questions or that I may at any point withdraw from taking part in the interview

I understand that all responses would be treated confidentially, and that the findings of the study will be presented without identifying me as the respondent. All collected data would be anonymised and de-identified.

I give / do not give permission for this interview to be recorded (strike through as appropriate)

…….

…….

Signature (Interviewee) Date

…….

…….

Signature (Researcher) Date

Appendix D: Interview schedule

1) **Do any funder requirements for research data management pose a challenge to you or your research group?** If **YES,** why is this a challenge?
   For example – are the requirements from funders clear?

2) **Does the volume/type of research data you produce pose a challenge to you?**
   If YES to the above – why is this a challenge?

3) **Have you ever used existing data to do your research?**
   If **NO**, do you have a specific reason for not using existing data?
   If **YES** to the above, where did you find this data?
   If **YES** to the above, did you experience any challenge in finding the data and reusing the data?
   If **YES** to the above, why was it a challenge?

4) **Do you think it is important to use metadata when capturing your data?**
   If **YES** to the above, why do you think it is important?
   If **NO**, why do you think the use of metadata is not important when capturing data?

5) **Do you think it is important to make use of metadata standards or guidelines?**
   If **YES** to the above, why do you think it is important to use metadata standards/guidelines?
   If **NO**, why do you think the use of metadata standards is not important?

6) **Have you used a metadata standard/s before**? If **YES**, which standard did you use and why?

7) **In the online questionnaire the question was asked if you have ever stored your research data for long-term preservation. Following onto this, do you agree that research data should be stored for long-term preservation and use?**
   If **YES** to the above, why?
   If **NO** to the above, why not?

8) **If the requirement to store data for long-term preservation is not mandated/required** by e.g. a funder, a journal publication or your research group or for your own personal use **would you consider storing the data for long-term preservation for any other reason and what would this be?**
   If **YES** to the above, why?
   If **NO** to the above, why not?

9) **Are you familiar with or have you ever used any repository for the long-term preservation of data?** If **YES** to the above, can you elaborate on why you used a certain repository? Also, did you have any challenges in finding a relevant repository to archive your data?

10) **How do you feel about sharing your research data with others?**

11) **Have you ever linked a dataset to a paper you published?**
If **YES** – have you noted a change/increase in your citation count after publishing?

12) **Are you aware of any policies/procedures for research data management at Wits or within your School or research group?**

13) **Do you think formal processes or policies would make a difference to the storing and management of research data in your research group/University?**
If **YES** to the above, why do you think so?
If **NO**, why not?

14) **Assistance with creating DMPs was identified as one of the most important training needs for researchers.**
Why do you think there is a need for training in creating DMPs?
Who do you think needs to be trained? **By whom**?

15) **What do you see as the major challenges you have in terms of RDM?**
Why do you see it as a major challenge?

16) **How do you think the University can assist towards best practice and efficient Research Data Management?**

Appendix E: Online questionnaire

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

LETTER OF INTRODUCTION TO PARTICIPANTS

The purpose of this questionnaire is to determine the Research Data Management (RDM) practices, needs and challenges of researchers in the Life Sciences at Wits University. The results from this study will be used for my mini-dissertation towards the Master of Library and Information Studies for which I am registered at the University of Cape Town.

Your answers will enable me to identify areas where training, services and infrastructure can be offered in response to identified needs and challenges in the area of RDM. The research is strictly for academic purposes and all collected data will be treated confidentially. This study has been approved by the UCT Humanities Faculty Research Ethics Committee - Reference: UCTDKIS201910-07, as well as the University of the Witwatersrand (Office of the Deputy Registrar).

A link to the survey is provided below. It should take you approximately 10-15 minutes to complete the questionnaire.

Should you have any questions regarding the research you may contact me at PTGSAL001@myuct.ac.za or Cell no. 082-3391547.

Thank you for your time and cooperation.

Researcher: Salomé Potgieter - PTGSAL001@myuct.ac.za

Supervisor: Michelle Kahn - Michelle.Kahn@uct.ac.za; Tel (021) 650 1851

INFORMED CONSENT FORM

By continuing with this survey, you are indicating that you:

- agree to take part
- understand that participation in this survey is completely voluntary and that you may at any point withdraw from taking part in the survey
- understand that all responses will be treated confidentially, and that the findings of the study will be presented without identifying you as a respondent. All collected data will be anonymised.

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART I: DEMOGRAPHIC INFORMATION

1. **Please indicate in which of the below researcher categories you fall** *(Select all that apply)*

- [ ] Masters student
- [ ] PhD student
- [ ] Postdoctoral fellow
- [ ] Academic staff
- [ ] Other (please specify) e.g. Non-academic/Technical staff

2. **Please indicate your highest qualification** *(Select one answer only)*

- ◯ Undergraduate degree (e.g. B.Sc.)
- ◯ Honours
- ◯ Masters
- ◯ PhD
- ◯ Other (please specify)

3. **How many scientific papers have you published?**

- ◯ <5
- ◯ 5-10
- ◯ 11-20
- ◯ 21-50
- ◯ 51-100
- ◯ 101-150
- ◯ >150
- ◯ Not applicable

4. **Which of the below life sciences sub-disciplines describe your area/s of research** *(Choose all that apply)*

- [ ] Agriculture
- [ ] Aquatic Biology (including Marine Biology)
- [ ] Biochemistry
- [ ] Biodiversity
- [ ] Biogeography
- [ ] Bioinformatics
- [ ] Biotechnology
- [ ] Botany
- [ ] Cell Biology
- [ ] Conservation
- [ ] Developmental biology
- [ ] Ecology
- [ ] Environmental Science
- [ ] Evolution
- [ ] Genetics
- [ ] Genomics
- [ ] Molecular Biology
- [ ] Proteomics
- [ ] Global change (including Climate change)
- [ ] Microbiology (including Virology, Bacteriology, Parasitology & Mycology)
- [ ] Remote sensing
- [ ] Soil biology
- [ ] Systems biology
- [ ] Zoology (including Entomology, Herpetology, Mammalogy, Ornithology etc.)
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

RESEARCH DATA MANAGEMENT FUNDING

5. **Is any of your research or your research unit's research publicly funded?**

- ◯ Yes
- ◯ No
- ◯ I don't know

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

RESEARCH DATA MANAGEMENT FUNDING

6. **Who is funding your research?** *(Choose all that apply)*

- [ ] National Research Foundation (NRF)
- [ ] Medical Research Council of South Africa (MRCSA)
- [ ] International Centre for Genetic Engineering & Biotechnology (ICGEB)
- [ ] Royal Society (UK)
- [ ] Medical Research Council (UK)
- [ ] National Institutes of Health (NIH)
- [ ] National Science Foundation (NSF)
- [ ] Andrew Mellon Foundation
- [ ] Carnegie Corporation of New York
- [ ] Ford Foundation
- [ ] My research is not publicly funded
- [ ] Other (please specify)

7. **Which of the below options does your funder require in terms of data management?***(Select all that apply)*

- [ ] Data Management Plan (DMP)
- [ ] Metadata (used to describe your data in order to make it easier to find, understand and use)
- [ ] Sharing of data in a repository
- [ ] Making your research data completely open
- [ ] My funder does not have any requirements for data management
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA COLLECTION & CAPTURE

8. **Which of the below options best equates to the volume of research data you produce from your research projects at Wits?** *(Choose one of the below options)*

- ◯ less than 1 GB
- ◯ 1-50 GB
- ◯ 51-100 GB
- ◯ 100-500 GB
- ◯ 501-999 GB
- ◯ 1-50 TB
- ◯ 51-100 TB
- ◯ more than 100 TB
- ◯ I don't know

9. **What types of research data do you create or collect for your research?***(Select all that apply)*

- [ ] Documents (text, Microsoft Word, PDF, PowerPoint etc.)
- [ ] Spreadsheets/tabular data (e.g. Excel)
- [ ] Images (JPEG, PNG, TIFF, GIF etc.)
- [ ] Databases (DBASE, MS Access, Oracle, MySQL etc.)
- [ ] Structured scientific & statistical data (SPSS, GIS etc.)
- [ ] Raw data (device specific data)
- [ ] Non-digital data (specimens, artefacts, paper, slides, laboratory notebooks & diaries)
- [ ] Electronic lab notebooks (ELNs)
- [ ] Web-based data (websites, e-mails, social media, blogs etc.)
- [ ] Audiovisual data (WAVE, MP3, MP4 etc.)
- [ ] Plain text (TXT in different encodings)
- [ ] Structured text (XML, SGML etc.)
- [ ] Structured graphics (CAD, CAM, VRML etc.)
- [ ] Archived data (ZIP, RAR, ZAR)
- [ ] Sequence data (e.g. EMBL, GenBank)
- [ ] Other (please specify)

10. **How do you usually generate the data for your research?***(Select all that apply)*

- [ ] Create/collect/produce new data
- [ ] Use data from research group
- [ ] Use existing data from an open archive/repository
- [ ] Other (please specify)

11. **Do you assign any additional information (metadata) to your research data?***(Metadata is used to describe your research data in order to make it easier to find, understand and use)*

- ○ Always
- ○ Sometimes
- ○ No
- ○ I don't know

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA COLLECTION & CAPTURE

12. **When adding metadata, do you use any metadata standards/guidelines/ontological structures to describe your data?**

- ◯ Always
- ◯ Sometimes
- ◯ No

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA COLLECTION & CAPTURE

13. **Which of the below metadata standards/guidelines/ontological structures do you use?***(Select all that apply)*

- [ ] ABCD (Access to Biological Collections Data) Schema
- [ ] DC (Dublin Core)
- [ ] DwC (Darwin Core)
- [ ] EML (Ecological Metadata Language)
- [ ] EnvO (Environment Ontology)
- [ ] Genome Metadata
- [ ] Gene Ontology Resource
- [ ] ISA-tab
- [ ] MIBBI (Minimum Information for Biological and Biomedical Investigations)
- [ ] NCBO BioPortal
- [ ] Observ-OM
- [ ] OMP (Ontology of Microbial Phenotypes)
- [ ] OME-XML (Open Microscopy Environment XML)
- [ ] PDBx/mmCIF (Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework)
- [ ] Repository-Developed Metadata Schemas
- [ ] UKEOF
- [ ] Metadata standardised within my research group
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA ANALYSIS

14. **Which of the below software or applications do you use to analyse or manipulate your data?**
(Select all that apply)

- [ ] Access
- [ ] Adobe Photoshop
- [ ] ArcGIS
- [ ] Eviews
- [ ] ImageJ
- [ ] Excel
- [ ] Labview
- [ ] Matlab
- [ ] MS Word
- [ ] Python
- [ ] R
- [ ] SigmaPlot
- [ ] SPSS
- [ ] Stata
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA STORAGE (short-term)

15. **Where do you store your research data during the active research phase**(short-term storage)? *(Select all that apply)*

- [ ] Hard disk drive of office computer/laptop
- [ ] Hard disk drive of home computer/laptop/tablet
- [ ] External hard drive/USB/flash drive
- [ ] Hard disk drive of instrument which generates data
- [ ] Cloud storage e.g. Dropbox, Google Drive, Microsoft OneDrive, Google Docs
- [ ] Shared drive
- [ ] Server in unit/School
- [ ] CD/DVD
- [ ] Paper/Paper Laboratory notebook
- [ ] Electronic laboratory notebook (ELNs)
- [ ] Discipline-specific repository
- [ ] Not applicable
- [ ] Other (please specify)

16. **How frequently is your research data backed-up whilst you are collecting & analysing data** (short-term storage)? *(Select one answer only)*

- ○ Daily
- ○ Weekly
- ○ Monthly
- ○ Every 6 months
- ○ Annually
- ○ Ad-hoc
- ○ Never
- ○ I don't know

17. **Where do you keep your data back-ups during the life of the project**(short-term storage)? *(Select all that apply)*

- ☐ Hard disk drive of office computer/laptop
- ☐ Hard disk drive of home computer/laptop/tablet
- ☐ External hard drive/USB/flash drive
- ☐ Hard disk drive of instrument which generates data
- ☐ Cloud storage e.g. Dropbox, Google Drive, Microsoft OneDrive, Google Docs
- ☐ Server in unit/School
- ☐ CD/DVD
- ☐ Not applicable
- ☐ Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA STORAGE (long-term)

18. **Have your research data ever been stored for long-term preservation?**

- ◯ Yes
- ◯ No
- ◯ I don't know

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA STORAGE (long-term)

19. **What were the reason/s for archiving your data for long-term preservation?***(Select all that apply)*

- [ ] Funder requirement
- [ ] Journal publication requirement
- [ ] Research group requirement
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA STORAGE (long-term)

20. **Which of the below repositories have you or your research unit used for the archiving of research data?** *(Select all that apply)*

- [ ] Dryad
- [ ] European Molecular Biology Laboratory (EMBL)
- [ ] Figshare
- [ ] GenBank
- [ ] Gene Expression Omnibus (GEO)
- [ ] Global Biodiversity Information Facility (GBIF)
- [ ] JSTOR Global Plants
- [ ] MetaboLights
- [ ] Movebank (animal tracking data)
- [ ] Protein Databank (PDB)
- [ ] Sequence Read Archive (SRA)
- [ ] South African Environmental Observation Network (SAEON)
- [ ] South African National Biodiversity Institute (SANBI)
- [ ] TAIR (The Arabidopsis Information Resource)
- [ ] WIReDSpace (Wits Institutional Repository)
- [ ] I don't know
- [ ] Not applicable
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA SHARING

Data sharing can include any of the following: sharing data informally with other researchers; formal sharing of data via an open archive repository; sharing your data as a dataset in support of a publication.

21. **With whom do you or your research group share your data after completion of a research project?** *(Select all that apply)*

- [ ] Researchers in your research group (informal sharing)
- [ ] Researchers you collaborate with (informal sharing)
- [ ] Other researchers in the discipline/field outside my research group or collaboration (informal sharing)
- [ ] Funders
- [ ] Journal publishers
- [ ] Open archives/open repositories (either disciplinary or institutional)
- [ ] I do not share my data
- [ ] Other (please specify)

22. **Under what conditions are you willing to share your data for future research?**(*Select all that apply)*

- [ ] After publication of my research
- [ ] If I am offered co-authorship
- [ ] If I have the opportunity to collaborate on a project
- [ ] On request at my discretion
- [ ] If my dataset would be acknowledged and cited
- [ ] I am not willing to make my data available for future research
- [ ] Not sure
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA SHARING

23. **What are the reasons for not making your data available for future use?***(Select all that apply)*

- [ ] Sensitivity of data
- [ ] Confidentiality of data
- [ ] Other (please specify)

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART II: DATA MANAGEMENT PRACTICES

DATA SHARING

24. **Have you ever linked a dataset to a paper you published?**(Select one answer only)

- ◯ Yes
- ◯ No
- ◯ I don't know

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART III: CHALLENGES

To what extent do you **agree/disagree** with the below statements?

25. **I perceive the below as being a challenge/barrier that hinders effective Research Data Management (RDM).**

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Lack of funding for institutional infrastructure & programs for RDM	○	○	○	○	○
Lack of organisational structures for the effective management of research data across different divisions such as ICT, Research Office & the Library	○	○	○	○	○
Lack of training provided in RDM	○	○	○	○	○
Lack of services/support provided for RDM	○	○	○	○	○
Lack of suitable metadata standards & ontologies in my subject discipline	○	○	○	○	○
Lack of sufficient storage during the life of the research project (short-term)	○	○	○	○	○
Lack of sufficient data storage beyond the life of the research project (long-term)	○	○	○	○	○
Lack of institutional mandates e.g. a university policy on Research Data Management	○	○	○	○	○

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART IV: RESEARCH DATA MANAGEMENT TRAINING & SUPPORT NEEDS

The following statements relate to the importance of training and support in Research Data Management (RDM) being offered by the university.

26. **Please rate the below RDM related training & services according to importance:**

	Very important	Important	Somewhat important	Not important
Training on best practices for data management	○	○	○	○
Assistance with creating Data Management Plans (DMPs)	○	○	○	○
Assistance with creating metadata to describe my data or datasets	○	○	○	○
Assistance on how to cite datasets	○	○	○	○
Assistance on the preservation of datasets e.g. advice on available archives/repositories for the long-term storage of data	○	○	○	○
Technical support and the necessary software tools for data management during the life of the project (short-term)	○	○	○	○
Technical support and the necessary software tools for data management beyond the life of the project (long-term)	○	○	○	○

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

PART V: GENERAL

27. Please state any other comment, concern or issue related to Research Data Management not already covered in this survey.

Research Data Management practices, needs and challenges of researchers in the Life Sciences at the University of the Witwatersrand

THANK YOU FOR PARTICIPATING IN THIS SURVEY!