Data archiving, management initiatives and expertise in the Biological Sciences Department, University of Cape Town.

Margaret Marie Koopman

KPMMAR003

Submitted in partial fulfillment of the requirements for the award of the degree of MLIS

February 2015

Faculty of Humanities

Supervisor: Associate Professor Karin de Jager

Library and Information Studies Centre

University of Cape Town


1
Handle With Care


2


3


4


5


6


7


8


9

Collecting biological field data on Robben Island 19-21 April 2014

The images in the Frontispiece are an example of the kind of activities which go into collecting biological field data. The PhD student (1) is Davide Gaglio from Sicily who is investigating population dynamics, distribution, foraging behaviour and food abundance of the Swift Tern (*Thalasseus bergii*). Davide has three supervisors, one based at UCT, another in France, and the third in the UK. The site is Robben Island, in Table Bay, Cape Town, South Africa (2). This field trip was organised in order to place colour rings (4) on newly fledged Swift Tern chicks (9). There were nine participants in the field data collecting exercise, and the ringing (banding) of the chicks took place in two sites, the village (1,2,3) and in the north of the island near the wreck of the *Sea Challenger* (6,8). The chicks were not yet flying, and could be herded (8) toward capture nets (1) and then transferred into aerated cardboard capture boxes (9), taken to a temporary field station (3,7) where they were ringed, measured (5), weighed and then released back to the location of their nursery. The temporary field stations were located adjacent to the nurseries.

This is part of an exercise which will enable the PhD student, and future researchers to assess the survival success of the Swift Tern by identifying the birds in the field using the colour rings. The author of this dissertation was a participant in the field trip in order to get first hand experience of the intricacies of collecting biological field data. The photos were taken by the author.

Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another person's work and pretend that it is one's own.
2. I have used the Harvard UCT convention for citation and referencing. Each contribution to and quotation in this thesis from the work(s) of other people has been attributed, cited and referenced.
3. I acknowledge that copying someone else's assignments or essays, or part thereof, is wrong and that this work is my own.
4. I have not allowed, and will not allow anyone to copy my work with the intention of passing it off as his or her own work.

Signature _____

Date _____

Abstract

Researchers produce large amounts of data during their research investigations and have a variety of interventions for the management of these data. It has not been the responsibility of academic institutions to manage research data, this responsibility has resided with the researchers and their research units.

This investigation attempted to understand how pre-digital, early digital and current digital research data in the Biological Sciences Department at the University of Cape Town had been and is being managed, if researchers had archived any of these data and what their opinions were on sharing their research data. Long-term ecological data are an important component of research in the Biological Sciences Department as researchers wish to understand ecosystem changes such as climate change, the spread of alien species and the impact of humans on land and marine exploitation. It is consequently critical that research data, past and present are properly managed for future research so that meaningful management decisions can be made.

Research Data Management and the Research Life Cycle are phrases that are very much in the literature at present as librarians and university administrators grapple with the task of implementing data policies and data repositories. The literature review revealed that although the University of Cape Town may be a somewhat behind other international institutions in engaging with Research Data Management and repositories, investigations have been ongoing in other parts of the world and in the international community the groundwork has already been done.

Research data have been the preserve of researchers and they are reluctant to give up control of their hard-earned data, usually the result of hours spent on funding applications, and field or laboratory work. Data sets of sufficient quantity and quality to answer research questions can take a researcher a lifetime to accumulate and they understandably do not wish to make these openly available without the insurance that their work will be acknowledged.

The findings of this research project revealed that in the absence of systematic data management initiatives at institutional level, researchers had preserved many long-term data sets and in some instances were archiving with international repositories specific to their data types. The findings resulted in a range of suggested interventions for the support of Research Data Management at the University of Cape Town.

Acknowledgements

I would like to thank the following individuals and groups of people:

The Percy FitzPatrick Insitute of African Ornithology, DST-NRF Centre of Excellence, in particular Prof Peter Ryan, for encouragement and for funding this degree.

The 2014 staff and students in Biological Sciences who generously gave me their time to participate in the anonymous survey which underpins the findings in this dissertation. You know who you are!

The senior, retired and emeritus staff and research associates who agreed to be interviewed about past data gathering exercises and where these data are now – Graham Avery, George Branch, Tim Crowe, Jenny Day, Richard Dean, John Field, Gerd Gäde, Charles Griffiths, Timm Hoffman, John Hoffmann, Sue Milton, Roy Siegfried and Les Underhill.

The technical support staff who gave up time to discuss data archiving past and present in Biological Sciences – Gonzalo Aguilar, Michael Brooks, René Navarro, Andrea Plos and Chris Tobler. As well as Dieter Oschadleus and Dane Paijmans who engaged with me in endless informal discussion about data and shared their SQL knowledge with me.

The Island Closure Task Team who contributed time to discuss the Case Study presented in chapter three – Astrid Jarre, Lorien Pichegru, Richard Sherley and Antje Steinfurth who read many drafts and sat with me making sure I had my facts straight!

Davide Gaglio for including the Niven Librarian on his field trip, and for discussion over glasses of *Monkey Shoulder* about differing measurement standards and field note recording and how these can affect data accuracy.

My supervisor Karin de Jager, invaluable and experienced mentor of students' work, thank you! Dale Peters for frank discussion about RDM and UCT Libraries initiatives in this connection.

Last but not least Patrick Morant who has been supremely patient with and supportive of this student at home, and Putzi and Moja who did not get quite as much attention as they would have liked.

Contents

Chapter Five

Conclusions and recommendations for institutional level support

Figures

Glossary

ADU – Animal/Avian Demography Unit, University of Cape Town

ARPANET - Advanced Research Projects Agency Network

BGIS – Biodiversity Geographic Information System, SANBI

CIB – Centre for Invasion Biology, Stellenbosch University

CRNS - Centre National de la Recherche Scientifique, France

CoE – Centre of Excellence

CSIR – Council for Scientific and Industrial Research, South Africa

CWAC – Co-Ordinated Waterbird Counts

DANS – Data Archiving and Networked Services, the Netherlands

DCC – Digital Curation Centre, United Kingdom

DMP – Digital Management Plan

DNA - Deoxyribonucleic Acid

DOI – Digital Object Identifier

DST – Department of Science and Technology, South Africa

EMBL – European Molecular Biology Lab

EML – Ecological Metadata Language

FRD – Foundation for Research Development, South Africa

FRU – Freshwater Research Unit, University of Cape Town

GBIF – Global Biodiversity Information Facility

GCDML – Genomic Contextual Data Markup Language

GPS-TD – Geographic Positioning System, Time-Depth

ICTS – Information Communication and Technology Services

IGBP – International Geosphere-Biosphere Programme

INSDC - International Nucleotide Sequence Database

IOC – Intergovernmental Oceanographic Commission

IUCN – International Union for the Conservation of Nature

JISC – Joint Information Systems Committee

JSTOR – a digital library, Journal Storage, founded in 1995 to digitize pre-digital era academic journal articles.

KNB – Knowledge Network for Biocomplexity

KNAW – Royal Netherlands Academy of Arts and Science

LTER – Long-term Ecological Research

MARAM – Marine Resource Assessment and Management group

MPA – Marine Protected Areas

Ma-Re – Marine Research Institute, University of Cape Town

NCBI – National Center for Biotechnology Information

NII – National Information Infrastructure, United Kingdom

NLMS – National Marine Linefish System, South Africa

NOW – Netherlands Organisation for Scientific Research

NRF – National Research Foundation, South Africa (successor to the FRD)

NSF – National Science Foundation, United States of America

OBIS – Ocean Biogeographic Information System

OECD – Organisation for Economic Cooperation and Development

ORI – Oceanographic Research Institute, South Africa

PAIA – Public Access to Information Act, 2 of 2000, South Africa

PCU – Plant Conservation Unit, University of Cape Town

PFIAO – Percy FitzPatrick Institute of African Ornithology, University of Cape Town

PLoS – Public Library of Science

SAAMBR – South African Association for Marine Biological Research

SABAP – Southern African Bird Atlas Project

SABIF – South African Biodiversity Information Facility

SADCO – South African Data Centre for Oceanography

SAEON – South African Environmental Observation Network

SAFRING – South African Bird Ringing Unit

SAIAB – South African Institute of Aquatic Biodiversity

SANBI – South African National Biodiversity Institute

SMRU - Small Mammals Research Unit, University of Cape Town

UCAR - University Corporation for Atmospheric Research

UCT – University of Cape Town

UNEP – United Nations Environment Programme

UNESCO – United Nations Educational, Scientific and Cultural Organisation

USGS – United States Geological Survey

US NESC – United States National Evolutionary Synthesis Centre

US NIH – United States National Institution of Health

US NSF – United States National Science Foundation

WRC – Water Research Commission, South Africa

XML – Extensible Markup Language

Chapter One

Setting the scene - The importance of Long-term data

1.1 Introduction

Data are the currency of academic research. Research staff and students investigate and gather data around a research topic, combine data sets, analyse and interpret the accumulated data, write up the results of their findings in research papers and publish these in the international research literature. Researchers attend international conferences to meet with other researchers in their field of expertise, present the findings of their data gathering initiatives and discuss their interpretations with their peers.

Research data thus gathered at the University of Cape Town (UCT) has not traditionally had a home in the university libraries or university archives, remaining the responsibility of research units, researchers, and in some cases archived in special collections associated with the research unit and their specialised focus (e.g. The Bolus Herbarium and The South African Bird Ringing Unit). Digital research data generated within academia has largely been an invisible resource utilised within the research unit and shared with a select group of trusted colleagues. Recent changes in international funding and grant applications require evidence that research data have been securely archived in an approved repository with protocols enabling access to the data (Wellcome Trust, 2010). This has been adopted by many of the international scientific journals that have made it mandatory to archive data underlying published research (Nature, 2014; Fairbairn, 2010).

The management of research data at UCT is poorly understood and only through focused discipline-specific investigations can this situation be remedied. Each academic discipline produces unique data, these require a range of specialised management and archiving interventions. This dissertation endeavours to understand research data management in the Biological Sciences Department at UCT and to make suggestions for a way forward.

1.2 Research questions

The overarching question this thesis sought to answer was the fate of the data behind the accumulated knowledge about the southern African environment gathered by researchers in Biological Sciences over the years. This was addressed by examining the data management practices of researchers in the past; interrogating the present; and investigating the strategic support available at the institutional level for data management initiatives.

The questions put to past and present researchers as part of this investigation were designed to answer four basic questions:

1. How are research data, past and present managed, archived and shared?;
2. What do researchers understand by the concept of metadata and how this aids data sharing and data re-use?;
3. How much public funding was supporting Biological Sciences research, and whether researchers were ready to make their data open?; and
4. How should institution level support for data management and archiving be configured?

1.3 Rationale for the investigation

Research data archiving may soon be mandatory in South Africa, as recommended by the National Research Foundation (NRF) for NRF funded research (2015). The survey questions were consequently designed to be both informative and investigative - informative in order to enable researchers who had not yet engaged with research data archiving and management to think about the implications; and investigative in order to document what had already been achieved. Data management and archiving are topics currently being discussed in a range of international biological, ecological and zoological journals making it important for UCT researchers and policy makers to engage with the issues and imperatives. Data have commercial and intrinsic value, and in both cases it is important that they are archived for future use, particularly as to re-collect data is expensive in both time and money. It is not possible to recreate long-term ecological data as human population growth and resource usage change ecological systems over time.

1.4 The importance of biological research data

1.4.1 Historical perspective

The southern African natural environment has been documented since the 1650s with valuable information recorded in the diaries of Jan van Riebeeck, the first commander of the Dutch East India Company settlement at the Cape (Thom, 1952-1958). The Cape of Good Hope became the staging area for many natural history expeditions to the interior of southern Africa. Travellers such as François Le Vaillant whose 1781-1784 expedition was recorded with illustrations and field notebooks, contributed considerably to our knowledge about the buildings, customs, landscape, natural history and people of the places he visited. Known for his lavishly illustrated *Histoire naturelle des oiseaux d'Afrique*, Le Vaillant could be called the father of African ornithology.

Figure 1.1 Cape Sugarbird (*Promerops cafer*) *Histoire naturelle des oiseaux d'Afrique* volume 6 Plate 287 (Le Vaillant, 1799-1808)

In 1963 the collection of illustrations which had been made during Le Vaillant's southern African expedition were sold at auction by Sotheby in London, nearly 180 years after the event. Le Vaillant published accounts of his travels and

produced multi-volumed and numerous editions of his famous ornithological treatise. The illustrations had found their way into the hands of a 19th century collector in Rotterdam, L.V. Ledeboer before coming into the ownership of the Library of Parliament in Cape Town (Quinton, Lewin Robinson & Sellicks, 1973:xvii).

Another famous naturalist, who passed through Cape Town on his return from his round the world trip aboard the *Beagle*, was Charles Darwin. Darwin was hosted in Cape Town by Dr Andrew Smith, the first Superintendent of the South African Museum, and together they went on field trips to collect botanical and geological specimens. Darwin kept field notebooks of his time in Cape Town, and from these we know where he went, what he saw and collected, and what his opinions were on a variety of topics including slavery (Darwin, 1836).

The botanical artist and traveller Marianne North painted her way up the east coast of South Africa between 1883-1884, recording her impressions of the plants of Table Mountain, Tulbagh, Ceres, Grahamstown, Port Alfred, Port St John's and Verulam. North kept journals of her travels and her paintings are now in the North Gallery at the South Entrance of Kew Botanical Gardens. Many of her paintings illustrate the association of plants with birds and insects in context with their geographical landscape (North, 1980) making these of scientific and historical importance.

Figure 1.2 Marianne North. Strelitzia and Sugar Birds [Sunbirds], South Africa (Trustees of the Royal Botanical Gardens, Kew, n.d.).

What these three visitors have in common is that they were natural historians, they had funding, they made notes, collected specimens and illustrated their findings, thus creating benchmarks to be emulated by future field researchers who now more commonly illustrate their findings with photography. These early travellers were also sufficiently influential and original to find prestigious homes for their collections, which have been carefully archived and made available for long- term research. The archiving of what amounts to the research data of these explorers was not systematic, and the security of their collections was often precarious with shipments of specimens and accompanying field notebooks being lost at sea, consumed by fire or devoured by insects. Alfred Wallace, contemporary of Darwin, lost field specimens and notes by all of these means, on one occasion watching from a life boat while the ship in which he and his field records were travelling was consumed by fire before sinking (Quammen, 1996:68-70).

This illustrates that the importance of collecting and preserving data had been the focus of past researchers in the natural sciences long before the digital age was imagined. Field researchers collected data in notebooks that were kept in offices and garages and if the researcher was well connected and organised, their data may have found a home in a museum collection or archives for the benefit of posterity. The book *Field Notes on Science & Nature* (Canfield, 2011) reports on the data collecting and archiving efforts of Joseph Grinnell, Director of the Museum of Vertebrate Zoology at the University of California at Berkeley. Grinnell was not satisfied with only archiving vertebrate specimens, he expected collectors to keep notebooks about behaviour, habitat and ecology and this information had to be submitted along with the specimens. Detailed baseline surveys were conducted across California between 1908 and 1939 and the notebooks and specimens were archived at the Museum. This archived survey and accompanying specimens enabled contemporary researchers between 2003 and 2006 to re-survey the areas visited by Grinnell and his associates to investigate changes in the communities studied a century earlier. It will be seen in on page 7 that a similar re-survey using historical data is currently being carried out by a Biological Sciences PhD student, utilising the fisheries survey data gathered 100 years ago.

1.4.2 A brief history of the Biological Sciences Department

In order to gain an understanding of the Biological Sciences Department and the past and present research undertaken at UCT, it was necessary to delve into the history of the department. This was facilitated by Alec Brown's (2003) history of the Zoology Department and by Diana Rex's (1985) history of the Botany Department.

The Biological Sciences Department was established in 2013 and is made up of the former Botany and Zoology Departments, the Animal Demography Unit (ADU), the Bolus Herbarium, the Marine Research Institute (Ma-Re), the Percy FitzPatrick Institute of African Ornithology (PFIAO), the Plant Conservation Unit (PCU), and the Small Mammals Research Unit (SMRU). The Freshwater Research Unit (FRU) was disbanded at the end of 2012.

Botany and Zoology were offered as courses at the South African College (now UCT) early in its formation, with a Department of Zoology created in 1903 (Brown, 2003:11). The teaching of Botany was introduced in the late 1850s when the Colonial Botanist, an appointment of the Cape of Good Hope Government, was engaged to teach botany at the College (Brown, 2003:11). This initiative was short-lived and Botany was only offered as a subject again in 1902 when Harry Bolus endowed a chair in Botany at the College (Rex, 1985:55), and in 1911 the herbarium and library of Harry Bolus were also bequeathed to the South African College (Rex, 1985:55-6). These collections are still curated in the Bolus Herbarium and Library, and the herbarium specimens form a component of the baseline data collection of the plants of the Western Cape, a very important collection for contemporary comparative study. Scanned images of the type specimens (the first specimens to receive scientific names) in the Bolus herbarium have been archived at the digital library Journal Storage (JSTOR), in their Global Plants Initiative (Bolus Herbarium, 2011).

The Zoology Department made its name initially as a marine studies centre with the part-time appointment in 1905 of John Gilchrist who was simultaneously an employee of the Department of Agriculture responsible for investigating the possibility of establishing a Fisheries Department at the Cape of Good Hope. With government funding and access to government vessels, Gilchrist conducted baseline

surveys of the marine fauna of South Africa, both vertebrate and invertebrate, to investigate whether it would be economically viable to set up a fishing industry (Brown, 2003:11). His baseline surveys are also important for contemporary comparative studies of what remains of marine fauna after 100 years of intensive exploitation by fisheries. Researchers in the Biological Sciences Department at the University of Cape Town are attempting to make a centennial assessment of the state of fish resources off the coast of the Western Cape, following in the footsteps of John Gilchrist, Zoologist at the Cape of Good Hope in 1905. Gilchrist was a prolific note-taker who published all his notebooks in a scientific journal *Reports of the Fisheries and Marine Biological Survey,* which were published between 1920 and 1936. The specimens collected during Gilchrist's numerous marine surveys are still stored at the Iziko Museum in Cape Town.

Figure 1.3 John Gilchrist's Marine Aquarium and Research Station situated at St James, Cape Town, 1902-1954, where all the type specimens and type descriptions of his field research were kept (Zandvleitrust, 2000-2015).

Gilchrist's seminal work was continued by his successors: T.A. Stephenson with his systematic surveys of rocky shores; J.H.O. Day who instigated intertidal and continental shelf surveys as well as conducting estuarine surveys; J.H. Field who looked at marine animal distribution patterns; G.M. Branch who researched the ecology of marine invertebrates; C.L. Griffiths, a taxonomist, who investigated marine invertebrates; and currently C.L. Moloney who researches marine plankton ecology; A. Jarre who is involved in modelling marine food webs; C.G. Attwood

whose research interests lie in coastal fish ecology and the various impacts of fisheries; and D. Pillay who is researching anthropogenic effects on estuarine systems.

The Gilchrist, Stephenson, Day, Field and Branch specimens were transferred to the South African Museum (SAM now Iziko) during the 1980s when the Zoology Department was relocated to the present John Day Building (Day, 2014). Keppel Barnard, Director of the SAM between 1946-1956 (Iziko, Biodiversity Explorer, n.d.) reports that the earliest ecological specimens were collected by Dr Andrew Smith, the first Superintendent of the museum 1825-1837 (Iziko, History of the South African Museum, n.d.). Smith's collection was handed over to W.S. McLeay in 1837 when Smith returned to England. McLeay in turn migrated to Australia and took the type specimens with him. They were only discovered 100 years later in the Australian Museum in Sydney (Barnard, 1950:6).

As the Botany and Zoology departments developed, new disciplines were introduced, each with their own contribution to the South African ecological knowledge base. In 1971 Behavioural Ecology of Small Mammals was introduced; in 1981 Freshwater Ecology was introduced as a teaching and research discipline; and in 1986 Entomology was introduced. Ornithology had been offered as a post-graduate research subject since 1960. The ADU, which started out as the Avian Demography Unit, was an offshoot of the Department of Statistical Sciences and of Ornithology.

Palaeobiology was introduced as a discipline in 1997. The PCU, which is involved in interdisciplinary research, was established in 1993 (Plant Conservation Unit, 2014). The social and economic conditions of people using natural resources and the consequent impact on biodiversity in rural area land use practices informs the research of this unit. The Unit is also involved in the investigation of land-use changes through studying early photographs and comparing these with contemporary photos of the same site. The early photographs are archived in the UCT Libraries digital collections in the Cowling Collection.

All of these researchers have generated large amounts of valuable data which have

been used for past and present academic dissertations and published scientific research papers, but much of the data remains stored on hard drives and are invisible and inaccessible beyond the research unit. Data have also been lost over the 100 years that the research has been carried out at UCT.

1.5 Long-term data series

Long-term data are important for a range of management and planning applications across many disciplines. Baseline information about the natural environment in which we live, and upon which we depend for economic and survival purposes, is essential to enable us to reflect on the level of exploitation of resources and the ability of these resources to renew themselves for the benefit of future generations.

South African research scientists had become sufficiently concerned about data to organise a conference in 1987, the *National Conference on Long-term Data Series Relating to Southern Africa's Renewable Natural Resources* (Macdonald & Crawford, 1988). The editors of this publication respectively were from the PFIAO at UCT and, the then Sea Fisheries Research Institute within the then South African Department of Environment. Funded by the Foundation for Research Development (FRD) of the South African Council for Scientific and Industrial Research (CSIR), the conference was attended by numerous UCT researchers who contributed significantly to the discussion of long-term data series. Names such as J. Cooper (PFIAO), C.L Griffiths (Zoology), G.M. Branch (Zoology), I.A.W. Macdonald (PFIAO), L.G. Underhill (ADU), P.A.R. Hockey (PFIAO), and W. R. Siegfried (PFIAO) are some of the UCT contributors, all now retired, or in the case of Hockey, deceased. In view of recent developments at UCT with the establishment of the African Climate & Development Initiative, it is interesting to note that the conference was the first South African meeting of the Study of Global Change (IGBP), "a programme of international cooperative research aimed at improving mankind's ability to model the global environment and the changes that are taking place within it" (Macdonald & Crawford, 1988:iii). This description mirrors some current UCT research concerns – collaborative research and global environmental change, which should be supported by effective data management and data

archiving.

Findings reported from this conference were "that data for environmental sciences in southern Africa were generally not well curated." (Phillips, 1988:467). Not much has changed in the ensuing 27 years, although three exceptions may be noted. The South African Data Centre for Oceanography (SADCO), an initiative established in the 1960s at the CSIR in Stellenbosch was made available nationally for the archiving of oceanographic data (SADCO, 2010) and the South African Bird Ringing Unit (SAFRING) was initiated in1948 in order to archive bird movement data (SAFRING, n.d.). The Southern African Bird Atlas Project (SABAP) was initiated in 1987 (SABAP, 2001) and is hosted by the ADU at UCT, which has in turn been in existence since 1991 (ADU, 2009).

Freshwater ecologists were also anxious to preserve what was known about research into South Africa's water resources. With the sponsorship of the Water Research Commission (WRC) a report, *The Freshwater Science Landscape in South Africa, 1900-2010* (Ashton et al., 2012) was compiled documenting information about South African freshwater research. B. Davies (Zoology), J.H.O. Day (Zoology), J.A. Day and J.M. King (FRU), were UCT researchers who contributed to the report who were associated with South African freshwater research in the past. Freshwater ecology is no longer a discipline represented on the UCT campus, the Freshwater Research Unit (FRU) was disbanded at the end of 2012 through lack of funding.

This dissertation attempted to uncover how much of this and other possibly less well-known long-term data have been preserved for use in future research.

1.6 Pre-digital data management in Biological Sciences

In order to explore the historical approaches to data management at UCT and to address Research question one, data management and archiving initiatives of emeritus and retired staff were investigated through interviews and augmented with information from the literature review to see how the UCT experience compared with international examples.

The results of these interviews showed that there are a number of physical long-term data sets in Biological Sciences, which are still in existence. As there has been no institutional support in the past for systematic data management, these data sets owe their existence to the foresight of responsible, concerned individual researchers, as is discussed on page 97. The lack of institutional support is of great concern, in particular to researchers who have recently retired or will retire in the near future and are holding data sets that are neither documented, nor have a future home. Recommendations to address the concerns of researchers are discussed in Chapter 5.

1.7 Data management and archiving initiatives at the PFIAO

The PFIAO received DST-NRF Centre of Excellence (CoE) status in 2004 and is primarily publicly funded, although individual collaborators bring their own funding to augment research projects. The mission of the Institute is "To promote and undertake scientific studies involving birds, and contribute to the practice affecting the maintenance of biological diversity and the sustained use of biological resources" (PFIAO, Mission Statement, 2014). With a research staff complement of Director and four academics, approximately 24 research associates, five additional CoE members based at collaborating departments and universities, approximately 12 postdoctoral fellows, 14 doctoral students, and 25 masters students, research at the Institute is very active and produces on average 85 scientific publications per annum.

During 2013 the PFIAO embarked on a pilot project (Koopman, 2013) to make research data archiving mandatory for its postgraduate students. This was initiated in response to the international climate whereby public funders and prestigious international journals increasingly require researchers to make their data publicly available. Because of the Centre of Excellence status of the PFIAO, articles are expected to be published in high impact journals that are likely to make such conditions. The Director of the Institute is of the opinion that it is part of the research training for the students to manage and archive their research data (Ryan, 2013). From the beginning of 2014 PFIAO students have been mandated to lodge their data along with their dissertations in order to graduate. A relationship was established with the South African Environmental Observation Network (SAEON) in order to archive appropriate data sets. SAEON is also an NRF-funded initiative and

was considered to be the most appropriate place to archive PFIAO ecological data.

In order to address Research questions two, three and four an online anonymous survey was compiled to find out how research was funded, how much research data required archiving, researcher opinions on data sharing, what their data management skills were and what they understood by the term 'metadata'. To validate the responses from the survey and to address research question 3 in more detail, a desktop study was undertaken to look at the nature of publication output of the researchers. It will be shown in the fourth chapter that while some responses concurred with the findings of the survey reported in the third chapter, other responses diverged from actual publication practice. UCT Libraries staff were interviewed in order to investigate UCT's RDM initiatives. A data archiving repository at UCT is in the developmental stage and as yet there is no data management policy at UCT (Peters, 2014).

1.8 Conclusion

This investigation endeavours to place researchers in the Biological Sciences Department at UCT within the context of international experience in order to learn from and compare with the initiatives of research institutions that have engaged with research data management for the past decade or more.

The following chapter reviews these international experiences and considers which interventions would be most suitable at UCT, looking specifically at studies which speak to the research questions posed for this investigation in order to find appropriate answers for local concerns. Knowledge gained from the literature review was useful in the formulation of the questions posed for the interviews and the survey that are reported in the third chapter. It will be seen that the concerns expressed by UCT researchers mirror the international findings from a range of similar studies.

Chapter Two

Literature Review - International context and imperatives for research data management and data archiving

2.1 Introduction

This literature review focuses on examining the way the global research community has approached the questions of data management, metadata creation, data archiving, and data sharing, in other words, the research questions posed in this dissertation with regard to the fate of data in the Biological Sciences Department at UCT.

The Advanced Research Projects Agency Network or ARPANET was established in 1969 specifically to enable researchers to share data between laboratories in geographically distant locations (Dasgupta, 2006:173). ARPANET was the template upon which the internet was subsequently built. There had also been a number of discipline-specific archiving initiatives around the world prior to the Open Access (OA) movement, but these had mainly been for the large national projects whose business it was to compare temporal data sets. Organisations such as National Weather Bureaus, National Censuses and National Oceanographical Surveys come to mind. But these organised initiatives did not make much impact in academia, even though census, oceanographic and weather data are used by academic researchers on a regular basis.

It is convenient to fix the time that academic research data management and archiving came under the spotlight to the OA movement, because the movement raised awareness resulting in a plethora of articles published on managing and archiving research data. A search using Google Scholar between 2002 and 2012 generates 15,800 articles on the topic while the same search for the period 1991 to 2001 generates 2,720 articles. The OA movement raised the question of *universal* access to research, particularly publicly funded research, through the medium of the Internet (Budapest Open Access Initiative, 2002; Max Planck Gesellschaft, 2003-2014). This does not however provide a time-scale for the emergence of digital research data management and archiving. There are a number of contributing factors which have led to the current preoccupation with research data archiving.

2.1.1 Factors contributing to research data archiving

- Global climate change research has alerted governments and researchers to the value of long-term ecological studies. As was seen in Chapter One, researchers in Biological Sciences at UCT were already engaging with the IGBP in 1987 (Macdonald & Crawford, 1988). The Anthropocene, the epoch of human driven environmental change, was suggested in 2000 by Paul Crutzen, Nobel Laureate and Vice Chairman of the IGBP as the next major geological epoch (IGBP, n.d.). To come to a conclusion such as this requires research, research data and collaboration - on the one hand data about human populations and associated socio-economic research data and on the other hand long-term ecological data. These are the kind of data collected by Joseph Grinnell and John Gilchrist discussed in Chapter One.

- Funding has become an extremely competitive exercise where funders want evidence that the research has not previously been undertaken, that the data collected are being preserved, and that the research is open to scrutiny. Articles on research funding are a constant presence in mainstream scientific journals such as *Nature*, *Science* and *Scientific American*. The latter publication reported that scientists spend 40% of their time writing funding proposals (Scientific American, 2011). An article in *Nature* discusses the initiative *Science Exchange* set up by Elizabeth Iorns as a new evaluation mechanism for researchers. One component enables funding agencies to see how much research costs and how research funding is spent (Iorns, 2013).

- The ubiquity of the Internet at all levels of society has made it possible for broad-scale sharing of data. This is the fundamental tenet of making published research and research data open and is discussed in forums such as the Royal Society of London report (Royal Society, 2012) discussed in more detail in 2.1.2 .

- Reviewers of research papers require underlying data for verification of research findings, a necessary but controversial requirement which could lead to data theft unless a clear ethics policy is in place. Providing underlying data is regarded as a way to prevent fraud in research, the findings in the publication are expected to

have robust scientific data underlying the research (Doorn, Dillo & Van Horik, 2013).

- Last but not least, there is a global awareness that digital records are in danger of being lost, or have already been lost because of inadequate preservation initiatives. *The Digital Dark Age: revolution preview* is a chilling video about the consequences of failing to archive digital artifacts (Computer History Museum, 2011).

2.1.2 The emergence of data archiving initiatives

In the United Kingdom (UK), the Digital Curation Centre (DCC) was launched in 2004 as an initiative of the "JISC Continuing Access and Digital Preservation Strategy" in order to support digital curation for UK Higher and Further Education. Over the past decade the DCC has developed into a centre of expertise in the provision of support for digital data management and archiving for higher education internationally (Digital Curation Centre, 2004-2015a). By 2007 the DCC had focussed on the archiving of research data with the SCARP Project, an initiative which produced a number of case studies reporting on the issues around managing, archiving and sharing research data. Projects such as *Digital Curation Approaches for Architecture* (Neilson, 2009), *Curated Databases in the Life Sciences: The Edinburgh Mouse Atlas Project* (Fairley & Higgins, 2009), and *Curation of Research Data in the Disciplines of Engineering* (Ball & Neilson, 2010) are three of the seven projects undertaken by SCARP. These projects informed the way the DCC provided support for their multiple-discpline stakeholders (Digital Curation Centre, 2004-2015b).

A year after the DCC was formed in the UK, a similar initiative, the Data Archiving and Networked Services (DANS) was formed in the Netherlands as an Institute of the Royal Netherlands Academy of Arts and Science (KNAW) and the Netherlands Organisation for Scientific Research (NOW) (DANS, n.d.). In an editorial written for a special issue of *Archival Science* on archiving research data, the DANS director Peter Doorn and his co-author Heiko Tjalsma, gave a succinct overview of the state of research data archiving in the social sciences and humanities, fields which they considered to be the earliest to archive electronic data (Doorn & Tjalsma, 2007:4).

One of the problems with research data archiving identified in this paper is that the researchers who are the orginators of the data do not take responsibility for long-term data archiving, and in turn, information technologists do not consider long-term data preservation their domain as they are at the cutting edge of producing the latest technology. This finding is not referenced in the article, but certainly tallies with the findings of the survey and interviews undertaken in Biological Sciences at UCT, where technical staff did not take responsibility for data produced by researchers as this was not within the parameters of their job. Research funders are considered to be the key influencial agents of long-term data preservation (Doorn & Tjalsma, 2007:9).

In 2012 the UK Government published the findings of a working group investigating Open Data that made the case for Open Data strategies (United Kingdom, Open Data White Paper, 2012). This led to the development of the National Information Infrastructure (NII) framework, that provides useful information on policies, standards, definitions and guidance for the delivery of open government data (United Kingdom, National Information Infrastructure, 2014). The framework discusses the concept of "strategically important data", supports a data list, and itemises all the components of the NII such as standards, licensing, quality of the data, governance, interconnectivity and usability, all of which are considerations that have to be taken into account in the development of an institutional repository for long-term data archiving and have commonality with the LTER policy on data management.

A report by the Royal Society of London which considered academic research data and how and why this should be made open access, was published during the same year as the UK Open Data White Paper. The findings discuss the importance of Open Data "for the sake of better science", how the pervasiveness of the internet makes data sharing possible, and the various considerations which should be taken into account before making data open, such as privacy and intellectual property. Chapter 4 of the Royal Society document – *Realising an open data culture: management, responsibilities, tools and costs* (Royal Society, 2012) – is particularly pertinent to this research and discussion as it apportions responsibility for management, for cost calculations and for the development of appropriate tools to

facilitate the Open Data culture. It will be seen in chapters three and four of this dissertation that researchers at the Biological Sciences Department at UCT are not far behind their international colleagues in facilitating access to their published data. They have availed themselves of the opportunities to archive data offered by the National and International Repositories pertinent to their field of research. Some of these repositories are discussed in 2.6 below.

2.2 Has lack of institutional support been the experience of researchers in other parts of the world? - Three case studies

International examples such as those discussed by Elliot (2008), Scaramozzino et al. (2012), and Diekmann (2012) which were consulted for this thesis, indicate that in each of these case studies there had been very little institutional support for RDM in the past.

The Otago Biodiversity Data Management Project confirmed that this was the case for researchers in New Zealand (Elliot, 2008:4). Elliot's year-long, funded project, demonstrated that undocumented data are invisible data and this situation can and does result in funding being spent collecting data that are already in existence, wasting both time and money (Elliot, 2008:5).

The article by Scaramozzino et al. (2012) discussed the role libraries could play in data management highlighting the necessity for librarians to understand individual researcher's needs and to understand research data. This investigation conducted at a state university in California gave a wish-list, generated from a survey of researchers, of the kind of support the library intended to provide to the researcher community. The findings of this investigation demonstrated that past data management was mainly in the hands of researchers, with minimal support from information technology (IT) personnel on the campus and no support from campus libraries (Scaramozzino et al., 2012:356). Such findings were identified by both the Elliot (2008:17) survey and the Diekmann (2012:24) survey.

In an investigation subtitled "results from an exploratory study", Diekmann (2012) examined how agricultural researchers managed their data at the Ohio State

University. The study used focused interviews with 14 researchers, all of whom were doctoral graduates of many years' standing. The topics chosen for the interviews sought to find out whether researchers re-used existing data, types of data they generated, how this was collected and analysed, what the arrangements were for security and storage of data, whether they shared data and what sort of institutional support was available to assist with data management. As with the Scaramozzino findings, data management was not fully developed and the author was of the opinion that initiatives which existed outside academia deserved closer scrutiny for adoption in local situations (Diekmann, 2012:30). Diekmann futher found that valuable data were constantly being lost to science because of poor data management and archiving practices at the university.

In another initiative Cornell University Library set up DataStaR (Data Staging Repository) to correct this omission on their own university campus. This campus repository was designed to be a temporary holding location for Cornell University researchers' data, to enable researchers to share their data, create robust metadata and ultimately to publish the data in appropriate repositories (Steinhart, 2007:34). Such initiatives are scattered throughout the literature but do not appear to have achieved systematic adoption. The literature shows that libraries have been slow to engage with research data as an alternative information resource and a new paradigm which requires management and institutional support.

2.3 Data Archiving

The imperatives for either publishing or archiving research data are multiple, and are discussed in numerous forums (Costello, 2009; European Union, 2013; Huang, Hawkins & Qiao, 2013; Marx, 2012; Steinhart, 2007; Van Noorden, 2014a; Vines, Andrew, Bock et al., 2013). In most cases the benefits are dependent on the research data being visible or discoverable, accessible or open and interpretable. Imperatives for open data are:

- Availability for further integrated research
  Mark J. Costello, professor of marine ecology at the University of Auckland has strong views about sharing data for the benefit of future research, going so far as to say that "scientists who do not publish or release their data are compromising scientific development" (2009:418)

- Contributing to global research initiatives e.g. natural resource use decision making

  Sharing scientific data for natural resources management purposes is promoted by Huang et al. (2013:5), but it will be seen from the case study presented in 3.4 that unless there are very strict ethical rules around such data re-use, data misappropriation can result. Costello (2009:421) reports this fear of "incorrect use of data" in his published findings.
- Improving researchers' international profiles

  Piwowar et al., (2007:1) reported a 70% increase in citations for cancer publications that shared their data. As the practice of data archiving becomes more mainstream and appropriate mechanisms such as data DOIs, make it easier to reference data, researchers who generate and share data will benefit (Van Noorden, 2013:244).
- Preventing expensive duplication of research

  A report undertaken by JISC in collaboration with the Centre for Strategic Economic Studies (Fry et al., 2008:12-13) undertook a hypothetical cost-benefit analysis of the costs of creating and sharing data versus the benefits of sharing data, calculating that the benefits were more than four times the value of the costs.
- Verification of research findings

  While sharing data to enable research verification is promoted as essential for good science, Borgman found that this was easier said than done. There are so many variables to be taken into account that it becomes almost impossible to reproduce research, (2012:1067) ecological fieldwork being a case in point. Weather conditions, equipment, observer bias are all variables that can result in a slightly different result.
- Sharing data to make research more efficient

  Although funders, publishers and research groups perceive that data sharing will make research more efficient, Piwowar found that this was difficult to measure, mainly because the levels of sharing remain so low (2011:1).
- Transparency in research

  Funders, publishers, and researchers promote transparency through publishing data openly, but in reality the levels of open data publication remain low. This is reportedly because researchers do not like to lose control of their data in case it is

misused and they want to be credited for their time, money and hard work (Molloy, 2011:1).

- Mandates from journal publishers and funders
  There are now many publishers' and funders' mandates for open data (Nature, 2014; Fairbairn, 2010; Borgman, 2012) and this is regarded as the incentive that is most likely to lead the way to open data.

Many of the 'benefits' discussed above have been discussed in other parts of the text as these are the main reasons given by those promoting open data. It does not appear from the literature that researchers are convinced however, and until their fears can be allayed many researchers will only partially participate in lodging open data, this will be the data they have already published.

It appears that unless data archiving is mandatory, researchers are slow to archive or make their data available (Steinhart, 2007; Van Noorden, 2014a; Vines, Andrew, Bock et al., 2013). A number of journals have made data archiving mandatory – *American Naturalist*, *Molecular Ecology, Nature, Public Library of Science* (PloS) journals, *Royal Society of London* journals, *Science*, to name a few. Avoidance of data archiving is inconsistent as some disciplines, such as molecular genetics, routinely archive their data in repositories which have long-term life-spans. GenBank, developed in 1982 and European Molecular Biology Lab (EMBL) developed in 1974 are two such repositories. These are discussed in more detail in Chapter 3.

The Ecological Society of America encourages the submission and archiving of data in their data archive, *Ecological Archives*, but at the time of writing this was only mandatory for articles published in *Ecological Applications* and *Ecological Monographs* (Ecological Society of America, 2014).

The DCC web page which supplies information on repositories recognises that different disciplines have different needs. An online presence to enable researchers and librarians to assess the quality and long-term reliability of repositories has been developed by the DCC. Repositories are scored by using criteria such as reliability,

assessment of risk, organisational mandate, and they provide an interactive tool, DRAMBORA, to assist in the assessment (Digital Curation Centre, 2004-2015e). They have made it their goal to assess the changing data repository environment and keep higher education informed of the latest thinking and initiatives.

2.4 Local, national and international repositories used for archiving ecological data

This sub-section presents a range of data repositories that either are used by Biological Sciences researchers or have the potential to be used to archive their ecological data. It will be possible to lodge the metadata for these data sets in a UCT data repository once this is established.

Dryad Digital Repository http://datadryad.org/

This data repository, "built upon the open-source DSpace repository software" (Dryad Digital Repository [Dryad], 2014), was suggested by a number of well-respected scientific journals (e.g. *Nature, American Naturalist, PLoS*) as a suitable place to deposit research data underlying published articles. Of the 85 journals using Dryad as their preferred research data repository, 55 are already integrated with Dryad, which makes data archiving a simple process for authors (Dryad, 2013). Dryad was launched in 2008 by the United States, National Evolutionary Synthesis Centre (US NESC) after four years of development (Dryad Data Repository Wiki, 2014). Dryad has a business plan whereby sustainability into the future will be ensured (Dryad Data Repository Wiki, 2013).

Figshare http://figshare.com/

Although this resource was not used by any of the researchers in Biological Sciences, it is a solution supported by a number of prestigious institutions such as Monash University and Loughborough University and is recommended by the Public Library of Science (PLoS). (Figshare Blog, 2014). Figshare advertises the repository as a place where you can "manage your research in the cloud and control who you share it with or make it publicly available and citable" (Figshare, 2014).

GenBank http://www.ncbi.nlm.nih.gov/genbank

By far the most highly utilised of the digital repositories investigated in this survey (see Figure 3.9), Genbank is the United States National Institution of Health (US

NIH) genetic sequence database (GenBank, 2014). GenBank is hosted by the National Center for Biotechnology Information (NCBI), US National Library of Medicine and is a collaborating repository in the International Nucleotide Sequence Database Collaboration (INSDC). GenBank was launched in 1982 (GenBank, 2014) and where the US Government is the data creator, these data are made available in the public domain (NCBI, Copyright and disclaimers, 2009). This proviso does not apply to all data sets submitted to GenBank, particularly where the authors of the data claim patent, copyright and other Intellectual Property rights over their data (Genbank, Overview, 2014).

GBIF/SABIF http://www.gbif.org http://www.sabif.ac.za/

The Global Biodiversity Information Facility (GBIF) is an international open data facility, funded by participating governments (GBIF, 2014a). GBIF was established in 2001 and was an initiative of the Organisation for Economic Cooperation and Development (OECD). South Africa became a participant in 2003 when the South African Biodiversity Information Facility (SABIF) was launched. The initiative is funded by the South African Department of Science and Technology. Data providers are the South African Institute of Aquatic Biodiversity (SAIAB) at Rhodes University; the Animal Demography Unit (ADU) at the University of Cape Town; Iziko Museums of Cape Town, the Albany Museum at Rhodes University; and the South African National Biodiversity Institute (SANBI). The GBIF incorporates three biodiversity data types: Metadata; Occurrences; and Checklists (GBIF, 2014b) which are linked by geographic coordinates to a world map.

OBIS/AfrOBIS http://www.iobis.org/ http://afrobis.csir.co.za/

The Ocean Biogeographic Information System (OBIS) hosts marine species data on a global scale. The OBIS was initiated in 1997 as a project of the Census of Marine Life with the mandate to create "an online, user-friendly system for absorbing, integrating, and accessing data about life in the oceans" (Grassle, 2000:5). It is now part of the Intergovernmental Oceanographic Commission (IOC) of UNESCO. As well as being the African presence in the OBIS, AfrOBIS has strong links with GBIF, SABIF (discussed above) and with SADCO (discussed below).

EMBL http://www.embl.org/

As with GenBank (discussed above), the European Molecular Biology Laboratory

(EMBL) is a participating institution in the International Nucleotide Sequence Database Collaboration. EMBL was founded in 1974 (European Molecular Biology Laboratory (EMBL), 2009-2014a). EMBL has a range of databases including the EMBL Nucleotide Sequence Database, as well as a range of Bioinformatics Services (European Molecular Biology Laboratory (EMBL), 2009-2014b).

KNB http://knb.ecoinformatics.org/

The Knowledge Network for Biocomplexity is a United States, National Science Foundation (US NSF) supported international repository for ecological and environmental data. It is a member node of DataOne, which in turn is a product of the United States Geological Survey (USGS). No one in Biological Sciences uses this repository. They are using the South African Environmental Observation Network (SAEON) data repository which similarly supports ecological and environmental data. The DataOne, KNB and SAEON data archiving initiatives all use Ecological Metadata Language (EML) for their metadata standard.

SAEON http://www.saeon.ac.za/data-portal-access

The South African Environmental Observation Network is a South African Department of Science and Technology (DST) funded initiative led by the National Research Foundation (NRF). It is "an institutionalised network of departments, universities, science institutions and industrial partners." (SAEON, background, 2009). SAEON also plays a role in the coordination of the Southern African Data Centre for Oceanography (SADCO) and AfrOBIS (SAEON, Developing information systems for Earth observation, 2009).

SADCO http://sadco.csir.co.za/data.html

The Southern African Data Centre for Oceanography (SADCO) has been in existence since the 1960s (SADCO, 2010). The initiative is entirely government funded, including funding from the Namibian Ministry for Fisheries and Marine Resources. The resource is a southern African initiative and data collected in Namibia is archived at SADCO as well. Namibia, South Africa and Mozambique can use all data stored at SADCO (Pillay, 2014).

Movebank https://www.movebank.org/

This resource is hosted by the Max Planck Institute for Ornithology. It is a "free, online database of animal tracking data", (Movebank, 2014) which integrates with

Argos data. "Argos is a satellite-based system that collects, processes and disseminates environmental data from fixed and mobile platforms worldwide." (Argos-System, n.d.). Biological Sciences researchers who use satellite telemetry in their research utilize the Argos system.

Global Plants http://plants.jstor.org/

Global Plants is supported and populated by herbaria. Plant type specimens are contributed to the database which is in turn utilised by researchers and students in the fields of botany, ecology and conservation biology (Global Plants, 2000-2014). Type specimens from the Bolus Herbarium have been contributed to this database.

UCT Libraries Digital Repository https://open.uct.ac.za

The Open UCT Initiative, a three year project tasked to provide open access to UCT "research, teaching and scholarly resources", handed over the OpenUCT Repository to UCT Libraries in December 2014 (OpenUCT, Farewell to the OpenUCT Initiative Team, 2014). This repository archives theses, images, gold open access publications, ePosters, slide shows, lecture series, and webpages. Although recently launched, this DSpace application with detailed Dublin Core coded metadata is already launching UCT research into the Google-sphere, which can be seen through the discoverability of UCT dissertations, e.g. Wright (2011).

National Marine Linefish System (NMLS)
http://www.seaworld.org.za/content/page/data-management
The NMLS is an initiative of the Oceanographic Research Institute (ORI) which is hosted by the South African Association for Marine Biological Research (SAAMBR). ORI holds a number of long term data sets which are archived on servers at ORI (ORI, Data Management, 2014). One of these is the NMLS, which is a 29 year data set. Data are provided by email on request in the form of a data report.

Animal Demography Unit (ADU) http://vmus.adu.org.za/
The ADU hosts a number of citizen science databases which can be interrogated online. The data are open but must be acknowledged if used. Data sets include the South African Bird Atlas Projects, the Nest Record Cards, the Virtual Museums made

up of photographs of a variety of species submitted with geographic coordinates by citizens, and the South African Bird Ringing Unit which archives bird movement and migration data (Animal Demography Unit, Virtual Museum, 2014).

BirdLife Seabird Tracking Database http://www.seabirdtracking.org
Contributors to this database include past and present Biological Sciences staff who conduct research on seabirds. BirdLife International curates the data but ownership remains with the contributor. Requests for data are made through the BirdLife site. The earliest, and longest data set available is one of Wandering Albatross breeding adults from Crozet 1989-2001 (BirdLife Seabird Tracking Database, 2014).

South African National Biodiversity Institute (SANBI) data archives
http://www.sanbi.org/biodiversity-science
Data initiatives based at SANBI include the DNA bank of South African plant genetic material which was initiated in 2005; and the Millennium Seed Bank Partnership, a collaborative archive which banks the seeds of indigenous plants, and was initiated in 2000. The data archived by these initiatives are not digital, they are biological data. The National Vegetation Map is another initiative which enables researchers to acess the data from the vegetation map on SANBI's Biodiversity GIS site (BGIS) (BGIS, 2007).

UvA-BiTS http://www.uva-bits.nl

The University of Amsterdam bird tracking system (UvA-BiTS), tracks bird movement and bird behaviour in space and time. The system is used to "study migration, navigation and foraging strategies on land and at sea" (UvA-BiTS, 2013). Access to the data is via a contact person and the data are under copyright to the data provider. Data on the Verreaux's Eagle in the Western Cape, South Africa have been provided by researchers in the Biological Sciences Department.

Scientific Data (Nature) http://www.nature.com/sdata/

This recently launched data resource is a data publication rather than a data repository. *Scientific Data* is an Open Access resource that publishes Data Descriptors of "scientifically valuable datasets". Submissions are peer-reviewed and

hosted by Nature.com (Nature, Scientific Data, 2014). The researchers have to identify a suitable repository for the data sets according to the data policy of Nature. Criteria specified for repositories include

- Expert curation;
- Stable identifiers;
- Unrestricted public access;
- Long-term persistence and preservation.

2.5 Data management

Although data archiving has been discussed at length in the scientific literature by the scientists doing the research and producing the data, the topic of data management and research support has not been that vigorously debated among this cohort. Internationally the topic is now being discussed at the strategic institutional level and by the library fraternity (Akers & Doty, 2013; Diekmann, 2012; Elliot, 2009; MacColl & Jubb, 2011; Patrick & Wilson, 2013; Reinhart, 2007:34; Scaramozzino et al., 2012; Tenopir et al., 2011) who are now working closely with researchers and paving the way for more systematic data management, metadata standards and assistance for post-graduate researchers and academics who are anxious about archiving their data and making these openly available to global research.

Tenopir et al. (2011) in their investigation into data sharing practices of scientists, discussed data management practices at the institutions where their respondents were based. Researchers expressed concern about the lack of long-term data storage available to them at their institutions, with only a third of respondents reporting that there was sufficient support for long-term data management (Tenopir et al., 2011:7). Akers & Doty (2013) focussed on disciplinary differences and how these required the library services to provide a range of approaches in data management support for the suite of academic disciplines.

The DCC have taken data management one step further by developing digital data management tools such as DMPonline, that are designed with UK funders' data management requirements in mind (DCC, 2004-2015d). With funders considered to

be the prime motivators for data archiving, tools like DMPonline are essential models for South African research funding applications. DMPonline has a set of templates for the major UK, US and EU research funders and can be customised to suit local situations. The plan walks the researcher through all the steps required to create a data management plan for a funder.

2.6 Metadata initiatives

Metadata is the detailed description or documentation of data which can be understood and shared or harvested by computers. Metadata is structured and has standardised languages according to the research discipline, e.g. Ecological Metadata Language is the standard which has emerged for ecological research (Berners-Lee, 1997). Taking ecological field data as an example, metadata includes numerous fields which may be broken down into sub-fields. Fields such as name of the data collector, when data were collected, where data were collected, what types of data were collected, what equipment was used to collect data, why data were collected are all metadata inputs which are standardised for the repository in which data will be archived. The title of the project is the most important field for data access and this should include who, what, where, when and the scale of the data set.

The description of data is the essential component that makes it possible to share data and to evaluate research findings. Without metadata, data are valueless to everyone other than the researcher who collected the data, and even then, as time passes, poorly described data become meaningless even to the originator. This was certainly the finding of Whitlock et al. (2010:145) and reported in their editorial in *The American Naturalist*. The authors believed that in the fields of ecology and evolutionary biology, legacy data were being permanently lost, because there were no data preservation strategies. The editorial laid down the policy developed for a suite of journals (itemised in 2.3 above) which would mandate that data underlying published research were to be archived in an appropriate repository, with metadata, described as "a short additional text document, with details specifying the meaning of each column in the data set." (Witlock et al., 2010:146).

The necessity to describe data with metadata has resulted in the development of a range of specialised and subject specific metadata standards, for example Ecological Metadata Language (EML) and Genomic Contextual Data Markup Language (GCDML). EML is the favoured standard of the LTER network discussed below, as this deals with ecological data, while GCDML is specific to genetic sequence data. In a paper discussing these two biological

metadata standards it was pointed out that in order for a metadata standard to be adopted, tools and training had to be developed in parallel (Gil, Sheldon, Schmidt, et al., 2008:152).

Ecological Metadata Language is made up of a set of XML (extensible markup language) schema, designed to result in structured metadata (Knowledge Network for Biocomplexity (KNB), n.d.). EML is used by a variety of ecological repositories such as DataONE, Knowledge Network for Biocomplexity (KNB) and the South African Environmental Observation Network (SAEON) (discussed on page 23). In order to make the creation of EML standardised, data management software known as Morpho was developed which prompts the researcher to enter data according to the EML structure. Such tools are designed to enable researchers to enter their own data into a repository without the specialist intervention of a data archivist.

The DCC web page on Disciplinary Metadata provides links to a range of metadata standards, tools and repositories (Digital Curation Centre, 2004-2015c), including EML. This is in response to the wide range of data formats and disciplines which were revealed by the SCARP case studies conducted by DCC in 2007. This DCC information enables the research community to evaluate and decide which metadata standard and data repository is the best fit for their data.

2.7 Data sharing

2.7.1 Data sharing costs

In a 1994 publication on environmental information management, Porter and Callahan attempted to quantify the cost of sharing ecological data and came to the conclusion that the cost to the person creating the data was much higher than the cost to the person using the data. Their calculations included:

- Time available for research;
- Time spent preparing publications using data;
- Time saved through using resources resulting from previous work;
- Time spent collecting data;
- Time between data collection and data availability (Porter & Callahan, 1994:194).

If one considers these time variables, researchers who do not share their data and researchers who do not create data but use data created by others, both have a time

advantage over a researcher who collects data, describes them and makes them available to other researchers. The way around this inequality, in the opinion of the authors, was to provide rewards to data contributors. The authors commented that ecological and environmental research was one of the only scientific fields to be without a "community-mandated data archiving and data sharing policy" (Porter & Callahan, 1994:195).

Pertinent to this investigation is the Long-term Ecological Research (LTER) Network in the USA which has been generating research data since its establishment by the National Science Foundation (NSF) in 1980 (LTER, 2013). The NSF insisted on active data management at all LTER sites. The LTER data management policy included ten rules to ensure that data creators were not penalised, that data were adequately described and that data were stored for the long-term. The policy has all the components of a contemporary academic institutional data management policy (Monash University, 2010; Edinburgh University, 2014):

- Timely availability of data to the scientific community;
- Adequate acknowledgment of researchers who contribute data and receipt of copies of publications using the data;
- Documentation of data sufficiently adequate to permit data re-use by researchers not involved in the original collection;
- Continued availability of data even when an investigator leaves a project;
- Adherence to quality assurance and quality control;
- Maintenance of long-term archival storage of data;
- LTER funded researchers obliged to contribute the data collected and publish the results in an open forum;
- Recovery of costs of making data available either directly or by reciprocal sharing or collaborative research;
- Data users may not sell or distribute LTER data sets;
- Investigators should have a reasonable opportunity to have first use of data they collected (Porter & Callahan, 1994:197).

2.7.2 Data sharing opinions - incentives and barriers

There is a growing collection of opinion papers on data sharing (Piwowar et al., 2008; Research Information Network, RIN, 2008; Borgman, 2012; Marx, 2012;

Doorn et al., 2013; Wallis et al., 2013; Nature Editorial, 2014) discussing topics such as increased citations through data sharing (discussed under 2.3 above), motivations for sharing data, etiquette for sharing data, technical and human barriers to sharing data, prevention of data fraud (discussed in 3.4 in relation to a case of data misappropriation), and reuse of data which is the whole point of archiving data. The motivations for sharing data are varied, they may be because of incentives such as improved visibility for a researcher, the researcher may have adopted an open science stance, or because there is reciprocal data sharing between researchers or research laboratories (Borgman, 2010). Data sharing etiquette is not mentioned much in the literature, but a few instances where authors have had to publish apologies and retractions related to unacknowledged use of data indicate that data sharing ethics and etiquette are underdeveloped (Jetz & Rubenstein, 2011; Said et al., 2007). In the absence of policies and rules to protect the generators of data in these instances, the threat to a researcher's reputation, and withdrawal of collegial approval or ostracism appear to be the interventions used for such cases. In the case presented by Van Noorden (2013) below about a PhD student who wanted to share her data, some of the technical and human barriers are discussed.

Richard van Noorden, an assistant news editor with *Nature*, has published his findings on research data sharing in a number of papers (Van Noorden, 2014a; 2014b; 2013). These findings indicate that opinions on data sharing, particularly open data sharing, are found at the extremes of the spectrum, i.e. some researchers will share their data, others won't share their data. In his article (Van Noorden, 2014b) discussing the Public Library of Science (PloS) mandate for open data, his findings indicated that only 40% of 20 papers sampled had lodged full underlying data. Prior to the PloS mandate, where underlying data was encouraged but not mandatory, only 12% of 51 papers sampled had complied. Van Noorden's 2013 report discussed the experience of a PhD graduate who wanted to share her data but found (between 2003-2009) that this was not encouraged, there were no incentives to share data and her peers only shared their data privately, if at all. By 2010 the data sharing climate had changed and she was able to lodge her data with the KNB, discussed above in more detail on page 23. Some of the reasons for not sharing were itemised in this paper – too much work for the researcher, no reliable or appropriate data repositories, funders had not made data archiving mandatory, no data

description standards and researchers received no credit for data sharing. These reasons are also used for not archiving data and are pretty standard findings in all of the published investigations consulted for this literature review.

Tenopir, Allard, Douglass, et al. (2011) reported on the findings of a survey of the data sharing habits of 1,329 international scientists. The investigation gave insights into data management habits of these scientists and demonstrated that lack of institutional support was a key reason for questionable data management, with long-term data storage being a major concern among respondents. Their reasons for not sharing data echo the findings of Van Noorden's (2013) investigation, namely time restraints and lack of funding mandates (Tenopir et al., 2011:9). The requirements for data sharing, or "fair use" of data as it is termed in this paper, are dependent on the data generator being given credit through proper citation; being given co-authorship on any publications using their data; or, an invitation to collaborate on projects using their data, dependencies identified in the LTER policy discussed above on page 29.

These systemic findings all concur with the findings of the survey conducted and reported on in Chapter 3. The link between funding, data archiving and data sharing is discussed more fully in the following two chapters.

2.8 Conclusion

This chapter has shown that Research Data Management and Research Data Archiving are keenly debated subjects and actively published topics among the full gambit of research roleplayers: the data generators, the data users, the research funders, the research policy makers, the libraries and the journal publishers. While some data gathering initiatives have been conscientiously attending to the long-term preservation of their data, many others have not. The range of topics discussed relating to the long-term archiving and sharing of research data indicate that there is still much to be done, but that there is sufficient interest and energy to expect that a reliable infrastructure will emerge to ensure that this is successful.

Chapter Three

Survey of Research Staff and Students

3.1 Introduction

This project originated from the FitzPatrick Institute pilot project on introducing mandatory data archiving for postgraduate students (Koopman, 2013), which was described more fully in chapter one. The current dissertation expands the original pilot in order to explore data management and archiving in the broader Biological Sciences Department at UCT and to answer the research questions posed in chapter one.

Three surveys were designed to answer the research questions. Two were interview-based surveys and the third was a digital survey.

- In order to address the first research question, how was research data managed, archived and shared in the past, an interview-based survey was designed for recently retired and Emeritus researchers who would be conversant with non-digital research and the history of data collection in the Biological Sciences Department.
- An interview-based survey was designed for technical staff in the Biological Sciences Department and was used to find out whether they had a role in supporting data management and archiving carried out in the department in the past and present. This group was interviewed in order to answer a component of the first research question.
- A multiple-choice digital survey was prepared to send to all 2014 research staff and students of the Biological Sciences Department that was designed to be informative as well as interrogative and was intended to answer all four research questions. The intention was that the 'don't know' respondents would benefit from the informative aspect of the questionnaire.

3.2 Research methodology

Three techniques were used to collect information about the data management and archiving initiatives and expertise in the Biological Sciences Department. These were

- Face-to-face interviews,

- an online survey, and
- a published article review. (The review will be discussed in Chapter 4.)

According to Babbie and Mouton (2001:249), face-to-face interviews are commonly used in South African surveys because of respondents' low level of literacy. This was not the reason face-to-face interviews were used in this survey as all interviewees are highly literate, but not all retired or emeritus interviewees are computer literate and the face-to-face interview was considered to be the most effective way to elicit information from them. A further reason for this interview technique was certainty of obtaining the information. As the project only had a short time-span, this was considered the fastest and most efficient way to acquire information. Only one interviewee failed to attend the appointment, and two other potential interviewees could not be contacted. This method resulted in a 76% response from the retired or emeritus group.

A standard questionnaire was used for each interview in two target groups of researchers, namely retired or emeritus, and technical staff. The questionnaire for the retired or emiritus research staff was designed to elicit information about historical data management and data archiving in the Zoology and Botany Departments, and to enquire about any long-term data sets which may have still been in existence. Informal discussion around the questions was encouraged if the researcher wished to discuss any of the questions further. All interviews were undertaken during August 2014 once ethics clearance had been granted.

A response sheet was completed for each interview, and notes were taken when the interviewee wished to provide additional information. The response sheets were used to populate a spreadsheet and the notes were immediately committed to a .doc file after the interview. The interview atmosphere was relaxed as all the staff who were interviewed were known to the interviewer. See Appendix A for the QUESTIONS FOR INTERVIEWS of Emeritus/Retired Biological Sciences researchers.

The reason for conducting face-to-face interviews with technical staff was also for reasons of efficiency as the staff in question were known to be extremely busy and unlikely to respond to an online survey. As it was, three of the appointments were

changed two to three times before the interview took place. The second questionnaire was designed in order to find out the level of support for data archiving offered in the past and offered currently, as stated in the first research question. These interviews were conducted during early September 2014. The response rate is given in Figure 3.1. (See Appendix B for the QUESTIONS FOR INTERVIEW of technical support Biological Sciences staff).

In order to sample the remaining estimated 286 research staff and students, an online survey was designed. The online survey was a variation of the Computerised Self-administered Questionnaire (CSAQ) (Babbie & Mouton, 2001:259). This was a Web Survey where the respondents went to an identified site and completed the questionnaire themselves without assistance. The survey was designed with Google Forms and consisted of 32 multiple-choice questions broadly similar to the questions posed for the face-to-face interviews. The set of 32 multiple-choice questions consisted of 12 questions requiring multiple responses and the remaining 20 questions requiring a single response. The request to respond to the survey was emailed to research staff and students with an outline of the project and a URL link to the survey. Responses were automatically saved to a spreadsheet and were anonymous. The survey was run from 9 September to 15 October 2014 with weekly reminders. There were seven different categories of researchers sampled with this survey:

- Academics
- Research Associates
- Postdoctoral Researchers
- PhD Researchers
- Masters Researchers
- Honours Researchers
- Other (research support staff of various categories)

Fourteen online surveys were undeliverable as the email addresses were no longer valid, and four respondents declined to participate. A total of 163 researchers completed the survey. Taking the undelivered and declined participants into account, the overall survey response was 51% This was due to sending out repeated emails about the survey, conducting face-to-face interviews and the high degree of interest

and concern about archiving data. (See Appendix C for the SURVEY QUESTIONS posed to Biological Sciences researchers).

3.3 Results and discussion

The following results are a combination of the digital survey and the face-to-face interviews with emeritus or retired researchers and technical researchers. The 'Other' category includes interviewed technical staff and respondents who indicated that they had a research support position in the online survey questionnaire. In some cases the questions asked in the digital survey were not asked of, or relevant to either the emeritus or retired researchers and/or the technical researchers. This accounts for their absence in some of the reported results in the following figures (Figure 3.1-3.32). Where the figures include all categories of respondent, the total number of respondents was 163, without the category of 'other/technical' the total number of respondents was 149, without the category of 'emeritus' or 'retired' the total number of respondents was 139. These are the totals which were used to calculate the percentage contributed by each category of respondent in Figures 3.1-3.32.

3.3.1 Researcher categories

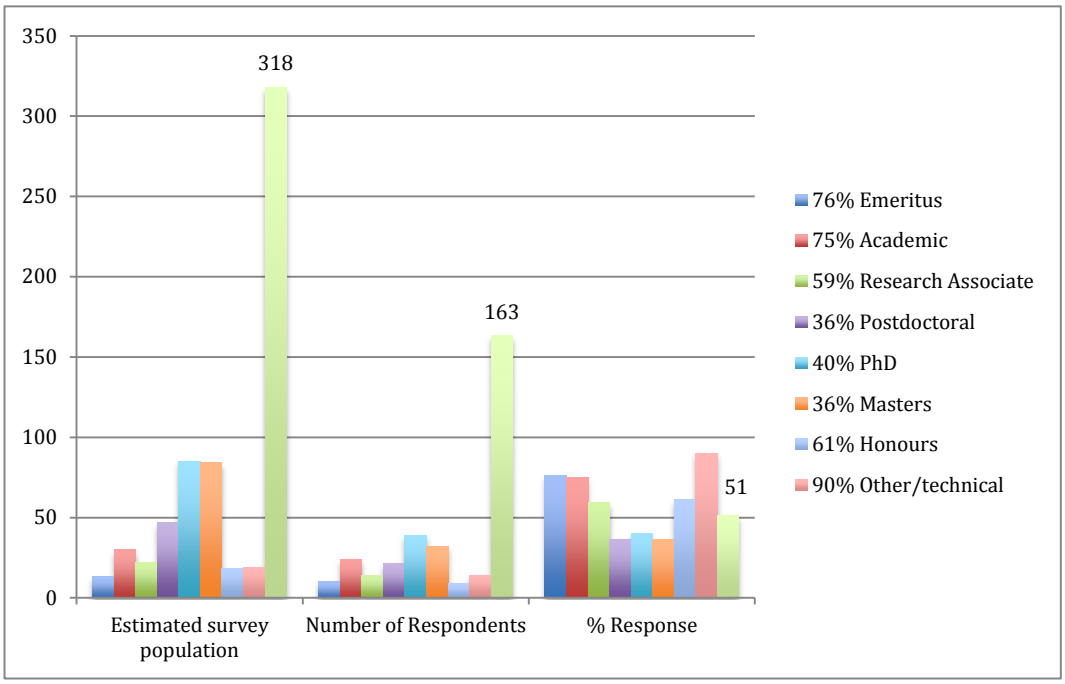

Figure 3.1 Which research categories describe you?

Discussion:

It was difficult to ascertain the exact number of research staff, research associates and students in Biological Sciences at UCT for 2014. The departmental web page was two years out of date at the start of the survey and the October 2014 upgrade of the web site remained out of date for many of the staff and students. The survey population numbers given in Figure 3.1 have been extrapolated from the 2013 Biological Sciences Department component of the UCT Annual Research Report and are only an estimate of the 2014 numbers. The response from Postdoctoral, PhD and Masters respondents was poor. It is probable that this was because the emails were sent to their official university accounts which were not being opened by students who may prefer to use an alternative email account such as gmail.

3.3.2 Researcher qualifications

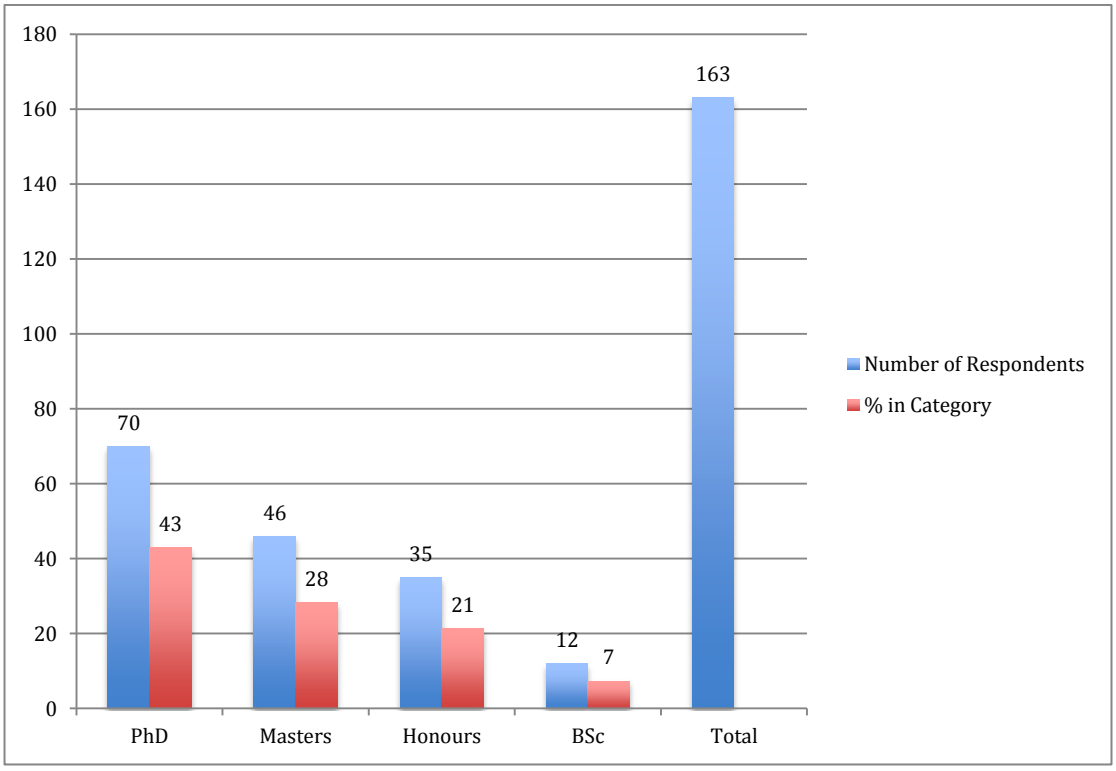

Figure 3.2 What is your highest academic qualification?

Discussion:

As the online survey was anonymous, this question was posed to make it possible to verify the number of student researchers in combination with question one. The number of academic staff was the only number that could reliably be estimated. If the

respondent had a PhD then this was either an academic, a research associate or a Post-doctoral student. If the respondent had a Masters degree then this was a PhD student, if the respondent had an Honours degree then this was a Masters student and if the respondent had an undergraduate degree then this was an Honours student. The Other category had a range of qualifications ranging from BSc to PhD. Within the Biological Sciences department 43% of researchers have a PhD, some of these highly qualified staff are in the other/technical category and although they are in the UCT 'support staff' category they contribute to the research output of the university in many ways including the publication of articles in subsidy attracting journals.

3.3.3 Publicly funded research in the Biological Sciences Department

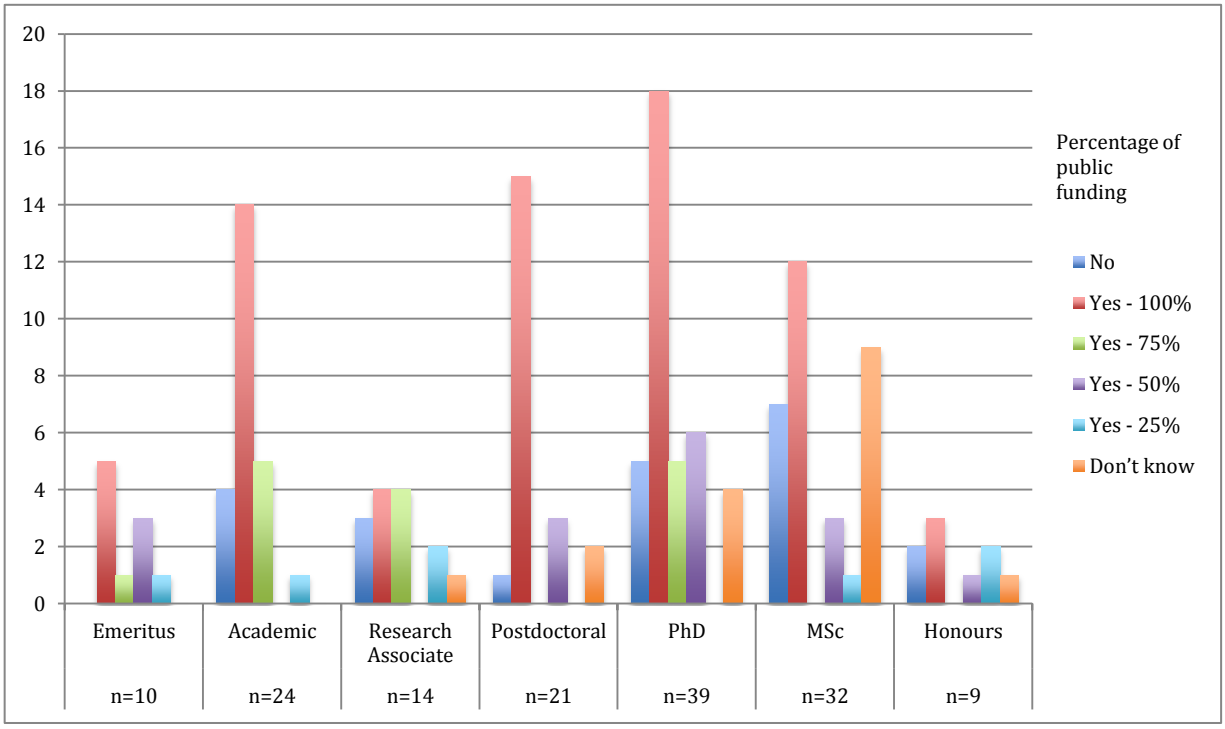

Figure 3.3 Is your research publicly funded?

Discussion:

This question was asked in order to assess how much research and research data could be expected to be OA in the future when public funders of research start to mandate OA and it is clear that a significant proportion of the research carried out in the Biological Sciences Department is indeed publicly funded. With regard to South African NRF funding the recommended date is from the 1st of March 2015 (NRF, 2015). This question in combination with Question 16 - 'Under what conditions

would you or your research unit make data available for further research?' identified the extent to which researchers were aware of the trend to make publicly funded research OA. It will be seen in Question 3.16 below that only 15% of the researchers were prepared to make their research data OA prior to publication of their findings. After publication 57% were prepared to make the underlying data OA.

3.3.4 Biological Sciences researchers' published research output

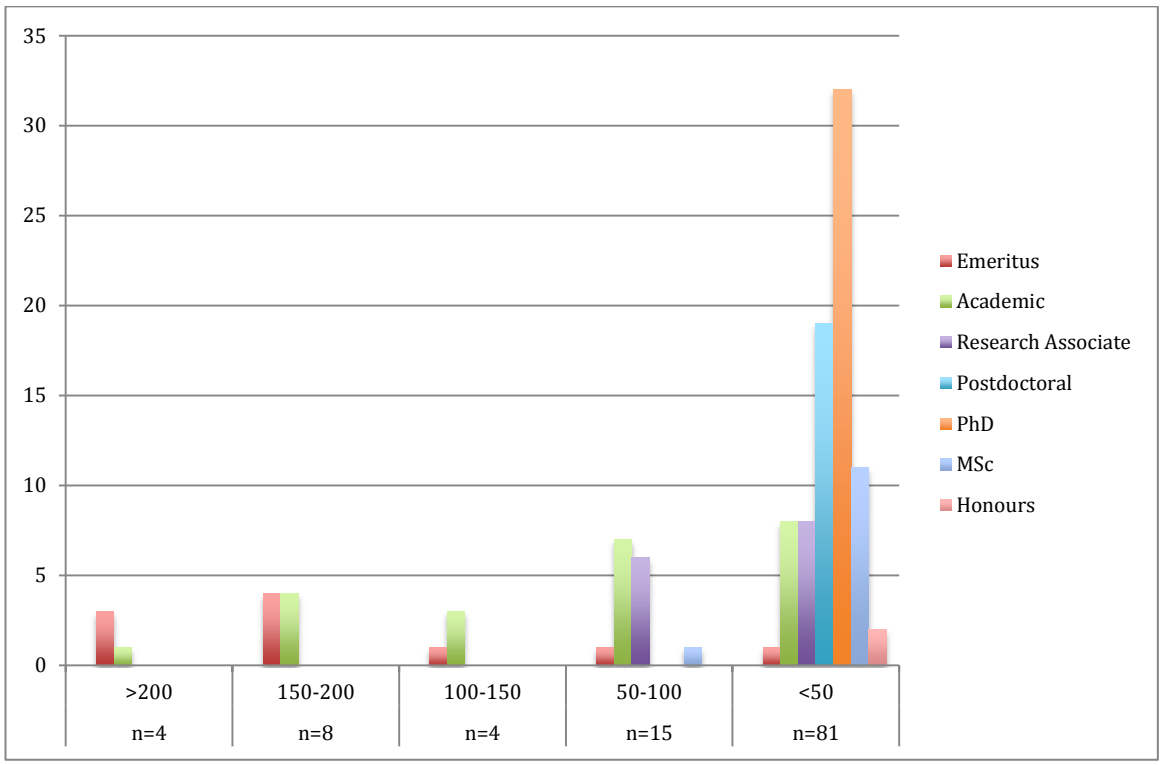

Figure 3.4 How many scientific papers have you published?

Discussion:

This question was asked in order to compare publication output with the amount of data reported. In an anonymous survey the number of research papers is also a good indicator of the seniority of the researcher. It can be seen in the above figure that the Postdoctoral, PhD and MSc researchers are clustered around the category of fewer than 50 papers, which is to be expected for researchers who are only just embarking on their research career. The Biological Sciences Department is the highest publishing department on the University of Cape Town campus. The number of scientific articles for 2014 reported by the Web of Science database was 224.

3.3.5 Researchers publishing supplementary data

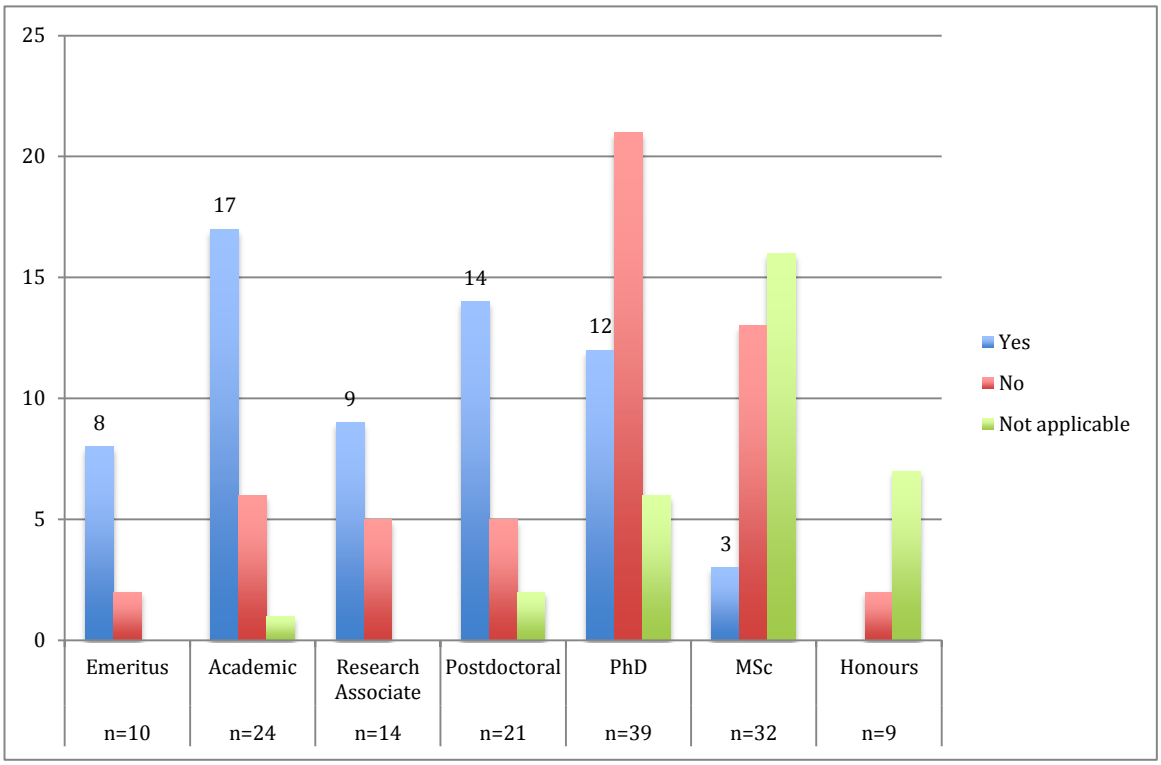

Figure 3.5 Have you published supplementary data with your published research?

Discussion:

The publication of supplementary information (SI) in the biological sciences was not as established as in the field of medicine and will be discussed in chapter 4. Supplementary information is supplementary to the published research and includes a range of material including data, methods, code, extended bibliographies, images, videos, audios, tables, discussion, equations, notes (Nature Publishing Group, 2015a). Answers to this question demonstrated that 42% (63 out of 149) of respondents had published SI files with their published research. Further discussion about SI files, what sort of information is submitted in supplements, and whether these are archived in reliable long-term archives will be discussed in chapter 4.

3.3.6 Reasons for publishing supplementary information

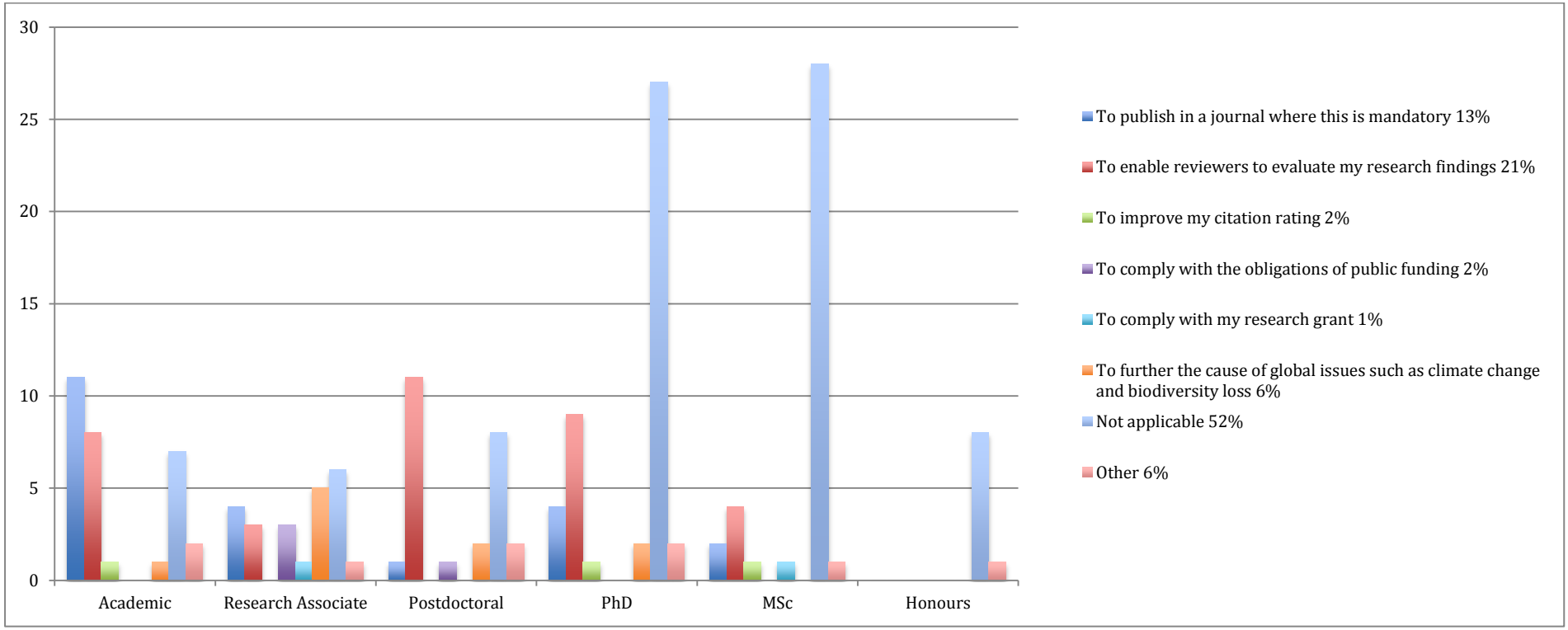

Figure 3.6 Why did you publish supplementary data?

Discussion:

Of the researchers submitting SI files 21% reported that this was done in order to enable reviewers to evaluate research findings, 13% to enable publication in a journal where this was mandatory, while only 2% reported this as an obligation of public funding. Sharing data with a published paper via SI files is a fairly new requirement with journals such as *The American Naturalist,*

Evolution, Journal of Evolutionary Biology, *Molecular Ecology* & *Heredity* leading the way. (Borgman, 2012). Other than *The American Naturalist* which publishes both ecological and evolutionary material, the other four titles are specialists in the field of phylogenetic studies. Underlying data in this sphere of research is usually sequence data where there are already two well-established repositories EMBL and GenBank and the discipline of deposting data is already well-established. It can be seen in Figures 3.9 & 3.25 that data are already being deposited in these repositories by staff and students in Biological Sciences and that there are a number of research projects producing sequence data.

Some of the additional comments submitted to this question about publishing supplementary data, were

- that some of the elements of the published research 'were likely to be of interest but weren't suited to the main body of the paper'
- that supplementary files allowed them 'to provide supporting information for which there was not space in the article' and 'to publish data expanding the 8000 words limit for research papers.'
- that where the data are too large to publish in the article they include evidence and figures in supplementary files
- that data were provided as supplementary information so that they could be re-used.
- 'To provide useful details to those interested in the paper to facilitate further research in the topic'

3.3.7 Researchers and Research Units with public funding

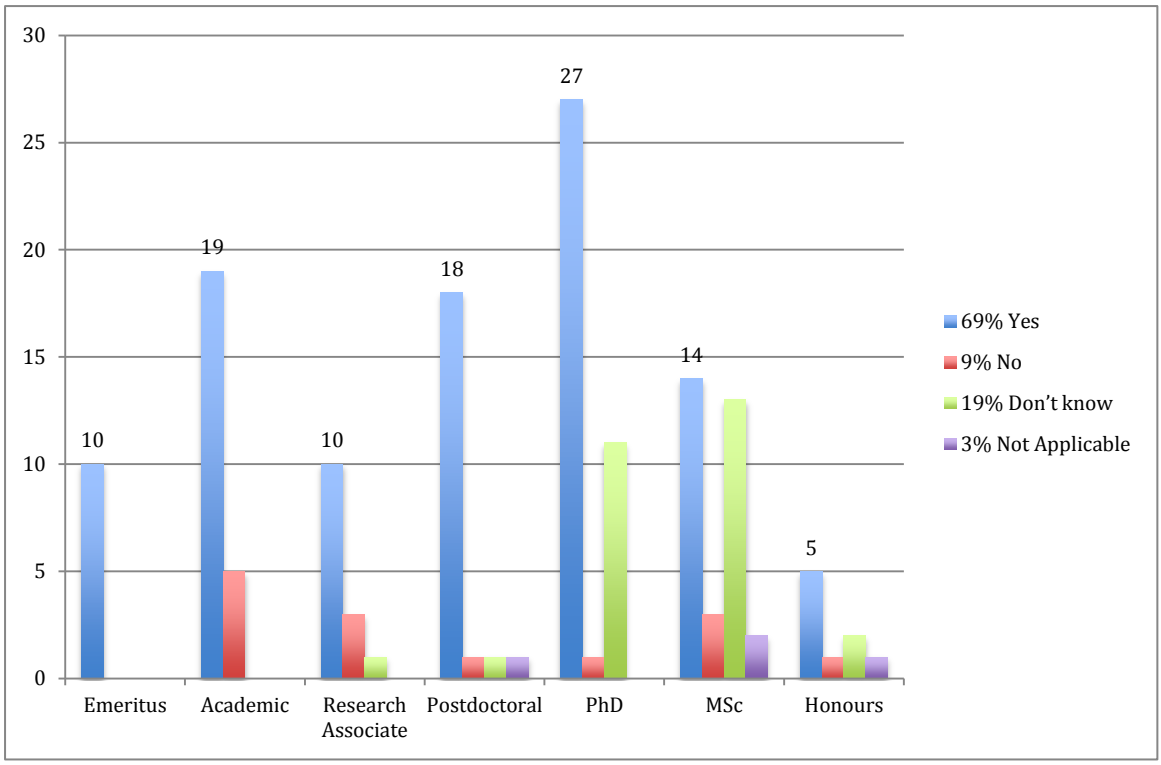

Figure 3.7 Do you or your research unit have public funding?

Discussion:

This question was asked in order to establish the level of public funding supporting research projects in the Biological Sciences Department. In the future it is likely that open access to research data generated by these project will be mandatory once the research has been published (NRF, 2015). The 'not applicable' group are probably those who do not require research funding as they are doing desktop studies or synthetic studies on existing data, either their own or from published research. It can be seen that public funding of Biological Sciences research is high at 69% in 2014.

3.3.8 Researchers with funding that required data curation

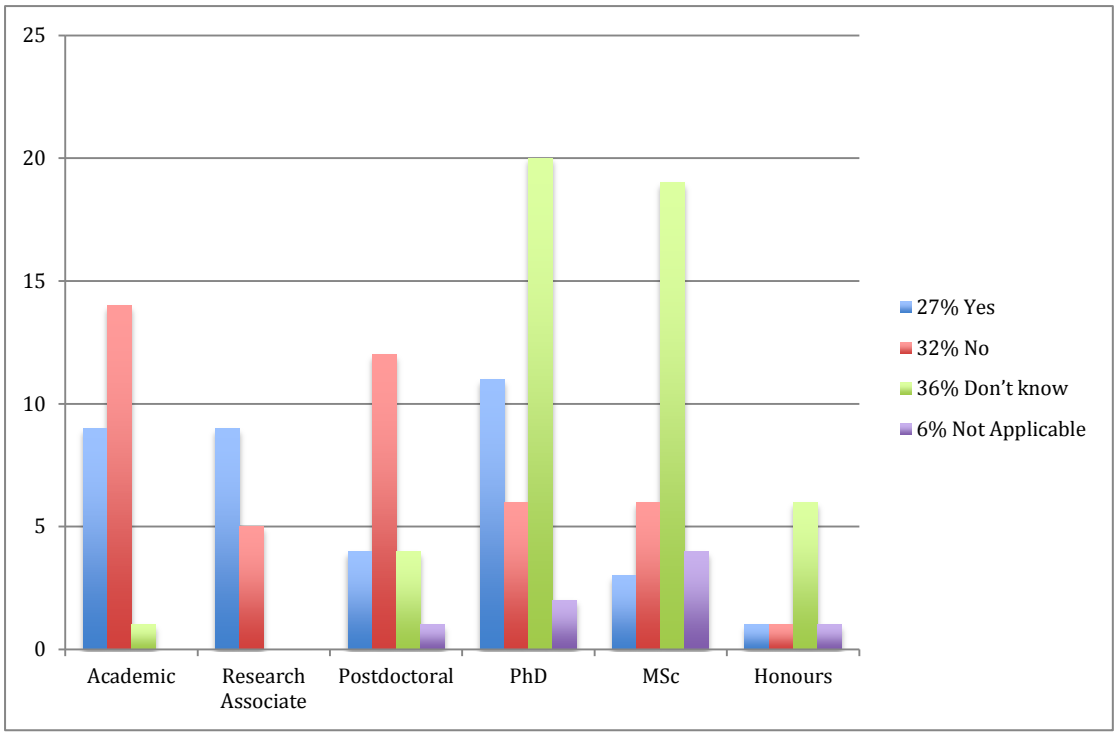

Figure 3.8 Has any of your funding or your research unit's funding required data curation?

Discussion:

South African public funding through the NRF has only recommended open data underlying published research from 1st March 2015, so it is interesting to see that 27% of researchers in Biological Science surveyed during 2014 responded that they were required to archive data. The most likely explanation for this is that they were collaborating with international researchers where funding requires data curation. It is concerning that academic researchers do not know whether their funding requires data curation, but understandable at the MSc and Honours level of research. Once funding proposals require evidence of data curation, researchers should be fully conversant with the requirements of funders.

3.3.9 Repositories used to archive data

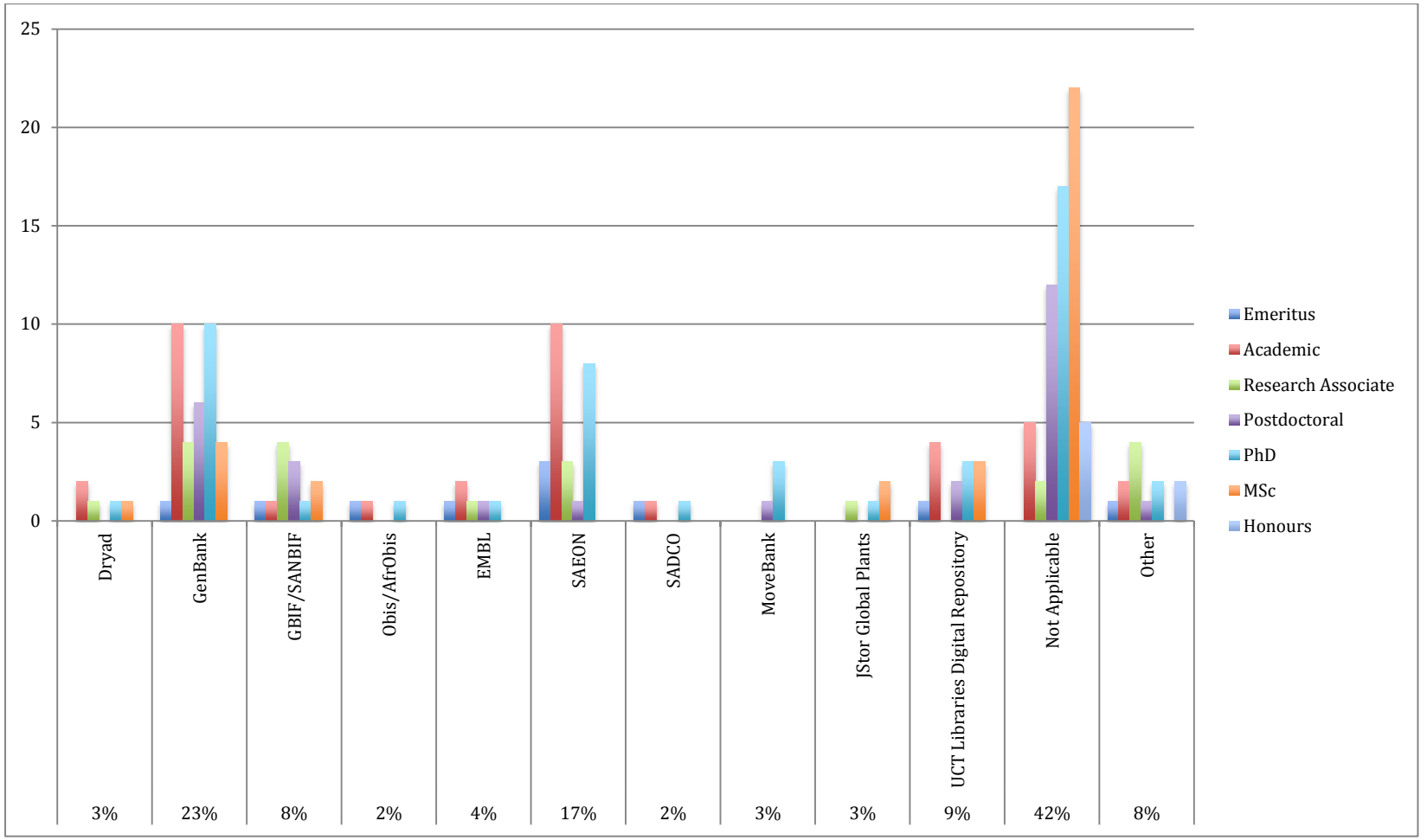

Figure 3.9 Have your data or your research unit's data been archived in any of the following repositories?

Discussion:

This was a multiple response question and the findings reported here demonstrate that research data are archived in a range of repositories according to the data type and the discipline of the repository. Of the respondents interviewed, 42% were not archiving data. The remaining 58% of respondents were archiving their data in multiple repositories, for example phylogenetic data would be archived in Genbank while geographic point data would be archived in one of the geographic suite of repositories, see below. The multiple-choice questionnaire also supplied KNB and FigShare as alternative repositories. As no one in Biological Sciences had used these repositories they were removed from the figure.

Respondents contributed the names of the following additional repositories where their data had been archived:

- The National Marine Linefish System
- The Animal Demography Unit
    - South African Bird Atlasing Project
    - Online Virtual Museum
    - South African Ringing Unit
- The BirdLife Seabird Tracking Database
- The South African National Biodiversity Institute data archives
- The British Library Sound Archive
- A departmental server used to archive remote sensing data
- The Plant Conservation Unit
- UvA-BiTS (University of Amsterdam Bird Tracking System [UvA-BiTS], [2013])
- Iziko Museums

Archiving digital research data for long-term accessibility is of great concern to the scientific community. Numerous new initiatives are emerging e.g. *Scientific Data (*Nature Publishing Group, 2015b*)* which make the selection of an appropriate repository easier for the researcher and suggest that a data staging repository, such DataStaR developed at Cornell University, may be the

correct level of repository appropriate at an academic institution which has not yet engaged with establishing an institutional repository. A staging repository enables researchers to share data and to create metadata prior to archiving in a subject specific repository (Steinhart, 2007:34).

The repositories in Figure 3.9 can be divided into broad subject categories

- Those providing access to genetic or molecular data – GenBank, EMBL
- Those providing access to geographic point data – AfrObis, SADCO, SABIF, SAFRING
- Those providing access to tracking data – Movebank, UvABiTS, BirdLife Seabird Tracking database
- Those providing access to images or sound as data – UCT digital repository, British Library Sound archive
- Those providing access to ecological data – Dryad, KNB, SAEON
- Those providing access to species data – AfrObis, Jstor Global Plants, ADU Virtual Museum

Information about the above data repositories are presented in chapter two.

3.3.10 Data ownership perceptions of researchers

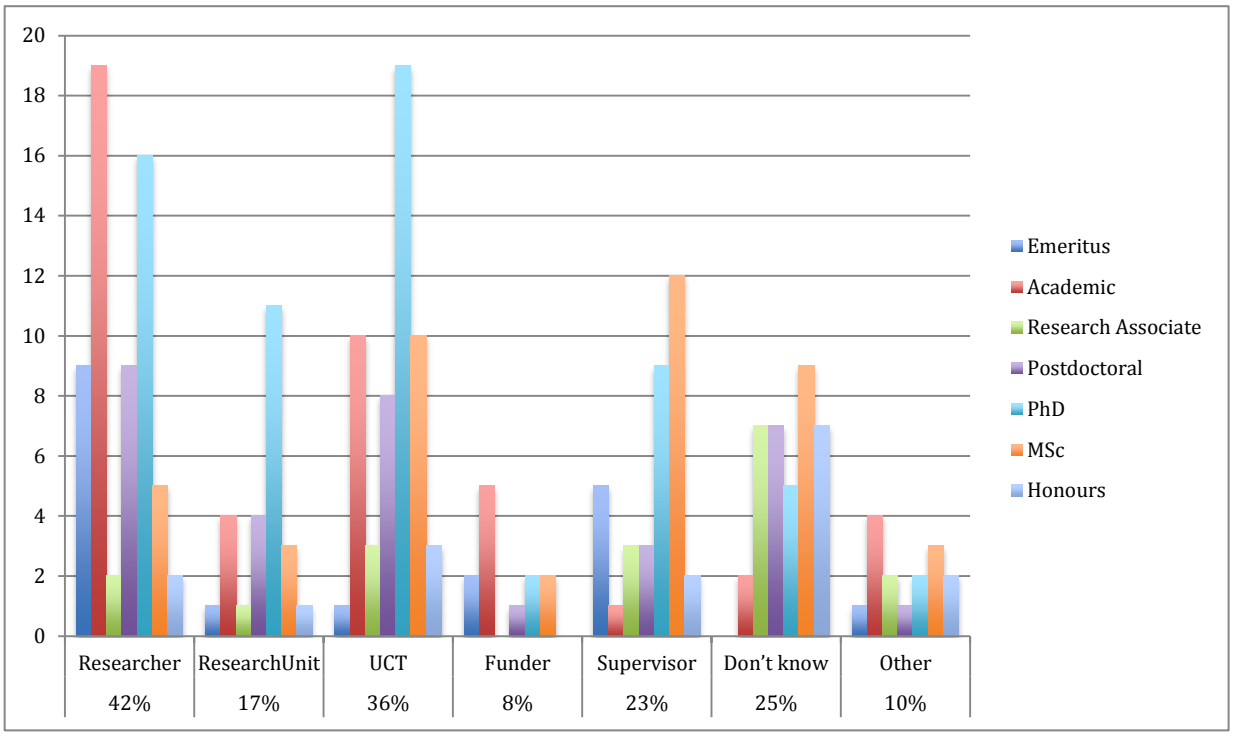

Figure 3.10 Who owns your data or your research unit's data?

Discussion:

This was a multiple-selection question allowing respondents to choose a variety of perceived owners of their data. In the 'Other' category there were various alternatives supplied by the owners of the data being used for research. Examples include:

- observers who contribute to the South African Bird Atlassing Project 'agree to their data being open access when they submit it' to the ADU.
- 'organisation I work for' or 'government or DAFF',
- 'some of them are archival data that have been digitized from government reports in the UCT library'
- [SA] Government
- SANBI
- Centre for Invasion Biology (CIB) at Stellenbosch University
- UVA-BiTS (see data repository descriptions in chapter two)

3.3.11 Data curation perceptions of researchers

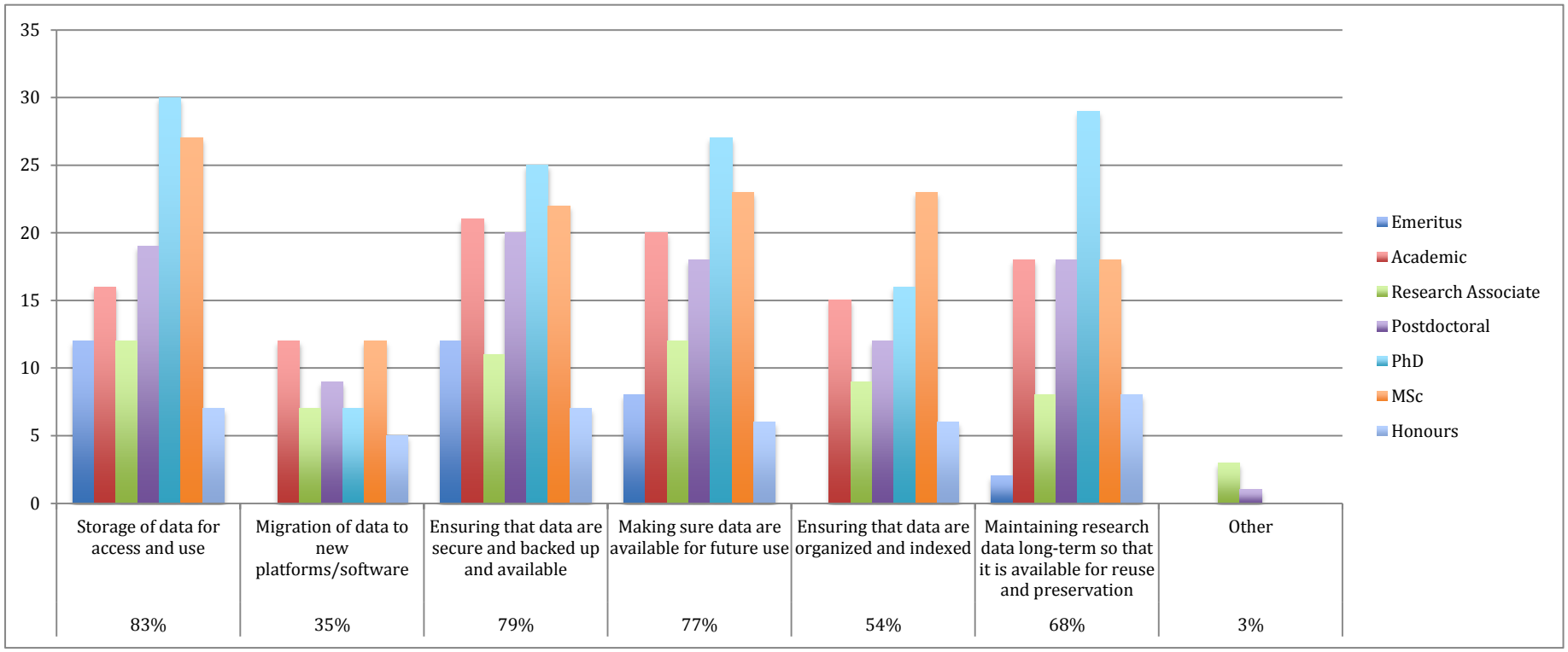

Figure 3.11 What do you think is the purpose of data curation?

Discussion:

This was one of the questions that was posed in order to be informative as well as interrogative and there was a level of duplication provided for the possible responses. As digital data curation is a relatively new concept to UCT researchers, the question was formulated to inform respondents who had not necessarily thought about curating their data. Set answers were also supplied in order to make it easier for respondents to work through the question more rapidly. Respondents could choose one or all of the answers and could also supply another response if they so wished. In this set of responses it can be seen that researchers value data storage, accessibility, security and availability.

The option 'migration of data to new platforms/software', which is a subset of 'making sure that data are available for future use' received the lowest score. It is surprising that this was not selected by emeritus/retired researchers as it is legacy data from their active research era that has become obsolete and inaccessible.

Three alternative responses were supplied:

- 'Maintaining research data long-term so that it is available for reuse and preservation' would allow 'others to verify research findings.'
- 'testing new hypotheses' and quoting a saying of Dan Janzen: 'Hypotheses come and go, but data are forever' – this citation could not be verified.
- 'ensuring that date are secure and backed up and available' facilitated 'reproducible research'.

3.3.12 Long-term data sets held or used by researchers or research units

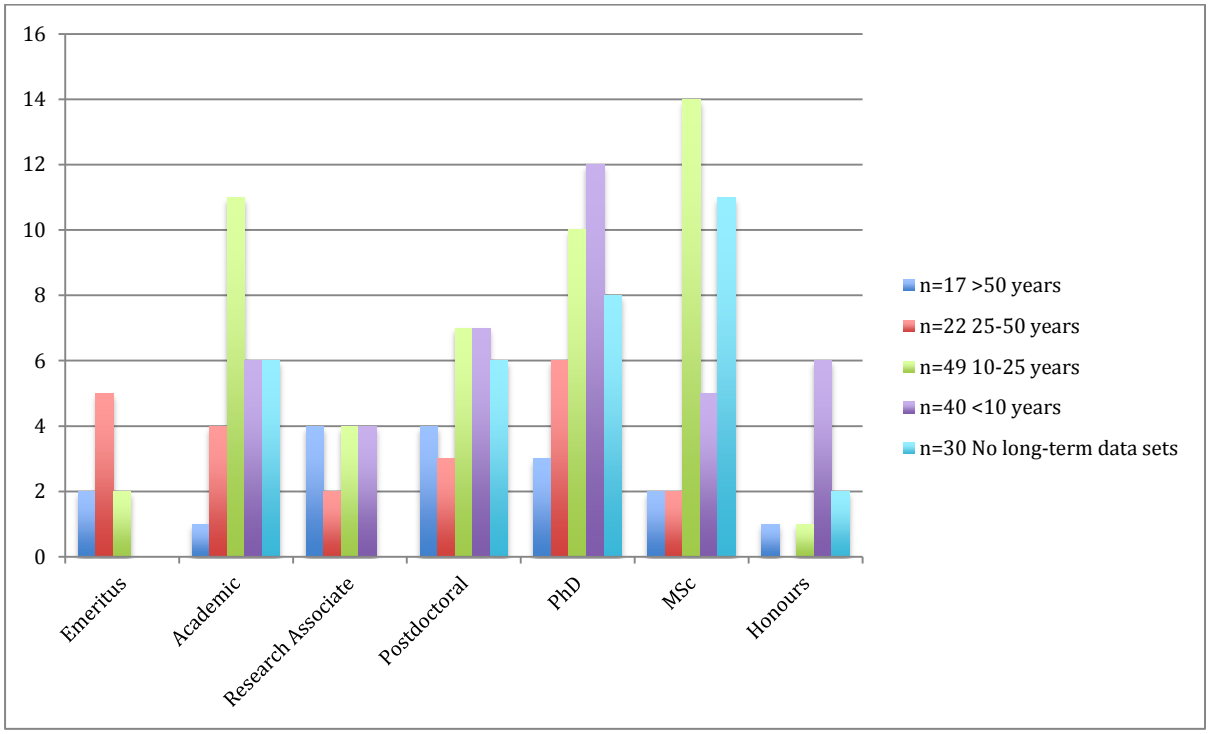

Figure 3.12 Do you or your research unit have long-term data sets?

Discussion:

As can be seen by the number of years next to the legend, there are a large number of long- and medium-term data sets. It is unlikely that the junior researchers had developed long-term data sets through their own research and were probably reporting on the data sets with which they were working for their research. Other than the interviews with retired or emeritus researchers, the fate of pre- and early-digital long-term data sets was unknown. It is reassuring to see that 10 junior researchers reported working with data sets of over 50 years in extent. This may however indicate only one or two over 50 year data sets on which collaborative research is being conducted. From the interviews it was found that a long-term data set existed for South African marine biology – the Southern African Coastal Ecological Survey, which was archived at Iziko Museums of South Africa. A pre-digital data set of freshwater ecosystems continues to be curated by an emeritus researcher. There are a number of pre-digital ornithological long- and medium-term data sets, some of which are archived in the Niven Library in the PFIAO, while others are curated by the data owners. The Bolus Herbarium represents a long-term data

set and there are also a number of medium-term data sets in the form of images, most of which are archived at the UCT Libraries.

3.3.13 Researchers' re-use of data

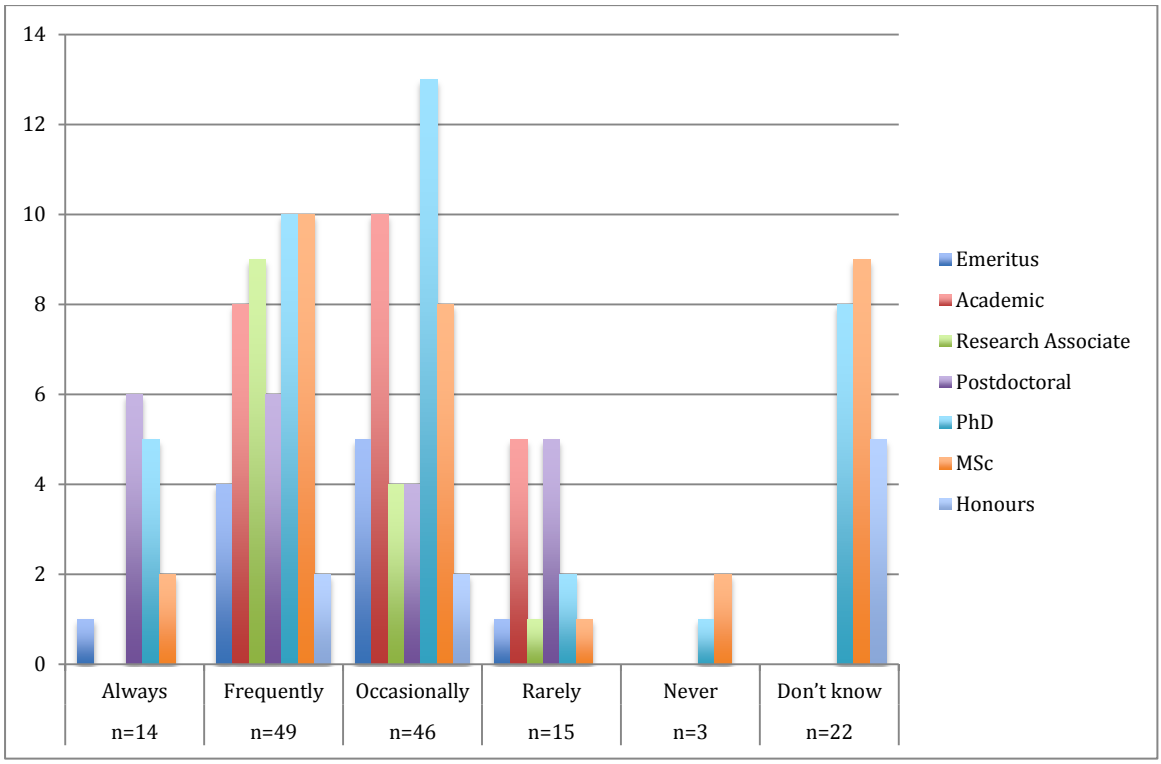

Figure 3.13 Do you or your research unit re-use your data?

Discussion:

Four out of the five academics who commented that they rarely re-used their data also responded that they did not have long-term data sets. The academic respondents who had data sets of less than 25 years responded that they occasionally re-used their data, while those with data sets of more than 25 years corresponded to those who re-use their data frequently. Should these data sets be archived in repositories in the future it is likely that the level of re-use would increase, which would make the funding of the research initiative more cost-effective.

3.3.14 Researcher willingness to make data available for future research

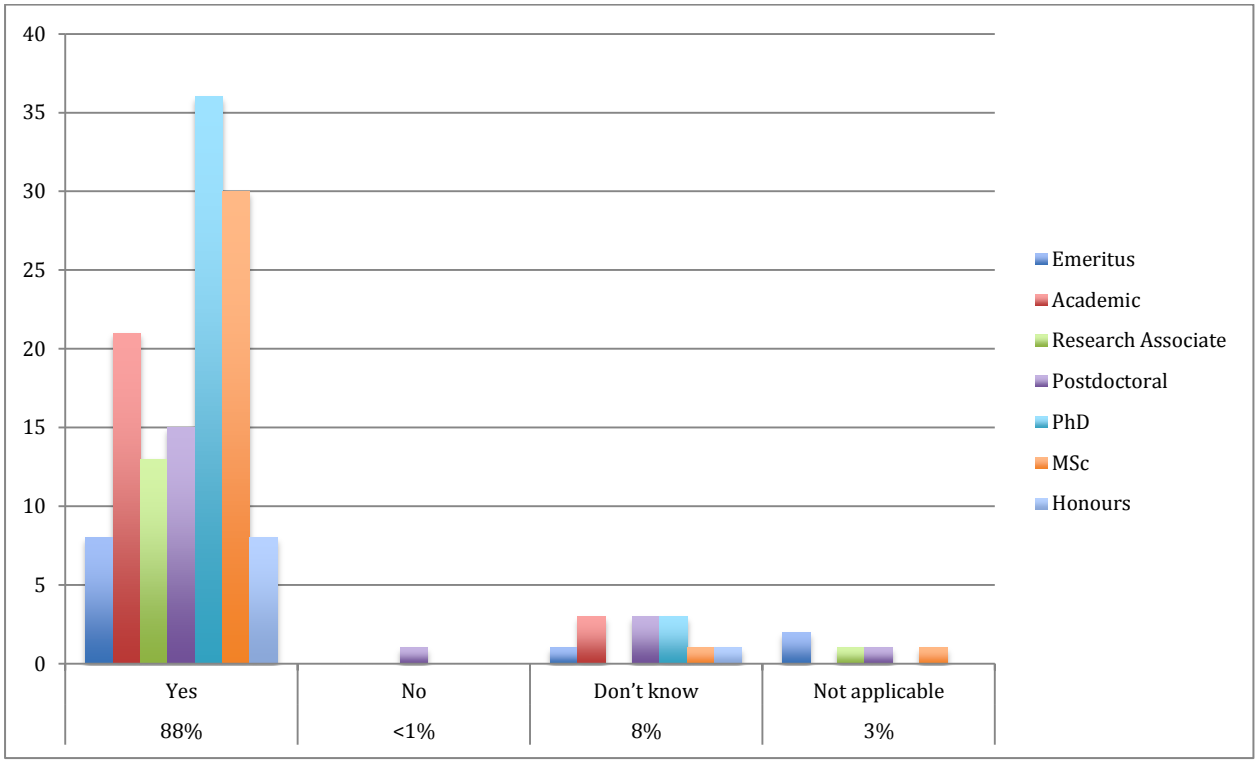

Figure 3.14 Should your/your research unit's data be made available for future research?

Discussion:

The large majority of respondents thought that their data should be made available for future research, although one qualified the 'yes' by stating that collaboration was a condition of making data available. The *conditions* under which the 88% will make their data available for future research are displayed in Figure 3.16 below.

3.3.15 Ways in which researchers share their data

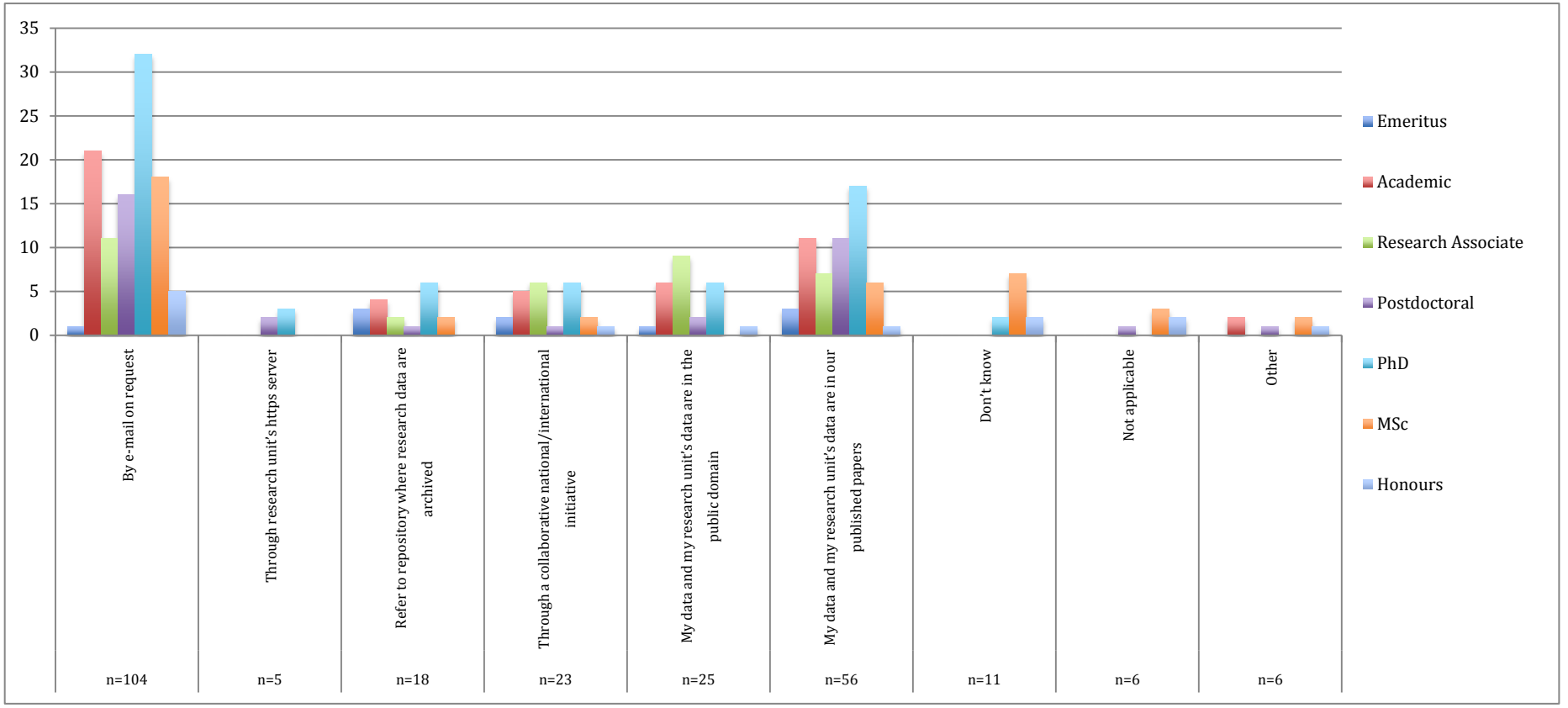

Figure 3.15 How do you share your research data or your research unit's data with other researchers?

Discussion:
Sharing data is one of the most controversial research data topics. Despite this, it can be seen in figure 3.15 that researchers do share their data, but the conditions for sharing are specific and intended to ensure benefits to the researcher who generated the data. The number of researchers who consider that their data is in their research papers is concerning as this is usually sythesised data and not raw data which may mean that it is not useful for futher research. It will seen in the next figure that researchers are very cautious about sharing their data, which is usually expensive to produce both in the amount of time spent on collecting and the cost of funding their field work. Making data available to another researcher is also expensive as was analysed by Porter & Callahan (1994) and discussed in chapter two.

Other methods of sharing data which emerged from the question that was illustrated in Figure 3.15 were:

- Dropbox
- Vula
- that 'some data are proprietary and only shareable in anonymised format'
- the research unit was obliged to respond to PAIA requests for data
- data were requested from DAFF & DEA and other UCT research units, which were then properly acknowledged.
- once the thesis had been accepted, the data would be archived in a repository in the public domain

3.3.16 Conditions under which researchers will make their data available for future research

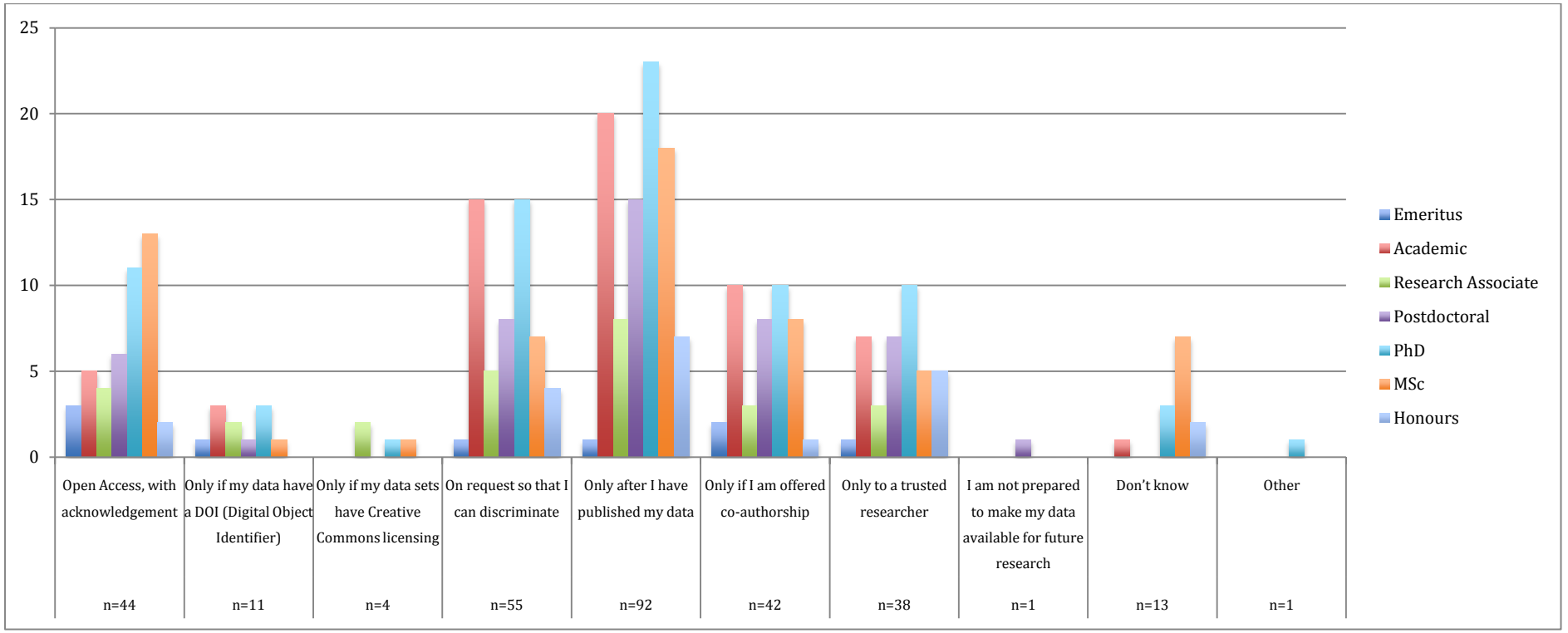

Figure 3.16 Under what conditions would you/your research unit make data available for further research?

Discussion:

The great majority of respondents treated this question, which allowed a multiple-selection response, as a Boolean 'AND', namely that sharing data was predicated on specific conditions, all of which had to be met and in a specific order; for example that data could be shared only after publication, and only if the data creator was offered co-authorship. The most common dependency was being able to interrogate the person who requested data 'so that I can discriminate' and the most dominant condition was 'only after I have published my data'. Being able to trust the person with whom data would be shared was clearly also an important consideration. Although this is not readily apparent from Figure 3.16, the raw data demonstrated these dependencies as the choices of individual respondents could be viewed.

Respondents who were prepared to share their data through acknowledgement, DOI or Creative Commons licensing were a different group of respondents, in all likelihood those working with big open data sets such as citizen science data which already have an open mandate.

Only one academic respondent was prepared to go the Open Access with acknowledgement route, with no further conditions. A PhD respondent said that as the data was secondary data, ownership and making data available was not a decision that could be made by the researcher.

The percentage breakdown for researchers who made publication a condition of data sharing is as follows:

- Academic – 17% was provisional on having published data
- Research Associate - 28% was provisional on having published data
- Postdoctoral – 23% was provisional on having published data
- PhD – 21% was provisional on having published data

The percentage of researchers who selected the category 'on request so that I can discriminate' was 34%, while the percentage of researchers who selected the category 'only to a trusted researcher' was 23%.

Despite this, it can be seen in Figure 3.15 that researchers do share their data, but the conditions for sharing are specific and ensure benefits to the researcher who generated the data.

Researchers internationally express anxiety that their research data may be misused by other researchers (RIN, 2008:28; PARSE Insight, 2010:19; Sayogo & Pardo, 2013:S24; Doorn, Dillo & Van Horik, 2013:237). They do not want their data used for unintended purposes, or with lack of integrity. In the case of endangered species, information about the geographic location of endangered species, is anonymised in a data archive. For example, the South African Bird Ringing Unit does not make the location of nests available for the Martial Eagle when sharing data (SAFRING, n.d.).

Discussion with researchers in the Department around the topic of data sharing revealed a situation where data was regarded as having been misappropriated by researchers in another academic department. Researchers affected by the data misappropriation had given permission for a case study to be presented in this dissertation, and were given the opportunity to edit the case study to ensure 100% accuracy of the facts. The situation was a good example of the need for a new ethical paradigm for digital data to be spelt out in policy documents. This case study is presented as 3.4 below.

3.3.17 Use of research data for desktop studies

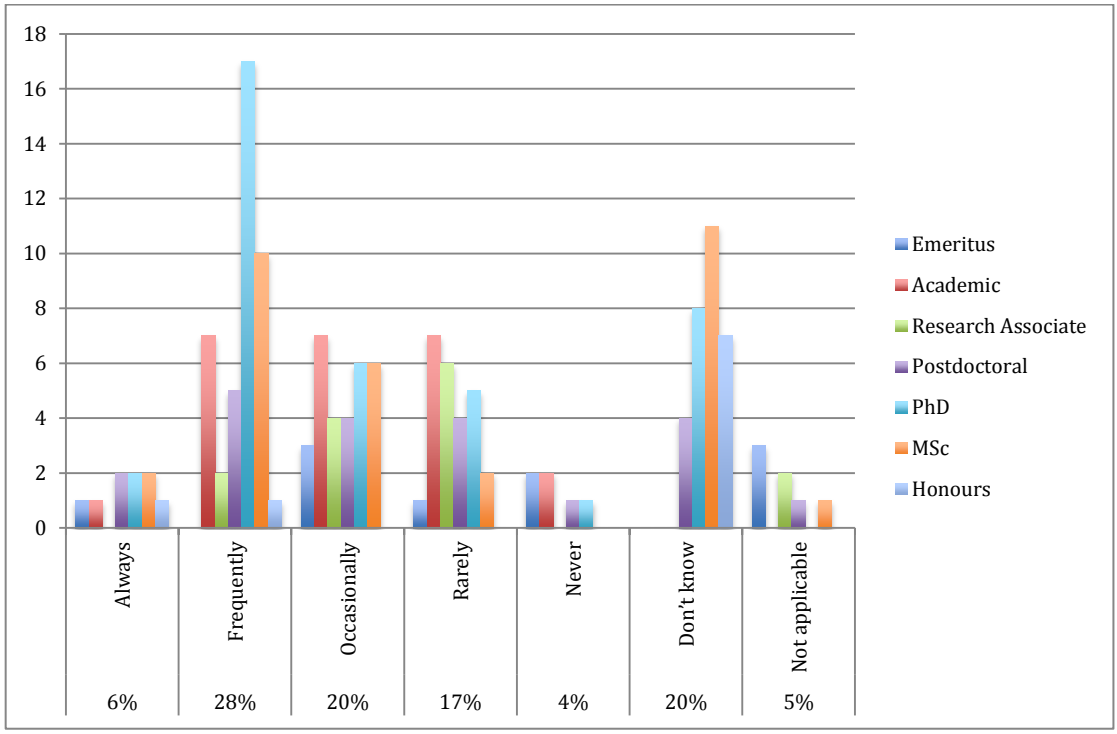

Figure 3.17 Do you or does your research unit conduct desktop studies using data?

Discussion:

A desktop study or desktop research, is office-based rather than field-based research. It analyses existing data in order to answer a research question not previously considered when collecting the data in the field or laboratory. Insufficient time, or a limited budget is a driver for desktop research projects. Such studies often require an extensive literature review. Figure 3.17 indicates that desktop studies were always or frequently undertaken (34%) or occasionally undertaken (20%) in the Biological Sciences as can be seen from the responses of Academics, PhD and MSc students.

3.3.18 Responsibility for data sets

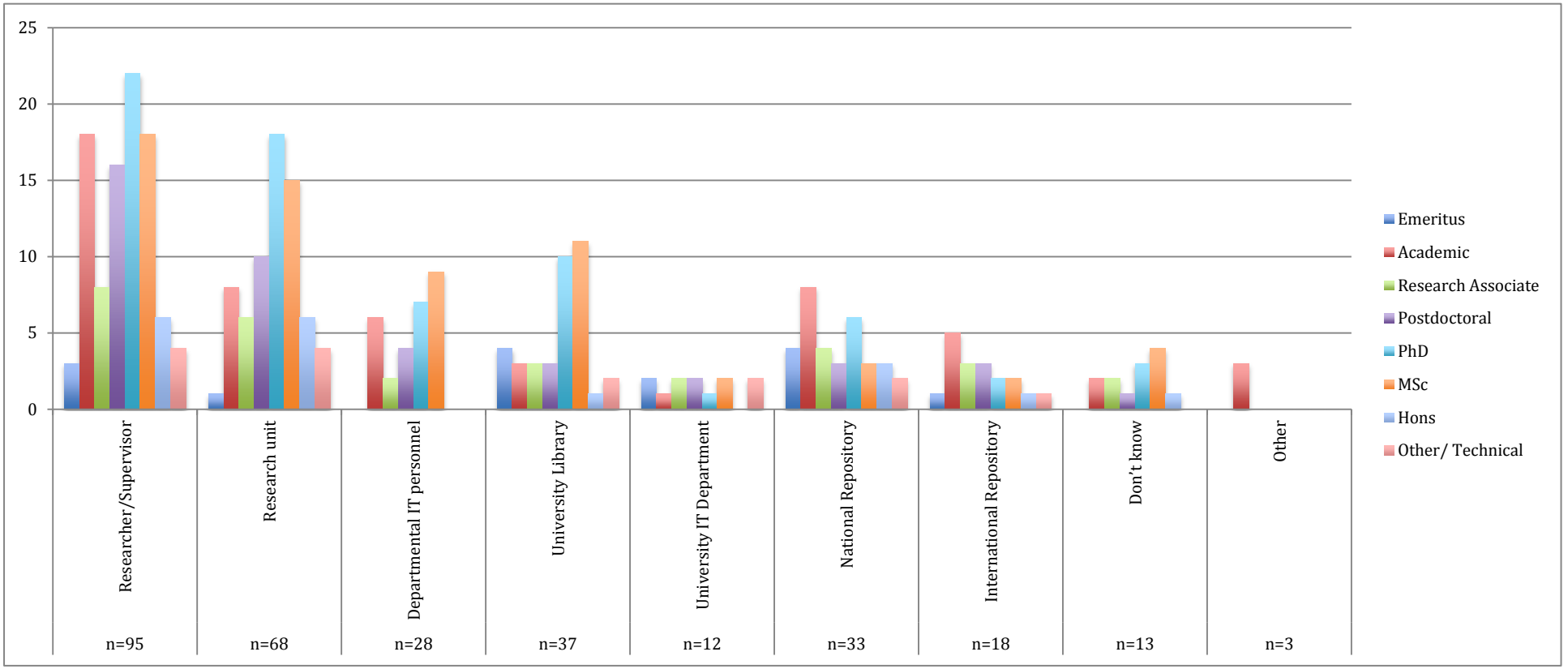

Figure 3.18 Who should be responsible for storage of data sets that are generated in this department?

Discussion:

It can be seen that the scores for the responsibility for data storage residing with the Researcher or Supervisor and the Research Unit were highest in this question. This could reflect a lack of confidence in other possible data storage sites or a lack of experience in using these sites.

Other suggestions for data storage responsibility were the Animal Demography Unit and the Funder.

3.3.19 Data Back-up frequency

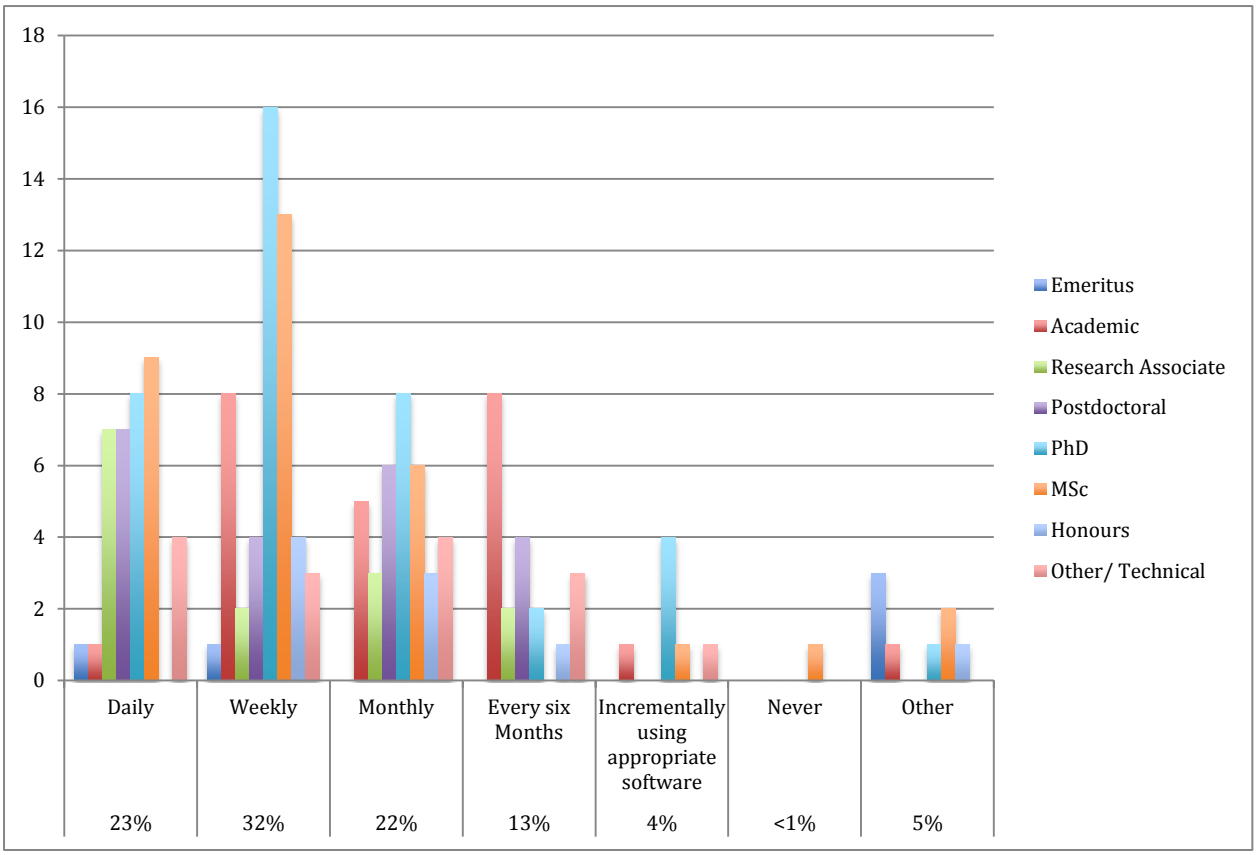

Figure 3.19 How often do you back-up your electronic data?

Discussion:

The first two alternatives, 'Daily' and 'Weekly' are pragmatic responses, because any researcher working with data would be expected to back-up constantly to avoid data loss. The other responses apply to longer-term back-ups.

Back-ups of digital data are an integral part of research data management among Biological Sciences Researchers.

Other responses to this question were

- 'As often as I need to - could be daily, sometimes weekly and other times longer.'
- 'storage [back-up] in the cloud via Dropbox automatically'
- 'irregularly'
- 'occasionally'
- 'cumulative'

3.3.20 Location of data back-ups

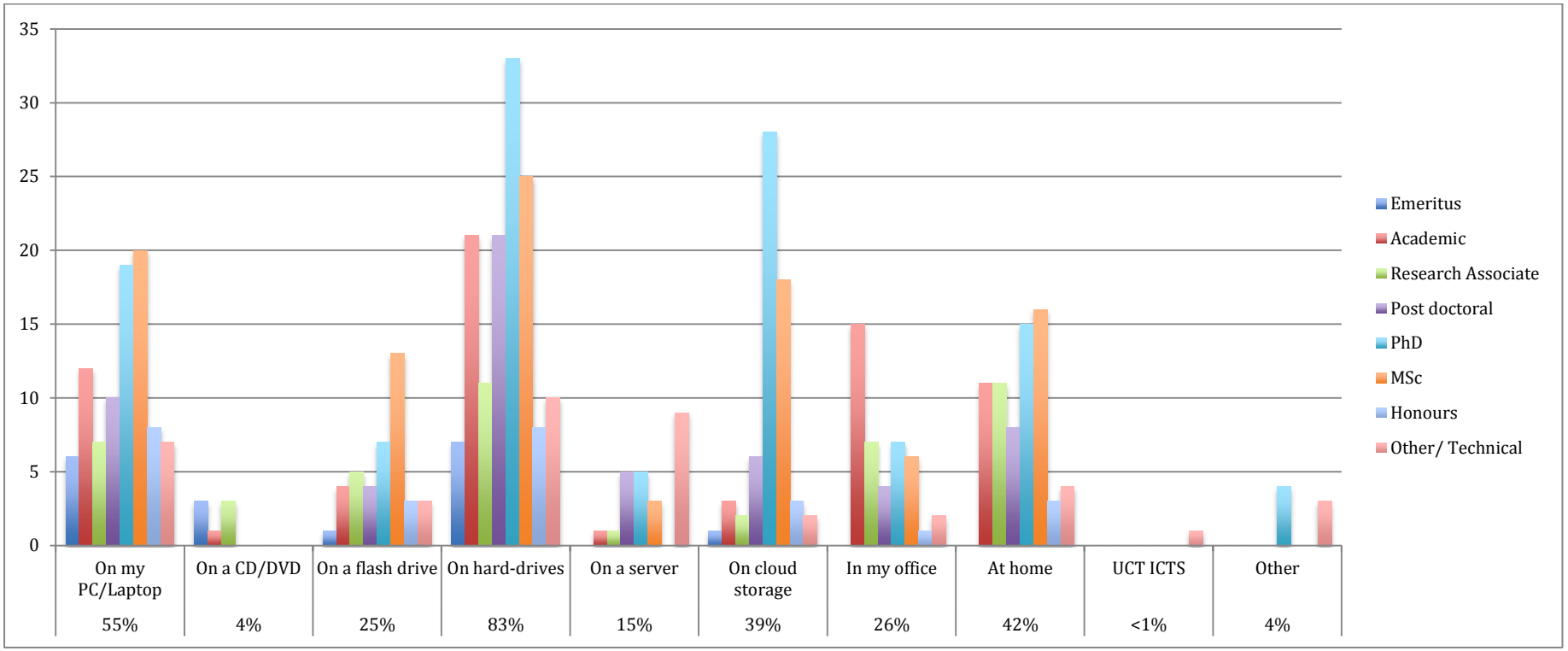

Figure 3.20 Where do you keep your data back-ups?

Discussion:

The problem here is one of where these back-ups are located for long-term data archiving. The majority of respondents keep their own back-ups on hard-drives, on their PC or Laptop, at home and on cloud storage. A data staging repository would be an appropriate place for more secure storage of data, but such a repository is not available at UCT. There is an opportunity for UCT ICTS to investigate such a solution. Github, a data staging repository where researchers can share their code (Github, 2015), is suggested as an interim cloud solution for researchers. Many researchers in the Biological Science Department use the statistical programme R to analyse their data sets, the code for these analyses can be shared with other researchers doing similar analyses, Github is an international solution for code sharing.

Other responses to this question were:

- Technical respondents – Dropbox, Google drive, data server. Dropbox and Google drive are both examples of Cloud Storage
- Postdoctoral respondents – Supervisor, Printed hard copies, e-mail inbox, Github data repository

3.3.21 Number of data back-ups kept by researchers

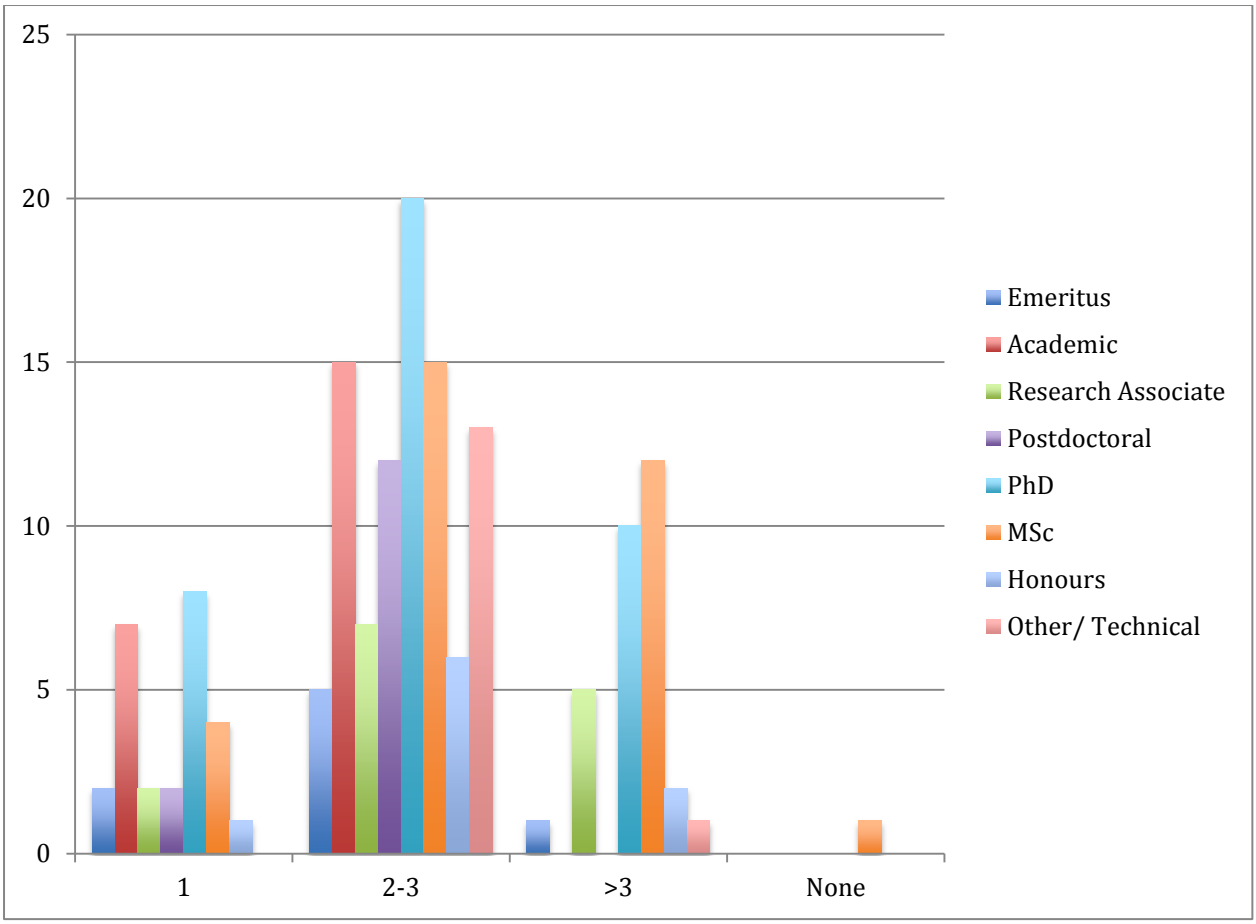

Figure 3.21 How many data back-ups do you have?

Discussion:

As can be seen in this figure the majority of respondents have two to three data back-ups. Three copies of data are the minimum number recommended by digital data management guides, namely original data, a copy stored externally, a copy stored externally in a remote location.

3.3.22 Types of metadata considered important to describe research data

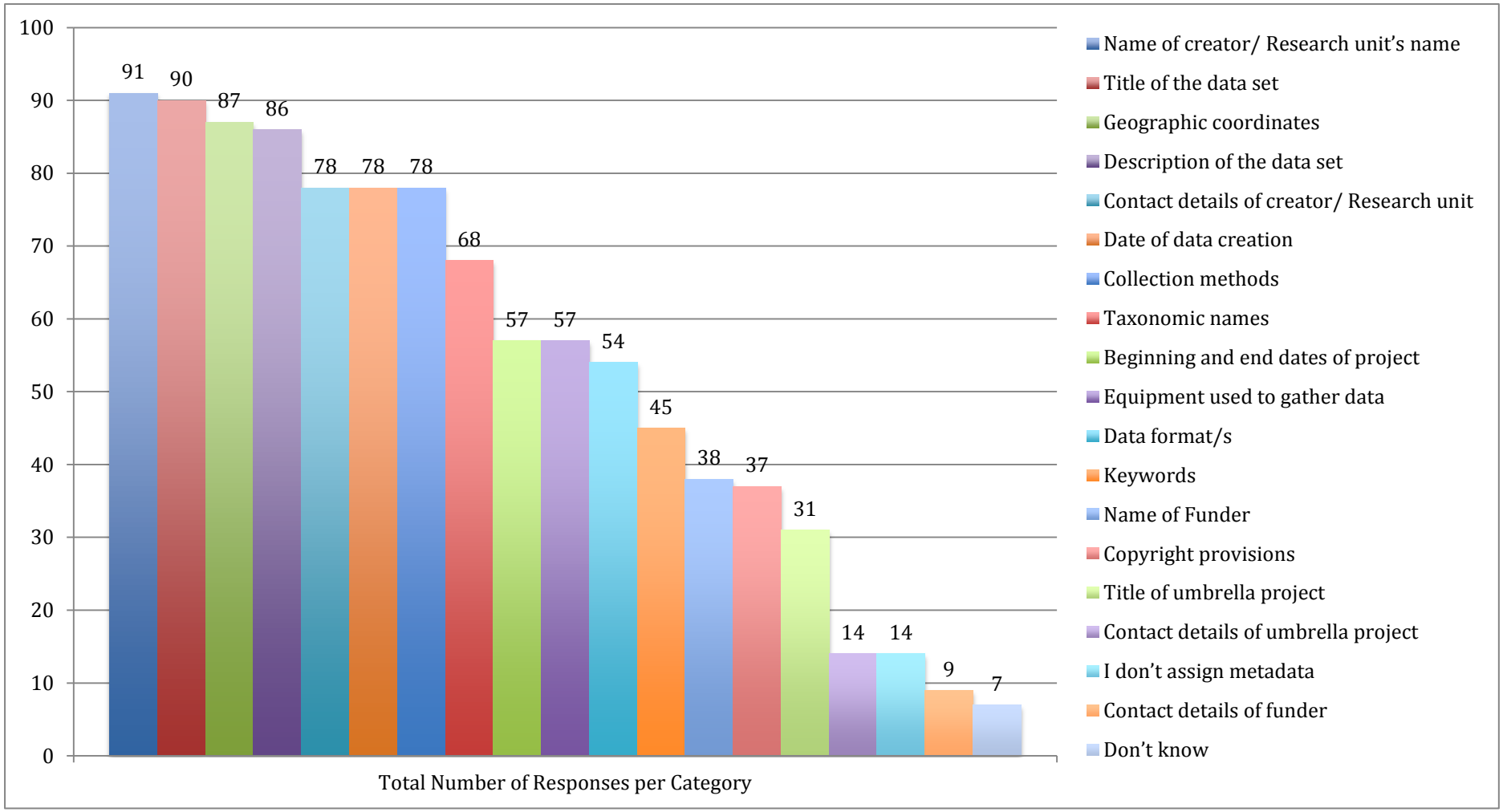

Figure 3.22 What types of metadata do you consider important to describe your data?

Discussion:

This question was designed primarily to provide information to the respondents rather than be interrogative. What is of interest is that 10% of respondents "don't assign metadata" and 5% of respondents who "don't know", showing that there is still some ignorance about metadata among the researchers. This may just be that the term metadata is not familiar and that "data description" may be the terminology used by researchers. The respondent numbers for the categories 'I don't assign metadata' and 'Don't know' can be seen above each column in the histogram.

3.3.23 Interest expressed for attendance of workshops to discuss metadata generation

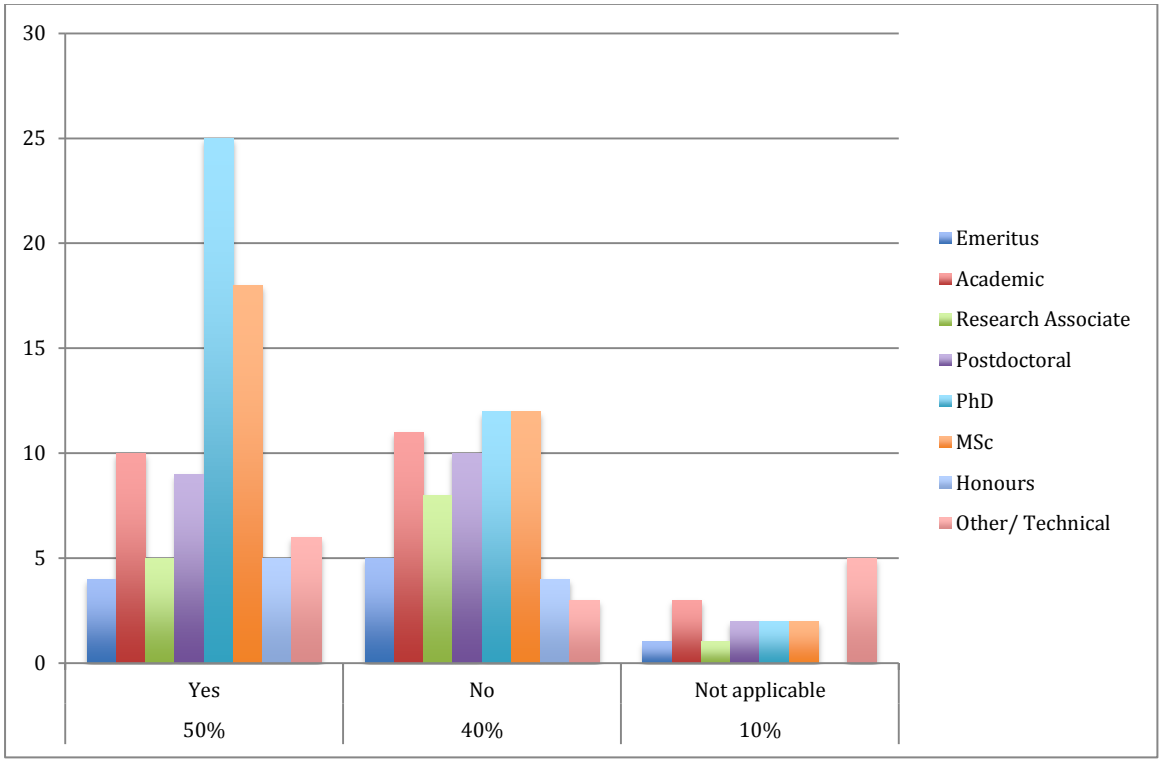

Figure 3.23 Would you attend a workshop to discuss metadata generation?

Discussion:

The previous question demonstrated that 15% of respondents did not assign or know about metadata. The remaining 85% of researchers seemed to understand what metadata was, but as can be seen in this figure only 50% of the researchers in

Biological Sciences would attend a workshop on metadata generation. This is an opportunity for UCT Libraries to develop an appropriate online document as well as offer workshops targeting metadata generation and assignment for science researchers. An investigation into the appropriate metadata language would have to be conducted prior to presenting workshops as there is not a one-size-fits-all solution.

3.3.24 Size of data sets held by researchers

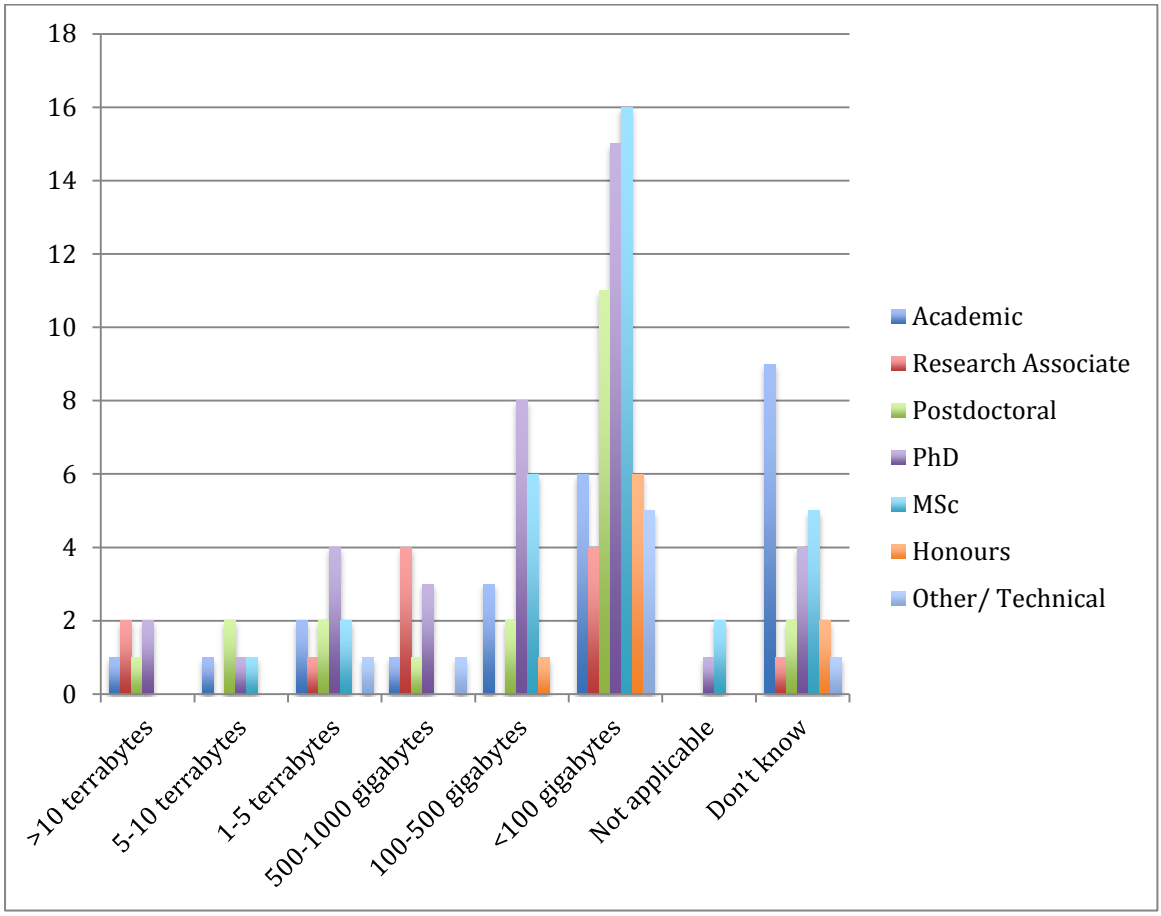

Figure 3.24 Approximately how much research data do you have?

Discussion:

As can be seen from the above figure, there is a lot of digital data that is generated in the Biological Sciences Department. If one compares the response to this question to the results in figure 3.25 below, it is possible to get an idea of the type of research which generate these large data sets. Image data is one of these data types, sound data is another. The research unit working on bats would generate very large data sets on each field expedition as they collect echo-location (sound) data. Streaming camera traps and other cam recorders gather large amounts of image data, and satellite loggers used to map the foraging range of seabirds, and the movement patterns of terrestrial animals collect large amounts of spatial data. Information about these data were provided by technicians in their response to question 6 in Appendix B. While technicians in the department may not be responsible for archiving data, they are usually called in to assist with the technical aspects of research.

3.3.25 Types of digital data generated by researchers

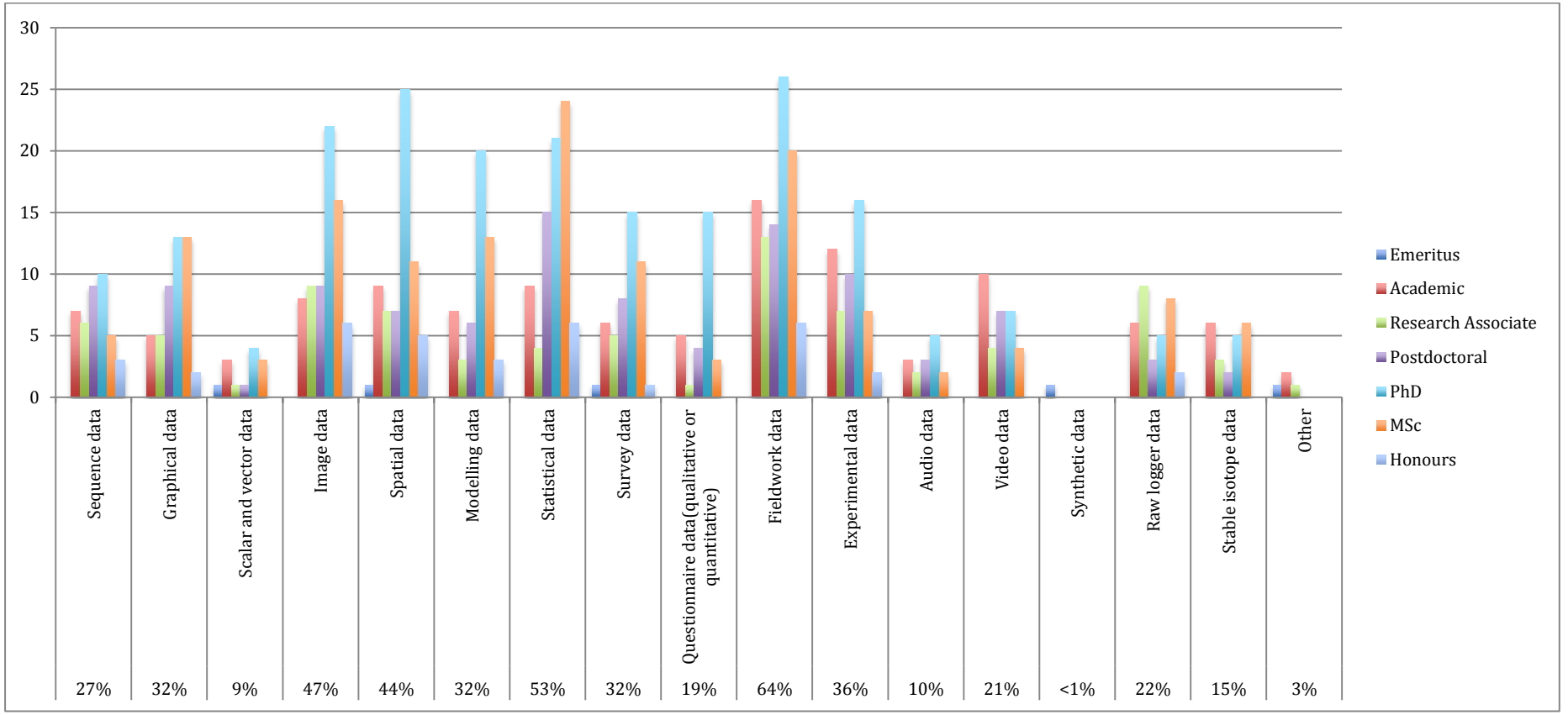

Figure 3.25 What types of digital data does your research generate?

Discussion:

It can be seen from this figure that fieldwork data form the largest data type. This is however misleading, as fieldwork can include many of the other data types such as image data, survey data, questionnaire data, audio data, video data, raw logger data and stable

isotope data, all of which are collected in the field. Statistical data is usually the interpretation of other types of data collected in the field. Other responses to this question were spectrophotometric data, trade data, molecular data, and synoptic data.

3.3.26 Formats of researchers' digital data sets

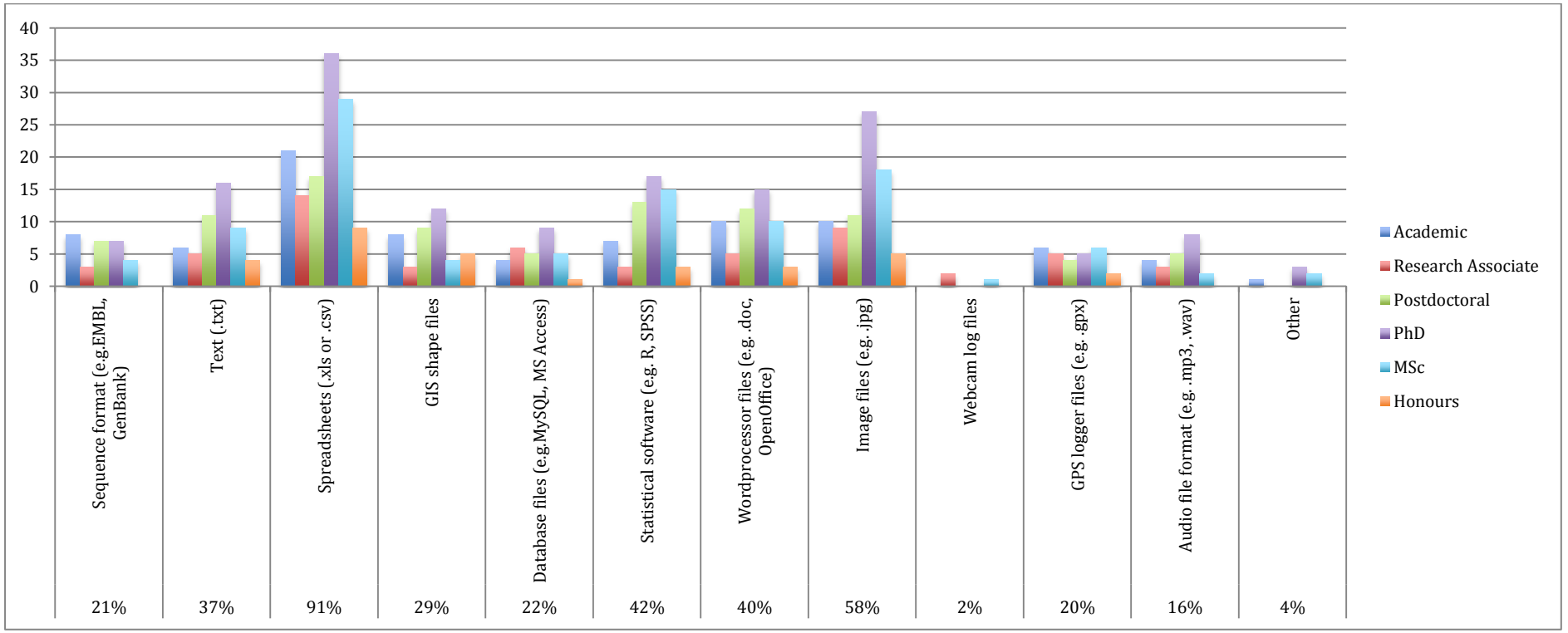

Figure 3.26 In what formats are these digital data sets?

Discussion:

Other responses to this question were Mp4 (which was grouped with the Audio file format), 3D images, NetCDF, and satellite data format. NetCDF data is scientific binary data generated by array-oriented data. The acronym stands for Network Common Data format and originated with the University Corporation for Atmospheric Research (UCAR) (Unidata, 2014). Video files and blend files were also mentioned. Raw video file format would have the extension .yuv, but it is more likely that a video file format would correspond with the type of equipment/programme used to create the video. The file extension is very important metadata as it indicates the programme that can be used to open the file (Arms et al., 2013). Blend files are used in 3D image creation. 'Blender' is a suite of free, open source, cross-platform tools which can be used to create a range of 3D products (Blender.org, 2015).

The majority of file types were spreadsheets (91%) as many digital data types can be exported as spreadsheets which makes the data easier to interpret.

3.3.27 Data loss among researchers

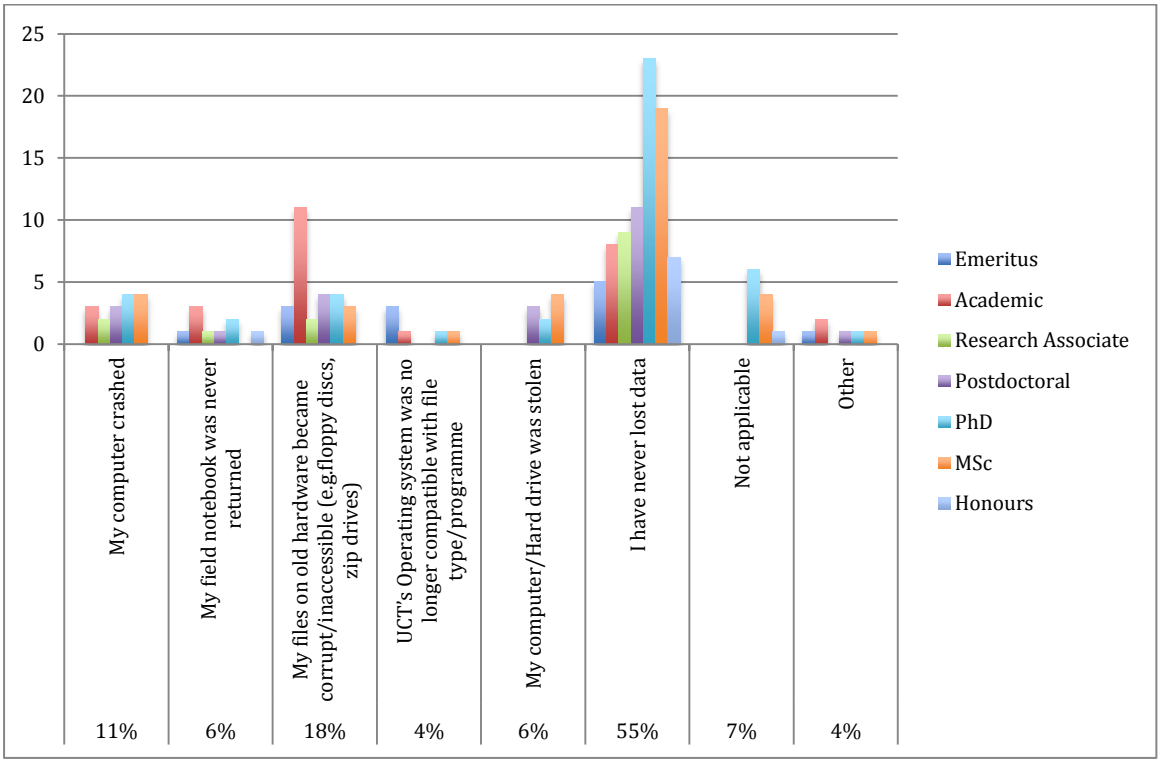

Figure 3.27 Have any of your data been lost?

Discussion:

A few of the emeritus/retired researchers had interesting stories about their data. One had a field notebook shredded by a mouse making a nest in a filing cabinet, another lost data because of degradation of audio tapes, a garage clean out and mainframe tapes which were not migrated by technical staff. One threw away his data because there was no interest shown by colleagues, the department or the University Archives.

Other responses were a lost lab notebook, software not migrated to the latest format, loss of data when the UCT network crashed and when the UCT server was hacked. A respondent lost figures and analysis when the latest version of proprietary software was unable to read old files, which is another version of software not migrated to the latest format.

It is reassuring that 55% of respondents have never lost data. Cloud storage can prevent data loss from theft and crashes, but some researchers do not trust resources such a Dropbox or Google-drive and feel that these resources are insecure from hacking or a lack of confidentiality.

3.3.28 Data migration to new software/operating systems

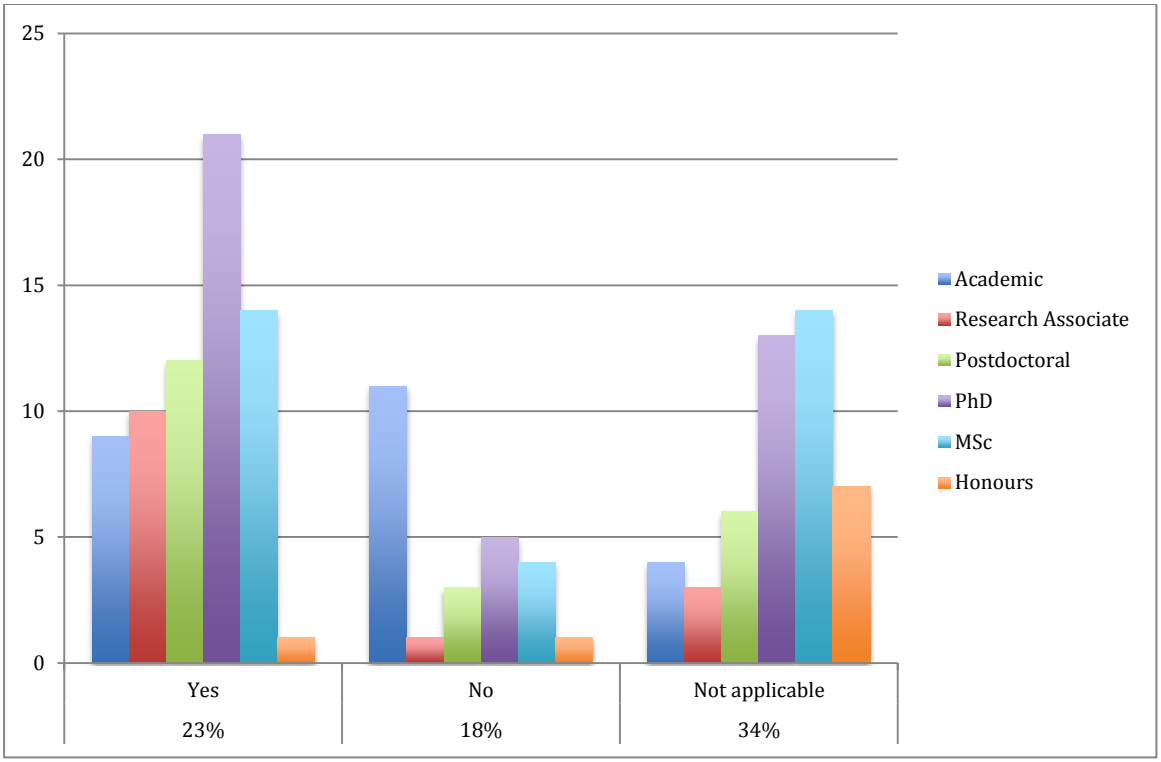

Figure 3.28 Do you migrate your data to new software/operating systems when the current system becomes obsolete?

Discussion:

The figure of 52% for the no/not applicable categories is interesting and can probably be attributed to the percentage of young researchers in the group of respondents who have not been in the research field for long enough to have encountered changes in operating systems and software that required data migration. There was a 58% response to the survey from Honours, Masters and PhD students.

3.3.29 Researchers requiring data management assistance

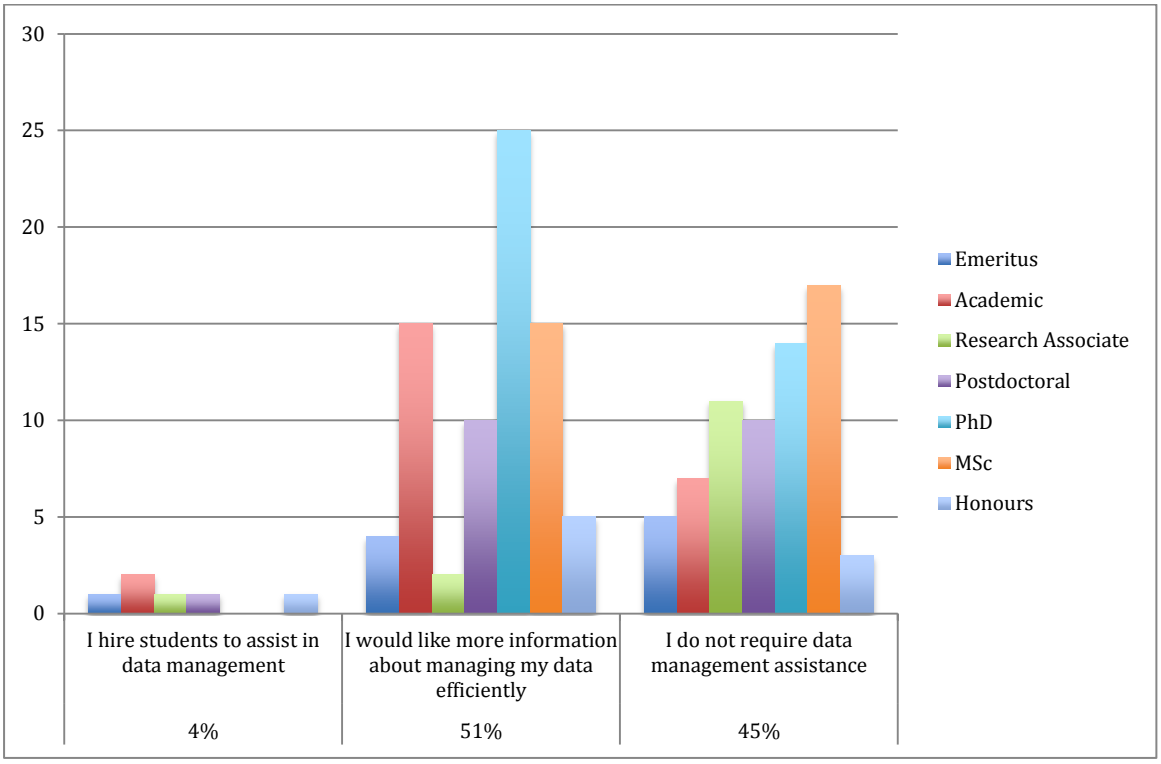

Figure 3.29 Do you require data management assistance?

Discussion:

From answers to this question it can be seen that data management means different things to different researchers. There are groups of researchers in the Biological Sciences Department who work with big data sets and are extremely data literate, these are the respondents who hire students to manage their data. The 45% who do not require data management assistance could be responding in the negative because they too are competent data managers, or do not have big data sets so are able to manage these themselves. This leaves the 51% who would like more information.

3.3.30 Interest expressed for attendance of workshops to discuss data management

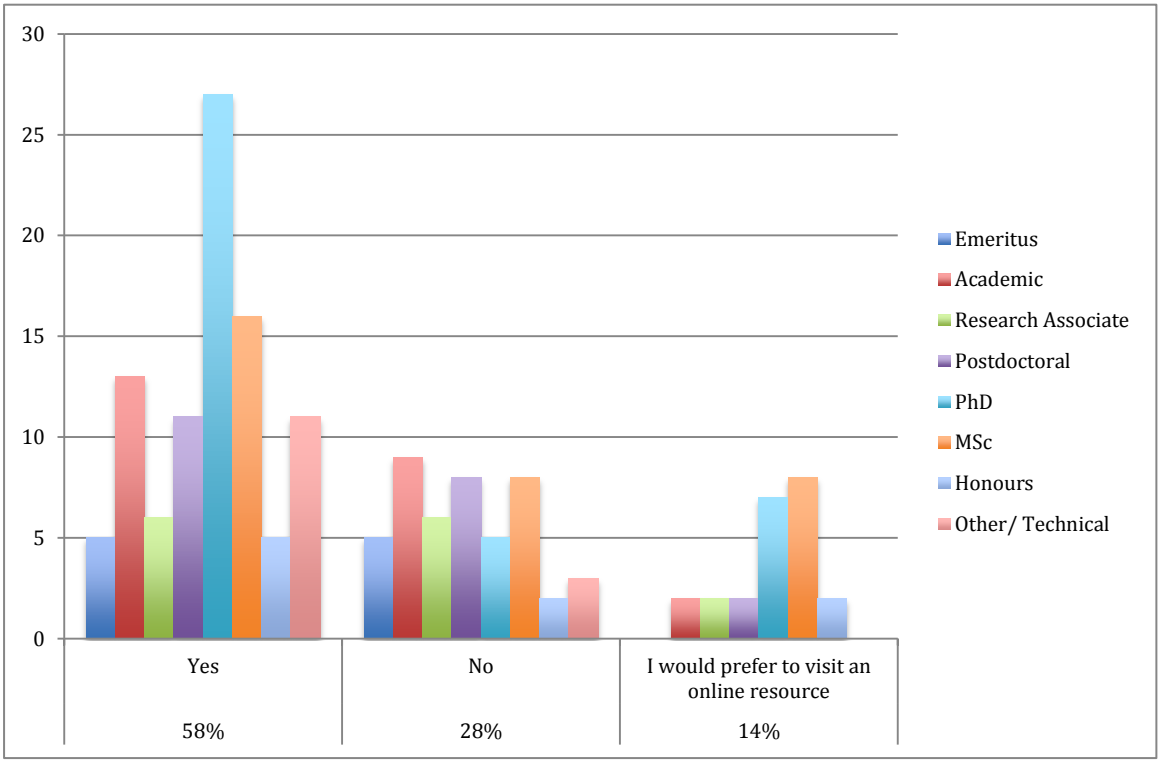

Figure 3.30 Would you attend a workshop to discuss data management?

Discussion:

As discussed under figures 3.23 and 3.29 there is an opportunity for UCT Libraries to create online documents and offer focussed data management workshops and assistance.

3.3.31 Budgeting for data management and data curation

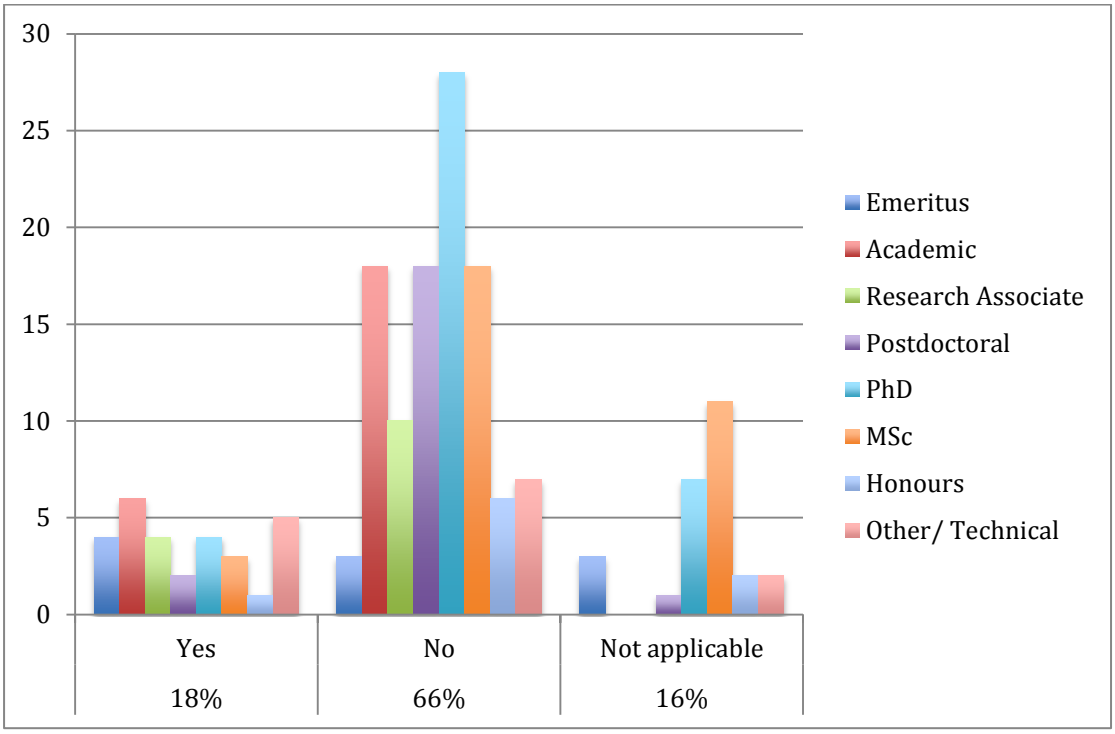

Figure 3.31 Do you budget for data management and data curation?

Discussion:

As can be seen in this figure a high percentage of respondents do not set aside a proportion of their research funds for data management and data curation. It is essential that researchers are aware of the costs of these routines ahead of funding mandates. Fortunately there are researchers in the Biological Sciences Department who have experience in data management budgeting (18%). If UCT Libraries intend to support this aspect of RDM it would be advisable to consult with staff in the ADU and the PCU, as both units have expertise in budgeting for RDM. Their research is closely involved with data management which forces them to budget successfully.

3.3.32 Evidence of data preservation plans

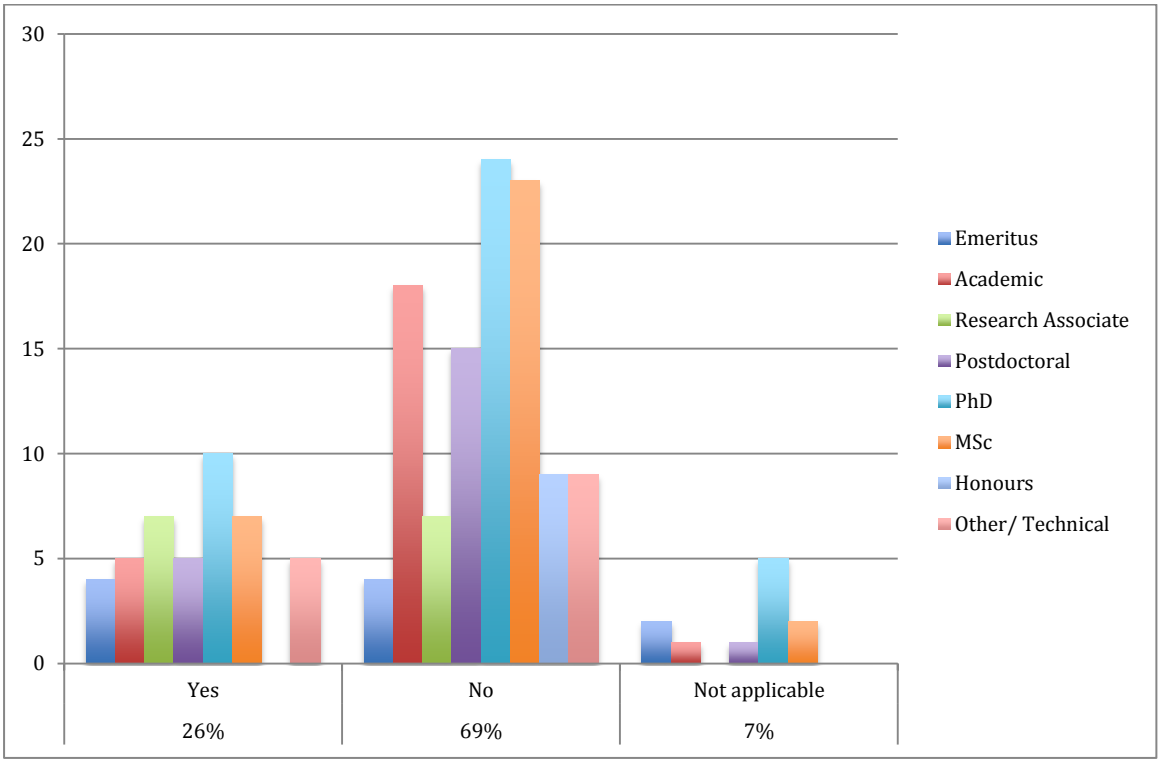

Figure 3.32 Do you have a data preservation plan?

Discussion:

The respondents to this final question demonstrated that data preservation is not being attended to by the majority of researchers in Biological Sciences. There are a number of interrelated possible reasons for this. Ideally, data preservation should be led by senior university staff such as the Deputy Vice Chancellor for Research and feed down through deans to departmental heads and academics. There is no evidence that this is being done at present. Data preservation requires time and funding, neither of which are available to the 69% of respondents who have indicated that they do not have a data preservation plan. Academics have heavy administrative and teaching loads which often take precedence over research. Funding is also a challenge and as will be seen in the following chapter, research is funded in multiple ways and unless funders require evidence of data management and data preservation plans researchers would rather spend their hard-earned research funds on the requirements for conducting their own field research and/or enabling students to engage in field research.

3.3.33 Conclusion

The results of this survey have shown that despite no evidence of systematic RDM at UCT, Biological Sciences researchers are to some extent engaged in data archiving as respondents are archiving some of their data in local, national and international repositories (see Figure 3.9). This is in all likelihood data underlying the publication of a dissertation or scientific paper. There is a lot of potential for the UCT Libraries to provide support for RDM at many levels for postgraduate and academic researchers. This will be discussed in more detail in chapter five.

Data preservation planning and budgeting for data preservation are neglected aspects of the research process at UCT as can be seen in Figures 3.31 and 3.32. The recent proposal by the NRF (2015), recommending open access to data underlying published research funded through an approved repository, demonstrates the urgency for RDM to be taken seriously at all levels of the UCT research community and UCT Libraries.

3.4 Data sharing case study

3.4.1 Introduction

Many publications discuss the importance of making research data available for verification purposes and in order to extend the boundaries of science by building on previous research (see 2.5 above). As can be seen in Chapter 3, Figure 3.16 above, the majority of responses from Biological Sciences respondents clustered around maintaining some sort of control over sharing their data. The literature attributes the reluctance to share data to fear of misuse of data or use for conflicting purposes (see 2.6 above), but there is very little evidence in the literature about the fear that data will be used without proper acknowledgement. There is also very little in policy documents specifically addressing research data plagiarism.

A recent case of misappropriation or unethical re-use of data in Biological Sciences at UCT is used to demonstrate the importance of a clear RDM policy and the need for more clarity on data re-use in policies such as the *UCT Authorship Practices Policy* (2010) and the *Policy & Procedures for Breach of*

Research Ethics Codes and Allegations of Misconduct in Research (2014) which are presented and discussed below.

3.4.2 Data misappropriation or poor data ethics? - a case study

3.4.2.1 Context of the case study - Experimental Fishing Exclusions for African Penguins in South Africa

African Penguins (*Spheniscus demersus*) have seen a 60% decline in numbers in the eight years between 2001 and 2009. Findings indicated that there was a strong relationship between this decline and competition for food resources, namely anchovy (*Engraulis encrasicolus*) and sardine (*Sardinops sagax*) which are caught by the purse-seine fisheries off the east and west coast of South Africa. It was decided by ornithological scientists to investigate whether the decline in the penguin population was related to paucity of suitable food resources by closing areas surrounding breeding colonies to fishing

An Island Closure Task Team (ICTT) was formed "under the auspices of the Pelagic Scientific Working Group of South Africa's Department of Agriculture, Forestry and Fisheries" (Crawford et al., 2011:150). Participating researchers in the ICTT would share their data in order to test hypotheses about sardine and anchovy fish stocks and how these relate to the food requirements for African Penguins.

In 2008 the first fishing exclusion zones for the benefit of African Penguins off the coast of South Africa were established. Two pairs of islands with African Penguin colonies were chosen for a feasibility study, Dassen and Robben islands on the West Coast and St Croix and Bird islands, in Algoa Bay on the East coast. The islands in each pair would serve respectively as a Control (open to fishing) and an Experiment (closed to fishing) in order to test the food resources hypothesis. Researchers collected field data to enable them to assess whether the fishery closure would assist the African Penguin in their foraging efforts, enabling them to raise their chicks with sufficient appropriate food, and to assess whether this would improve the recruitment (chick

survival). The findings in the experimental area on the East Coast demonstrated that even small no-take zones, as represented by a 20 km radius closure, benefitted the foraging efforts of the African Penguin (Pichegru et al., 2010:498). Research on the Robben Island African Penguin population on the West Coast found that the availability of appropriate food resources was the most important driver of African Penguin population viability (Weller et al, 2014:42).

3.4.2.2 Context for the sharing of the data from the field research

The members of the Island Closure Task Team are divided into Group A and Group B for this discussion. The Group A researchers are those studying the population dynamics of the African Penguin, a mixed group of UCT, BirdLife South Africa, DEA and independent researchers who gather extensive and expensive field data. The Group B researchers, who consult for the Department of Agriculture, Forestry & Fisheries (DAFF), are from the Marine Resource Assessment and Management group (MARAM) in the Department of Mathematics and Applied Mathematics at UCT. They are mathematical modellers who do not collect data, they model data to test hypotheses.

Group A made their field data on the African Penguin available to the Island Closure Task Team to use for evidence-based decision making for the island closure feasibility study presented above. One of the members of Group B used the data for a modelling exercise which resulted in the award of a PhD by UCT. Group A researchers had provided data to the student in good faith for management purposes through the ICTT and had emailed additional data on request to the student. The data collectors were recognised in the general acknowledgement at the beginning of the thesis, but not acknowledged in the individual figures and tables contained within the body of the thesis, whereas data from the fishing industry were acknowledged in tables and figures. The PhD student then submitted his findings for publication to a local scientific journal based on these data without any acknowledgment. Group A cried data misappropriation and refused to contribute futher data to the working group and the Group B researchers. The Group B researchers were of the opinion that as the data had been provided to the ICTT and used to complete a PhD

thesis, these data were now in the public domain and the leader of Group B could not see that there was a problem.

3.4.2.3 Comments on data sharing ethics

The etiquette and ethics for the sharing of quantitative field data are undeveloped globally although the ethics for the re-use of qualitative data in the fields of medicine and social sciences are well developed. There is an awareness of the occurrence of field data plagiarism, but it is difficult to pinpoint documented cases in the literature. The wording of ethics and data management policies is vague and field research data are not specifically identified as forms of intellectual property requiring copyright protection.

UCT has a number of policies in place which have clauses which could be interpreted to protect the ownership of data and to penalise the unethical appropriation of data. These are:

- *UCT IP Policy* (University of Cape Town, 2011)
    - 2.13 "**Intellectual Property (IP) means all outputs of creative endeavour in any field that can be protected either statutorily or not, within any jurisdiction, including** but not limited to all forms of copyright, design right, whether registered or unregistered, patent, patentable material, trademarks, know-how, trade secrets, rights in databases, information, **data**, discoveries, mathematical formulae, specifications, diagrams, expertise, techniques, research results, inventions, computer software and programs, algorithms, laboratory notebooks, business and research methods, actual and potential teaching and distance learning material, UCT's name, badge and other trade marks associated with the operations of UCT, Tangible Research Property, and such other items as UCT may from time to time specify in writing;"
- *UCT Policy & Procedures for Breach of Research Ethics Codes and Allegations of Misconduct in Research* (University of Cape Town, 2014)

    - 3.3 "**Plagiarism – misappropriation or use of someone else's work**, ideas, results, methods or intellectual property **without acknowledgement or permission**"
    - 3.4 "**Abuse of confidentiality – taking or releasing the ideas or data of others which were shared with the legitimate expectation of confidentiality"**
- *Authorship Practices Policy* (University of Cape Town, 2010)
    - Page 3 **"In the case of interdisciplinary and inter-institutional research, the senior researcher(s) have a special responsibility to ensure that discussions about authorship matters and possible differences in conventions are initiated early and with all researchers that are involved"**
- *UCT Policy for Responsible Conduct of Research* (University of Cape Town, 2012)
    - Preamble to Policy

    "In keeping with the emphasis on excellence in research, **UCT has a Responsible Conduct of Research framework of policies that govern research at the university, all of which are designed to promote ethical research conduct, integrity in research and related relationships** and to provide procedures to guide decision makers or persons who wish to raise concerns"
    - Implementation

    "**All UCT-based or affiliated researchers bear responsibility for ensuring that these policies are implemented properly and are adhered to."**

The highlighted portions of the above policies are regarded as appropriate rules for the sharing of data, even though quantitative field data are not specifically itemised. In particular

- point 3.4 of the *Policy & Procedures for Breach of Research Ethics Codes and Allegations of Misconduct in Research*, "**Abuse of confidentiality**

– taking or releasing the ideas or data of others which were shared with the legitimate expectation of confidentiality"

- the statement on page three of the *Authorship Practices Policy* **"In the case of interdisciplinary and inter-institutional research, the senior researcher(s) have a special responsibility to ensure that discussions about authorship matters and possible differences in conventions are initiated early and with all researchers that are involved"** and
  the statement in the implementation of *UCT Policy for Responsible Conduct of Research* "**All UCT-based or affiliated researchers bear responsibility for ensuring that these policies are implemented properly and are adhered to."**

all have direct bearing on the correct rules and procedures for sharing research data.

3.4.3 Conclusion

There is a clear conflict of interest in this data sharing case study with, on the one hand, a group of field ecologists working towards the conservation of a now endangered South African bird, and on the other hand, mathematical modellers who consult for fisheries and who are employed to support the fishing industry. Compounding the conflict of interest in sharing the field data was the lack of ethics on the part of the fisheries consultants, the absence of data sharing policies within DAFF, and the lack of a stated memorandum of understanding for the sharing of data between the ICTT members prior to sharing any data with the fisheries consultants.

The case study demonstrates is that there should be a very clear policy for data sharing, for ethical research behaviour, respect for data provided in good faith and rules ensuring data confidentiality in the same way as this is done for qualitative data. Data should always be acknowledged, and data that is not yet published by the data gatherer should not be placed in the open domain without provisos about time frames to enable the data gatherers to publish their data. Where such data are utilised in a third-party dissertation, the

dissertation should be embargoed until such time as the data generators have completed their own dissertations and published their data.

Chapter Four

Investigation of Biological Sciences Supplementary Information files, OA publishing and research funding streams.

4.1 Introduction

This chapter reports on an additional investigation of the number of papers published with SI files, the level of OA publishing and the level of public funding supporting Biological Sciences Department research. This links to the first question into the investigation of research data management and archiving initiatives in Biological Sciences, and which was further explored in chapter one, point 1.2; relating to the fact that research data management may become mandatory, and that making underlying data open will be part of the equation. The additional investigation was undertaken to verify some of the responses to the survey, such as

- Question 3 'Is your research publicly funded?' (see Figure 3.3 for responses),
- Question 5 'Have you published supplementary data with your published research?' (see Figure 3.5 for responses), and
- Question 16 'Under what conditions would you/your research unit make data available for further research?' (see Figure 3.16 for responses)

The link between public funding and open data publishing is the trend which is unfolding in the international research arena, and SI data files are used to make these data accessible.

Supplementary or Supporting Information (SI) files which accompany published research have a number of purposes. These may be to provide additional information which has to be excluded because the journal specifies a maximum paper length, or data files provided to enable reviewers to evaluate the research in the article. SI files can contain a variety of information types – tables, figures, data, images, extended bibliographies, videos, protocols. The Public Library of Science (PloS) provides guidelines to authors for Supporting Information and hosts author's SI files on their server (PLoS, 2015).

Science articles are typically short, with extremely abbreviated references. An article by Baldwin et al. (2014) which was published in the journal *Science* for example, has a number of supplementary materials files, including references 26-44 as only

references 1-25 could be accommodated in the published article. The sections on materials and methods were not included in the article, nor were figures 1-6, tables 1-4 and movies 1 and 2. Of these SI files materials and methods and the two movies can be considered data files. Like *PLoS*, *Science* provides information elucidating the policy about contributing underlying data. The instructions are very specific: "All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*. All computer codes involved in the creation or analysis of data must also be available to any reader of *Science*. After publication, all reasonable requests for data and materials must be fulfilled." (American Association for the Advancement of Science (AAAS), 2015). Furthermore

> "*Science* supports the efforts of databases that aggregate published data for the use of the scientific community. Therefore, appropriate data sets (including microarray data, protein or DNA sequences, atomic coordinates or electron microscopy maps for macromolecular structures, and climate data) must be deposited in an approved database, and an accession number or a specific access address must be included in the published paper. We encourage compliance with MIBBI guidelines (Minimum Information for Biological and Biomedical Investigations)." (AAAS, 2015).

Wiley-Blackwell SI files carry a disclaimer about content and functionality and are only available in the online version of the journal (Ackerman & Bishop, 2010). Ackerman & Bishop's (2010) article was published prior to the 2011 mandate for the provision of underlying data reported by Witlock et al. (2010:146) and demonstrates that the trend to supply SI has not yet settled down. This may be why earlier SI files have not necessarily been a reliable way to archive data for the long-term, as reported by Anderson et al. (2006). Each Scientific journal publisher provides guidelines for authors and these are increasingly providing guidelines and standards for the provision of SI files.

The publication by Anderson et al. (2006) motivated the current investigation into the fate of the Biological Sciences Department's researchers SI files. Anderson et al.'s investigation looked at the persistence of information supplementary to published

research over the period 1998-2005 in the field of biomedical sciences and found that only between 71-92% were still accessible. The types of information relegated to supplementary files are recorded as being "raw data, experimental design specifications, specific software, statistical models and experimental protocols." (Anderson et al., 2006:1)

An investigation similar to that of Anderson et al. was conducted by Vines et al. (2014) where the investigation found that the age of the article was directly linked to the availability of underlying research data. In this case the research was not investigating SI files, instead authors of papers were contacted to find out if they still had copies of data underlying their published research. The findings of this investigation suggested that archiving underlying data with the publication as SI or in an approved repository would improve the life span of research data. Vines et al. (2014) attempted to contact the authors of 516 articles published between 1991-2011, to find out if the underlying data of these articles were still available and 23% confirmed that they still held the underlying data. This latter study excluded articles that had supplementary data archived with the published article as this was no longer the responsibility of the researcher. This investigation sought instead to evaluate whether researchers could be relied upon to archive their own data and whether their interventions had been successful.

4.1.2 Linking SI files to publicly funded research

In chapter three it was shown that 42% of 2014 Biological Sciences researchers responded that they had published SI files with their published research (see Figure 3.5 for responses) . It was found that 75% of researchers in Biological Sciences reported that they had between 25-100% public funding (see Figure 3.3 for responses), and that 29% reported that they provided their data either open access or through a repository (43 0f 149 respondents, see Figure 3.15 for responses). When asked whether they would be prepared to publish their data files open access, 30% replied in the affirmative (44 of 149, see Figure 3.16). The additional investigation was undertaken to verify these responses.

4.2 Methods

A preliminary investigation, which attempted to emulate the example of Anderson et al. (2006) revealed that there were no SI files accommpanying the 442 articles published by Biological Sciences Department researchers between 1998-2000, although 11 article had published data within the article. There were no OA articles among these 442 publications, although four articles had been published in journals which were freely available online.

The lack of OA articles among the 442 Biological Sciences articles is understandable as the OA movement only started in 2002. The lack of SI files among the Biological Sciences publications was not that easy to resolve and may be explained by the difference between biomedical sciences journals and biological sciences journals, where it appears that SI files were only introduced around 2007. This was established by sampling a range of journals in which Biological Sciences researchers had published between 2000 and 2007. It was decided to restrict the investigation to articles published in 2007, 2010 and 2014, as it was observed that SI files become more common in biological sciences journals from 2007. The 2014 publication set was included because this would be used to compare the responses from 2014 Biological Sciences researchers who responded to the survey, and 2010 was chosen as an intermediate set to measure changes between 2007 and 2014.

In order to investigate how robust the archiving of past SI files has been, what the level of public funding has been and whether research has been published OA, a collation of all the scientific papers published by Biological Sciences for the years 2007, 2010 and 2014 was undertaken.

The collations were undertaken using Web of Science (WoS), limiting to the University of Cape Town affiliation, and limiting in turn to the departmental addresses of what is now the Biological Sciences Department. From 2007 to 2014, and the earlier years 1998-2000, this required investigating the following addresses:

- Avian/Animal Demography Unit
- Bolus Herbarium
- Botany Department

- FitzPatrick Institute
- Freshwater Research Unit
- Marine Biology Research Institute
- Plant Conservation Unit
- Small Mammal Research Unit
- Zoology Department

To investigate publication output for 2014 all the above addresses were used with the addition of Department of Biological Sciences, as the lead-time for articles can be as much as two years.

The investigation was done in annual sets, as this was more manageable. The WoS link to SFX via Full Text Options enabled the link to the journal website where evidence of supplementary information could be found as this information is not reported in the WoS database. Evidence of OA publishing and evidence of funding streams are available on the WoS database, although the latter only from 2008 onward. The 2007 set in most cases required looking at the acknowledgements on each paper, although some publishers provided a link to acknowledgements which linked direct to that part of the article. The results of the investigation are presented in Figure 4.1.

4.3 Results

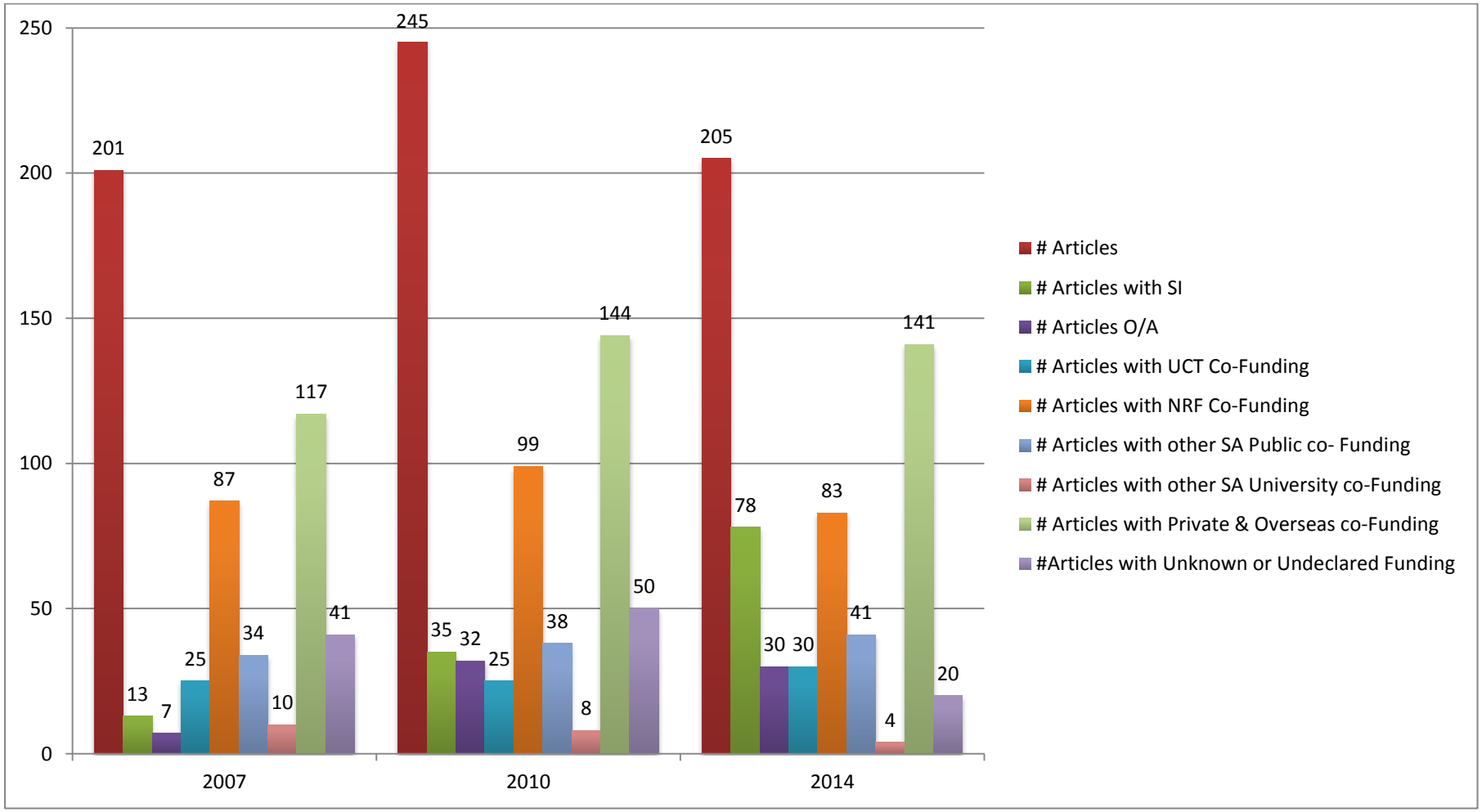

Figure 4.1 Publishing trends in the Biological Sciences Department for the years 2007, 2010 and 2014 by number of articles.

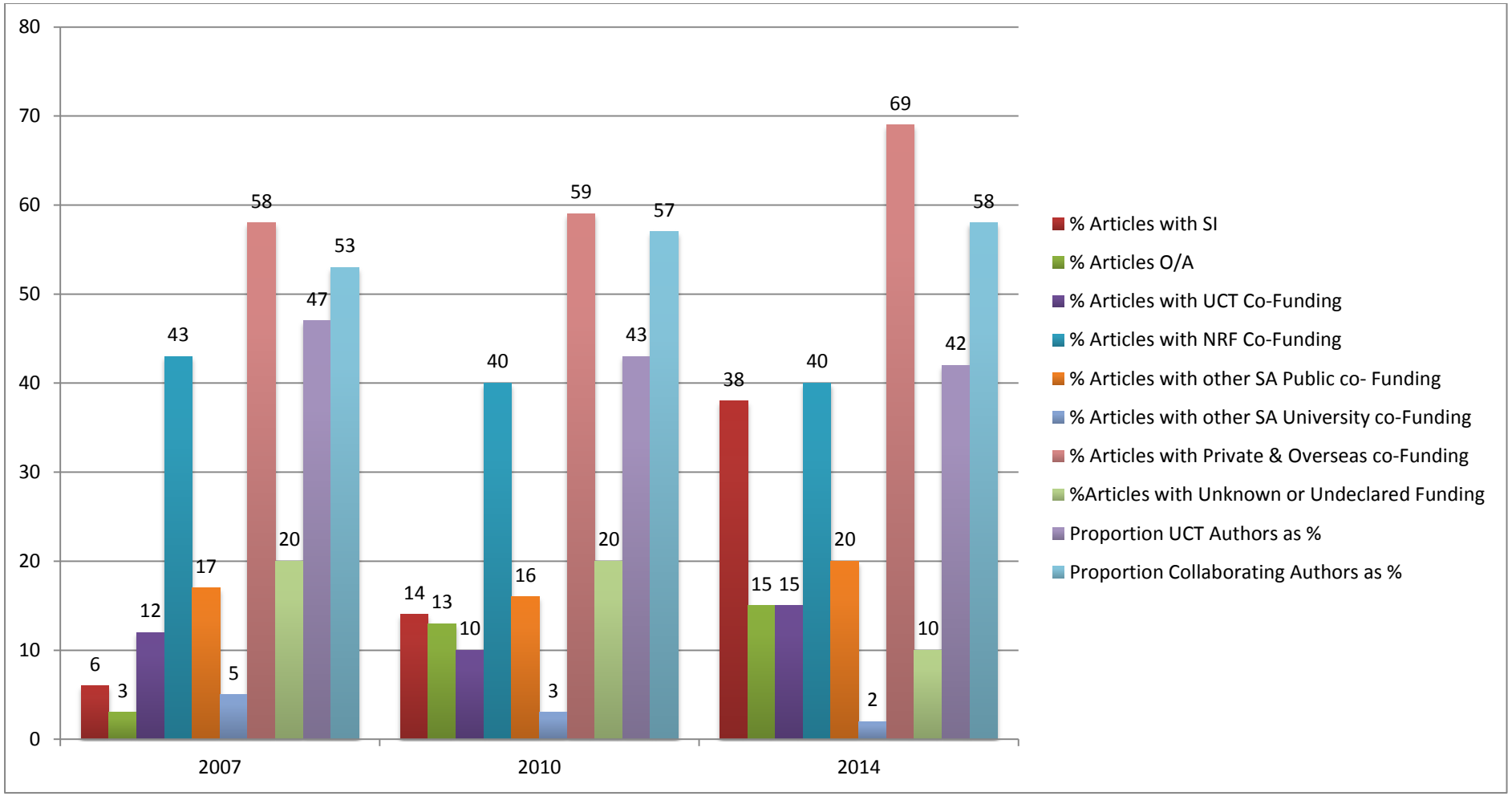

Figure 4.2 Publishing trends in the Biological Sciences Department for the years 2007, 2010 and 2014 by percentage.

The findings presented in Figures 4.1 and 4.2 are discussed in 4.4 below.

4.4 Discussion

A number of interesting trends emerged from this investigation. The fluctuation in the number of articles between 2007 and 2014 is probably insignificant, although the closure of the FRU in 2012 must have contributed to the drop between 2010 and 2014. What is of interest is the increase in percentage of articles with SI files, which climbed from 6% in 2007 to 14% in 2010 to 38% in 2014. This figure corresponds well with the survey responses to publishing with SI files reported in chapter three which yielded a figure of 42% of researchers submitting SI files. The fate of the SI files was good, as each 2007 and 2010 SI file URL was tested, but because the SI files are a fairly recent phenomena in biological sciences journals, this is to be expected, as insufficient time has elapsed to result in significant problems with software and hardware incompatibility. Only one set of SI files from 2007 was found to be missing when the URL was tested. This was reported to the publisher who admitted to be working on a more stable solution for SI files and once the missing SI files had been made available, the publisher reported the reinstatement of the SI via an email.

The number of articles published as OA remained fairly static between 2010 and 2014, but did increase between 2007 and 2010. That articles are being published OA is of interest in relationship to public funding, as the NRF has recommended OA publication with underlying data from 1st March 2015 for South African publicly funded research (NRF, 2015). The survey reported that 75% of Biological Sciences research was supported by between 25% to 100% public funding. That should result in a corresponding percentage of OA articles in the 2014 data set, but was not the case as only 15% of articles were published OA in 2014. There are a number of reasons contributing to this anomaly, primarily that South African funders have not yet made OA publishing mandatory. The other reason is that publishing OA is expensive for South African researchers because of the exchange rate between the ZAR and stronger currencies such as the USD, the Euro and the GBP which are the currencies of the major biological sciences journal publishers. Authors may not want to make their data OA, but they are very keen to have their published research OA as this improves the visibility of their research and consequently their research impact.

The funding streams are also interesting, with 75% of respondents to the survey reporting that they were supported by South African public funding in 2014, while the figures which emerge from the investigation of the three data sets reveal 60% South African public co-funding in 2014, 56% in 2010 and 60% in 2007. One has to take into account that the 75% of respondents who claim that they are currently supported by public funding, have not all published their research which could account for the differences in the reported and the actual 2014 figures. UCT research co-funding acknowledged in published articles for the three years sampled amounts to 15% (2014), 10% (2010) and 12% (2007).

The figures which emerged from the data set with respect to UCT authors vs UCT collaborators, (researchers located at other national and international universities and research institutes), probably account for the high levels of external co-funding. The proportions of UCT authors to UCT collaborators in 2014 was 42:58, in 2010 the ratio was 43:57 and in 2007 the ratio was 47:53. These figures were calculated using the publication output method of the Department of Education (DoE) where each article represented one unit and the UCT authors and UCT collaborators were calculated as a fraction of one. The total figure for each year was rendered as a percentage.

The percentages for overseas and private funding accounted for 69% (2014), 59% (2010) and 58% (2007) of the research conducted by UCT authors and their collaborators. Percentages of funding from other South African universities were low at 2% (2014), 3% (2010) and 5% (2007) respectively. This would be funding brought to the research by collaborators from other South African universities, but the analysis did not attempt to drill down any further on the nationality of collaborators.

A figure has been calculated for the number of articles without funding information. These fell into two categories, articles which could not be checked because the journal was not available electronically through the UCT Libraries ejournal portal, and articles where researchers did not acknowledge their funders. As was mentioned in the methods section, WoS introduced a section on funders in their database from 2008 onwards, which is extracted from the journal articles and where possible grant numbers were included. A paper by Sirtes (2013) an employee of the German

Institute for Research Information and Quality Assurance, was highly critical of the WoS results for the German Research Foundation, a large and diverse funder of global research. Sirtes research shows that acknowledgement of funding streams in journal articles have increased from 47% (2009) to 54% (2010) and 57% (2011). The figures from the three data sets analysed for this study for which there was no funding information are 10% (2014), 20% (2010) and 20% (2007) which is much lower than the Sirtes figures, although his study analysed approximately 1.2 million journal articles. His 2011 funding table which breaks down the articles into subject categories, gives a figure of 80% for Ecology articles in 2011. This figure correlates well with the percentage given above for the 2010 articles analysed from WoS. It appears that this information has been provided by databases such as Scopus and WoS for the purposes of "research evaluation and funding policy evaluation" (Rigby, 2011:366).

The use of published research output was used in this chapter to both verify the information given by respondents to the survey and to tease out information about levels of public funding versus levels of open access publishing of both research articles and accompany SI files. It can be seen that Biological Sciences Department researchers do not publish many OA articles in relation to their public funding streams, but it is expected that this will change in the near future.

4.5 Conclusion

Although the research questions were framed around the fate of research data underlying published research, and the more general management of research data in the Biological Sciences Department, the rationale for the investigation was the global trend to make publicly funded research data open and to investigate how this would affect UCT researchers.

It was seen in the literature review in 2.3 that a number of biological sciences journals have mandated the archiving of underlying data and the investigation into the research output for 2007, 2010 and 2014 demonstrated that Biological Science researchers are already publishing in many of these journals and supplying underlying research data either in the Dryad Repository or as SI files.

Chapter Five

Conclusion and recommendations for institutional level support

5.1 Review of research questions

This study was conducted in order to understand how Biological Sciences research data were managed in the past and in the present and to answer the research questions presented in chapter one. The only way to elicit this information was to approach the researchers, and to pose questions either in face-to-face interviews or by a self-administered survey and then to examine the responses. In order to verify the responses to the survey, an additional study was undertaken to find out how much Departmental research was openly available. These questions had to tease out aspects of the research life cycle (Digital Curation Centre, 2004-2015f), such as the creation, preservation, storage and re-use of research data in the Biological Sciences Department at UCT.

The first research question sought to find out how research data were managed, archived and shared. To get appropriate responses the research question had to be broken down into multiple questions, to find out how much research data is held by researchers, how they store the data, how many back-ups they have, how they share their data and whether any of the data had been archived in repositories.

The second question on the surface appeared less complex as it dealt only with the description of data, but as data repositories are not yet that familiar to all researchers, this too had to be broken down into multiple components. To complicate the discussion of metadata are the topics of standards and metadata languages appropriate to each discipline. The link between metadata and sharing had to be made so that researchers could see the importance of comprehensive data descriptions.

The third question about public funding of research tied in to the rationale for the investigation, this was the recommendations from funders to make published research and data underpinning the publications openly available. At the start of the research on the project, this had already occurred in the US,

Europe and the UK and is discussed in chapters one and four. The major South African public funder, the National Research Foundation made their recommendation in this connection in a document released to their staff on 19 January 2015 and subsequently posted on their web page for more general information (NRF, 2015). Questions were posed to researchers in the Department to enquire how much research was dependent on public funding and the responses demonstrated that this figure was 75% of researchers were dependent on a level of public funding. The survey questions linked funding to data archiving and OA publication, but it was found in the responses to these imperatives that implementation remained low. In order to verify the responses to the survey a desktop study was undertaken to find out how much research was published OA by Biological Sciences researchers and look at the provision of research data (such as code, methods, tables, image and videos) that were provided as SI files.

The fourth research question dealt with institutional support, and in the survey researchers were asked who owned their data, who should be responsible for their data and where their data should be stored. Few researchers felt that the institution had a role to play in managing research data, but expressed a willingness to attend workshops to discuss RDM and metadata generation. International best practice was consulted in the literature review in order to make suggestions for appropriate institutional support at UCT.

5.2 Research Data Management preparedness at UCT

As this investigation has shown, the University of Cape Town is in the early stages of grappling with the complexities of Research Data Management in its mission to support the multiple disciplines which contribute to UCT research. Since the establishment in 2014 of an eResearch Centre at UCT (University of Cape Town, eResearch Centre, 2015a) a number of activities have been initiated to support research data generators at UCT. Various workshops and conferences have been hosted by UCT ICTS and UCT Libraries and a webpage has been established which offers lists of resources for researchers to consider. Collaborators at UCT in this initiative are UCT Libraries, UCT ICTS

and the UCT Research Office. This group is in the process of "developing new policies for research data management, improving support for preservation and dissemination of research outcomes, and collaborating with the eResearch Centre" (University of Cape Town, eResearch Centre, 2015b). This study of data management and archiving initiatives in the Biological Sciences Department was synchronous to the development of the envisaged eResearch centre but were independent of each other.

Data archiving at UCT Libraries Manuscripts and Archives remains limited to an archival service for "the political, social, cultural and economic history of Southern Africa" (UCT Libraries, LibGuides, 2015); a digitization service targeting theses and special collections; and data curation information provided through links to the UK, DCC (University of Cape Town, eResearch Centre, 2015c).

The UCT Libraries have recently launched a Research Data Management presence on their web page with a link to the proposed UCT Research Data Management Plan (University of Cape Town, Libraries, 2015), and during 2014 launched the Savvy Researcher Series which provided support for aspects of data management for postgraduate students. A scan of the UCT Libraries LibGuides – the virtual guides used by libarians to share information with their subject communities – did not reveal a libguide on Research Data Management and the only libguide with a section on this topic is the Libguide for ornithology (UCT LibGuides, Ornithology, 2015).

5.3 What are the requirements for providing RDM support?

Findings

It has been shown in the preceding chapters that, even within the Biological Sciences Department, research is varied and field data collection requires a range of specialist skills, equipment and tools. The same applies to the synthesis of those data in order to produce research outcomes in the form of theses and published articles. There have been no systematic interventions for supporting researchers with data management or data storage facilities, and an ad hoc situation with varying success in the preservation of research data has been the norm.

Suggested interventions

For UCT libraries to give appropriate support to researchers, librarians with specialised backgrounds or experience to interface with researchers would be required. Much of the advice needed at postgraduate level is however generic, such as file naming conventions, data back-up habits, keeping records of the how, when, where and why data were gathered (metadata), and types of metadata protocols required for archiving specific data types. UCT Libraries have a role to play in providing such support and during 2014 the Savvy Researcher series hosted a range of workshops on some of these topics. The UCT libraries also have a role to play in directing researchers to other divisions on campus where information can be found, such as research funding, ethics support, IP support and temporary data storage. Some of these links have already been put in place in the lists of resources for researcher on the eResearch Centre web presence.

5.4 Past pre-digital and early digital research data

Findings

At present there is no strategy in place for the management or archiving of pre-digital or early digital research data and these data in Biological Sciences are still in the hands of the retired and emeritus staff who were interviewed for this study. Some of the data have been lost or discarded because of the lack of appropriate data archiving interventions. Some physical data sets are already archived in Biological Sciences, e.g. the historical Nest Record Card data of southern African breeding birds in the Niven Library, and the plant specimens deposited as herbarium vouchers and archived at the Bolus Herbarium.

Where early digital data are concerned, all of the retired and emeritus researchers who participated in this study had boxes of 8 inch and 3.5 inch disks containing data, much of them past student data which cannot be read with current computers, operating systems or software. Only one researcher among the retired and emeritus respondents to the survey had consistently migrated her data to contemporary platforms. Some of her data are archived on the SAEON data portal, but she does not have a data preservation plan for the remainder of her data.

Figure 5.1 Field notebooks

Figure 5.2 Field notes on cards

Suggested interventions

To make pre-digital data discoverable, the data will have to be digitised and interpreted by their generators, most of whom can still be contacted, so that adequate metadata can be provided. It is suggested that an inventory of these data should be made, and funding sourced for digitisation of the data. A secure repository for the physical data should also be established. An archive or a museum would be the traditional place to store such material.

Most of the early digital data sets will require a specialist to open the files and migrate the data to a current format in order to be useable. This requires now obsolete computers with the appropriate drives, the availability of operating systems which can open the directories and the availability of the software that was used to create the data. Such services exist but are commercial

initiatives and would require appropriate funding to take advantage of their availability.

5.5 Current digital research data management

Findings

As was seen in the findings of the survey presented in chapter three, 58% of Biological Sciences researchers were archiving some of their published research data in local, national and international repositories (Figure 3.9). Interim data management is undertaken on either a personal level or a research unit level (Figure 3.18) with the majority of back-ups made on hard drives which are stored on-site in offices and off-site at home (Figure 3.20). But these routines are not systematic and do not comply with any RDM plan, as there has not been a plan in place at UCT.

There are numerous long-term data sets in the possession of researchers, and some of the respondents expressed anxiety about not having a more secure place to archive these data or a long-term data preservation plan.

Suggested interventions

Systematic research data management and archiving at UCT will only come about when policies have been established in consultation with researchers. Research data management education of the new cohort of researchers is a pre-requisite for establishing systematic data archiving and initiatives in this regard should be introduced at fourth year or honours level. Because RDM is a very new concept in South Africa, support should also be offered to senior and medium level academic researchers so that they are sufficiently informed to ensure that student data are properly managed and archived.

Ensuring that long-term data sets are preserved is urgent and important as ecological data cannot necessarily be re-collected and it would be expensive and time consuming to do so if this was possible.

5.6 Sharing digital research data

Findings

The literature review in chapter two and the survey presented in chapter three demonstrated that sharing data was the most contentious aspect of research. It was shown that researchers do share their data (Figures 3.14 and 3.15), and that 88% believe that their data should be made available for future research (Figure 3.14), but that 82% of researchers regard ownership as being the preserve of the researcher or research unit (Figure 3.10). This was corroborated by the responses to question 16 of the survey that enquired about conditions for data sharing. Respondents indicated a number of pre-conditions before they would share data and most of these responses indicated that they wished to retain a level of control of their data (Figure 3.16).

Suggested interventions

Evidence from the literature review suggested that research funders would be the most likely implementers of research data sharing through mandating long-term data preservation (Doorn & Tjalsma, 2007:9). The very recent mandate from the NRF (2015) demonstrates that this will become the case in South Africa where publicly funded research is concerned.

There are however policies which need to be in place at UCT to ensure that researchers who generate data are protected from data misappropriation, and example of which was discussed in chapter 3, and that there is sufficient support for systematic RDM and appropriate interim repositories to archive data until data have been published and can be openly archived in a suitable discipline specific repository.

5.7 Understanding metadata or providing data descriptions

Findings

Question 22 of the survey asked about types of metadata and provided the respondents with informative answers to choose from, but despite this 15% of respondents answered that they don't assign metadata or that they did not

know what to do. As was discussed in chapter two, metadata is essential as without detailed descriptions data have no value. It was found that discipline-specific metadata standards have been developed and that repositories have tools and instructions which make it easier for researchers to supply all the detailed data descriptions required by the repository to make the data re-useable (Gil, Sheldon, Schmidt, et al., 2008:152). The response to the potential offer of workshops to discuss metadata generation received a positive response from 50% of respondents and suggestions are made below for implementation.

Suggested interventions

Findings from this study therefore suggest that UCT Librarians should become familiar with the variety of metadata standards appropriate to their field of research so that guides and teaching tools can be developed to support researchers. These will also enable junior postgraduate students to develop metadata for their research data, in order that these can be archived. There are many examples available at other international libraries, and the DCC has developed a Disciplinary Metadata web page in order to provide support for libraries and researchers (Digital Curation Centre, 2004-2015c). It is also suggested that metadata generation could be a regular feature of the UCT Libraries Savvy Researcher series for postgraduate students. The eResearch Centre could provide guidance for interested senior researchers.

5.8 What sort of institutional support for research data management should be provided at UCT?

A temporary repository for data or data staging repository was the solution provided on the Cornell University campus and discussed in chapter two, point 2.2 above. It was found in the literature review that the international data repository environment has become well established over the past decade, making it unnecessary to duplicate what is already available to Biological Sciences researchers. What the researchers are missing is a local repository where their data is secure and which would enables them to share their data with collaborators prior to publication. Discussion with postdoctoral students revealed that they were reluctant to use a cloud solution

as they thought that this was insufficiently secure, although 39% of respondents indicated that they were using cloud storage as a back-up location (Figure 3.20).

It was found that UCT is interested in retaining research data which have commercial value and information about the retention of IP for this category of data can be found in point six of the UCT IP Policy (2011). Much of the data generated by Biological Science researchers do not fall into this category as they do not produce patents as is done in the Molecular Biology Department, so that the value in archiving the data lies in preventing duplication of research and making data available for research verification.

Suggested interventions

The eResearch Centre discussed in 5.2 appears to be the intervention UCT collaborators have decided is the most appropriate level of support for RDM. This is a very new development and does not currently have much substance other than a declared intention. RDM should be routine in research units and departments and have the support of Departmental Heads and Deans, as this will ensure that there is systematic implementation at every level of the research process.

5.9 Who will be responsible for archiving research data?

Findings

The findings of the literature review and the survey indicate that researchers archive their own data (Doorn & Tjalsma, 2007) and this is certainly the case with regard to data archived by researchers in Biological Sciences. Unless UCT intends to establish research data archivist posts, which seems unlikely, this will put the onus on the proposed eResearch Centre to ensure that researchers are sufficiently skilled and supported to enable them to archive their own data. Fortunately local, national and international data repositories have guides and tools which enable researchers to do the archiving themselves. Data repositories do not unfortunately archive non-digital data which was discussed in 5.4, and these data sets require different interventions suggested below.

Suggested interventions

Providing a home for the non-digital resources and digitising the same in order to make the data visible will be an expensive undertaking, but one with which UCT should be engaged. Although the data do not have commercial value per se, recollecting the same data is expensive, is a waste of research funding and loss of temporal data sets impoverishes our knowledge of our environment.

5.10 Open data and research funding

Findings

OA publishing and open data underlying published research has been recommended by the NRF from 1st March 2015 for all publications "generated from research either fully or partially funded by NRF" (NRF, 2015). This will have a considerable impact on UCT research funding and UCT Libraries during 2015. UCT Libraries have already taken on the role of funding the page charges for OA publishing, and additional funding required to fulfill the NRF recommendation will have to be allocated.

Suggested interventions

Many other academic institutions (e.g. University of Pretoria, Stellenbosch University, Nelson Mandela Metropolitan University; from information communicated to the author by research associates of the PFIAO) allocate a proportion of the DOE funding from research output to researchers. It is not easy to find out what this amount is, as this seems to differ from university to university. UCT has firmly resisted following this trend, but allocating a percentage of DOE funding from research output may be a way of awarding researchers additional funds to cover the cost of OA publication of their research. How this could be done would have to be discussed collaboratively with research administrators, UCT libraries, Deans, Heads of research generating departments and researchers.

5.11 Conclusions

Although open data and research data management initiatives have been gaining momentum in the international arena for the past seven to ten years, South African universities have been slow to respond. Even UCT with the status of top university in Africa has not had the capacity or initiative to engage with these moves. UCT has been particularly slow in implementing an institutional archive, but the OpenUCT repository has already demonstrated the benefits of making UCT material available in the open domain.

UCT will have to allocate funds to employ competent staff to support their open data ambitions, it is not sufficient to support proposals in principle but not with implementation. The NRF recomendation for open publicly funded research will increase the pace as was demonstrated by Borgman (2012) in his article entitled *The Conundrum of sharing research data* where he discussed the mandatory initiatives of *The Welcome Trust*, *The National Science Foundation*, *The Economic and Social Research Council* and other public funders in ensuring that published research data is archived and made open.

This investigation has therefore shown that there are numerous biological sciences data archiving initiatives for researchers to utilise to openly archive their data underlying published research that will ensure their compliance with funding and journal mandates. The investigation showed further that researchers require information about appropriate metadata standards and languages and training in metadata generation. It is however critical that UCT ensure that policies are in place to protect data generators and that researchers receive the necessary support for interim data management, such as safe data storage facilities while data are generated and analysed. Researchers should feel secure in the knowledge that by openly archiving their unique research data that these will be acknowledged in perpetuity through Digital Object Identifiers or other identifier schemes.

References

Ackermann, R. & Bishop, J.M. 2010. Morphological and molecular evidence reveals recent hybridization between Gorilla taxa. *Evolution* 62:271-290. Wiley-Blackwell SI disclaimer.

Akers, K.G. & Doty, J. 2013. Disciplinary differences in faculty research data management practices and perspectives. *The International Journal of Digital Curation* 8:5-26.

American Association for the Advancement of Science (AAAS). 2015. *Science general information for authors*. Available: http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml [2015, January 11]

Anderson, N.R., Tarczy-Hornoch, P. & Bumgarner, R.E. 2006. On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics* 7:260. Available: http://www.biomedcentral.com/1471-2105/7/260 [2014, August 8]

Animal Demography Unit (ADU). 2009. Available: http://adu.org.za/about.php [2014. September 16]

Animal Demography Unit (ADU). *Virtual Museum*, 2014. Available: http://vmus.adu.org.za/ [2014, December 16]

Argos-System. n.d. *Argos*. Available: http://www.argos-system.org/?nocache=0.5800009497907013 [2014, December 12]

Arms, C.R., Fleischhauer, C. & Murray, K. (Compilers) 2013. *Sustainability of digital formats: planning for Library of Congress collections*. Available: http://www.digitalpreservation.gov/formats/fdd/fdd000052.shtml [2014, December 7].

Ashton, P.J., Roux, D.J., Breen, C.M., Day, J.A., Mitchell, S.A., Seaman, M.T. & Silverbauer, M.J. 2012. *The Freshwater Science Landscape in South Africa, 1900-2010: overview of research topics, key individuals, institutional change and operating culture*. WRC Report No. TT 530/12. Gezina: Water Research Commission.

Babbie, E. & Mouton, J. 2001. *The practice of social research*. Cape Town: Oxford University Press.

Baldwin et al. 2014. Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. *Science* 345:929-933.

Ball, A. & Neilson, 2010. *Curation of research data in the discipline of engineering*. SCARP Case Study No. 7. Available: http://www.dcc.ac.uk/sites/default/files/documents/publications/case-studies/SCARP_B4812_EngCase_v1_2.pdf [2015, February 1]

Barnard, K.H. 1950. Descriptive catalogue of South African decapod crustacean (crabs and shrimps). Descriptive List of South African stomatopod crustacean (Mantis shrimps). *Annals of the South African Museum* 38: 1-837.

Berners-Lee, T. 1997. Metadata architecture. Available: http://www.w3.org/DesignIssues/Metadata [2015, February 11]

Biodiversity GIS (BGIS), 2014. Available: http://bgis.sanbi.org/vegmap/map.asp? [2014, December 16]

BirdLife International, Seabird Tracking Database, 2014. Available: http://www.seabirdtracking.org [2014, December 16]

BirdLife Seabird Tracking Database. About, 2014. Available: http://seabirdtracking.org/mapper/about.php [2014, December 16]

Blender.org. 2015. Blender 2.73a. Available: http://www.blender.org/ [2014, December 7]

Bolus Herbarium, 2011. Available:http://web.uct.ac.za/depts/bolus/col&db.html [2014, September 16]

Borgman, C.L. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63:1059-1078.

Brown, A.C. 2003. Centennial history of the Zoology Department, University of Cape Town, 1903-2003: a personal memoir. *Transactions of the Royal Society of South Africa* 58(1):11-34.

Budapest Open Access Initiative, 2002. Available: http://www.soros.org/openaccess/read.shtml [2015, January 10]

Canfield, M.R. (Editor) 2011. *Field notes on science & nature*. Cambridge, Massachusetts: Harvard University Press.

Computer History Museum. 2011. *Digital Dark Age: revolution preview*. Available: https://www.youtube.com/watch?v=PSlMzirvsFc [2015, February 1]

Conservation Biology, 2014. Available: http://onlinelibrary.wiley.com/doi/10.1111/cobi.2014.28.issue-1/issuetoc [2014, December 16]

Costello, M.L. 2009. Motivating online publication of data. *Bioscience* 59:418-427.

Crawford, R.J.M., Altwegg, R., Barham, B.J., Barham, P.J., Durant, J.M., Dyer, B.M., Geldenhuys, D., Makhado, A.B., Pichegru, L., Ryan, P.G., Underhill, L.G., Upfold, L., Visagie, J., Waller, L.J. & Whittington, P.A. 2011.

Collapse of South Africa's penguins in the early 21st century. *African Journal of Marine Science* 33:139-156.

Darwin, C. 1836. *Darwin's Beagle diary*. Available: http://darwin-online.org.uk/EditorialIntroductions/Chancellor_fieldNotebooks.html [2015, January 13]

Dasgupta, S. 2006. Arpanet. In *Encyclopedia of Virtual Communities and Technologies*. London: Idea Group Reference. 173.

Data Archiving and Networked Services (DANS), n.d. *About DANS*. Available: http://www.dans.knaw.nl/en/content/about-dans [2015, January 8]

Day, J.A. 2014. Personal comment. Interview 19 August 2014.

Diekmann, F. 2012. Data practice s of agricultural scientists: results from an exploratory study. *Journal of Agricultural & Food Information* 13:14-34.

Digital Curation Centre. 2004-2015a. History of the DCC. Available: http://www.dcc.ac.uk/about-us/history-dcc/history-dcc [2015, January 8]

Digital Curation Centre. 2004-2015b. SCARP. Available: http://www.dcc.ac.uk/projects/scarp [2015, January 8]

Digital Curation Centre. 2004-2015c. Disciplinary Metadata. Available: http://www.dcc.ac.uk/resources/metadata-standards [2015, January 8]

Digital Curation Centre. 2004-2015d. Data Management Plans. Available: http://www.dcc.ac.uk/resources/data-management-plans [2015, January 10]

Digital Curation Centre. 2004-2015e. Repository audit and assessment. Available: http://www.dcc.ac.uk/resources/repository-audit-and-assessment/repository-audit-and-assessment [2015, January 10]

Digital Curation Centre. 2004-2015f. DCC Curation Lifecycle Model. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model [2015, February 9]

Doorn, P. & Tjalsma, H. 2007. Introduction: archiving research data. *Archival Science* 7:1-20.

Doorn, P., Dillo, I. & Van Horik, R. 2013. Lies, damned lies and research data: can data sharing prevent data fraud? *International Journal of Digital Curation* 8:229-243.

Dryad Digital Repository, Dryad. 2014. Available: http://datadryad.org [2014, November 29]

Dryad Data Repository Wiki, 2014. Available: http://wiki.datadryad.org/Repository_History [2014, November 29]

Dryad Data Repository Wiki, 2013. Available: http://wiki.datadryad.org/Business_Plan_and_Sustainability [2014, November 29]

Dryad. Integrate your journal, 2013. Available: http://datadryad.org/pages/journalIntegration [2014, November 29]

Duke, C.S. & Porter, J.H. 2013. The ethics of data sharing and reuse in biology. *Bioscience* 63:483-489.

Ecological Society of America. [2014]. ESA Data Policy. Available: http://esapubs.org/esapubs/DataReg.htm [2015, February 12]

Edinburgh University. 2014. Research Data Management Policy, 2011. Available: http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy [2015, January 17]

Elliot, G. 2008. *Otago biodiversity data management project report. Part 1. Questionnaire Report*. University of Otago Library, New Zealand. Available:http://otago.ourarchive.ac.nz/handle/10523/198 [2014, July 17]

European Molecular Biology Laboratory (EMBL). 2009-2014a. *About Us*. Available: http://www.embl.de/aboutus/general_information/index.html [2015, January 7]

European Molecular Biology Laboratory (EMBL). 2009-2014a. *Services*. Available: http://www.embl.de/services/index.html [2015, January 7]

European Union. 2013. *Guidelines on Open Access to scientific publications and research data in Horizon 2020*. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf [2014. September 3]

Fairbairn, D.J. 2010. The Advent of mandatory data archiving. *Evolution* 65:1-2.

Fairley, E. & Higgins, S. 2009. *Curated databases in the life sciences: the Edinburgh Mouse Atlas Project*. SCARP Case Study No. 4. Available: http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP_EMAP_Final13Jul09A.pdf [2015, February 1]

Figshare, 2014. Available: http://figshare.com [2014, November 29]

Figshare Blog, 2014. Available: http://figshare.com/blog [2014, November 29]

Fry, J., Lockyer, S. and Oppenheim, C. 2008. *Identifying benefits arising from the duration and open sharing of research data produced by UK Higher Education and research institutes*. Available: http://repository.jisc.ac.uk/279/ [2015, February 2]

GBIF, 2014a. *What is GBIF?* Available: http://www.gbif.org/whatisgbif [2014, November 30]

GBIF. 2014b. *Publishing Data*. Available: http://www.gbif.org/publishingdata/summary [2014, November 30]

GenBank, 2014. Available: http://www.ncbi.nlm.nih.gov/genbank/ [2014, November 29]

Genbank, Overview, 2014. Available: http://www.ncbi.nlm.nih.gov/genbank/ [2014, November 29]

Gil, I.S., Sheldon, W., Schmidt, T., et al. 2008. Defining linkages between the GSC and NSF's LTER program: how the Ecological Metadata Language (EML) relates to GCDML and other outcomes. *OMICS: a Journal of Integrative Biology* 12:151-156.

GitHub, 2015. About. Available: https://github.com/about [2015, January 24]

Global Plants, 2000-2014. Available: http://about.jstor.org/content/global-plants [2014, December 12]

Grassle, J.F. 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing,, modeling and mapping marine biological data in a multidimensional geographic context. *Oceanography* 13:5-7.

Huang, X., Hawkins, B.A. & Qiao, G. 2013. Biodiversity data sharing: will peer-reviewed data papers work? *BioScience* 63:5-6.

International Geosphere-Biosphere Programme (IGBP). n.d. Anthropocene. Available: http://www.igbp.net/globalchange/anthropocene.4.1b8ae20512db692f2a680009238.html [2015, February 1]

Iorns, E. 2013. Research 2.0.3: the future of research communication. *Nature* 14 June 2013. Available: http://blogs.nature.com/soapboxscience/2013/06/14/research-2-0-3-the-future-of-research-communication/ [2015, February 1]

Iziko, Biodiversity Explorer, n.d. Available: http://www.biodiversityexplorer.org/people/barnard-kh/ [2014, September, 16]

Iziko, History of the South African Museum, n.d. Available: http://media1.mweb.co.za/iziko/sam/muse/hist/smith.html [2014, September 16]

Jetz, W. & Rubenstein, D.R. 2011. Environmental uncertainty and the global biogeography of cooperative breeding in birds. Current Biology 21:438. DOI:10.1016/j.cub.2011.02.025

Knowledge Network for Biocomplexity. n.d. Ecological Metadata Language. Available: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html [2015, January 7]

Koopman, M.M. 2013. Digital data archiving at the Percy FitzPatrick Institute of African Ornithology: an investigation and pilot study. Unpublished LIS6013S project.

Le Vaillant, F. 1799 – 1808. *Histoire naturelle des oiseaux d'Afrique*. Paris : J.J. Fuchs Delachaussée.

Le Vaillant, F. 1799 – 1808. *Histoire naturelle des oiseaux d'Afrique*, Volume 6, Plate 287. Paris : J.J. Fuchs Delachaussée. Available: http://www.biodiversitylibrary.org/item/129431#page/221/mode/1up [2015, January 31]

LTER, The Longterm Ecological Research Network, 2013. Available: http://www.lternet.edu [2015, Janary 9]

MacColl, J. & Jubb, M. 2011. *Supporting research environments, administration and libraries*. Dublin, Ohio: OCLC Research. Available : http://www.oclc.org/research/publications/library/2011/2011-10.pdf [2014, June 8]

Macdonald, I.A.W. & Crawford, R.J.M. Eds. 1988. *Long-term data series relating to southern Africa's renewable natural resources*. South African National Scientific Programmes Report No. 157. Pretoria: CSIR. Available: http://researchspace.csir.co.za/dspace/handle/10204/2282 [2014, August 8]

MARAM, 2015. *Research focus*. Available: http://www.mth.uct.ac.za/maram/ [2015, January 16]

Marx, V. 2012. My data are your data. *Nature Biotechnology* 30:509-511.

Max Planck Gesellschaft. 2003-2014. *Open Access. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. Available: http://openaccess.mpg.de/Berlin-Declaration [2014, September 16]

Molloy, J.C. 2011. The Open Knowledge Foundation: Open Data means better Science. PloS Biology 9(12):e1001195. Available: http://dx.plos.org/10.1371/journal.pbio.1001195 [2015, January 9]

Monash University. 2010. *Research Data Management Policy*. Available: http://policy.monash.edu.au/policy-bank/academic/research/research-data-management-policy.html [2015, February 1]

Movebank, 2014. Available: https://www.movebank.org [2014, December 12]

National Research Foundation (NRF). 2015. *Statement on Open Access to Research Publications from the National Research Foundation (NRF)-*

Funded Research. Available: http://www.nrf.ac.za/media-room/news/statement-open-access-research-publications-national-research-foundation-nrf-funded [2015, February 4]

Nature. 2014. *Scientific Data: data policies*. Available: http://www.nature.com/sdata/data-policies [2014, December 16]

Nature. 2014. Editorial: Share alike. *Nature* 507:140.

Nature Publishing Group. 2015a. Supplementary Information. Available: http://www.nature.com/nature/authors/submissions/final/suppinfo.html#check [2015, February 4]

Nature Publishing Group. 2015b. About *Scientific Data*. Available: http://www.nature.com/sdata/about [2015, January 20]

NCBI, Copyright and disclaimers, 2009. Available: http://www.ncbi.nlm.nih.gov/About/disclaimer.html [2014, November 29]

Neilson, C. 2009. *Digital curation approaches for architecture*. SCARP Case Study No. 6. Available: http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP_Architecture.pdf [2015, February 1]

North, M. 1980. *A vision of Eden: the life and work of Marianne North*. Exeter, Devon: Webb & Bower.

OpenUCT, 2014. Available: http://openuct.uct.ac.za/about-open-uct [2014, December 13]

OpenUCT, Farewell from the OpenUCT Initiative team, 2014. Available: http://openuct.uct.ac.za/blog/farewell-openuct-initiative-team?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+openuct_blog+%28OpenUCT+Blog%29 [2014, December 16]

ORI, Data management, 2014. Available: http://www.seaworld.org.za/content/page/data-management [2014, December 16]

PARSE Insight. 2010. *Permanent Access to the Records of Science in Europe*. Version 3.6. Available: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf [2015, January 17]

Patrick, M. & Wilson, J.A.J. 2013. *Getting data creators on board with the digital curation agenda: lessons learned in developing training for researchers*. Oxford: University of Oxford, DaMaRO Project. Available: http://93.63.166.138:8080/dspace/handle/2012/140 [2014, June 9]

Percy FitzPatrick Institute of African Ornithology (PFIAO). 2014. *Mission Statement*. Available: http://www.fitzpatrick.uct.ac.za/docs/mission.html [2014, December 31]

Peters, D. 2014. Interview held at UCT Libraries on 22 December 2014, 11:00.

Phillips, M. 1988. Data storage. In *Long-term data series relating to southern Africa's renewable natural resources*. I.A.W. Macdonald & R.J.M. Crawford, Eds. South African National Scientific Programmes Report No. 157. Pretoria: CSIR. 467-468.

Pichegru, L., Grémillet, D., Crawford, R.J.M. & Ryan, P.G. 2010. Marine no-take zone rapidly benefits endangered penguin. *Biology Letters* 6:498-501. Available: http://rsbl.royalsocietypublishing.org/content/roybiolett/6/4/498.full.pdf [2015, January 16]

Piwowar, H.A., Day, R.S., Fridsma, D.B. & Ionnidis, J. 2007. Sharing detailed research data is associated with increased citation rate. *PloS ONE* 2:e308.

Available: http://dx.plos.org/10.1371/journal.pone.0000308 [2014, December 14]

Pillay, P. 2014. Email correspondence about SADCO funding. [2014, December 12 09:17 AM]

Plant Conservation Unit. 2014. Available: http://www.pcu.uct.ac.za [2014, September 16]

PLoS. 2015. Supporting Information Guidelines. Available: http://www.plosone.org/static/supportingInformation [2015. January 11]

Porter, J.H. & Callahan, J.T. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. In *Environmental information management and analysis: ecosystem to global scales*. W.K. Michener, J.W. Brut & S.G. Stafford, Eds. London: Taylor & Francis. 193-202.

Quammen, D. 1996. *The song of the Dodo*. London: Hutchinson.

Quinton, J.C., Lewin Robinson, A.M & Sellicks P.W.M. 1973. *François Le Vaillant: traveller in South Africa and his collection of 165 water-colour paintings 1781-1784*. Cape Town: Library of Parliament.

Research Information Network (RIN). 2008. *To share or not to share: publication and quality assurance of research data outputs*. Available: http://eprints.soton.ac.uk/266742/1/Published_report_-_main_-_final.pdf [2014, December 14]

Rex, D. 1985. The Bolus Herbarium and Library, 1865-1985. *Jagger Journal* 5:53-63.

Rigby, J. 2011. Systematic grant and funding body acknowledgement data for publications: new dimensions and new controversies for research policy and evaluation. *Research Evaluation* 20:365-375.

Royal Society. 2012. *Science as an open enterprise*. London: The Royal Society. Science Policy Centre. Available: https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf [2014, July 15]

Ryan, P.G . 2013. Personal communication in discussion about implementing data archiving at the Percy FitzPatrick Insititute of African Ornithology. [2013, July]

Said, Y.H., Wegman, E.J., Sharabati, W.K. & Rigsby, J.T. 2007. Social networks of author-coauthor relationships. *Computational Statistics & Data Analysis* 52:2177-2184. DOI:10.1016/j.csda.2007.07.021

South African Bird Atlassing Project (SABAP). 2001. Available: http://adu.org.za/sabap1.php
[2014, December 10]

South African Bird Ringing Unit (SAFRING), n.d. Available: http://safring.adu.org.za/content.php?id=1
[2014, December 10]

South African Bird Ringing Unit (SAFRING). N.d. Martial Eagle, *Polemaetus bellicosus* Available: http://safring.adu.org.za/search_public.php?type=species&spp=142
[2015, Janary 17]

South African Environmental Observation Network (SAEON). 2009. *Background*. Available: http://www.saeon.ac.za/saeon-background [2014, December 02]

South African Environmental Observation Network (SAEON). 2009. *Developing information systems for Earth observation*. Available: http://www.saeon.ac.za/developing-information-systems-for-earth-observation [2014, December 02]

Southern African Data Centre for Oceanography (SADCO). 2010. Available: http://sadco.csir.co.za [2014, December 02]

Sayogo, D.S. & Pardo, T.A. 2013. Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly* 30:S19-S31.

Scaramozzino, J.M., Ramírez, M.L. & McGaughey, K.J. 2011. A study of faculty data curation behaviors and attitudes at a teaching-centered university. *College & Research Libraries* 75:349-365.

Science, General Information for Authors, 2014. Available: http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail [2014, September 26]

Scientific American. 2011. Editorial: Dr No Money: the broken science funding system. 19 April 2011. Available: http://www.scientificamerican.com/article/dr-no-money/ [2015, February 1]

Sirtes, D. 2013. *Funding acknowledgements for the German Research Foundation (DFG). The dirty data of the Web of Science database and how to clean it up*. Institut für Forschungsinformation und Qualitätssicherung. Available: http://www.forschungsinfo.de/publikationen/Download/SIRTES_ISSI_2013_DFG_FUND_ACK.pdf [2015, January 11]

Steinhart, G. 2007. DataStaR: an institutional approach to research data curation. *IASSIST Quarterly*, Fall & Winter:34-39.

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:e21101. Available: http://dx.plos.org/10.1371/journal.pone.0021101 [2014, December 14]

Thom, H.B. Ed. 1952-1958. *Journal of Jan Van Riebeeck*. Cape Town: Balkema for the Van Riebeeck Society.

Thomas, G. 2014. Email correspondence about Research Data Management. [2014, September 19 07:41 AM]

Thomson-Reuters. 2013. Journal Citation Reports.

Trustees of the Royal Botanical Gardens, Kew. n.d. Marianne North. Strelitzia and Sugar Birds, South Africa. Available: http://prints.kew.org/art/469812/365-strelitzia-and-sugar-birds-south-africa [2015, January 31]

Unidata. 2014. NetCDF 4.3.2. Available: https://www.unidata.ucar.edu/software/netcdf/docs/ [2014, December 7]

United Kingdom, National Information Infrastructure, 2014. Available: http://data.gov.uk/consultation/national-information-infrastructure-prototype-document/what-national-information#1.1 [2014, December 13]

United Kingdom, Open Data White Paper, 2012. Available: http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf [2014, December 13]

University of Amsterdam. 2013. *Bird Tracking System, UvA-BiTS*. Available: http://www.uva-bits.nl [2014, October 29]

University of Cape Town. 2010. *Authorship Practices Policy*. Available: http://www.uct.ac.za/about/policies/ [2014, October 13]

University of Cape Town. 2011. *Intellectual Property Policy*. Available: http://www.uct.ac.za/downloads/uct.ac.za/about/policies/intellect_property.pdf [2014, October 13]

University of Cape Town. 2014. *Open Access Policy*. Available: https://www.uct.ac.za/downloads/uct.ac.za/about/policies/UCTOpenAccessPolicy.pdf [2014, December 13]

University of Cape Town, 2014. *UCT Policy and Procedures for Breach of Research Ethics Codes and Allegations of Misconduct in Research*. Available: http://www.uct.ac.za/downloads/uct.ac.za/about/policies/Research_Misconduct_Policy.pdf [2014, October 13]

University of Cape Town, 2012. *UCT Policy for Responsible Conduct of Research*. Available: http://www.uct.ac.za/downloads/uct.ac.za/about/policies/UCTresearchconductpolicy.pdf [2014, October 13]

University of Cape Town, eResearch Centre. 2015a. *How it started*. Available: http://www.eresearch.uct.ac.za/how-it-started [2014, December 15]

University of Cape Town, eResearch Centre. 2015b. *Who we work with*. Available: http://www.eresearch.uct.ac.za/who-we-work [2014, December 15]

University of Cape Town, eResearch Centre. 2015c. *Preserving research outcomes*. Available: http://www.eresearch.uct.ac.za/who-we-work [2014, December 15]

University of Cape Town, LibGuides. 2015. *Ornithology*. Available: http://libguides.lib.uct.ac.za/content.php?pid=242411&sid=4478160 [2015, January 25]

University of Cape Town, LibGuides. 2015. *Manuscripts and Archives*. Available: http://libguides.lib.uct.ac.za/mss/ [2015, February 7]

University of Cape Town, Libraries. 2014. *Savvy Researcher Series*. Available: http://www.lib.uct.ac.za/lib/savvy-researcher-series [2015, February 7]

University of Cape Town, Libraries. 2015. *Research Data Management Plan*. Available: http://www.lib.uct.ac.za/uct-research-data-management-plan [2015, January 20]

Van Noorden, R. 2014a. Confusion over open-data rules. *Nature* 515:478.

Van Noorden, R. 2014b. Online collaboration: scientists and the social network. *Nature* 512;126-129.

Van Noorden, R. 2013. Data-sharing: everything on display. *Nature* 500:243-245.

Vines, T.H., Andrew, R.L., Bock, D.G., Franklin, M.T., Gilbert, K.J., Kane, N.C., Moor, J.-S. , Moyers, B.T. et al. 2013. Mandated data archiving greatly improves access to research data. *The FASEB Journal* 27: 1304-1308.

Vines, T.H., Albert, A.Y.K., Andrew, R.L., Débarre, F., Bock, D.G., Franklin, M.T., Gilbert, K.J., Moor, J.-S., et al. 2014. The availability of research data declines rapidly with article age. *Current Biology* 24:1-4. http://dx.doi.org/10.1016/j.cub.2013.11.014

Wallis J.C., Rolando, E., Borgman, C.L. & Nunes Amaral, L.A. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PloS ONE* 8:e67332. Available: http://dx.plos.org/10.1371/journal.pone.0067332 [2014, December 14]

Wellcome Trust. 2010. *Data management and sharing*. Available: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm [2015, February 8]

Weller, F., Cecchini, L.-A., Shannon, L., Sherley, R.B., Crawford, R.J.M., Altwegg, R., Scott, L., Steward, T. & Jarre, A. 2014. A system dynamics approach to modelling multiple drivers of the African penguin population on Robben Island, South Africa. *Ecological Modelling* 277:38-56.

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. 2010. Data archiving. *The American Naturalist* 175:145-146.

Wright, D. 2011. Evaluating a citizen science research programme : understanding the people who make it possible. MSc. University of Cape Town. Available: https://open.uct.ac.za/handle/11427/10904 [2015, February 12]

Zandvleitrust. 2000-2015. Photograph in the Posthuys, Muizenberg, of the Aquarium at St James. Available: http://www.zandvleitrust.org.za/art-ZIMP%20history%20tinus%20de%20iongh%20etching.html [2014, September 26]

Appendix A

QUESTIONS FOR INTERVIEWS of Emeritus or Retired Biological Sciences researchers

1. When did you publish your first scientific article, where?
2. How was your PhD research funded?
3. Is/was any of your research publically funded?
4. Do you have any physical research data e.g. notebooks or similar?
5. Where are these data sets?
6. What do you understand by data curation?
7. Have any of your data been lost?
8. Have your data been migrated to new technologies so that they are still available for use?
9. Do you have long-term data sets?
10. Do you re-use your data?
11. Have your data been archived physically or electronically anywhere?
12. Should your data be made accessible for future research?
13. On what terms would you make your data available to other researchers?
14. Have you published any papers with Supplementary Data files?
15. Who should be responsible for storage of data sets?
16. How do you manage your electronic data?
17. Do you back up your data? How often? How many backups do you have? Where do you backup your data?
18. Do your students conduct desktop studies using data?
19. What happens to your student's data?
20. Who owns the data you generate?
21. Do you require data management assistance?
22. Would you attend a workshop to discuss data management?
23. Would you attend a workshop to discuss metadata generation?
24. Has any of your funding required data curation?
25. Have any of the journals in which you have recently published required you to submit the underlying data?
26. Do you budget for data management and data curation?
27. Do you have a data preservation plan?

28. Do you share your data?
29. Does the department have a formal agreement in place with researchers for the storage of data?

Appendix B

QUESTIONS FOR INTERVIEWS of technical support Biological Sciences staff

1. What do you understand by data curation?
2. Does your department have a policy for data management and archiving?
3. Does your department budget for data management and data archiving?
4. Do you have a data preservation plan?
5. Does the department have a formal agreement in place with researchers for the storage of data?
6. Do you manage research data for your department?
7. Is archiving part of your job description?
8. Do you consider that research data curation should be mandatory?
9. Who should be responsible for storage of data?
10. Have any physical research data e.g. notebooks or similar been lodged with you?
11. Where are these data sets?
12. Have any of these data been lost?
13. Have you migrated data to new technologies so that they are still available for use?
14. Do you have long-term data sets?
15. Have you made data available for re-use?
16. Should data be made accessible for future research?
17. Have you archived any physical or electronic data with a repository? If so, where?
18. On what terms should data be made available to other researchers?
19. How do you manage electronic data?
20. Do you back up data in your care? How often? How many backups do you have? Where do you backup these data?
21. What happens to students' data?
22. Do you require data management assistance?
23. Would you attend a workshop to discuss data management?
24. What do you understand by metadata?
25. Would you attend a workshop to discuss metadata generation?

Appendix C

SURVEY QUESTIONS posed to Biological Sciences researchers

Covering Letter

I am conducting an investigation into data management and archiving expertise and initiatives in the Biological Sciences Department. Analysis of the responses will be used for my mini-dissertation towards the Masters in Library and Information Studies for which I am registered. Ethics clearance has been granted to undertake this study.

Please will you respond to the survey, which can be found at this URL
https://docs.google.com/forms/d/15yyp62VhNUc5SXvufKzRttlWjuUTtW-Ga12jnKGLakM/viewform?usp=send_form
The survey will take 5-10 minutes to complete and responses, which are anonymous, will be available to everyone who completes the questionnaire. I hope that you will find the survey interesting and informative. Please let me know when you have completed the survey.

Many thanks for your time and cooperation

Margaret Koopman
Niven Library Manager
Percy FitzPatrick Institute
University of Cape Town

Digital data management & archiving, Biological Sciences Department, University of Cape Town

1. **Which** research category describes you? Please make multiple selections if appropriate:

a. Academic
b. Postdoctoral
c. Postgraduate
d. Research Associate
e. Research Assistant
f. Technical Staff
g. Research support
h. Other

2. **What** is your highest academic qualification?

a. Doctorate
b. Masters
c. Honours
d. Undergraduate degree
e. Other

3. **Is your research** publically funded?

a. 100%
b. 75%
c. 50%
d. 25%
e. My research is not publically funded
f. Don't know

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

a. >200
b. 150-200
c. 50-100
d. <50
e. Not applicable

5. **Have you published supplementary data** with your published research?

a. Yes
b. No
c. Not applicable

6. **Why did you publish supplementary data?** Please make multiple selections if appropriate:

a. To publish in a journal where this is mandatory
b. To enable reviewers to evaluate my research findings
c. To improve my citation rating
d. To comply with the obligations of public funding
e. To comply with my research grant

f. To further the cause of global issues such as climate change and biodiversity loss

g. Not applicable

h. Other

7. **Do you** or your research unit have public funding?

a. Yes

b. No

c. Don't know

d. Not applicable

8. **Has any of your funding** or your research unit's funding required data curation?

a. Yes

b. No

c. Don't know

d. Not applicable

9. **Have your data** or your research unit's data been archived in any of the following repositories? Please make multiple selections if appropriate:

a. Dryad (Ecological data) http://datadryad.org/

b. Figshare http://figshare.com/

c. GenBank (Genetic sequence data) http://www.ncbi.nlm.nih.gov/genbank

d. GBIF/SABIF (Global Biodiversity Information Facility) http://www.gbif.org http://www.sabif.ac.za/

e. Obis/AfrObis (Ocean Biogeographic Information System) http://www.iobis.org/ http://afrobis.csir.co.za/

f. EMBL (European Molecular Biology Laboratory) http://www.embl.org/

g. KNB (Knowledge Network for Biocomplexity) http://knb.ecoinformatics.org/

h. SAEON (South African Environmental Observation Network) http://www.saeon.ac.za/data-portal-access

i. SADCO (South African Data Centre for Oceonography) http://sadco.csir.co.za/data.html

j. MoveBank (Animal Tracking Data) https://www.movebank.org/

k. JStor Global Plants (Plant Type Specimens) http://plants.jstor.org/
l. UCT Libraries Digital Collections
m. Not applicable
n. Other

10. **Who owns your data** or your research unit's data? Please make multiple selections if appropriate:
a. Researcher
b. Research unit
c. University of Cape Town
d. Funder
e. Supervisor
f. Don't know
g. Other

11. **What do you think is the purpose** of data curation. Please make multiple selections if appropriate:
a. Storage of data for access and use
b. Migration of data to new platforms/software
c. Ensuring that data are secure and backed up and available
d. Making sure data are available for future use
e. Ensuring that data are organized and indexed
f. Maintaining research data long-term so that it is available for reuse and preservation
g. Other

12. **Do you or your research unit** have long-term data sets? Please select multiple responses if appropriate:
a. >50 years
b. 50-25 years
c. 10-25 years
d. <10 years

13. **Do you**/your research unit re-use your data?

a. Always
b. Frequently
c. Occasionally
d. Rarely
e. Never
f. Don't know

14. **Should your**/your research unit's data be made available for future research?
a. Yes
b. No
c. Don't know
d. Not applicable

15. **How do you share** your research data or your research unit's data with other researchers? Please make multiple selections if appropriate:
a. By e-mail on request
b. Through the research unit's https server
c. I refer queries to a repository where the research data has been archived
d. Through a collaborative national/international initiative
e. My data and my research unit's data are in the public domain
f. My data and my research unit's data are in our published papers
g. Don't know
h. Not applicable
i. Other

16. **Under what conditions** would you/your research unit make data available for further research?
a. Open Access, with acknowledgement
b. Only if my data have a DOI (Digital Object Identifier)
c. Only if my data sets have Creative Commons licensing
d. On request so that I can discriminate
e. Only after I have published my data
f. Only if I am offered co-authorship

g. Only to a trusted researcher
h. I am not prepared to make my data available for future research
i. Don't know
j. Not applicable
k. Other

17. **Do you** or does your research unit conduct desktop studies using data?
a. Always
b. Frequently
c. Occasionally
d. Rarely
e. Never
f. Don't know
g. Not applicable

18. **Who should be responsible** for storage of data sets that are generated in this department? Please make multiple selections if appropriate:
a. Researcher/Supervisor
b. Research unit
c. Departmental IT personnel
d. University Library
e. University IT Department
f. National Repository
g. International Repository
h. Don't know
i. Other

19. **How often do you back-up** your electronic data
a. Daily
b. Weekly
c. Monthly
d. Every 6 months
e. Incrementally using appropriate software
f. Never

g. Other

20. **Where do you keep your data back-ups?** Please make multiple selections.

a. On my PC/Laptop
b. On a CD/DVD
c. On a flash drive
d. On hard-drives
e. On a server
f. On cloud storage
g. In my office
h. At home
i. UCT ICTS
j. Not applicable
k. Other

21. **How many data back-ups** do you have

a. 1
b. 2-3
c. >3
d. None
e. Other

22. **What types of metadata** do you consider important to describe your data? Please select multiple responses if appropriate:

a. Name of creator/Research unit's name
b. Contact details of creator/Research unit
c. Copyright provisions
d. Name of Funder
e. Contact details of funder
f. Title of the data set
g. Description of the data set
h. Geographic coordinates
i. Date of data creation

j. Beginning and end dates of project
k. Collection methods
l. Equipment used to gather data
m. Data format/s
n. Keywords
o. Taxonomic names
p. Title of umbrella project
q. Contact details of umbrella project
r. I don't assign metadata
s. Don't know
t. Other

23. **Would you attend** a workshop to discuss metadata generation?
a. Yes
b. No
c. Not applicable

24. **Approximately how much** research data do you have?
a. >10 terrabytes
b. 5-10 terrabytes
c. 1-5 terrabytes
d. 500-1000 gigabytes
e. 100-500 gigabytes
f. <100 gigabytes
g. Not applicable
h. Don't know

25. **What types** of digital data does your research generate? Please make multiple selections if appropriate:
a. Sequence data
b. Graphical data
c. Scalar and vector data
d. Image data
e. Spatial data

f. Modelling data
g. Statistical data
h. Survey data
i. Questionnaire data (qualitative or quantitative)
j. Fieldwork data
k. Experimental data
l. Audio data
m. Video data
n. Synthetic data
o. Raw logger data
p. Stable isotope data
q. Not applicable
r. Other

26. **In what formats** are these digital data sets?
a. Sequence format (e.g. EMBL, GenBank)
b. Text (.txt)
c. Spreadsheets (.xls or .csv)
d. GIS shape files
e. Database files (e.g. MySQL, MS Access)
f. Statistical software (e.g. R, SPSS)
g. Wordprocessor files (e.g. .doc, OpenOffice)
h. Image files (e.g. .jpg)
i. Webcam log files
j. GPS logger files (e.g. .gpx)
k. Audio file format (e.g. .mp3, .wav)
l. Not applicable
m. Other

27. **Have any data of your data been lost?** Please make multiple selections if appropriate:
a. My computer crashed
b. My field notebook was never returned

c. My files on old hardware became corrupt/inaccessible (e.g. floppy discs, zip drives)
d. UCT's Operating system was no longer compatible with file type/programme
e. My computer/Hard drive was stolen
f. I have never lost data
g. Not applicable
h. Other

28. **Do you migrate your data** to new software/operating systems when the current system becomes obsolete?
a. Yes
b. No
c. Not applicable

29. **Do you require** data management assistance?
a. I hire students to assist in data management
b. I would like more information about managing my data efficiently.
c. I do not require data management assistance

30. **Would you attend** a workshop to discuss data management?
a. Yes
b. No
c. I would prefer to visit an online resource

31. **Do you budget** for data management and data curation?
a. Yes
b. No
c. Not applicable

32. **Do you have** a data preservation plan?
a. Yes
b. No
c. Not applicable

Thank you for your participation!