DEPARTMENT OF COMPUTER & INFORMATION SCIENCES

University of Strathclyde Science

SURVEYING THE STATE OF DATA CURATION: A REVIEW OF POLICY AND PRACTICE IN UK HEIs

AMY PHAM

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Information and Library Studies

DEPT. OF COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF STRATHCLYDE

AUGUST 2018

DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself. Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

Yes [√] No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices) is 19,519.

I confirm that I wish this to be assessed as a Type 1 2 3 [✓ 4] 5
Dissertation.

Signature:

Date: 17/08/2018

ABSTRACT

As the value of research data has become increasingly recognized in the United Kingdom by public funding organizations, pressure has been placed on higher education institutions to provide access to research data or risk future funding. As a result, research data services have emerged rapidly over the past few years. However, it is not clear whether these services effectively ensure the long-term preservation of research data or apply appropriate data curation measures.

Through a three-part methodology, the research aimed to provide a clear picture of the current state of data curation in UK HEIs, including adherence to best practices and the existence of provisions for data curation efforts. A survey questionnaire was disseminated as the primary method of data collection, and additional information was gathered through a literature review and an analysis of online resources and institutional policies.

Data curation practices were found to be mostly inconsistent with best practices and were largely focused on facilitating access to research data. However, there was an awareness of the underdeveloped areas of data curation, especially preservation, and efforts are being made to improve these areas. Institutional policies were found to be mostly documents that defined roles and responsibilities and provided little guidance for follow-through. The role of researchers was repeatedly emphasized in both policy and practice and was essential in understanding the current state of data curation.

ACKNOWLEDGEMENTS

Many thanks are due to my supervisor, Dr. Diane Pennington, for her guidance throughout the dissertation process. I would also like to extend my gratitude to the ROADS (Repository, Open Access and Datasets) team at the University of Strathclyde, particularly George Macgregor and Alan Morrison, for their assistance and expertise.

Finally, thank you to Mike and Bobby, and to my family and the Balcony Book Club.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In recent years, there has been increasing recognition of the long-term benefits of sharing and preserving research data. Providing access to research data can increase the visibility of publications, expedite verification of final results, and facilitate reuse for future research (Akers, *et al.*, 2014; Buys and Shaw, 2015; Higman and Pinfield, 2015; Locher, 2016; MacMillan, 2014; Olendorf and Koch, 2012). As academic institutions begin to recognize the importance of properly maintaining research output that has been produced at their establishments, data curation has become a significant subject of interest. Looking beyond access, curation considers the whole of the research data lifecycle, from conceptualization to reuse and transformation (DCC, 2018a; Lavoie, 2012).

In support of these values, government mandates and grant funding policies now include the sharing, deposit, and management of data. Many funders' policies also necessitate the inclusion of an accompanying data management plan (DMP) (DCC, 2018b). In the United Kingdom, these policies are a requirement for almost all research publications. Spurred by these mandates, UK higher education institutions (HEIs) have sought to establish data management programs, with libraries acting as the primary mediators. However, as the needs and requirements for these programs are relatively new, institutions have struggled to implement standard policies and practices. As a rapidly emerging and evolving field, data curation currently lacks clear answers to several fundamental questions, including a clear distinction between "data curation" and "data management". The field also faces administrative and operational obstacles, such as a lack of resources and expertise.

Several publications and functional models have been released to address these issues, but the effective implementation of standards and recommended practices remains ambiguous. Solutions to these issues remain complex, with data curation being described variously as a "digital curate's egg" (Knight, 2012) and "wicked problem" (Cox, Pinfield and Smith, 2016).

1.1. Research questions and rationale

Currently, librarians lack both the technical and historical context of curation. A framework is necessary for understanding the holistic relationships involved in data management and to encourage discussion about the impact of poor practices in the current day and on the future.

The following research seeks to focus on and define “data curation” and to answer the following questions:

- What is the state of “best practice” in data curation?
- Do existing policies make provisions for standard practices?
- What are the connections between policy and practice?

While these questions are applicable globally, I will be concentrating specifically on finding answers to these questions within the context of the United Kingdom. With national requirements for data management and curation, the UK provides a unique case study for examining these issues. The resulting findings will hopefully contribute to a holistic understanding of the curation lifecycle, inform recommendations for future practice, and identify key underdeveloped areas in current research.

1.2. Overview

This dissertation is divided into five sections, including the present introductory chapter. The next section covers a literature review of data curation, including an overview of research data in the UK, an explanation of existing frameworks, and a brief examination of current practices and solutions in the field.

The third section describes the three-part methodology used to collect and analyze data. An evaluation framework was first established through a comprehensive two-cycle literature review. A content analysis was conducted in the first cycle to enable a coded thematic analysis in the second cycle. The framework produced from the literature review was then used to inform the design of the research. Primary research was conducted in the form of a survey questionnaire distributed within the HE sector in the UK. The survey findings were used to quantify current practices and examine textual input from respondents regarding the context of their answers. Thirdly, an analysis of online resources was conducted to supplement the survey and supply information that was conspicuously missing from the literature review. The online resource analysis included a count of services and a review of institutional policies that was then compared to the survey findings.

The fourth section of this dissertation presents an analysis of the raw data. Data curation practices were found to be mostly inconsistent with best practices and were largely focused on facilitating access to research data. However, there was an awareness of the underdeveloped areas of data curation, especially preservation, and efforts are being made to improve these areas. Institutional policies were found to be mostly documents that defined roles and responsibilities and provided little guidance for follow-through. The role of researchers was repeatedly emphasized in both policy and practice and was essential in understanding the current state of data curation.

The final section discusses the final results, determines key areas for development, and suggests recommendations for future research.

2. LITERATURE REVIEW

2.1. What is data curation? Definitions, frameworks, and stakeholders

2.1.1. Definition of data

While data seems to lie outside the wheelhouse of traditional librarianship, Rice and Southall (2016) argue that as an "archived resource", data is "becoming normalized as just another information resource" (p.2). Typically, data refers to a form of "scholarly output" that is a by-product of research (MacMillan, 2014, p.542). Data can be generated through both quantitative and qualitative techniques, although certain types of data, especially those produced qualitatively, may not be traditionally perceived as data (Rice and Southall, 2016, p.19). In a study conducted by Mohr *et al.* (2015), they discovered that the definition of data differed by discipline, and certain fields, particularly the Humanities, eschewed the term entirely. Differing perceptions of what constitutes "data" have resulted in inconsistent definitions of the term and, consequently, have affected the adoption of standard research data management (RDM) practices. In this case, RDM refers to the data practices of researchers and includes planning, storing, and recording metadata, activities which are mirrored in the processes of data curation (Akers et al., 2014; Buys and Shaw, 2015; Higman and Pinfield, 2015).

2.1.2. Definition of data curation

Data curation, often the responsibility of institutions or organizations rather than researchers, suffers the same inconsistency of terminology. At times referred to as "data management" and "data preservation" as well as "data curation", there is no clear agreement as to whether management, curation, and preservation are separate processes (Berman, 2008; Locher, 2016; Mohr et al., 2015; Pinnick, 2017; Shen and Varvel, 2013). Attempts to unite these processes under umbrella terminology have been made by organizations such as the Digital Curation Centre (DCC, 2018c) and CASRAI (2015), with "data curation" recognized as the generally accepted term to describe processes and activities related to the long-term management of data. Despite the call to adopt a standard term, consistency and clarification of related activities and processes remains an issue.

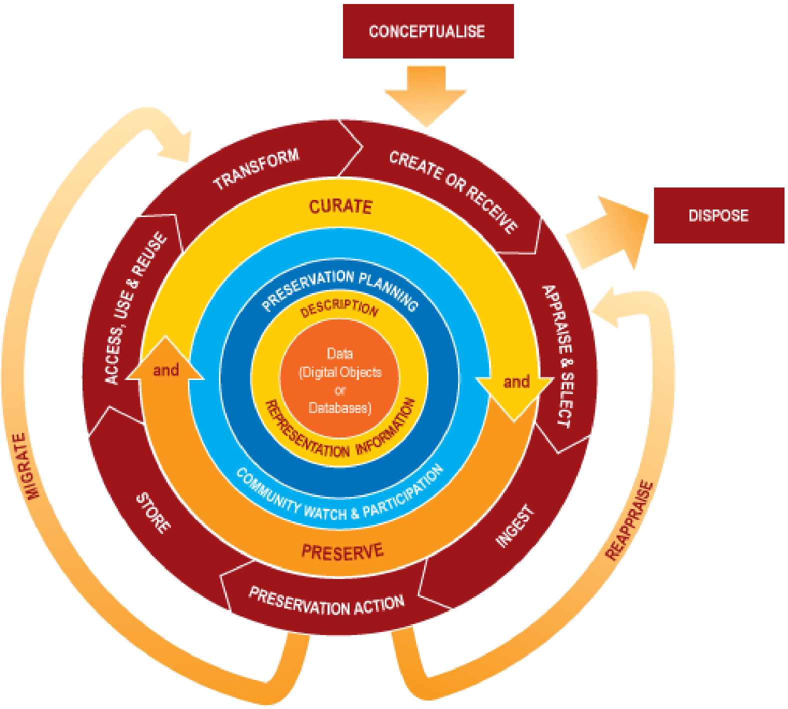
2.1.3. Frameworks

Further attempts to establish a baseline for data curation and encourage development of standard practices have resulted in two primary frameworks: the DCC Curation Lifecycle Model (fig. A) and the Open Archival Information System (OAIS) Reference Model (fig. B-D). Both were developed in the absence of equivalent frameworks and have gone on to inform the discussion and development of current practices.

2.1.3.1. DCC Data Curation Lifecycle Model

The Curation Lifecycle Model (Figure 1) is a "high-level overview" of data curation that is meant to provide a practical framework for describing activities and actions associated with each "stage" of curation (DCC, 2018b).

Figure 1. **DCC Curation Lifecycle Model**



While the model provides a necessary and concise summary, its strength lies in the depiction of the relationship between data and curation, particularly in "the longer lifespan that data has outside of the research project" (Perrier et al., 2017, p.4). The lifecycle model places long-term “preservation planning” at its core, thereby defining curation beyond basic maintenance processes and emphasizing the importance of data beyond initial usage (DCC, 2018). In addition, the Lifecycle Model “can be used in conjunction with relevant reference models, frameworks and standards” such the OAIS Reference Model to create more robust guidelines for roles and responsibilities. (Higgins, 2008, p.135)

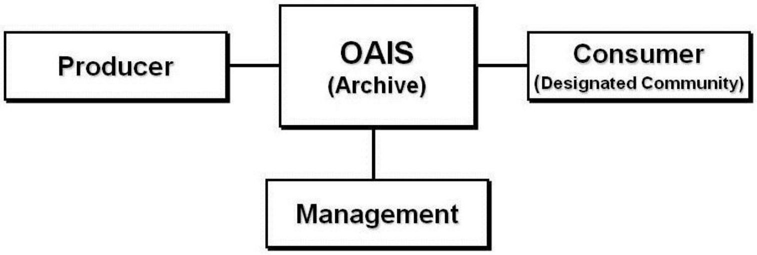
2.1.3.2. OAIS Reference Model

Unlike the functional specificity of the Curation Lifecycle Model, the Open Archival Information System (OAIS) model is a conceptual framework that is meant to act as a “starting point” for building sustainable strategies for long-term preservation (Lavoie, 2012, p.3). The OAIS model is recognized as providing the foundational archival framework for “serious digital archives and repositories” (Lee and Tibbo, 2007) and Lavoie (2012) described the model as “the *lingua franca* of digital preservation” (p.3). More than a framework, the model also includes key responsibilities that an “OAIS-type” or OAIS-compliant archive is expected to eventually fulfill (Lavoie, 2012, p.7).

The reference model illustrates “three separate but related parts”: environment, function, and information (Lavoie, 2012, p.8).

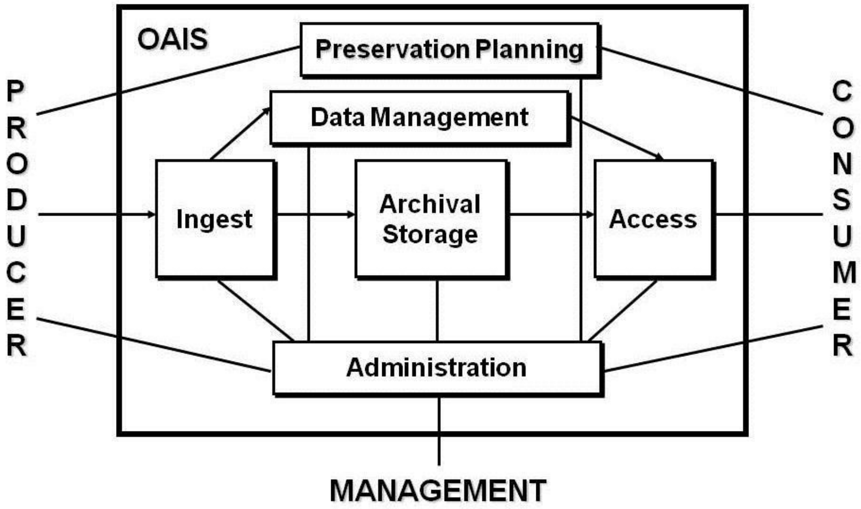
An OAIS environment (Figure 2) refers to the external stakeholders that impact its operations and functions (Lavoie, 2012, p.9).

Figure 2. **OAIS environment**



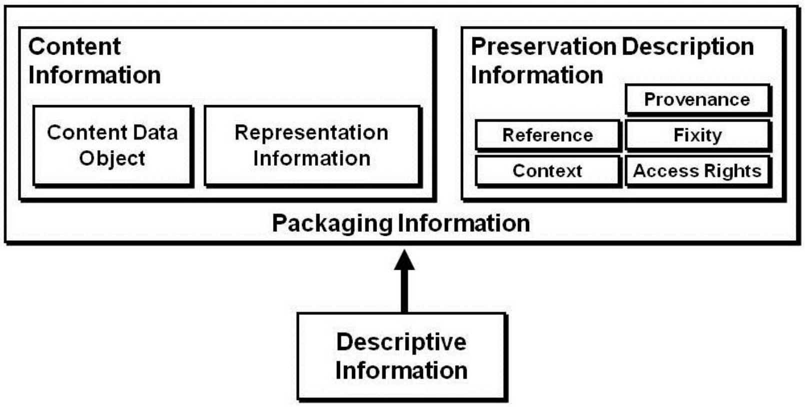
OAIS functions (Figure 3) include six key components (Lavoie, 2012, p.11), similar to the stages depicted in the Lifecycle Model.

Figure 3. **OAIS functions and relationship to environment**



Information handled within the OAIS environment is conceptualized in the form of a “package” (Figure 4) that includes the primary object being preserved and its supporting metadata (Lavoie, 2012, p.14).

Figure 4. **OAIS information package**



Despite the comprehensiveness of the model, due to its conceptual nature, “very few of its concepts have been directly and formally operationalized as standards” (Lavoie, 2012, p.3). Use of its concepts is also not protected, and as a result, multiple derivative and tangential definitions and applications exist.

2.1.4. Stakeholders

However, in order to effectively apply a framework, the role of stakeholders needs to be considered. Researchers and research funders, as primary stakeholders, have formed the basis for current concerns in data curation.

2.1.4.1. Researchers

Despite being essential stakeholders in data curation, as both producers and consumers of data, researchers have been historically sidelined. The lack of collaborative partnership with researchers can be seen in the separation between data services, data curation, and research data management (RDM).

Data services tend to refer to traditional library assistance and are often modeled off pre-existing services for digital information resources, such as assisting with “discovery, use, preservation, and dissemination” (Akers et al., 2014, p.183) or more “traditional services”, such as consultations, “reference support”, and “web guides” (Koltay, 2016, p.97). Existing need for data services arose with the development of quantitative methods of research in the social sciences (Rice and Southall, 2016, p.3). Data curation efforts developed concurrently. The resulting production of datasets naturally led to a need to store, archive, and provide access solutions.

However, as institutions have assumed responsibility over long-term data management, they have developed policies with little or no input from researchers (Fox, 2013). This disconnect has culminated in a mutual lack of understanding between researchers and libraries and, until recently, the neglect of support for RDM. New government and funding mandates, however, have increased the need for closer involvement with the research workflow and prompted an assessment of currently offered services.

2.1.4.2. Funders

Motivations for improving or increasing support for data-related services has been primarily influenced by new requirements from research funders. Higman and Pinfield (2015) identify “requirements of large research funders” as a key factor influencing policy development (p.377). Citing a need for accountability, cost savings, and public sharing and access (Higman and Pinfield, 2015; Lee et al., 2017; Rice and Southall, 2016), research funders have taken the initiative to ensure data is properly managed and “available in a useable form” (Olendorf and Koch, 2012). This is reflected in the requirement of a data management plan (DMP) with submission of funding applications. A DMP, to varying degrees, formally documents how data will be managed and shared, including details about collection, security, and storage.

Of the seven funding bodies in the Research Councils UK (RCUK), six require DMPs (DCC Funders’ Requirements), and for the seventh, DMPs “are encouraged” (University of Cambridge, 2018a). This requirement also extends to non-RCUK and international funders, including Cancer Research UK (CRUK), the Wellcome Trust, the European Commission (EC), and the US National Science Foundation (NSF) (DCC, 2018c; UK Data Service, 2018).

In more recent years, funding organizations have attempted to provide “clear and practical principles” regarding the sharing and management of research data by co-publishing a Concordat on Open Research Data (UKRI, 2016). The concordat principles formally recognize the complexities inherent to data management and emphasize the importance of supporting and complying with open research data requirements. In particular, expectations for data curation are characterized in an exclusive principle as “vital” but caution for the “reasonable” application of resources, especially in regards to researchers (UKRI, 2016).

As concerns for data management converge with data curation responsibilities, libraries are being pressured to furnish “reasonable” solutions for the many problems of data curation while being limited by available resources.

2.2. What are the problems of data curation?

Identifying and understanding the primary issues of data curation is equally as fundamental and complex as the problem of adopting a standard definition. There are many factors that contribute to the difficulty of addressing these issues, including the interdisciplinary nature of data curation, the lack of evidence-based decision-making, and the highly contextual requirements for solutions.

One of the principal issues is the need for implementation of standard practices and operations. Although the presence of legacy data provides an exhibit of preexisting preservation practices, previous issues are being perpetuated with current data practices. These include the use of proprietary file types and services (Locher, 2016; Pinnick, 2017) and poor metadata quality (MacMillan, 2014; Perrier et al., 2017; Pinnick, 2017). Proprietary file types especially result in issues with “versioning capabilities” (Locher, 2016, p.32). Both of these issues are primarily affected by a lack of standardization and a “lack of restrictions” (Pinnick, 2017, p.176). While proprietary software may potentially be unavoidable depending on the field of research, more troubling is the commercialization of digital data storage and services, such as Amazon cloud storage (Berman, 2008, p.53). One of the main struggles for data preservation is a need for storage (Buys and Shaw, 2015; Locher, 2016; Rosenthal, 2017), and reliance on commercial solutions could lead to a problematic dependency in the future. There is a fear, dismissed by Rosenthal (2017) as misrepresented, that “more digital data is being created than there is storage to host it” (Berman, 2008, p.51). Yet, storage costs are frequently cited as decreasing (Berman, 2008; Rosenthal, 2017). Instead, when discussing the issue of rising costs, several authors mention the financial requirements associated with maintaining a knowledgeable staff (Fox, 2013; Pinnick, 2017; Rosenthal, 2017) with the necessary “expertise” (Berman, 2008, p.53).

Knight (2012) compares these issues to a “curate’s egg”. A curate’s egg is “a mix of good and bad”, wherein the bad has spoiled the whole. In this way, data curation is a “digital curate’s egg”, where despite efforts to engage in good practices, their “inconsistent” application results in the perpetuation of bad practices, thereby jeopardizing the overall integrity of the preserved data (Knight, 2012, p.229). Part of the problem, he claims, is the incidental and “ad hoc basis” for most data management practices (Knight, 2012, p.229). An absence of centralized investment, neglect of formalized training, and limited engagement with stakeholders has resulted in the current patchwork state of data curation.

These issues were previously and formally identified in UK institutions through the Data Audit Framework (DAF) Implementation Pilot Project in 2008, which initially sought to “test the effectiveness of the new [auditing] tool created by the DCC” and instead discovered “not-so-good data management practices” (Rice and Southall, 2016, p.72). The project highlighted a startling “lack of” several necessities, including an “awareness and understanding of research data management”, formal plans, training, or guidance, and “clarity about roles and responsibilities” for staff (Rice and Southall, 2016, p.72). Although the project findings pertained to the United Kingdom, these problems are universally relevant.

In a workshop hosted in the United States in 2006, two years before the findings from the DAF Pilot Project, participants gathered to discuss how to best address “emerging principles” and how to formally incorporate standards and policies into data management procedures:

> Lee began by presenting the audience an image of an emergency checklist for an airplane. He then posed the question: Is availability of the checklist a sufficient condition for making it safely onto the ground? After members of the audience answered that it would not be sufficient, he asked, "What else would you need?" (Lee and Tibbo, 2007)

The ongoing dialogue around data curation, despite frequent engagement, has failed to produce viable solutions. Requiring more than a checklist, the deep-rooted complexities of data management and curation have been described by Cox, Pinfield and Smith (2014) as a “wicked problem”. A wicked problem has no “definitive formulation” or solution and bears “multiple value conflicts” (Cox, Pinfield and Smith, 2014, p.4-5). Examining data curation with this lens provides an explanation for the persistence of central issues. No simple solutions can exist because even basic problems are magnified by the scope of the situation. Instead, effective change can only occur with a “culture change”, which requires time and the concerted effort of all stakeholders (Cox, Pinfield and Smith, 2014, p.10).

For a new conversation to begin, however, all participants must first be informed. Cox, Pinfield and Smith (2014) further found a “lack of information about the problem” (p.15), confirmed by Perrier et al. in a scoping review of the field literature (2017). The scoping review revealed a significant gap in “empirical evidence that demonstrates the impact of interventions related to research data management" (Perrier et al., 2017, p.11). Without a foundation of shared knowledge and evidence-based strategies, discussion about data curation will continue aimlessly.

2.3. What are the solutions for data curation? Purpose, practices, and policies

With limited resources, proposed solutions must be workable and applicable for the problems of data curation, with an emphasis on understanding the historical context. By reframing the discussion, we can begin to direct our efforts productively. Rather than asking "what is data curation?", the question should be: "what is the purpose of data curation?" The answer to this question will help establish an objective that can inform practices and policies.

2.3.1. Purpose

The fundamental purpose of data curation is reflected in the founding of the FAIR Data Principles (Wilkinson et al., 2016). Under these principles, data should be "Findable, Accessible, Interoperable and Reusable (FAIR)" (Wilkinson et al., 2016). The reuse of data and research materials in the long term has long been recognized as the primary goal of data curation (Berman, 2008; Lee and Tibbo, 2007; Mohr *et al*., 2015). The CASRAI dictionary of research administration information (2015) defines curation through its goal: "to manage and promote the use of data from its point of creation to ensure it is fit for contemporary purpose and available for discovery and re-use." This includes for the purposes of "longitudinal research" (Locher, 2016, p.29), validation (Dürr et. al, 2008), examination and reproduction (MacMillan, 2014). There is an emphasis on making data not only discoverable but also shareable (Fox, 2013; Shen and Varvel, 2013), and as Mohr *et al*. (2015) summarize, libraries need to be "preparing for sharing" (p.53). This may include making data mineable as part of the process (Dürr *et al.*, 2008). According to Pinnick (2017), the most important consideration is "usability, trustworthiness and future interoperability" (p.176). Ultimately, data curation is a form of "stewardship" committed to securing data for future users (Lee and Tibbo, 2007).

2.3.2. Practices

To accomplish this vision, literature surrounding best practices calls for standardization and documentation to ensure consistency for future use (Dürr et. al, 2008; Rasmussen and Blank, 2007). The literature then stresses the necessity of quality control of metadata during ingest (Berman, 2008; Dürr et. al, 2008, Lee and Tibbo, 2007; Locher, 2016; Pinnick, 2017). Where the technology is concerned, Pinnick (2017) calls for a "technology watch" (p.11) that would keep an eye on file and storage requirements, resulting in frequent "format migration" and "storage media refreshment" (Locher, 2016, p.31). In addition, the "use of open standards for file formats and data encoding; and the promotion of information management literacy" would help mitigate reliance on proprietary and commercial solutions and increase knowledge of available alternatives (Lee and Tibbo, 2007). There is also a move towards storing multiple copies for security (Berman, 2008; Dürr et. al, 2008; Rosenthal, 2017).

Collaboration, including "standardization efforts on a global scale" (Locher, 2016, p.39) and "cross-sector partnerships" (Berman, 2008, p.53), would help offset issues of cost, questions about storage, and provide a platform for sharing knowledge and skillsets. MacMillan (2014) highlights existing subject repositories for data as a viable option.

Institutions may also consider seeking certification, such as the "Data Seal of Approval" (Data Seal of Approval, 2018) to assure compliance with current best practices. Furthermore, as the technology has improved, platforms and tools have become increasingly available to assist in implementing best practices (Amorim, et al., 2017; Austin, et al., 2015; Sallans and Donnelly, 2012).

2.3.3. Policies

To ensure effective implementation, though, a strong strategy is vital. Although there is an abundance of literature available discussing the importance of policies, very little is available on the impact of existing written policies. Key articles on this topic from Briney, Goben and Zilinski (2015) and Dressler (2017) have pointed out a real need to study the connection between policy and practice, and a DPC Technology Watch Report (Lavoie, 2014) has expressed a similar concern in the context of the OAIS framework.

Briney, Goben and Zilinksi (2015), examining data policies in the US, found that 44% of the universities studied “had a data policy of some sort” (Briney, Goben and Zilinski, 2015, p.19). Data policies, when they were “standalone” rather than a consequence of intellectual property policies, did “broadly cover different areas of data management” but tended to focus on legal topics such as “data ownership” (Briney, Goben and Zilinski, 2015, p.20). A review of data policies in the United Kingdom by Horton and DCC (2016) revealed similar coverage. Of the 162 higher education institutions recognized in 2016 (Universities UK, 2018), 57 policies (about 35%) were included in the study. While UK policies covered similar areas of data management, such as defining “support” and DMP requirements, they were less concerned with legal issues and more concerned with issues of ethics, access, and open availability (Horton and DCC, 2016).

Another US-based study, Dressler (2017) evaluated digital preservation policies. 26% of the responding universities held a relevant policy (p.152). Distinct from data policies, digital preservation policies supplied a “template” for preservation work but mainly covered the “challenges and risks of digital media”, including “increasing volume”, “staff expertise and cost” and advocated for formal education and training (Dressler, 2017, p.152). There was a lack of clear, specific guidelines and expectations, such as “*how* this work would be addressed and *who* would be completing such work” (Dressler, 2017, p.152).

Data curation policies would, ideally, include elements of both types of policies and provide clear directions for implementation. In order to ensure compliance with best practices, effective policies must also be established. While data curation may currently be a “wicked problem”, steps can be taken now to prevent it from being one in the future.

3. METHODOLOGY

3.1. Brief overview of components of research design

The research undertaken for this dissertation involved three components:

- A coded analysis and synthesis of standard field models, relevant literature and “best practices” resources
- A survey evaluation of practices and policies at HEIs in the UK
- An appraised count of research data management services and storage options and a content review of accompanying institutional data policies

The survey questionnaire was the main component of the dissertation research, but a mixed methodology was chosen to complement the survey results and enhance analysis and discussion of the findings. The literature review was intended to provide a holistic perspective to evaluate both current policies and practices. During the literature review, very little was found that satisfactorily addressed institutional services or policies, therefore an analysis of institutional websites and policies was also conducted to ensure adequate evidence was available to answer the research questions. This third component was also meant to supplement information in the event that the survey returned too few responses but was eventually utilized to assess the survey results against the institutional perspective.

3.2. Definition of data curation for this dissertation

As previously mentioned, the overarching processes, activities and tasks associated with the long-term management of data is most often labeled as “data curation” or “data management”. In the interest of consistency and in concurrence with international standards, the label of “data curation” has been selected to encompass data workflows, from receipt to storage to preservation for future purposes. In addition to consistency, the archival background of the term “curation” implies a holistic perspective that reflects the motivations of this project. “Curation” suggests a view of data management that is concerned with the future impact of research and extends the care of data beyond basic maintenance. Furthermore, “data management” as a label is problematic because of its close association with RDM literature. “Data management” is often conflated with researchers’ data management practices, and by identifying a distinct term, the hope is to avoid confusion.

3.3. Literature review

A literature review was conducted for the first portion of the methodology. The strength of a literature review lies in its use to “organize, integrate, and evaluate the state of research” (Onwuegbuzie, Frels and Hwang, 2016, p.130). Combined with a textual analysis of the information, a literature review provides a foundation for further analysis and identification of trends or topics within the text.

Literature was sourced for an introductory review through two principal databases: Library and Information Science Abstracts (LISA) and Library, Information Science and Technology Abstracts (LISTA). LISA was chosen as a search platform due to its self-evident focus on the library perspective, and LISTA was also utilized to find articles related to the more technical aspects of data curation. Broad search phrases, including “data curation” and “data preservation”, were used to ensure an extensive range of information. The first batch of articles were selected for review based on their direct relevance to the topic of practicing data curation. Subsequent articles were then selected by frequency of citation in related articles and by the usefulness of the information provided. An article was considered useful for the additional depth or context that could be gained from its review. While there was a slew of literature that described best practices, these resources were often either too high-level or theoretical to provide functional, practical guidance, or they were case studies that characterized unique, individual experiences of implementing practices. In order to identify the most recommended practices, these resources would need to be consolidated to produce a “checklist” of practices that could be applied in a general situation. Therefore, a combination of approaches was used to perform a comprehensive textual analysis of the literature.

Content and coded thematic analyses of relevant field resources were sequentially conducted to establish a framework as the basis for the survey questionnaire and for evaluation of final results. The content analysis distinguished important themes through frequency of coverage, while the thematic analysis specified greater patterns within each theme. The analyses were conducted in two stages: a first and second cycle. The first cycle discovered initial themes that were then expanded and explored further in the second cycle (Onwuegbuzie, Frels and Hwang, 2016).

In the first cycle, thematic codes were generated through a sequenced consolidation of two primary frameworks:

- OAIS Reference Model
- DCC Curation Lifecycle Model

These two models were selected because of their integral role in shaping current practices of data curation and for their intended purpose as high-level mapping modules. The DCC Curation Lifecycle Model "enables granular functionality to be mapped against it", while "the OAIS functional entities can be implemented and configured in any way appropriate to an archive's particular circumstances and technology" (Lavoie, 2012, p.11). The models were analyzed for repeated and thematically parallel content.

In the second cycle, the generated codes were validated and finalized into a series of categories and subcategories through synthesis of relevant literature and "best practices" resources. Themes from the literature were identified and summarized by recurrence, with the addition of subcodes to distinguish between repeated topics where necessary.

3.3.1. First cycle

Provisional and descriptive coding methods were used to map a series of thematic codes and baseline categories from a consolidation of the OAIS Reference Model and DCC Curation Lifecycle Model (Onwuegbuzie, Frels and Hwang, 2016). The categories were closely based off these frameworks and incorporated all individual functions or actions from each respective model.

Provisional, predetermined codes were sourced from these models to provide an initial structure to facilitate comparison of each model's functions and actions. "Key words" from the description of these functions was then matched to produce an initial evaluation framework of *six baseline categories* (Table 1) (Onwuegbuzie, Frels and Hwang, 2016, p.135). These categories were later referred to as "procedural categories" during analysis as a descriptor of their function as categories defining curation procedures.

These models are founded on archival theory, and as a result, certain terminology may be unfamiliar to information professionals without a background in archives. In order to situate archival practices within a working library context, certain aspects were coded using terms more commonly related to traditional librarian duties (ex. “acquisition”).

Table 1. **Initial evaluation framework**

Category	OAIS Model	DCC Curation Lifecycle Model
Acquisition	Ingest	Create or receive Appraise & select Ingest Preservation action
Metadata	Data management Description information Representation information	Description and representation information Create or receive Preservation action
Storage	Archival storage	Store
Administration	Administration	Preservation planning Community watch participation
Preservation	Preservation planning	Preservation planning Dispose Reappraise Migrate
Access	Access	Access, use and reuse Transform

- *Acquisition*: The selection, receipt, and processing of information resources. A term more familiar to traditional librarianship than archives, “Acquisition” as a coding label reflects the explicit recognition of data as an information resource, despite the lack of obvious or direct language in either the Curation Lifecycle model or the OAIS model. This category includes the “ingest” function of OAIS and the “create or receive”, “appraise & select”, “ingest”, and “preservation action” stages of the Curation Lifecycle.

- *Metadata:* Identifying information about the data resource and its related files. Neither model labeled metadata as a separate section, however, due to its prominence in both models and its impact on later stages of the data lifecycle, “Metadata” was included as an individual category. In the OAIS, metadata is referred to under “data management” in the functional model and under “description information” and “representation information” in the information model. In the Curation Lifecycle, metadata is referred to as a product of various actions, including “description and representation information”, “create or receive”, and “preservation action”.
- *Storage:* The technological infrastructure necessary for short- and long-term deposit of data files. This category combined the “archival storage” function of the OAIS model and the “store” stage of the Curation Lifecycle.
- *Administration:* High-level management of components that contribute to the curation workflow; administration does not tend to be in direct contact with data curation tasks. This category integrated the “administration” function of the OAIS model with aspects of the “preservation planning” and “community watch participation” elements of Curation Lifecycle.
- *Preservation*: The long-term maintenance of data. “Preservation” as a category closely aligns with the “preservation planning” function of the OAIS model. “Administration” and “preservation” overlap in some instances, including in the incorporation of “preservation planning” from the Curation Lifecycle model. “Preservation” is distinguished from “administration” through operational actions found in the curation lifecycle: to “dispose”, “reappraise”, and “migrate”.
- *Access*: Facilitating the discovery and use of data. This category aligned with the “Access” function in the OAIS and the “access, use and reuse” element of the Curation Lifecycle, as well as the “transform” stage.

3.3.2. Second cycle

In the second cycle, the provisional codes were refined into a comprehensive evaluation framework through the addition of subcodes. Content from practical resources were reviewed through focused and axial coding methods to ensure relevancy in the finalized coded categories. Focused coding identified the most repeated tasks or recommendations from the resources within each baseline category (Onwuegbuzie, Frels and Hwang, 2016). Then axial coding was applied to merge these duplications into corresponding subcodes. The results of the final evaluation framework produced from the merger are presented and discussed in the Analysis section of this dissertation (Table 2).

The following resources were used in the second cycle:

- OAIS Functional Model: Part of the OAIS Reference Model, each function was scoped for specific tasks and activities.
- DCC Curation Lifecycle Model: The DCC provides a generalized checklist for each stage of the model. These checklists did not include specific criteria were consulted as a guideline for defining processes.
- Data Asset Framework (2018): Previously known as the Data Audit Framework (DAF), DAF is a self-auditing framework to evaluate data curation practices at any given higher education institution, “identify any risks”, and assess “researcher’s attitudes towards data creation and sharing” (CITE). The DAF methodology is based on the DCC Curation Lifecycle Model to identify key stakeholder roles. DAF was included as a resource for review due to its examination of these roles and its clarification of aspects of the curation lifecycle. In particular, the DAF identifies the responsibilities of information professionals as encompassing “appraise & select”, “ingest”, and “access, use & reuse” tasks, as well as all actions related to preservation.
- FAIR (Findable, Accessible, Interoperable, Reusable): FAIR was included for its recent emergence as a driving philosophy behind data curation. The primary philosophy behind FAIR is “supporting discovery through good data management” (Wilkinson *et al*., 2016). Therefore, each stage of the curation lifecycle should contribute to data ultimately being FAIR.

- Data Seal of Approval (2018a): The Data Seal of Approval (DSA) is another self-auditing procedure that provides officially recognized certification of any data repository. The DSA lists 16 “core trustworthy data repository requirements” with a focus on accessibility, usability, reliability, and persistence (2018b). The DSA was chosen for inclusion over other available certification measures because of its abridged requirements. Unlike the complexity of other certification measures, the DSA provides a general overview of more comprehensive auditing frameworks that allows for more flexible applications (Knight, 2012).
- DCC Curation Reference Manual: The DCC’s manual was included because of the organization’s integral role in advancing the conversation around data curation. Completed chapters were combed for “advice, in-depth information and criticism on current digital curation techniques and best practice” (2018d). Since this is a general reference manual for digital curation, chapters were only selected and reviewed when chapter titles were clearly related to data curation and matched the pre-coded baseline categories.

In addition to the field resources, a literature scan was also conducted to incorporate authoritative knowledge from published findings. Literature for the literature scan was ultimately sourced from the Research Data Curation Bibliography (Bailey, 2018) due to its topical relevance. A search in generalized databases such as LISA yielded less relevant results. The inconsistency in field terminology meant searches for specific phrases were too restrictive. In addition, searches for “best practices” literature often resulted in articles related to RDM, and other current literature placed heavy emphasis on researcher needs, both of which fall outside the scope of the intended research.

Within the Bibliography, the search term “practice” was used to specifically target articles that address curation practices. Narrower search terms were avoided due to the aforementioned inconsistency in field terminology. A more restrictive search, although eliminating unrelated articles, may have also excluded relevant articles. Due to the rate and number of publications, the search was restricted to within the last 5 years (2013-2018) to incorporate only the most current guides and recommendations. The literature was scanned for a general overview or implementation of practices. Literature was discarded if it was too contextual or discipline-specific. However, this did not preclude discipline-specific literature. Some articles, despite their origins in a specific discipline, stated a possibility for extrapolation to general purposes.

The literature scan produced seven usable texts, including one framework, four case studies, and two surveys. The full list of included texts is provided in Appendix 1.

The results were recorded in an Excel spreadsheet in two stages identified through separate tabs. The first tab was used to keep track of “raw data”, which consisted of the title of the publication, the associated URL or DOI, and each thematic code that resulted from the first cycle of analysis. Direct quotes from each publication were recorded under a relevant thematic code. Both the publications and thematic codes were then assigned numbers as unique identifiers. These numbers were used to format the “final code review” in the second tab. The final code review was organized with the thematic codes as rows and the publications as columns. This allowed the addition of subcodes and enabled easier browsing of the final results. Codes could either be compared across publications to determine frequency of usage or a single publication could be evaluated against the code hierarchy to decide its value as a comprehensive source. The direct quotes recorded in the first tab were recorded again under the most closely corresponding subcode in the second tab. In certain instances, the second cycle of analysis produced text that was unique and could not be sorted under a subcode. Actions, processes, or recommendations that were not recurring throughout the literature was instead coded under a main thematic code. If analysis did not immediately correspond to any codes, the text was categorized under an ad hoc “unsorted” code.

During final synthesis, axial coding was applied to the unsorted text to assign a “weighted” value (Onwuegbuzie, Frels and Hwang, 2016, p.136). Text was appraised using the following criteria:

- Is the action described by the text task-oriented? (if not, the text was discarded)
- Does the action conform to, or is the action similar in sentiment, to the parameters of an existing category?

This second closer reading resulted in the elimination of text that was too vague, too specific, or too researcher-centric and led to the creation of new subcodes.

Certain subcodes related to Administration and Access were discarded during the final survey design. In an effort to provide in-depth and comprehensive analysis of central preservation actions most likely to be performed by information professionals, the scope of the survey was narrowed to focus on primarily technical tasks and solutions. As a result, certain categories no longer fell within the scope of the intended research. Therefore, specific questions about these categories were not included in the final survey questionnaire. Instead, questions in remaining categories were augmented or revised to incorporate information from the discarded subcodes where possible and when deemed useful for understanding the context behind certain procedures.

For subcodes related to Administration, individual questions about official policies were included when relevant and as they related to each of the main baseline categories. These questions assessed the availability of policies and briefly appraised key content. Where topics could not be thoroughly explored in the survey, questions about policies and standards were occasionally substituted to assess the presence of a defined workflow.

The subcodes related to Access were all merged into a single question. Their close reflection of the FAIR principles and values resulted in a single, cumulative question directly addressing the institution's achievement of FAIR practices. Additionally, questions about Access subcodes were concluded to be unnecessary and repetitive. Questions in the remaining categories already addressed key procedures or actions related to the facilitation of access.

3.4. Survey

The survey was chosen as a research methodology for its capability to identify trends and analyze data both quantitatively and qualitatively. As a "systematic process of data collection", a survey provides the opportunity to gather data about broad concepts in a measured format (Aiman-Smith and Markham, 2004, p.12). In particular, their "reach", "flexibility", and "speed and timeliness" of distribution supports an expedited process between the time of dissemination and the final reporting of results (Evans and Mathur, 2005, p.197). In addition, the "convenience" of the survey format allows participants the option of responding without the difficulties of location or travel (Evans and Mathur, 2005, p.198). The national focus of the dissertation research requires recruitment of a geographically dispersed participation pool, and an online survey provides a greater possibility of wider reach in a shorter amount of time.

Although surveys have many advantages, there are also two major risks associated with the methodology: the possibility of a skewed representative sample and, most of all, a "low response rate" (Evans and Mathur, 2005, p.201-202). Surveys are self-selecting and rely on "self-reports", meaning only certain groups may be represented, and results cannot be immediately verified (Perrier *et al*., 2017, p.1). However, the risk can be mitigated with a targeted distribution focus, and the response rate may be higher, given a demonstrated desire for quantified research about the surveyed topic (Perrier *et al.,* 2017).

3.4.1. Survey design

The final survey format (Appendix 2) was divided into thematic sections that reflected the final coded categories. The sections were arranged according to their placement within the curation lifecycle, and each survey section was intended to mimic an ideal workflow by representing a natural progression to the next phase in the curation lifecycle. Topics within each section followed the same structure. However, where processes were not clearly prioritized or differentiated within the workflow of the respective section, question placement did not adhere to a particular pattern. Questions generally focused on reviewing specific tasks or procedures related to data curation. Questions not within this scope, such as those regarding "administration" and "access", were placed at the end of the survey.

Due to the risk of a low response rate, a careful balance was struck between the length of the survey and the content of each question. The number of questions was limited to 25, and the survey was "no more than 15–20 minutes" long to avoid survey fatigue in respondents (Aiman-Smith and Markham, 2004, p.13). Questions were also formatted to encourage complete participation. Questions were primarily closed-ended but included open-ended opportunities. Closed-ended questions were a mixture of "categorical-nominal" and a couple of "interval" type questions (Aiman-Smith and Markham, 2004, p.13). In all instances, efforts were made to ensure that "one question should equal one idea" (Aiman-Smith and Markham, 2004, p.13).

To ensure more comprehensive results, closed-ended questions were typically multiple choice and allowed selection of all applicable options. Open-ended questions were asked when a procedure or process could result in a wide variety of unique solutions or workflows and an open-ended question would be more conducive to a free text explanation. Interval questions, represented by rating scales, were included for questions where topics fell outside the parameters of typical data curation responsibilities (i.e. “administration” and “access”), but a personal perspective was desired for holistic context.

3.4.2. Survey dissemination

Potential participants were recruited from a wide pool of information professionals situated within or associated with UK HEIs. There is no clear indication or documented gauge of the number of data services or repositories offered by UK HEIs, therefore making it difficult to determine a targeted respondent group. According to the Open Research Data Taskforce (ORDT) (2017), there is an estimate of “upwards of 30 UK universities” that provide repository services for data, although this number is obscured/ambivalent (p.28). However, the DCC (2018e) currently lists 80 institutional data policies on their site, suggesting a number of institutions may offer data services without operating a data repository. In addition, many institutions may not currently offer research data services or engage in data curation. As responses from these institutions are equally valuable to gauging an accurate state of data curation in the UK, and due to the discrepancy of reported numbers, the survey was not restricted to a set list of institutions.

A “background” section was included in the survey to filter respondents for their association with HEIs and data curation. The research is seeking broad generalizations, and consequently, the survey does not inquire for specific details, such as name of university or job title.

Surveys were distributed through email and circulated via Twitter. Four JISC mailing lists were selected for their relevance to the topic and intended audience of information professionals: Research Data Management, JISC repositories, UK Research Repository Administrators, and LIS ARLG (CILIP’s Academic & Research

Libraries Group). On twitter, the survey was circulated via the University of Strathclyde's RDMS (Research Data Management and Sharing) department account (@StrathRDMS).

3.4.3. Analysis of survey results

The results of the survey were collated through reports generated by Qualtrics. The reports were used to quantify and visualize response rates through percentages. To avoid irrelevant results, reports were filtered to exclude any respondents that were not affiliated with HEIs. As most questions stood independently, responses to one question did not affect the outcome of responses to another question. Therefore, partial responses were included to provide a larger sample size for analysis. The inclusion of partial responses ensured a distribution of responses that was more representative of the target audience (i.e. all UK HEIs) and allowed for a more detailed examination of trends both within and between questions. To ensure accurate interpretation of the results, the response rate for each question was taken into consideration when calculating percentages. Response rates did not fluctuate greatly between questions, and for that reason, results were compared even when response rates did not match. The findings were then discussed as a general comparison of trends rather than a direct comparison of responses.

For closed-ended questions, percentages were calculated and accompanying visualizations were also produced through Qualtrics in the form of comparative bar charts. Results were displayed in the bar charts from most to least responses. For open-ended questions, answers were analyzed for content and were summarized into the most common themes. Outliers were noted and included in the findings for thoroughness. Additional comments collected at the end of the survey were considered for unique insight into major factors currently affecting the field of data curation.

3.5. Online resources

The final portion of the research was aimed at establishing a concrete picture of offered services and policies related to data curation. An assessment of UK HEI websites was conducted to determine the overall total of universities that offer data management services, to ascertain responsibility for the data management services, and to calculate the subset that hosts data repositories. Then, a content scan was performed on the institutional research data policies attached to each university.

As a heavily textual source, web sites are a modern form of “documentary source” that are able to yield both “direct” and “indirect” content (Finnegan, 2006, p.143). Direct analysis of a document is useful for straightforward fact gathering, while indirectly, a document can reveal the motivations and intentions of the document creator--in this case, the university (Finnegan, 2006).

According to gov.uk (2018), there are 169 recognized higher education bodies in the UK, not including those “that can only award foundation degrees”. As foundation degree courses are vocational, the assumption was made that these higher education bodies are not research-intensive and therefore not likely to host data services or policies. A list of the 169 officially recognized HEIs was sourced from the gov.uk (2018) website, and data about each institute was collected and recorded in a master spreadsheet. A total of *16 categories* was produced: 4 related to research data services and 12 related to associated policies.

Each officially recognized HEI was assessed for the presence of research data management services through the university website. Commonly, funders and institutions require that research data is retained for a minimum of ten years after publication (University of Cambridge, 2018b). The presence of RDM services would suggest at least a minimum consideration of data curation needs to satisfy existing policy requirements.

An exhaustive search of each institution’s offerings was conducted through a combination of Google and university websites. An initial search was performed in Google with the search phrase “[full university name] research data service”. If no satisfactory page links were yielded within the first 10 results on Google, the following alternative search phrase was used: “[full university name] research data management”. If the second alternative search was equally unsuccessful, the university website was then searched with phrases such as “research data management” and “research data”. In the course of navigating each website, a pattern emerged from information about RDM services. Information about these services tended to be located either under pages about “Research” or “Research support” or within the library’s online resources. These trails were followed for institutions whose websites did not provide directly apparent search results. If no relevant information was finally produced after this process, a “no” was recorded to indicate that an institution does not currently have an existing RDM service.

If a relevant webpage was found, institutions had to, at minimum, offer online data management resources, such as handbooks, guidelines, or tutorials, as well as a department help contact to constitute an existing service. The following criteria were recorded to indicate the extent to which an institution satisfied the conditions required:

- Yes: Conditions were satisfied fully.
- Partial: Online resources were available but there was no clear departmental contact.
- Not public: Only contact or department information was offered.
- In development: An announcement declared current or future progress in developing appropriate resources for research data management.
- Unknown: Institutions that did not fall within any of the above parameters. This included references to services without accompanying online resources or contact information. A full explanatory note was attached to these institutions.

Once the presence of an RDM service had been established, the informational pages were searched to discover the department overseeing the service. A department was recorded as the “responsible department(s)” if the department or associated staff were explicitly listed as help contacts. If the department was a division of a larger organization within the university, the main organization was listed. For example, if a department was part of the library, the library was recorded as the “responsible department”. If multiple departments were cited as sharing equal responsibility, each department was listed with a forward slash separating each one. A forward slash was utilized to avoid confusing departments with names that included an “and”, such as “research and innovation”. In the event that there was no clear service administrator, responsibility was inferred from the department hosting the content on the university website. “N/A” was recorded for institutions without an RDM service.

Concurrently, HEI websites were also searched for existing repositories to determine the storage options available for research data. The lack of an RDM service did not preclude an institution from hosting a repository service.

For universities with available RDM services, the related informational pages were searched for "deposit" or "storage" options to locate repositories that housed research data. For institutions without RDM support pages, a search was conducted in Google for "[full university name] repository" or on the university website for "repository". Repositories were identified by their commitment to long-term, post-project/post-publication storage, with a distinction being made between data repositories and institutional repositories (IRs) that host datasets. While repositories were the most common storage options available, a handful of other options were presented/discovered during this process. If no relevant repository could be identified, these alternative options were recorded in lieu. The following classifications were used to indicate storage options:

- Data repository: A repository solely devoted to research data and datasets.
- Institutional repository (IR): A repository that jointly hosts research outputs and supporting research data. An IR had to explicitly accept "research data" or "datasets", either mentioned on the RDM webpage or on the repository "About" page, to qualify as a data storage option.
- Both: A data repository and an IR were presented as equally viable/available options
- In development: An announcement declared current or future progress in developing a data repository
- Special notes: Notes about other storage options were recorded when an in-house repository was not listed but alternative guidance or specific recommendations were offered. This included:
    a. Recommendations for external subject repositories
    b. Data catalogues
    c. Shared repositories
- None: No repository options or recommendations were provided.

Finally, each HEI was checked for existing research data policies. Where a relevant policy was not present or accessible on RDM informational pages, a combination of Google searches and website searches was used for “research data management policy”. If these searches were unsuccessful, publicly available “Policies” pages were checked directly. A final check against policies listed on the DCC site (DCC, 2018e) was conducted to ensure no policies were missed. RDM policies existed in various states, and the following criteria were recorded to indicate the condition of the policy:

- Yes: Institution provided a final policy with an accessible link to the online document. The link was then embedded in the corresponding field in the master spreadsheet. If multiple policies existed, or a policy was described alternately (e.g. as a strategy or “roadmap”), all relevant links were included.
- Draft: Institution provided a policy labeled “draft” with an accessible link to the online document. The link was then included in the spreadsheet.
- Not public: Institution indicated a final policy was available but access required user credentials. Where possible, a link to the policy was embedded in the spreadsheet.
- In development: Institution indicated a commitment to producing a policy, but a full document had not yet been released.
- Partially: Institution provided an RDM policy as a subsection of another policy. A link to the parent policy was included in the spreadsheet.
- No: Institution did not have a current or future-planned data policy.

3.5.1. Document analysis

After the data had been compiled for each institution, a document analysis was performed on the institutional data policies using the evaluation framework developed previously. Document analysis is a common research technique used in many fields both for primary data gathering and for supporting evidence (Oczkowski *et al*., 2018; Finnegan, 2006). The benefit of a document analysis is retrieving answers from the data directly, and analysis of policies is a useful tool for considering the context of practice and for directing the development of guidance (Briney, Goben and Zilinski, 2015; Dressler, 2017; Oczkowski *et al.*, 2018).

An initial sample scan of the first 10 policies quickly revealed an absence of specific direction. Policies tended to cover roles and responsibilities rather than offer practical guidance. Rather than attempt a comparative analysis of policy content to best practice guidelines, policies were instead reviewed for their scope.

Five components were selected for review:

- Date: The month and year of the most current review or last date of approval were recorded. If neither dates were listed in the document, a date was recorded from the file name. If no date was available, the document was noted as "undated".
- Attached procedures or guidance: If practical guidance (e.g. procedures, checklists, etc.) or a direct link to university-provided guidelines was included within the policy document, a "yes" was recorded. Otherwise, a "no" was recorded.
- Use of "curation" terminology: The synonymous use of "curation" or "curate" to refer to data management was logged with a "yes" or "no". As described earlier, these terms are often conflated, and this was an attempt to quantify and compare their official usage.
- Commitment to Open: References to open access or open sharing of data within each policy was logged with a "yes" or "no".
- Presence and assignment of responsibility for key procedures: Each of the 6 main categories from the evaluation framework were checked for, along with 2 additional subcategories, for a total of 8 categories. These categories were assessed by the presence of related terms within the document and for the parties responsible for their support.

3.5.2. Evaluation framework for policies

Certain categories, such as "metadata", were often only obliquely mentioned through vaguely described requirements or mechanisms. To establish a broader net for data collection, the following key words were accepted for each category in addition to the criteria established earlier in the research process:

- *Acquisition*: data collection, data capture

- *Metadata*: descriptive information
- *Storage*: deposit, security
- *Preservation*: retention, curation, disposal, archiving, deposit, assessment of data
- *Administration*: references to structures or systems that support data management, including training, guidance, and support
- *Access:* sharing, publication

In the absence of any met criteria or keywords, a “none” was recorded. Umbrella statements of responsibility regarding the whole data management process were not considered sufficient for meeting the set criteria.

Once the presence of a category was confirmed, the responsible party was ascertained. Predominantly, responsibility was assigned to the following parties and were recorded as such:

- Researcher: Principal Investigators (PI) and all individuals on the research team
- Department: Heads of faculty or department heads overseeing/supervising researchers.
- Institution: University administrators
- Specific departmental support: If individual departments were mentioned, each department was listed (e.g. IT, library, research data services, etc.)

In addition to the six main categories, the policy document was scanned for two subcategories: “DOI” under metadata and “security” under storage. According to the survey responses, “metadata” and “storage” received the most coverage in institutional policies, with DOIs and security being the primary concerns for the respective categories. Due to the high report rate of “DOIs” and “security” in institutional policies, documented evidence of these requirements was investigated. Policy documents were searched for “doi” and “secur*”, including “security”, “secure”, and “securely”. The presence of these terms was recorded with a “yes” or “no”. Both requirements had to be explicitly stated and within the context of data curation to be counted. References to legal or access compliance were not considered part of the data curation process.

If any category was missed in the initial scan, the document was then searched using in-built search mechanisms.

3.5.3. Analysis of data collected from online resources

The data was cleaned and analyzed through OpenRefine, an open source data transformation application. Data was cleaned by merging individual text strings with duplicate content, and missing data points were investigated and resolved. Basic numerical counts for each category were gathered through the “text facet” function of OpenRefine, which counts instances of each textual phrase within that category (OpenRefine, 2018). Text facets can be sorted by “name” or “count”, and “count” was chosen most often to provide a quick overview of trends within that category (OpenRefine, 2018). As an exception, policy dates were sorted by name due to their wide count range. However, policy dates were originally entered in a “Month-Year” format, resulting in dates being sorted by their month rather than their year. To resolve this issue, dates were converted to a “YYYY-MM” format using the “transform” function in OpenRefine.

More than one category could be sorted under a facet at one time, and the results from each subsequent category would be filtered through the first category. This function was used to easily analyze connections between categories and determine the impact of one category on another. Responsibilities in particular were quantified through a sum total of all policy-related categories to determine overall rates of responsibility for the lifecycle of data curation.

Counts were used to calculate corresponding percentages in Excel. A comparative analysis was then performed between the results to uncover connections between each category. Final visualizations were produced through Excel.

3.6. Conclusion

Finally, an analysis of all collected data was performed to compare practice and policy. Augmented by useful input from those in the field, the survey illustrated prevailing practice, while information gathered through documentary sources detailed differing levels of administrative support. The combination of these research methods provided a thorough examination of the current state of data curation.

4. ANALYSIS

The three-part research methodology allowed for a comparison between standard practice and best practice, as well as between the survey results and established documentation. Achievement of best practices was varied, and the difference between reported policies and actual policies illustrated a discrepancy between support at the institutional level and services at the departmental level. In addition, open-ended comments from the survey questionnaire provided further insight into the current state of data curation.

4.1. Final outcome of literature review

The literature review produced an evaluation framework of 21 activities associated with the six main procedural categories (Table 2). Six of the 21 activities were discarded due to either repetitiveness or irrelevance to the research questions, resulting in a final list of 15 activities. A majority of the activities, 12 in total, were related to the first 3 procedural categories: Acquisition, Metadata, and Storage. In particular, Storage was predominantly discussed in the literature and was therefore represented by the highest number of activities (5). Full definitions of each activity were developed through a consolidation of the literature, and for clarification have been provided below Table 2.

Table 2. **Final evaluation framework**

Acquisition	Receipt Appraisal & selection Validation Ingest
Metadata	Standards Description information Representation information
Storage	Documentation Security Format Migration Recovery
Preservation	Long-term strategy

	Risk assessment
Administration	Systems operations Stakeholder interaction (discarded) Policies & standards (discarded)
Access (discarded)	Access Reuse Interoperability Access controls

- *Acquisition*

    a. <u>Receipt</u>: Procedures for receiving data files, including “defined criteria” (DSA, 2018) for submission

    b. <u>Appraisal & selection</u>: Processes to evaluate value of data for future use before transfer into repository (DCC, 2018d; Laughton and du Plessis, 2013)

    c. <u>Validation</u>: Inspection of data files to ensure “authenticity” of information (DSA, 2018) and that content is “uncorrupted and complete” (Lavoie, 2014, p.12).

    d. <u>Ingest</u>: Tasks associated with transferring data into repository. Tasks are typically meant to prepare data for storage, and examples include metadata extraction and file conversion (DCC, 2018d; Laughton and du Plessis, 2013; Lavoie, 2014; Lee and Stvilia, 2017)

- *Metadata*

    a. <u>Standards</u>: Essential requirements to ensure sufficient quality control of metadata records (DCC, 2018d). Strategies may range from “satisficing” (Lee *et al.*, 2017) to choosing an “optimal set of metadata elements” (Lee and Stvilia, 2017).

    b. <u>Representation information</u>: Identified as half of the metadata integral to an OAIS Archival Information Package (AIP), relevant metadata elements should supply “structure” and “semantic” information (Lavoie, 2014, p.16). Structure information refers to metadata about the technical aspects of the data files, such as format or software information. Semantic information refers to metadata that assists in correct interpretation of the content, such as a “glossary” or “user documentation” (Lavoie, 2014, p.16). Representation metadata is intended to support successful rendering of files during future usage and is also referred to as “auxiliary information” (DCC, 2018d).

c. Description information: The other half of the metadata necessary for an AIP, elements cover background information about the content and history of the data files, such as “reference”, “context”, “provenance”, “fixity”, and “access rights” (Lavoie, 2014, p.18). These elements should provide unique identifiers, facilitate proper citation, and map relationships between other metadata elements (Chao, Cragin and Palmer, 2015; DSA, 2018; Friddell, LeDrew and Vincent, 2014; Helbig, Hausstein and Toepfer, 2015; Laughton and du Plessis, 2013; Van Zeeland and Ringersma, 2017)

- *Storage*

  a. Documentation: Policies for managing storage protocols, including individual “processes and procedures” (DSA, 2018)

  b. Security: Protection of data and associated assets to mitigate risk of loss or corruption and prevent potential of mishandling. Examples of appropriate actions include encryption and duplication (Laughton and du Plessis, 2013; Lee and Stvilia, 2017; Van Zeeland and Ringersma, 2017).

  c. Format: Containment of data in “constant and stable” structures to ensure persistence (Friddell, LeDrew and Vincent, 2014).

  d. Migration: Transfer mechanism to ensure long-term stability of data in case of format or media degradation (Lavoie, 2014; Lee and Stvilia, 2017).

  e. Recovery: “Safeguard mechanisms” and “disaster recovery policies” (Lavoie, 2014, pp.12), such as environment checks, to ensure preservation of data in the event of physical or technical issues (Laughton and du Plessis, 2013; Van Zeeland and Ringersma, 2017).

- *Preservation*

a. Long-term strategy: Plans and policies developed to ensure long-term preservation and future functionality of data, as well procedures and actions undertaken in anticipation of change (Chao, Cragin and Palmer, 2015; DCC, 2018d; DSA, 2018; Lavoie, 2014).

b. Risk assessment: An integral part of long-term strategy, procedures to evaluate potential risks and prevent future disaster (Lavoie, 2014; Laughton and du Plessis, 2013)

- *Administration*

    a. Systems operations: Support and maintenance of technological infrastructure, including updates, appropriate hardware and software software solutions, and performance monitoring (DCC, 2018d; Lavoie, 2014; Lee *et al.*, 2017)

    b. Two activities were partially or completely discarded: stakeholder interaction and policies & standards. The high-level nature of these categories did not typically lend themselves to centralized activities or tasks. Administrative priorities were instead distributed throughout each stage of the data lifecycle and tended to relate to organizational support and oversight of curation functions.

        i. Stakeholder interaction: Communications with stakeholders to facilitate smooth operations and essential collaborations within the curation lifecycle. This category was discarded entirely.

        ii. Policies & standards: Management guidelines and documentation to ensure “compliance” with established practices (DSA, 2018). Although this category does not explicitly involve direct interaction with data files, the literature emphasized the importance of “defined workflows” for each stage of the curation lifecycle (DSA, 2018). Therefore, “policies and standards” was only partially discarded as a category.

- *Access*: Four activities emerged from the second analysis of resources and literature: access, reuse, interoperability, and access controls. These categories generally involved processes related to the end user experience and were all eventually discarded in favor of a single category related to FAIR principles and practices.

The final evaluation framework was used fully to collect information about current practices through the survey questionnaire. Then, the framework was used partially to evaluate institutional policies.

4.2. Results of survey

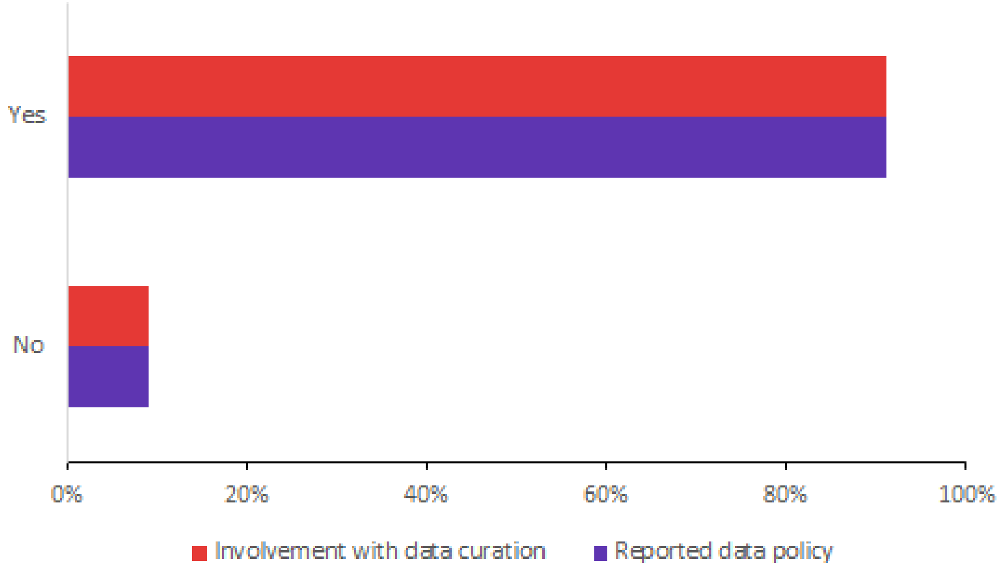
As partial responses were included for analysis, the response rate for each question has been included, where necessary, for clarity and context. The findings have been structured to parallel the evaluation framework, and a brief summary of the overall findings for each main procedural category has been provided.

4.2.1. Background of survey respondents

The survey garnered 23 complete responses and 40 partial responses. Out of the total responses, 2 were filtered for their lack of affiliation with a Higher Education Institution. Respondents with institutional support for data curation were more likely to participate.

30 out of 33 respondents (91%) were directly involved in data curation or data management, and the same amount of respondents worked at an institution with an existing data policy. Although a majority of respondents were directly involved with data curation or data management, 3 out of the 33 respondents were *not* involved, either through their own positions or through their departments. Again, this matched the number of institutions without a reported data policy; 3 out of the 33 respondents reported that their institution did *not* have a data policy. This would indicate a strong connection between those institutions with established data curation practices and an existing data policy (Figure 5).

Figure 5. **Percentage of respondents involved in data curation with a data policy**



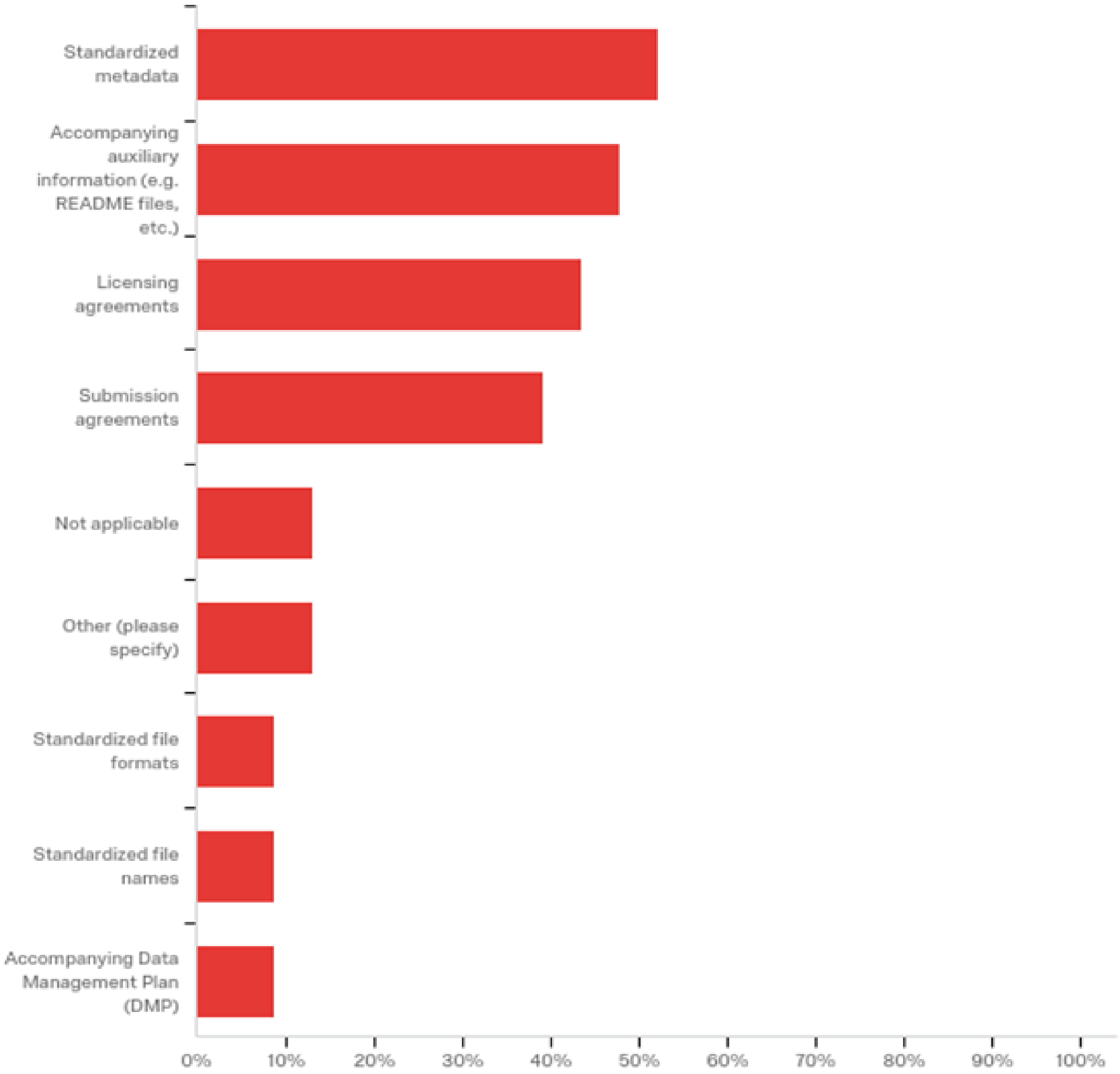
4.2.2. Acquisition

Activities during Acquisition generally encompassed assessment and preparation of files for long-term storage and preservation. According to the survey, about half of recommended activities were being performed, mostly related to storage. Preservation was largely neglected. In addition, file preparation was typically entrusted to researchers, with submission being facilitated by an institutional unit.

4.2.2.1. Receipt

During receipt of a file submission, the most common requirements for submission of data files (Figure 6) were related to expediting files for access and use, indicating many institutions are investing in a culture of shared research data.

***Figure 6.* Ranked percentages of requirements for file submissions**



Standardized metadata, accompanying auxiliary information, licensing agreements and submission agreements all provide crucial information about how to access or use files and about any ongoing restrictions that may exist.

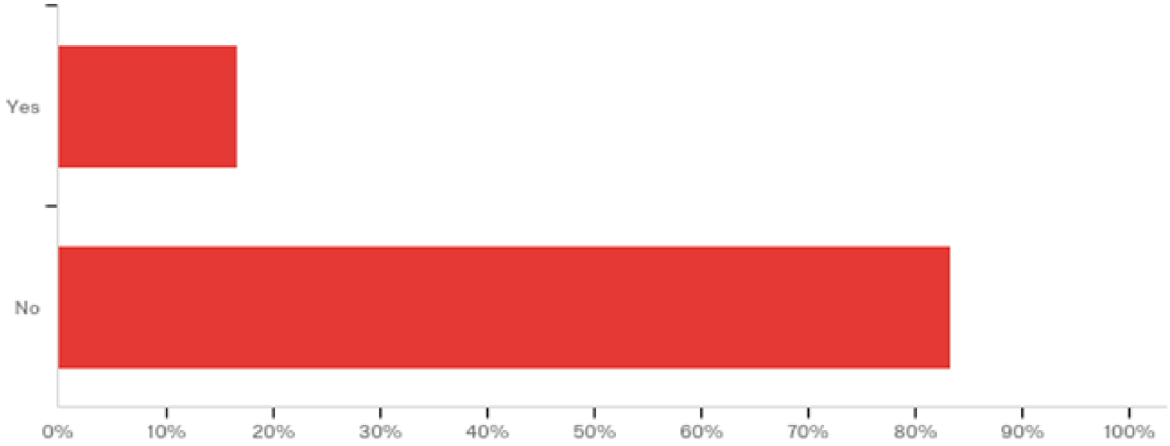
The remaining three conditions: standardized file formats, standardized file names and an accompanying DMP, were mostly related to internal procedures. These conditions were rarely required, and only by less than 10% of respondents. At 13%, submission requirements were more likely to be "not applicable" to respondents.

A similar percentage of respondents specified "other", and text comments clarified that rather than enforce a "formal requirement", the minimum was acceptable and additional provisions were encouraged. One comment noted that “We see most of the above as 'good practice' but getting submissions is difficult enough without insisting on specific requirements”. Although receipt is only the first step in the curation lifecycle, compromise was a recurring theme throughout the survey findings.

4.2.2.2. Appraisal and Selection

The second recurring theme was the neglect of actions related to preservation procedures. A majority of respondents reported that their institution did not have a process in place for appraisal and selection. 20 out of the 24 responses (83%) were recorded as having *no* process (Figure 7).

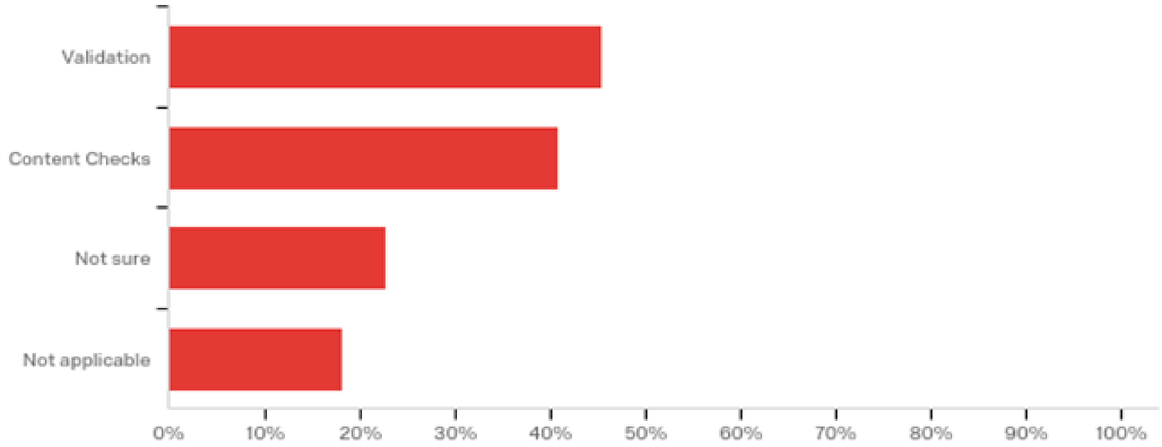
Figure 7. **Percentage of respondents with appraisal and selection processes**



4.2.2.3. Validation and Content Checks

Validation and content checks were more prevalent than appraisal and selection, with validation being slightly more likely to be conducted than content checks (Figure 8).

Figure 8. **Percentage of validation and content checks performed**



However, 23% of responses claimed uncertainty about the status of these processes, and for 18%, validation and content checks were not applicable to their institutions. This question was flagged as difficult to answer by an open-ended comment due to a "non-restrictive deposit policy" at their institution; data tends to be accepted *as is*, and researchers are expected to carry out the work of validating and checking their own content. If similar deposit policies are in place at other institutions, respondents may have been "unsure" or believed the activities were "not applicable" because these tasks would have been in the domain of researchers rather than the repository or related department. Likewise, this type of policy may also provide an explanation for the lack of appraisal and selection processes.

4.2.2.4. Ingest

In a collection of open-ended answers about preparing files for ingest, the main topic of discussion was the role of researchers. Briefly discussed in the survey results for validation and content checks, a summary of the open-ended answers revealed that the bulk of responsibility for file preparation lay with researchers. Those in data curation roles or departments more commonly acted as facilitators for researchers and rarely interacted with the data deposit directly. "Self-deposit" emerged as a regular standard. During deposit, files were either prepared by researchers and accepted as submitted, or submissions were checked for "required elements" or "general review" and spot checked for additional criteria.

Exceptions were not uncommon. One institution, still in the early stages of establishing deposit guidelines, currently prepares files on a “case by case basis” after a discussion with the researcher about necessary information, such as “retention, licensing etc.”. Another respondent mentioned a similar process, and cited their preparations were “varied between different disciplines”. In contrast, a third institution reviewed files not only for content and metadata but also confirmed the desired presentation and arrangement of data with the researcher.

Responsibility for metadata requirements, however, was shared between repository staff and the researcher. Metadata that was outside the typical purview of a researcher, such as “preservation and technical” details, were handled by staff, while “discovery and administrative metadata” were supplied by the researcher. Although staff were not responsible for originating administrative metadata, they were often responsible for ensuring that necessary legal and ethical documentation, such as consent forms or “third party material”, was present.

File type and size were also a concern. Compressed files were often requested, in an open format if possible. However, open formats were not a requirement.

An additional notable theme that emerged from the comments was the distinction between active storage and long-term storage. Several respondents noted a difference and included indicators to clarify their responses, such as “active storage” and “published data”. One respondent described answering from the “repository perspective” to indicate the following comments would only be about research data that was stored in the institutionally-based repository. It was unclear whether respondents were responsible for all stages of data storage, but the distributed storage of data indicated the complexity of the research lifecycle and the multitude of provisions that are necessary for complete data management.

Although comments were largely focused on long-term storage, many did not address prepping for preservation or explicitly mentioned there are no curation policies currently in place at their institution. At least one person mentioned that while preservation standards would be ideal, their current repository software does not support preservation efforts.

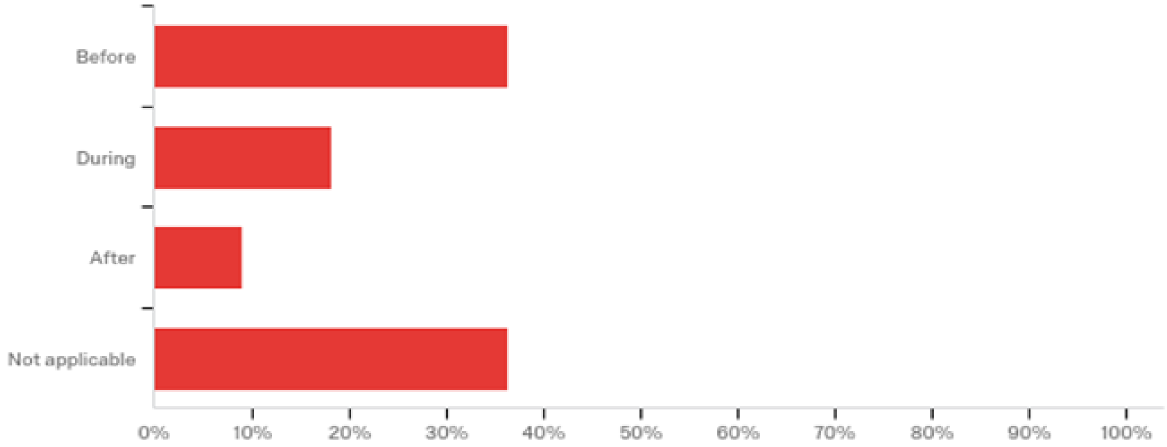
4.2.3. Metadata

Similar to Acquisition, activities related to Metadata were performed half by researchers and half by the relevant institutional department. Efforts focused on facilitating storage and access, but standard practice was inconsistent.

4.2.3.1. Standards

In equal measure, quality control (QC) of metadata either occurred before ingestion, which is considered best practice, or was "not applicable" (Figure 9).

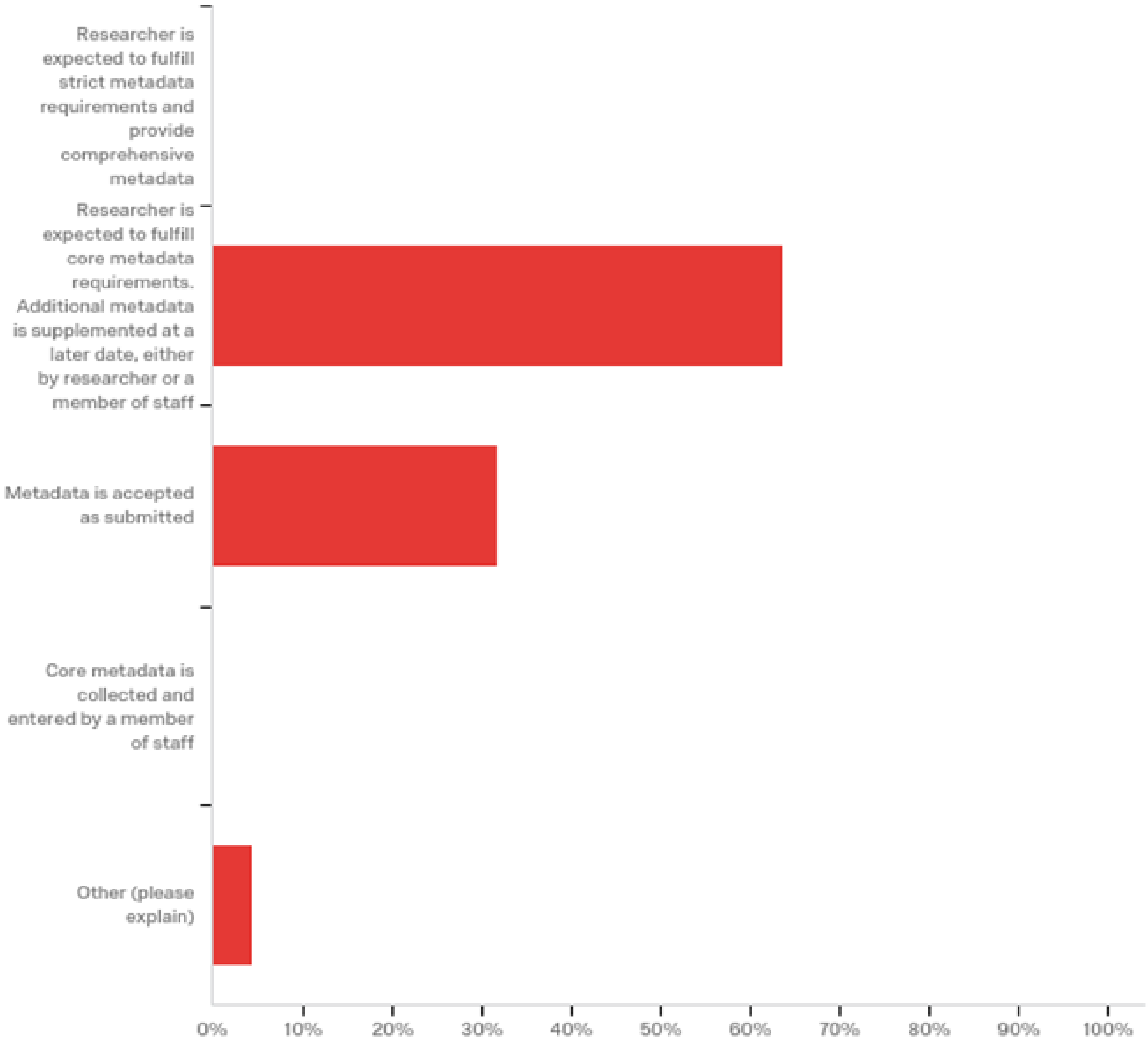
Figure 9. **Percentage of QC during ingest**



"Not applicable" could mean institutions did not have QC procedures in place, especially considering that data files tended to be self-deposit. A sizable 27% of respondents stated that QC is performed either during (19%) or after (9%) ingest.

Although standards for QC procedures were inconsistent amongst respondents, there was a majority consensus about metadata requirements for researchers (Figure 10).

Figure 10. **Comparison of metadata requirements for researchers**



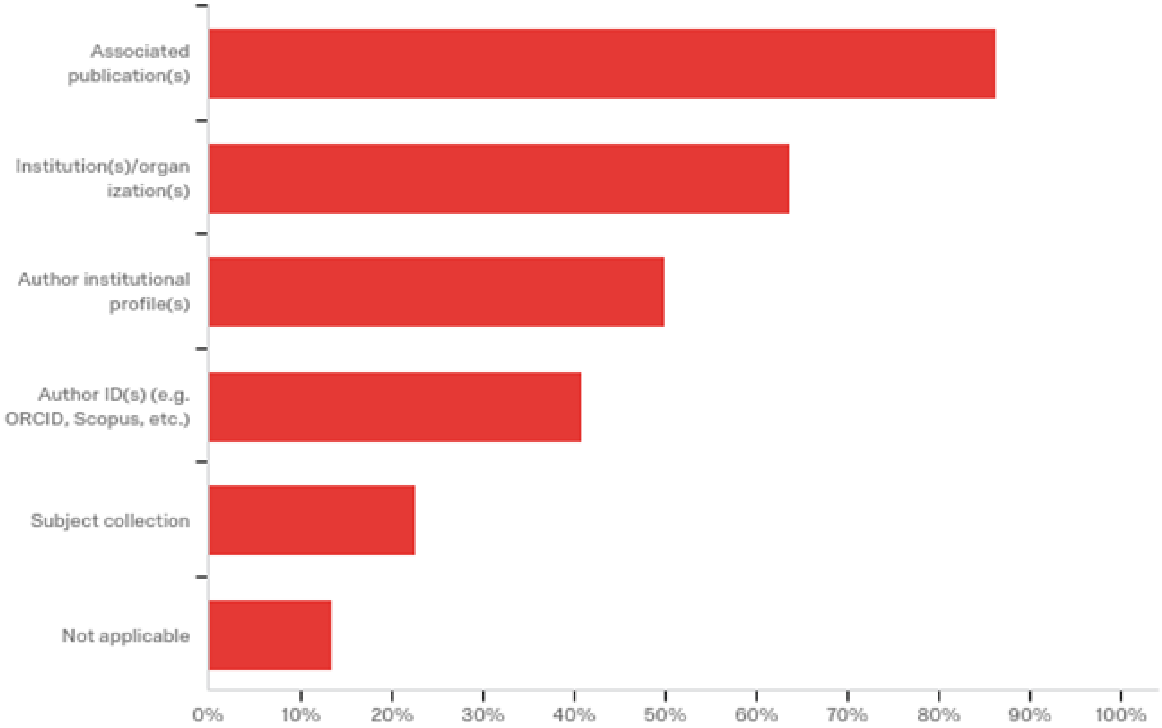
Primarily, 64% of respondents expected the researcher to fulfill *core* metadata requirements. Additional metadata is, or may be, supplemented at a later date, either by the researcher or a member of staff. Otherwise, 32% accepted metadata as submitted. One response (5%) stated their metadata requirements were "not yet known" under the "other" option. In no cases was metadata the sole responsibility of either the researcher or a member of staff.

4.2.3.2. Descriptive information

Information related to context, provenance, and access rights were all expected as standard metadata.

For contextual metadata, direct relations to the host institution seemed to be favored over author records. Deposited datasets were mostly linked to records related to associated publications and the corresponding institution or organization (Figure 11).

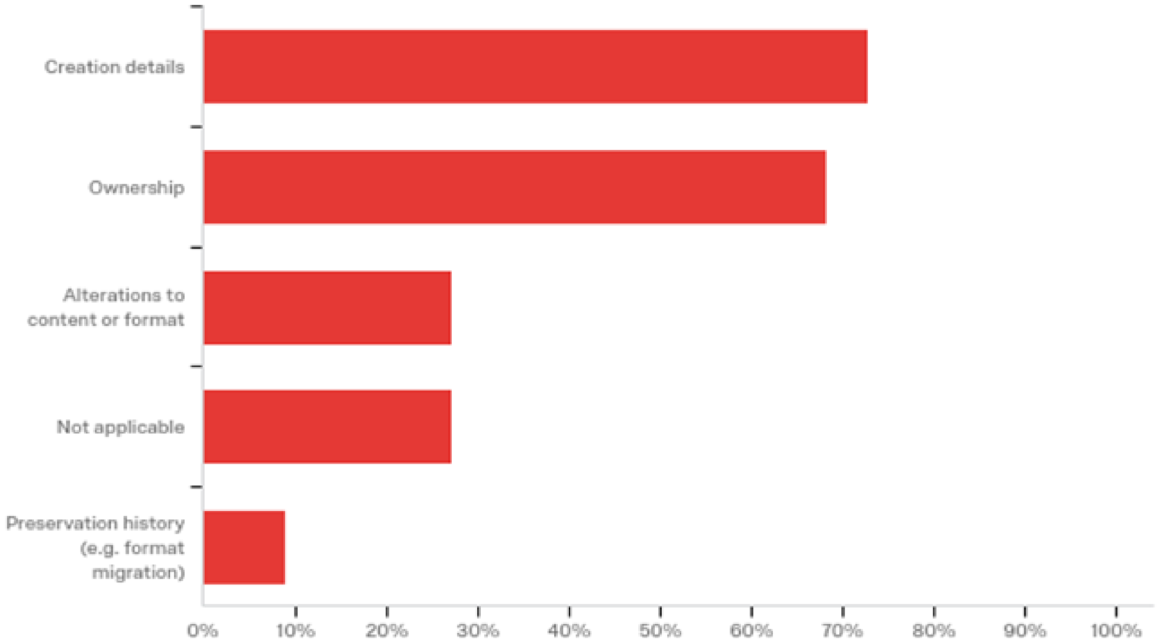
Figure 11. **Comparison of context metadata**



Associated publications were linked to almost all datasets (86%), while over half were connected to institutions or organizations (64%). Author institutional profiles were included to a lesser extent, with only half of respondents (50%) linking the related records. Author IDs, such as ORCID or Scorpus, were linked to even less than author institutional profiles (41%), although unlike author institutional profiles, author IDs are unique and persistent. Nearly a quarter of datasets were linked to a related subject collection (23%). This was a small margin in comparison to the other responses but was more than expected.

While the inclusion of provenance metadata was also standard, compared to contextual metadata, provenance metadata was less common (Figure 12).

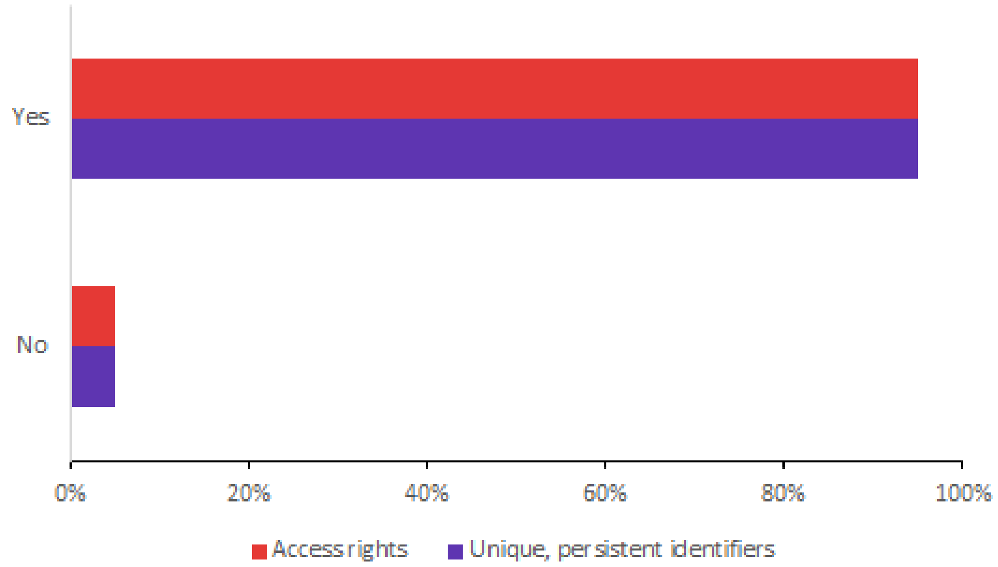
Figure 12. **Comparison of provenance metadata**



13% of responses claimed that provenance information was “not applicable”. When provenance was included, the information tended to relate to the origination and status of a file. 73% included “creation details”, and 68% included “ownership”. Information about content or format alterations was occasionally included (27%), but information about preservation history was the least likely to be included (9%). However, given the lack of current preservation practice, metadata related to preservation would be less necessary.

Finally, both access rights and unique, persistent identifiers (e.g. DOIs) were included in records by the majority of respondents (Figure 13).

Figure 13. **Percentage of respondents that record access rights and DOIs**



In both cases, 95% of respondents confirmed the inclusion. Both questions also received an equal amount of responses, and along with the high rate of affirmation, there was the suggestion of a strong connection between the two results.

4.2.3.3. Representation information

A review of the auxiliary information required by different institutions again revealed a lack of consistency. 3 out of the 14 open-ended responses (20%) stated auxiliary information was “encouraged” but entirely optional. When auxiliary information was required at all, the most commonly requested information was related to three categories:

- Content
- Context
- Access rights

Content information could be characterized as details about what is contained within the submission files and the tools necessary to read or render that content. For the most part, this included file details such as:

- Number of files
- File names
- File formats

- Size of files
- Summary of the research

This also included technical details, such as:

- Operating systems
- Software requirements, including version

For at least two institutions, this information was considered beyond the scope of a researcher and was instead supplied through a preservation system that automatically extracted the information.

Other content information that was mentioned but seemed unique to specific institutions included:

- “Time periods”
- “Geographic location”
- Information tailored “for tabular data”

This information would ideally be contained in the form of a readme.text with a “how-to” or user guide attached.

Context information was mostly related to the associated publication information, such as the research project title and its assigned DOI. In one instance, this included “related resources”, although further clarification was not provided.

Access rights specifically involved embargoes and other restrictions on access.

Notable information that was requested but did not fit into one of the three categories included: “Twitter handles” and information related to funding, such as funder and “grant numbers”.

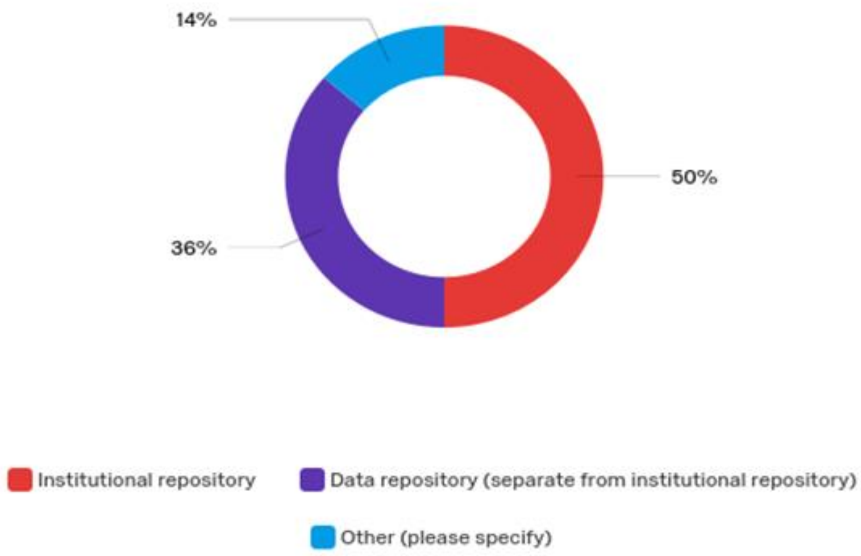
In addition, a couple of respondents commented that metadata requirements “depends” or “varies” based on the discipline.

4.2.4. Storage

Questions about storage returned the highest ratio of responses that conformed to best or recommended practices. The response rate for these questions was consistent with previous questions, confirming that storage was either prioritized over other areas of data curation, or storage practices were more straightforward to accomplish. Despite the purported success, storage had similar issues as other areas with neglecting preservation.

In terms of physical location for storing datasets, institutional repositories were more frequently used than data repositories (Figure 14).

Figure 14. **Percentage of reported storage options**

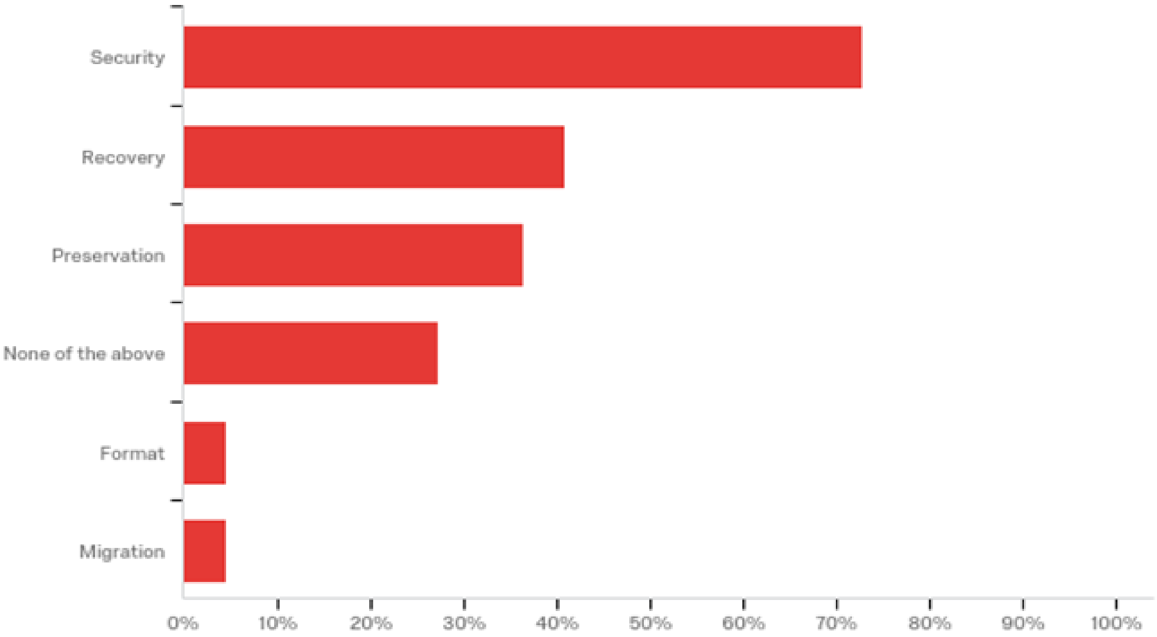


50% of respondents reported IRs as their hosting platform, while 36% used data repositories. 14% reported “other” options, including a data catalogue and availability of both an IR and a data repository. One response stated that their options were “not yet determined” due to the newness of their services.

4.2.4.1. Documentation

Storage policies were most likely to cover issues related to Security, Recovery, and Preservation (Figure 15).

Figure 15. **Comparison of aspects of storage policies**

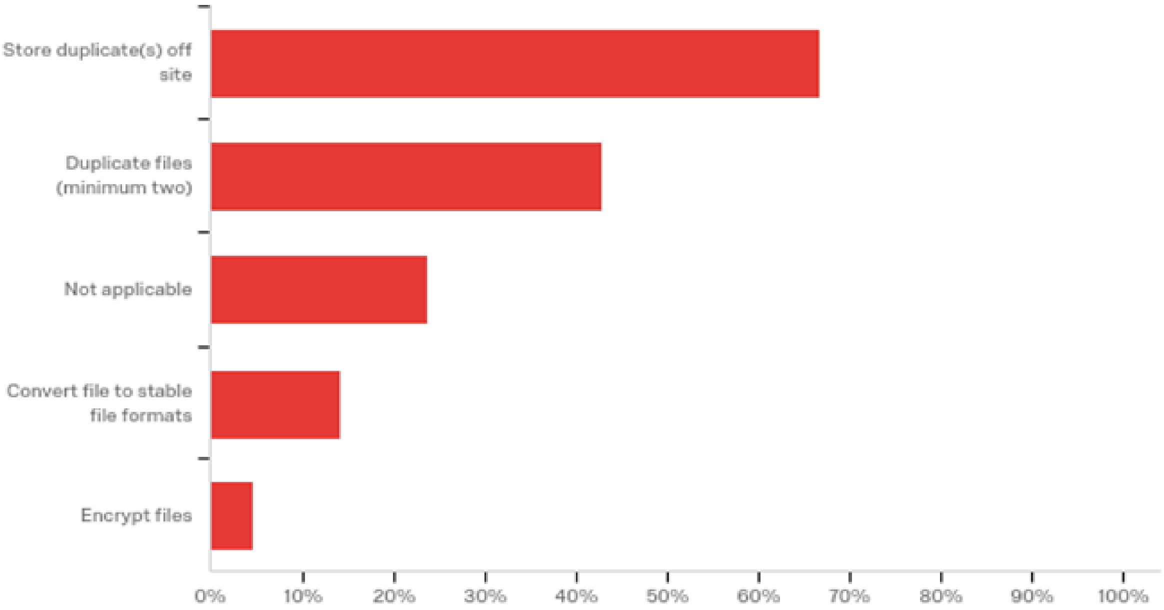


Security was the most prevalent in institutional policies, with 73% of respondents confirming its coverage. In comparison, Recovery (41%) and Preservation (36%) were included in less than half of institutional policies. Otherwise, 27% of institutions covered none of the typical aspects of storage in their policies. Finally, Format and Migration were only included in 5% of policies. However, file formatting and hardware migrations are specific preservation actions and may have been less likely to be included in policies due to the specificity.

4.2.4.2. Security

In accordance with recommended practices, over half (67%) of institutions duplicate their files and store the duplicates off-site, and less than half (43%) duplicate their files without storing the duplicates off-site (Figure 16).

Figure 16. **Comparison of actions to secure files**



Respondents were allowed to select all options that applied, therefore the disparity could be explained by poor question design. Otherwise, if the results are taken at face value, 43% of respondents recognize the recommended course of action but are either unable or unwilling to complete the full course of action. This could indicate multiple issues, such as inadequate storage infrastructure or an inability to enforce storage policies.

Files were less likely to be converted to stable file formats (14%) and not likely to be encrypted (5%), both of which are processes related to preservation. For nearly a quarter of institutions (24%), securing data files was “not applicable”.

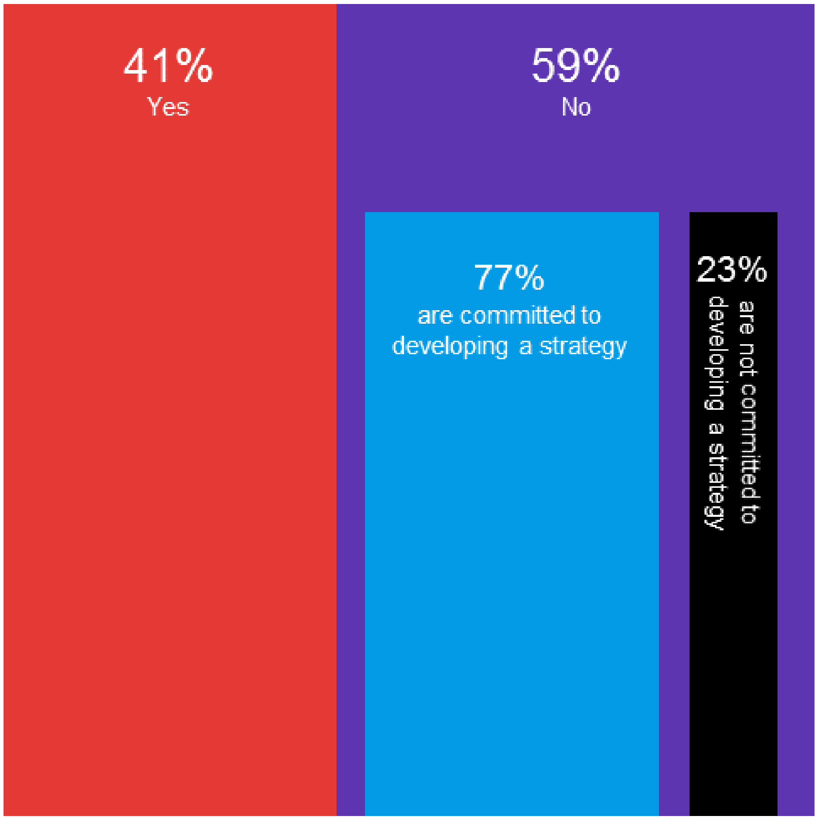
4.2.5. Preservation

The survey found that while preservation is not a priority currently, actions are being taken to establish preservation as a priority in the future, including the development of long-term strategies. When preservation solutions were enacted, they were typically uncomplicated and required limited support or infrastructure.

4.2.5.1. Long-term strategy

The numbers for institutions with a strategy for long-term preservation of data were almost evenly split, with 41% of respondents answering “yes” to having a long-term strategy and 59% answering “no” (Figure 17).

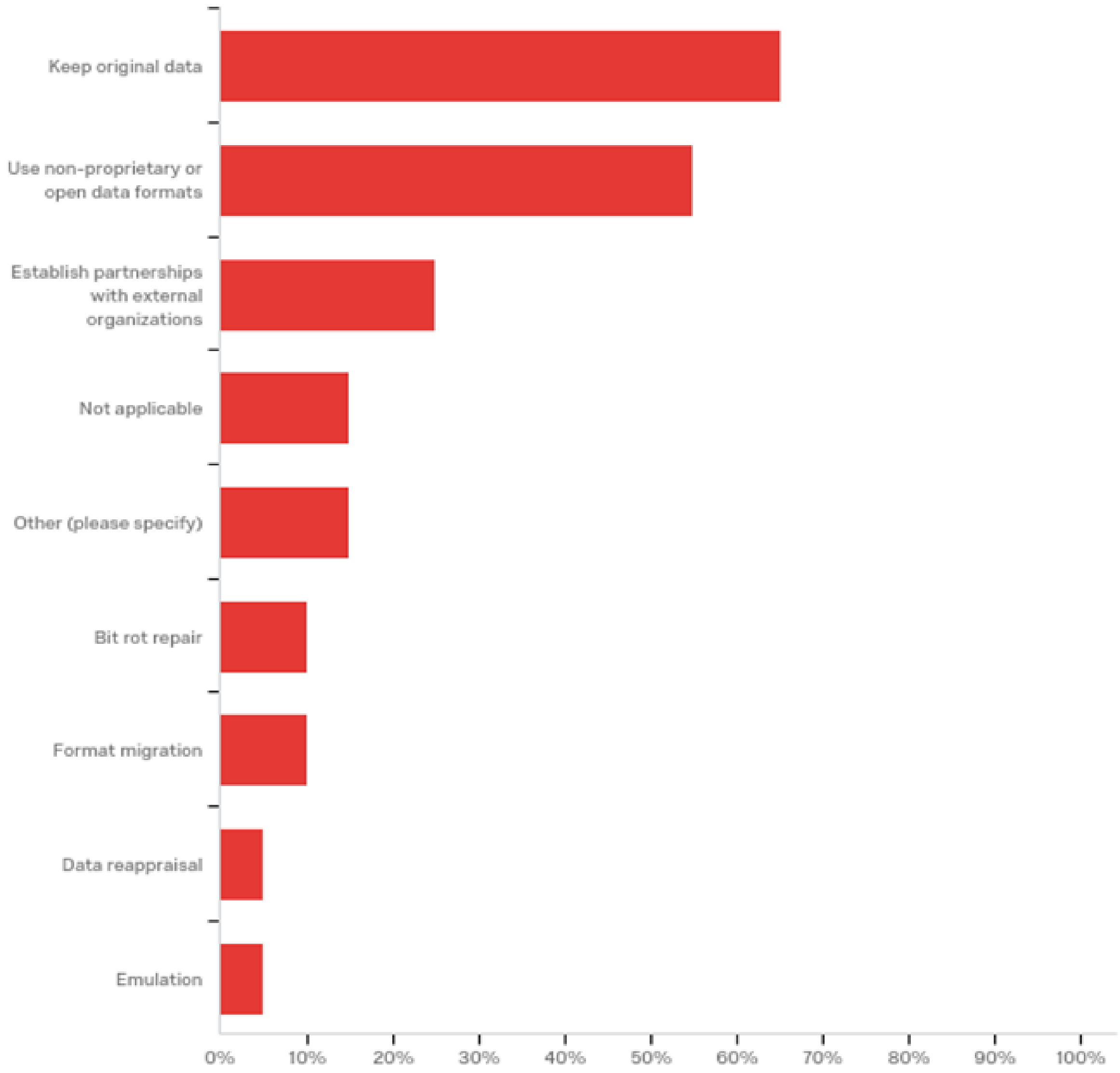
***Figure 17.* Percentage of respondents with a preservation strategy**



However, the majority of those without strategies, 77%, were committed to developing one. Despite this, nearly a quarter (23%) have no long-term strategy and are not committed to developing one at all.

Out of the available options listed as viable preservation solutions in the survey (Figure 18), the two most popular solutions were also the most straightforward: keeping original data and using non-proprietary or open file formats.

Figure 18. **Comparison of preservation solutions**



A majority of institutions (65%) keep their original data, and over half (55%) adopt open solutions. 25% establish partnerships with external organizations. 10% or less perform bit rot repairs, format migrations, data reappraisals, or emulations.

15% employ “other” alternatives, such as the use of digital preservation software. Two systems were named as primary tools, Archivematica and Preservica, for their in-built format migration mechanisms.

One response also made a distinction between *institution* and *repository* solutions:

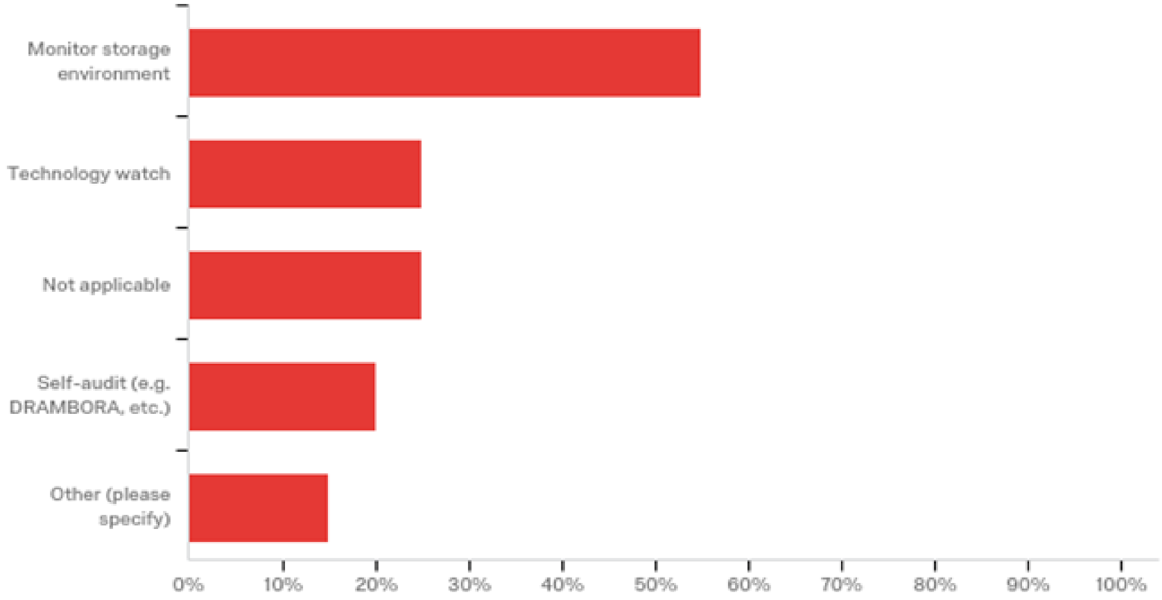
> *There is no institution wide solution. What applies to data deposited in the repository and data managed elsehwere* (sic) *on the University network*.

This response implied that solutions for research data could be inconsistent across an institution and dependent on the managing department. Similar sentiments were expressed in open-ended comments regarding Acquisition, signifying both storage and management of research data are often not unified under standard procedures even on the same campus.

4.2.5.2. Risk Assessment

When performing risk assessment procedures, at 55%, respondents most frequently chose to monitor the storage environment (Figure 19).

Figure 19. **Comparison of risk assessment procedures**



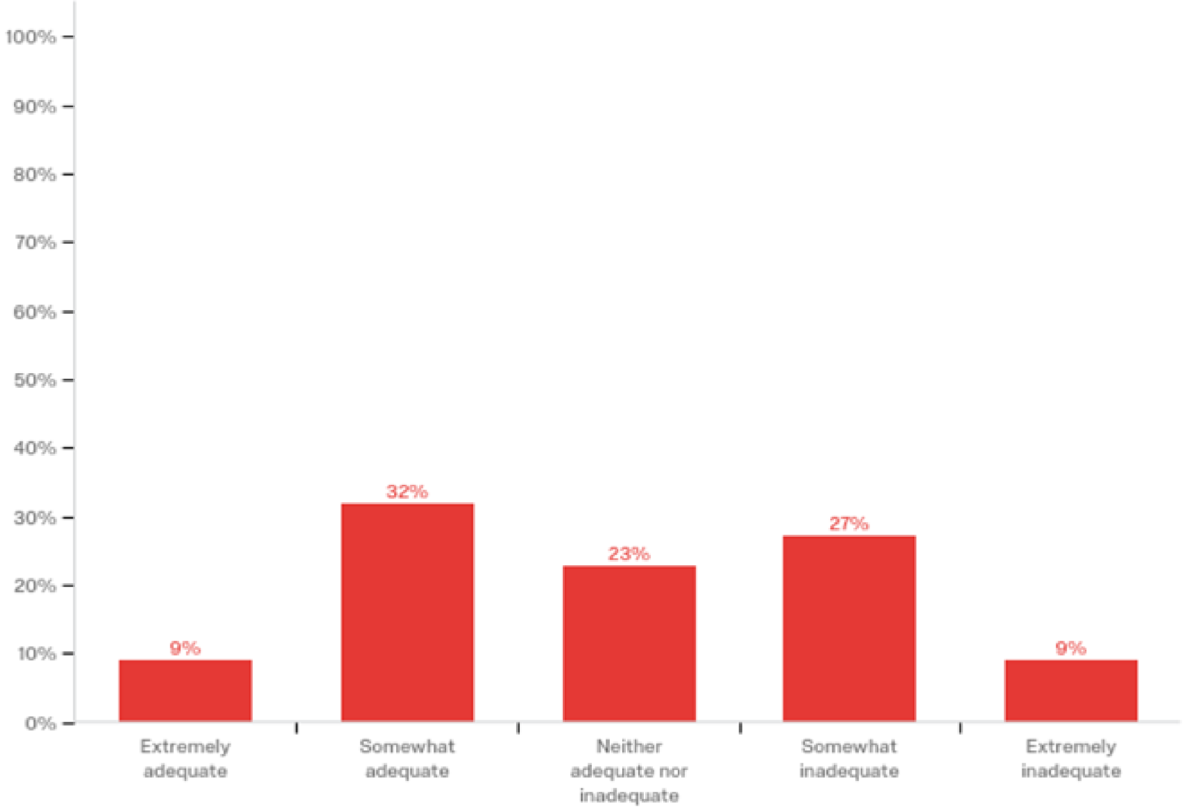
In addition, a technology watch was incorporated into a quarter (25%) of the procedures. However, for the same number of respondents (25%) the question was "not applicable", meaning none of the listed procedures or no procedures at all are undertaken to assess risk of data loss or degradation. Closely followed were self-audits, conducted by 20% of respondents.

Clarification comments were left under the "Other" option by respondents who were setting up preservation systems that included risk assessment features and by a respondent who wished to indicate their answer constituted the institutional repository but not “the institution as a whole”.

4.2.6. Administration: systems operations

In response to the level of software support received from their respective institutions, respondents’ ratings ranged broadly (Figure 20).

Figure 20. **Ratings for institutional software support**

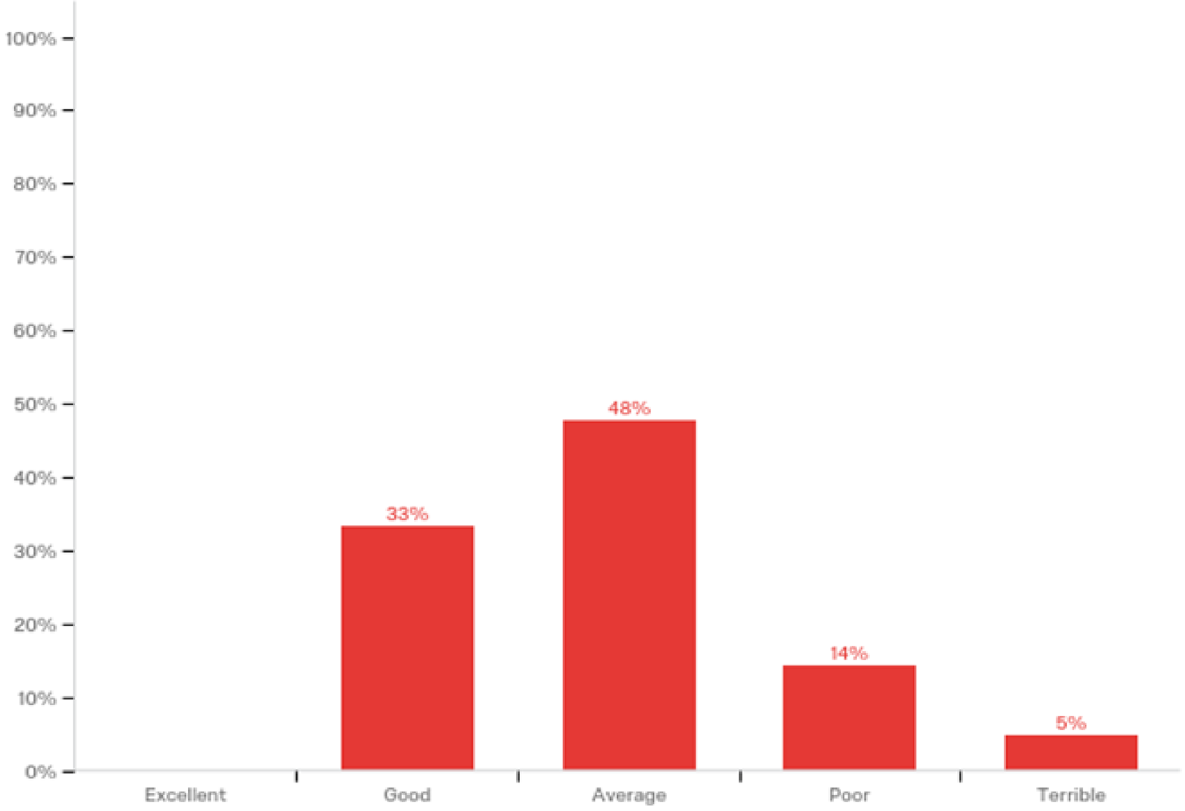


A higher percentage of respondents (32%) rated their support as “somewhat adequate” compared to those who received “somewhat inadequate” support (27%). However, a closer look at the count between each rating revealed an almost even divide: 7 out of 22 responses and 6 out of 22 responses, respectively. Closely, and in the same range, 5 out of 22 (23%) believed their level of support was “neither adequate nor inadequate”. This could indicate an absence of support, with no systems or operations available to rate, or support that exists but does not evoke strong opinions. An even number of respondents, 2 out of 22 (9%) for each rating, believed their institutions were either “extremely adequate” or “extremely inadequate”.

4.2.7. Access: FAIR

Most respondents felt that their institution’s fulfillment of FAIR Data Principles was “average” (48%) or “good” (33%) (Figure 21).

Figure 21. **Ratings for fulfillment of FAIR**



A combined 17 out of 21 responses (81%) rated their institution as one or the other. 3 (14%) believed their institution was poorly achieving FAIR, and 1 (5%) strongly believed their institution was “terrible”. The favorable majority reflected the emphasis that was placed on access throughout the survey results, particularly during Acquisition and Metadata stages.

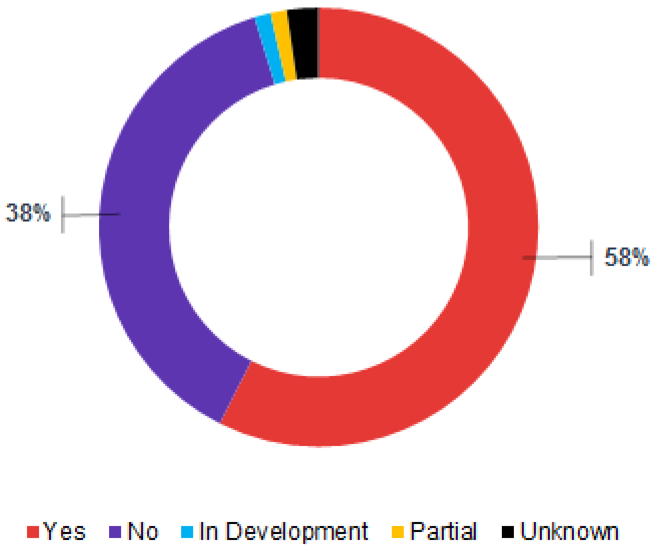
4.3. Appraisal of research data infrastructure

Establishing a count of data services, policies, and repositories revealed that 50-60% of institutions offer some form of support for research data management. In addition, the count and subsequent appraisal provided context for the survey results, and in certain cases, such as with the repository count, the survey results did not accurately reflect the current situation.

4.3.1. Data services

Out of 169 recognized higher education institutions, 97 (57%) offer apparent services related to research data management (Figure 22).

Figure 22. **Percentage of research data services**



Of these 97 institutions, 91 institutions (94%) share their resources with the public. 64 institutions (34%) provide *no* data services. For the remaining institutions:

- 2 (1%) are developing services
- 2 (1%) offer partial services
- 4 (2%) have an unknown service status

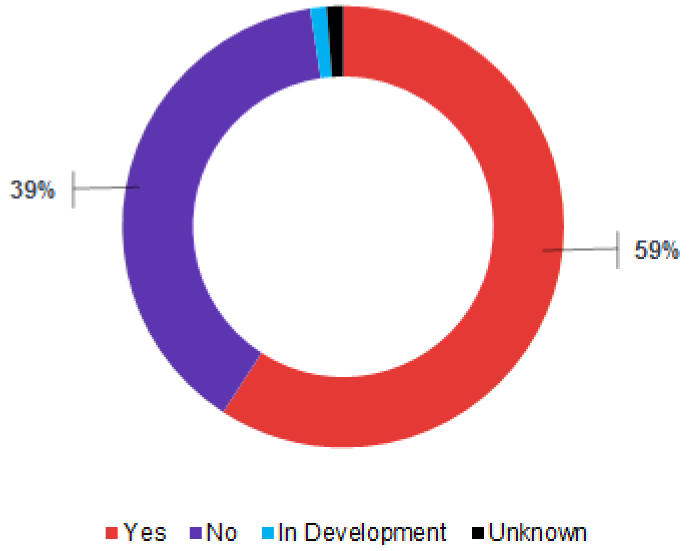
The survey did not gather data on the provision of data services, therefore a direct comparison between the survey results and the service count is not feasible. However, a close approximation may be found in the percentage of respondents who worked closely with data curation or management. Presumably, the existence of a role or responsibilities related to data would imply the availability of data services, if only partially. The survey results (91%) were significantly higher than the percentage of actual services (58%), but both reflect a majority.

Responsibility for data services at each institution primarily resided with the library, while the research office often acted as a secondary alternative. At 59 out of 97 institutions (61%), the library appeared as either a sole or primary host for data services, while a research office occupied the same role at 21 institutions (22%). In one particular instance, the research office was represented by “multiple research support departments” that were assigned to different faculties. Partnerships between departments were also common, with 19 RDM services being operated under the umbrella of both the library and research office or through shared responsibility between the library, research office, and several other departments, such as IT or faculty. One service was a unique case that was run exclusively by the “doctoral academy” of the institution.

4.3.2. Policy count

A total of 100 policies (59%) governing research data management were located out of 169 institutions (Figure 23).

***Figure 23.* Percentage of institutional policies**



This total included:

- 90 public policies
- 6 non-public policies
- 3 drafts
- 1 “partial” policy that was subsumed within a larger policy.

Of the 90 publicly available policies, 2 policies were exceptions:

- 1 policy was shared between institutions
- 1 policy was shared publicly but was unavailable for review due to a broken link

Out of the remaining 69 institutions, 65 institutions (38%) had no policy. Two institutions (1%) had policies in development, and the remaining two institutions had an unknown policy status.

The total count of actual policies corroborated the majority reported by the survey results. However, the high percentage of reported policies in the survey (91%) compared to the relatively lower percentage of actual existing policies (59%) suggested a skewed demographic of survey respondents. This may indicate respondents were more likely to engage with the survey if there were institutional policies or support already available for research data.

The percentage of existing institutional policies (59.1%) corresponded closely to the percentage of data services offered (57.3%), suggesting a connection between administrative interest and departmental support for data curation. Exceptionally, eight institutions offered services but have no obviously related or attached data policy.

4.3.3. Policy dates

There has been an obviously steady increase in either the creation or review of policies in the last four years (Table x).

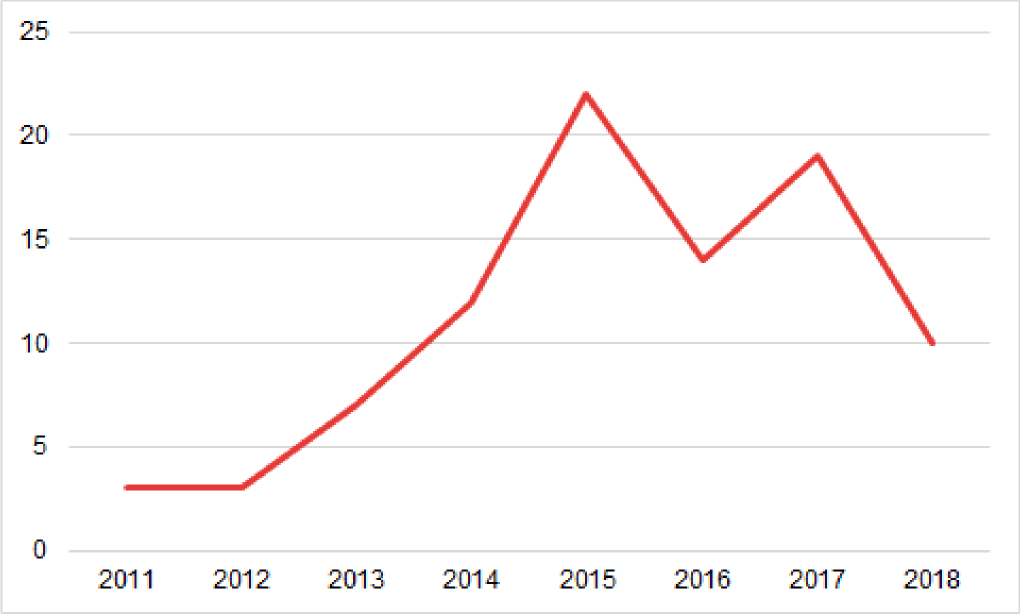
Table 3. **Institutional policy dates**

Year	Policy Count
2011	3
2012	3
2013	7
2014	12
2015	22
2016	14
2017	19
2018	10

Of the 94 public policies, four were undated, and these policies were not included in trend calculations.

The most significant growth occurred between 2014-2015, with a 45% increase in institutional policies (Figure 24).

Figure 24. **Trend in policy growth by year**



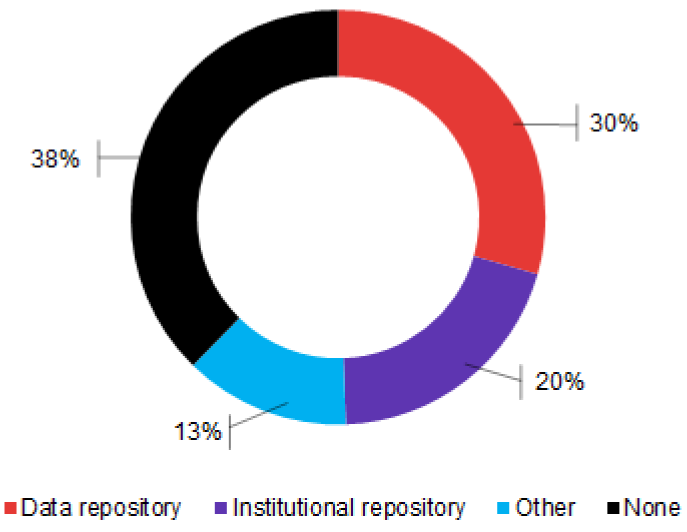
Between 2015-2016, however, there was a sudden decrease (57%). A smaller spate of policies occurred between 2016-2017 (26%), and as of the current day, there has been another reduction in the number of policies that have been generated. However, the numbers for 2018 may continue to increase as the year develops.

Beyond the trend of policy growth, there did not seem to be further connections between policy date and other aspects of institutional support or coverage of research data. For example, policy dates did not have an impact on the existence of data services. There were only four examples of institutions with a policy and without an apparent RDM service: one from 2016, one from 2018, and two that were undated. In addition, attached procedures or guidelines were not less likely to be included in policies depending on the policy date. The only noticeable connection was that undated policies tended to be more lacking, although this was not the case for one out of the four undated policies.

4.3.4. Repository count

Contrary to the survey results, more data repositories were available to host research data than institutional repositories (Figure 25).

Figure 25. **Percentage of storage solutions**



45 institutions currently provide data repositories, and five additional institutions are developing data repositories. 34 institutions provided institutional repositories as a primary platform for research data storage. One institution offered both, and one other was developing a shared repository. 20 institutions suggested alternative options, such as a subject-based repository or a data catalogue. 64 institutions offered *no* options or solutions for storage of research data.

The discrepancy between reported repository usage and the final repository count could be due to several reasons, including inactive data repositories, the misrepresentation or dual use of IRs as data repositories, or a skewed ratio of respondent demographics. The presence of a repository also did not guarantee the existence of publicly apparent RDM services or policies. Ten HEIs had repositories but had neither services nor policies.

4.4. Results of policy analysis

Although most of the main categories from the evaluation framework were covered by the scope of most policies, there were certain incongruities between reported practice and institutional policies. Where policy and practice matched, such as with the assignment of responsibilities or in support of open data, a case could be made for the impact of policy on practice.

4.4.1. Scope

Both the survey and the policy analysis assessed the presence of the 6 main procedure categories from the evaluation framework: Acquisition, Metadata, Storage, Preservation, Administration, and Access.

The survey gathered information about documented procedures or standards from 27 respondents. Of the listed categories, respondents were asked to select all that applied. Ranked from most to least included, respondents reported documentation for the following at their institutions:

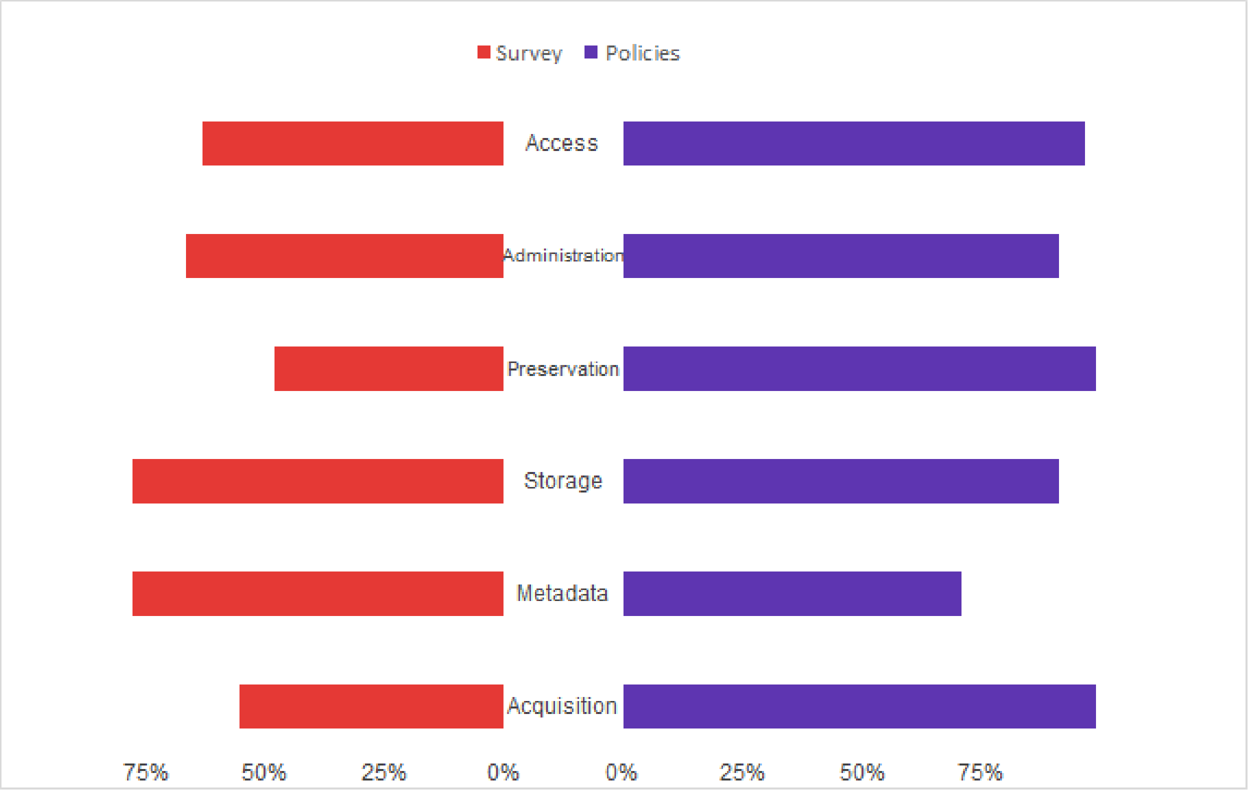
- 21 (78%) included Metadata
- 21 (78%) included Storage
- 18 (67%) included Administration
- 17 (63%) included Access
- 15 (56%) included Acquisition
- 13 (48%) included Preservation

Of the 100 policies found, there were 92 unique, publicly accessible policies, including drafts and a partial policy. Ranked from most to least included, actual documentation mentioned the following categories:

- 91 (99%) included Acquisition
- 91 (99%) included Preservation
- 89 (97%) included Access
- 84 (91%) included Storage
- 84 (91%) included Administration
- 65 (71%) included Metadata

There was a noticeable contrast between the survey results and the findings from the policy analysis (Figure 26).

Figure 26. **Comparison between survey results and policy analysis**



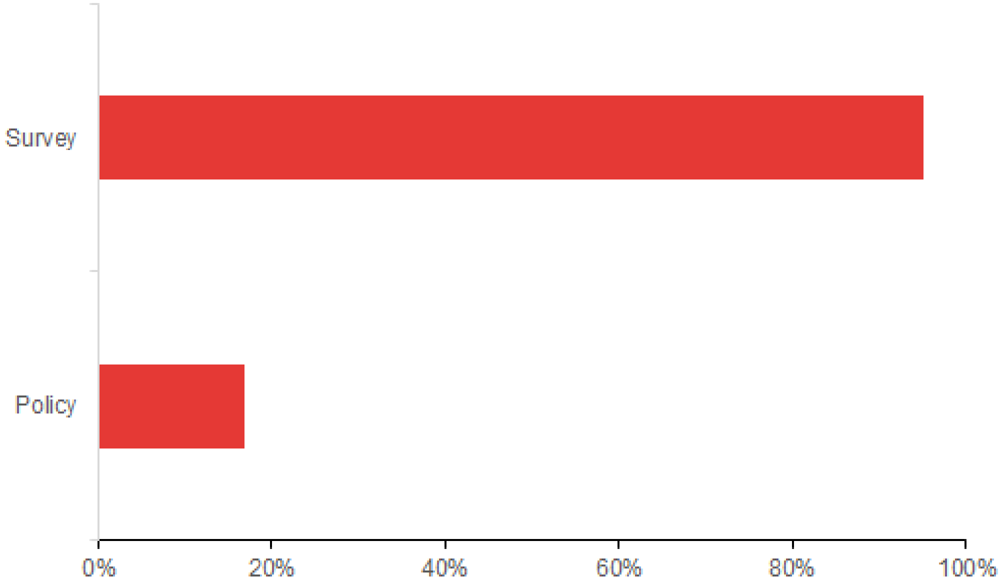
The lack of a connection between a category’s coverage within policies and its reported coverage in the survey could be explained by a couple of reasons.

The policy analysis only looked at public policies, and these policies tended to be short documents that assigned responsibility rather than provide guidance. Subsequently, the policy analysis also focused on determining the scope of policies rather than examining specific procedures or standards. Therefore, the disparity could be explained by additional, internal documentation that is not available to the public. For example, according to the survey, universities are most likely to have documented procedures or standards for metadata. However, the policy analysis conveyed that metadata received the least attention within institutional documentation.

The results could also imply that institutional documentation does not accurately represent current practice. For example, although preservation was included in almost all institutional policies, survey respondents reported preservation received the least amount of documentation. This is further reflected by the current lack of a long-term strategy for preservation at many universities and the current adoption of only basic preservation solutions, despite the inclusion of preservation in 99% of institutional policies.

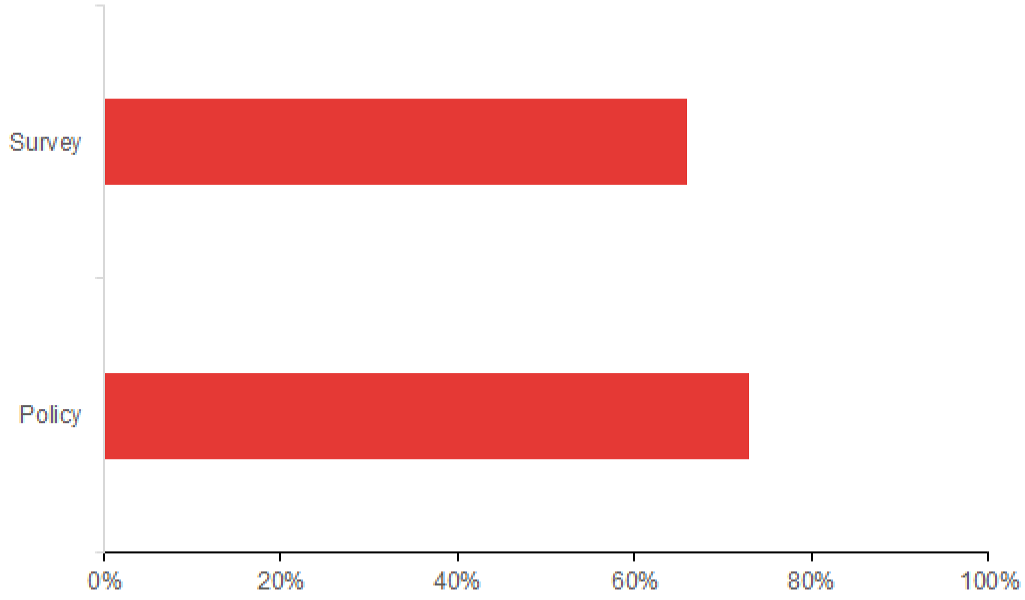
However, a directed look at specific aspects of policy coverage did not clarify the disconnect between the scope of policies and standard practice. In both recommended practice and standard practice, unique, persistent identifiers, commonly in the form of DOIs, are necessary for each record. Compared to the 95% of survey respondents who employ DOIs, only 16 institutional policies (17%) include DOIs as a requirement (Figure 27).

Figure 27. **Comparison of DOIs in practice vs. in policy**



In a different instance, the policy analysis was more closely aligned with reported policy. Security, as reported by survey respondents, was included in 73% of storage policies. Security was a present concern in 62 institutional policies (66%) (Figure 28).

Figure 28. **Comparison of security as reported vs. in policy**

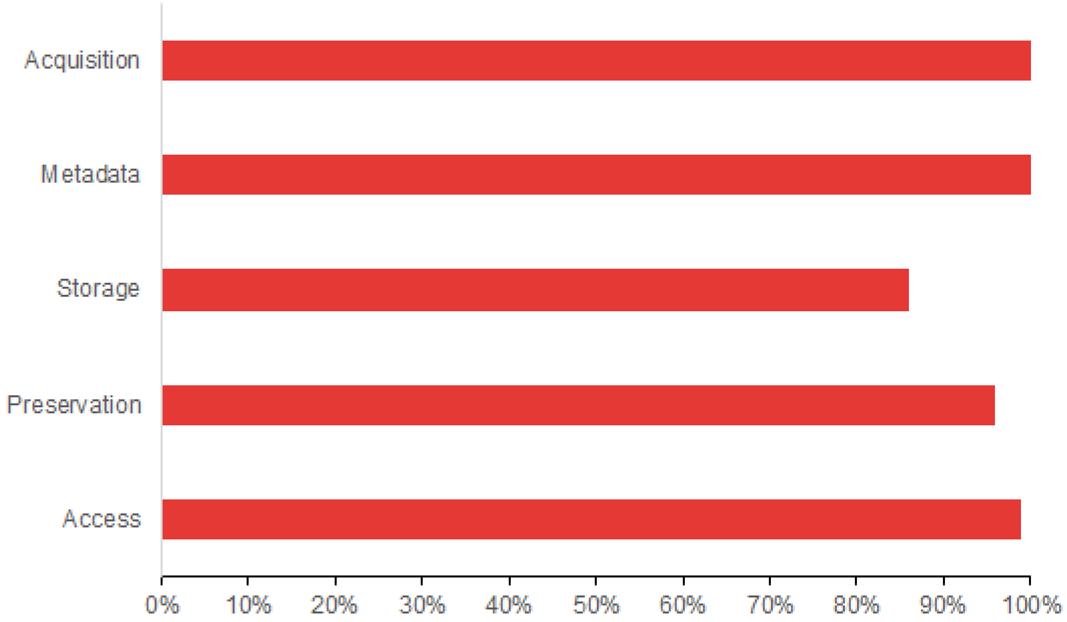


4.4.2. Assignment of responsibility

Responsibility for data curation or management procedures was primarily assigned to researchers. From the public policies that were analyzed, researchers were responsible for 5 out of 6 procedural categories. As each category was not equally represented in every policy, percentages of responsibility were calculated per category, rather than using the total number of policies. For example, Acquisition was included in 91 out of 92 policies, therefore responsibility could only be assigned within 91 policies, and the percentage was then calculated out of 91 policies. For each category, chief responsibility was assigned to researchers as follows (Figure 29):

- Acquisition: 91 out of 91 policies (100%)
- Metadata: 65 out of 65 policies (100%)
- Storage: 72 out of 84 policies (86%)
- Preservation: 88 out of 91 policies (96%)
- Access: 88 out of 89 policies (99%)

Figure 29. **Comparison of responsibilities by procedural category**



Administration was the only category where researchers were not held accountable, with only 4 out of 89 policies (4%) attaching responsibility to researchers. Administrative duties were instead supported by the Institution at 72 locations (80%).

Not including Administration, researchers were, on average, responsible for 96% of the overall duties involved in data curation and management. These findings were supported by the survey responses, which described "self-deposits" as the norm.

Storage was a category where researchers were less likely to hold complete or sole responsibility. Responsibility for storage was more likely to be a cooperative effort between the researcher(s) and multiple departments, including the library and IT. The diffusion of responsibility is likely due to storage involving additional infrastructure in the form of an institutional or in-house repository.

4.4.3. Terminology

Despite the clear line of responsibility laid out by institutional policies, survey respondents expressed frustration with the execution of data curation duties. The recurring disconnect between institutional expectations and actual curation functions was inherent in the terminology used to refer to data curation. As previously stated, "curation" and "management" are often conflated as synonymous terminology when discussing research data. While the term lacked majority usage, nearly half of institutional policies used “curation” interchangeably with “management”. 37 out of 92 policies (40%) used “curation” to refer to data management activities. The remaining majority either made no mention of “curation” or only used the term in a preservation context. The absence of “curation” as a managing descriptor highlighted the shortsighted mindset of institutional policies. In instances when “curation” is used to describe management duties, the confusion between “curation” and “management” is indicative of the confusion about how to describe responsibilities and functions and, consequently, how to effectively assign responsibility for different aspects of research data.

4.4.4. Support and guidance

The institutional assignment of research data responsibilities also highlighted a detachment from the departments responsible for hosting research data services. There was an apparent lack of connection between responsibility for providing data services and and the coverage of institutional policies. Providing data services did not signify an inclusion in institutional policies, even in relation to assigned responsibilities. For example, although a library or research office may be responsible for providing RDM services, their role is not more likely to be included in the policy.

In addition, although 57% of institutions provide data services, only 39% of policies included attached procedures or guidance in the document. 36 out of 92 policies included guidance, while 55 out of 92 (69%) policies did not include any form of practical support; this included not linking to existing institutional data services. One policy (1%) indicated attached procedures were in development.

4.4.5. Commitment to Open

82 out of 92 institutional policies included a reference to open data or open access and sharing of data. At 89% coverage, institutional support for Open scholarship reflected the combined 81% “average” and “good” ratings cited by survey respondents for their institution’s fulfillment of FAIR values.

4.5. Further context and future of data curation

A call for additional comments from the survey provided context about the current setting of data curation and supplied commentary about the actions necessary to ensure a successful future. There was a wide range of demographics that commented on the survey and on data curation, from those with well established repositories to those currently in development. This range of experience was also evident in flux of policy growth and appraisal of data services.

Respondents especially demonstrated a concern for the current distribution of responsibility and the difficulty in aligning motivations. As demonstrated through findings from both the survey and the policy analysis, present workflows place a heavy reliance on researchers. However, one respondent noted that researchers must be “willing to deposit”, but convincing researchers to deposit is only part of the issue. Another respondent stated that: “Getting researchers to actually engage with data management is STILL the biggest hurdle we face”. A couple of reasons were cited for the difficulty of engaging researchers, including skepticism about the open sharing of data and the already heavy workload of researchers. Although open data is supported by both repositories and institutions, researchers do not share the same sentiments. This may persist unless there is a “culture of change”, or there may be a shift “with time, and more evidence of the benefits of openly sharing research data”.

Respondents also discussed the obstacles of implementing good practices. Many commented that while they would prefer to improve their practices, they either lacked the funding, the staff, or the skills necessary to achieve that goal, in addition to relying on researchers to perform most of the work. In particular, the procedures and tasks described in the survey were considered to be a “wish list”.

Optimistically, there is a demonstrated development in the data curation field as more institutions invest in research data and related services.

5. CONCLUSIONS AND RECOMMENDATIONS

Conclusions to the research questions were reached by considering both the immediate and broad implications of the research findings, and recommendations have been proposed through a gap analysis of the research findings as examined through the lens of the literature review. Lastly, suggestions for future research have been presented after reflection on the dissertation and its possible future uses.

5.1. Standard definition of “data curation”

A definition of “data curation” was designated specifically for this dissertation to ensure clarity and to narrow the scope of the research. However, over the course of the research, the use of “data management” as a preferred term over “data curation” remained an issue and only further highlighted the “wicked problem” of data curation (Cox, Pinfield and Smith, 2016). Even with an issue as basic as terminology, there is no simple solution.

Implementing a standard definition or usage of “data curation” would be difficult because there is not a majority usage of the term in UK HEIs. Nevertheless, standardization would be a worthwhile effort considering nearly half of institutions already conflate “curation” and “management”. To start, reaching a consensus about a standard term, either “data curation” or “data management”, would help direct further conversations about a standard definition for the preferred term. The commonly preferred term in the UK is “data management”, perhaps due in part to the primary role of researchers. Although it would be more convenient to maintain “data management” as a standard term for research data activities, there are fundamental issues with its present usage that would make “data curation” a better alternative.

The present use of “data management” epitomizes the current culture at UK HEIs that prioritizes maintenance over long-term preservation and assigns primary responsibility for research data to researchers, who are often not equipped with the proper skills. Use of “data curation” would place an emphasis on preservation and on distributing responsibility to more knowledgeable staff, as skills related to curation are distinct from skills related to management. Promoting its usage might encourage a more consolidated discussion about sharing responsibilities and concentrate ongoing efforts to develop key skills related to curation. The use of terminology may shift automatically with time as preservation efforts increase. As the conversation around preservation starts to occur before the ingest stage, the term “curation” is more likely to be applied to the whole process, where right now preservation is a by-product, and as a result, curation is an afterthought.

An analysis of and comparison to global standard terminology is also recommended to confirm that UK term usage aligns with international standards. Considering the increasingly shared nature of research data through the growth of repositories and open data, conformance to an international standard would ensure smoother communications, especially in discussion of data curation practices.

5.2. RQ1: What is the state of “best practice” in data curation?

Six years after Knight (2012) described existing data management practices as a “digital curate’s egg”, the current state of “best practice” remains largely the same. On average currently, achievement of “best practice” is varied, and standard practices are irregular and not necessarily shared between all institutions. There is a focus on the the first half of the curation lifecycle model, with priority given to receiving files for storage. Again, present actions are less about curation and more about management of research data.

There is a demonstrated awareness of what constitutes “best practice”, but for many, best practices are a “wishlist”, and the reality involves compromising between available resources and institutional priorities. As a result, “satisficing” appears to be a popular strategy in order to achieve minimum, acceptable standards in lieu of fruitlessly pursuing optimal standards (Lee *et al*., 2017). As the conversation about data curation progresses, satisficing should be regarded as a viable short-term strategy. Considering the limitations being faced, endeavoring to establish standard *good* practices should take priority over concern for *best* practices. Best practice is untenable without a consensus on current practices and a joined effort in implementing standard practices.

Standard practices were more common in areas where satisficing was similarly typical, such as metadata and storage. Metadata and storage were also more likely to involve more mediation with researchers than other areas. While satisficing can be utilized as one solution, increased interaction with data providers seems to be equally important for good practice, especially given the primary role of researchers in the data curation process.

5.2.1. Key underdeveloped areas and recommendations

Three key underdeveloped areas of data curation emerged from the research findings: quality control, auxiliary information, and preservation actions. Developing these areas would contribute to improved access and reusability of research data.

Quality control is especially important at institutions where self-deposits are expected as part of the institutional policy. Although researchers possess disciplinary expertise and can provide detailed metadata, they lack the cataloging skills to provide good or useful metadata. Quality control is currently partially mitigated through shared responsibilities for metadata, however, further consideration needs to be given to incorporating appraisal processes, as well as validation and content checks. These processes ensure that available research data is useful for purposes outside the original project. The current failure to appraise data or apply validation and content checks could mean that research data available now is in danger of being or becoming unusable.

Auxiliary information can be utilized as a stopgap solution for insufficient quality control. Requirements for auxiliary information should have a minimum standard where possible. Even though research data is so contextual, and therefore accompanying metadata will need to be specific, there should at least be a minimum, *required* standard for auxiliary files and file information. A good standard to promote is the current expectation of a readme.txt consisting of file and technical details accompanied by a user guide, where necessary.

Preservation is already recognized by the data curation community as a neglected area, and many promisingly indicated a commitment to investing in its progress and development. In addition, many institutions also encourage the use of open and non-proprietary file formats, which is recommended for ensuring long-term, continued access to data files beyond current software. However, more comprehensive preservation actions appear to be underutilized, such as greater promotion of subject repositories and use of open-source preservation software. Neither option was heavily employed or highlighted, despite their clear benefits. Subject repositories are dedicated to hosting specialized research data and would help to decrease the burden at institutions that may not have the staff, skills, or resources to maintain an institutional or data repository. Open-source preservation software, such as Archivematica (https://www.archivematica.org/en/), performs and automates much of the preservation workflow, including validating file formats, checking content, and preserving the integrity of the original data. However, before a full recommendation can be endorsed, further research would need to be undertaken to accurately assess the viability of these options and to under the motivations and situations of those who currently utilize these resources and those who do not.

There is a small number of institutions without a dedicated commitment to preservation, but these remaining institutions most likely represent a contingency that is not research-intensive and therefore less concerned with preserving research data.

5.2.2. Reliance on researchers

Although lack of resources is a fundamental issue, the most limiting factor on improving data curation practices is the reliance on researchers. The specificity and contextual nature of research data requires the full participation of researchers, but there are two obstacles hindering their cooperation. Firstly, researchers are too previously burdened or preoccupied to fulfill more than the minimum requirements. The “Concordat on Open Research Data” expressed similar concerns about the burden placed on researchers (UKRI, 2016). Secondly, researchers are uninterested in sharing their data. Both obstacles can be addressed through heightened communication with researchers. Briefly covered in the literature review, there has been a continuous lack of communication with researchers (Fox, 2013)

One area of immediate interest that could bridge the gap is engaging with researchers during the creation of DMPs. Although a DMP is required of most researchers, DMPs are not regularly submitted with research data files. It is unclear where DMPs are stored after the completion of a research project. However, if developed in detail, a DMP would include provisions for the curation of research data. These plans would provide answers to future curation actions for submitted files.

Furthermore, this would involve the repository in the earliest stages of RDM, when it would be easier to make important provisions for later stages of data curation.

Providing greater support to researchers is paramount to improving data curation practices, whether by promoting existing infrastructure or redistributing responsibilities. Although assumed to be outside the scope of this dissertation, researchers' data management habits are inseparable from data curation practices. There is already a healthy area of interest devoted to RDM, and the emerging findings will be integral to the ongoing conversation.

5.2.3. FAIR practices

The ultimate goal of data curation is the eventual reuse, sharing, and transformation of research data. As institutions work to implement the FAIR Data Principles, they are also working towards an ideal state of data curation. While a majority of institutions fulfilled FAIR values on an "average" scale, an "average" rating is contextual and relies on a overall comparison of performances between institutions. A more nuanced examination suggests that "average" efforts are currently insufficient to accomplish all FAIR requirements, although a concerted effort is being made.

Institutions were the most successful in regards to *findability* of research data. DOIs were used by nearly all, and datasets were linked to related publications and institutions, presumably the institution hosting the data. However, further attention can be paid to including author information, especially author IDs.

In regards to *accessibility*, although there was an obvious emphasis throughout institutional practices and policies, the research did not comprehensively explore this principle, and a full conclusion cannot be accurately drawn.

As with accessibility, the research did not cover interoperability of data, and additional research would need to be conducted before a conclusion could be reached. However, considering the minimal expectation of metadata and the lack of a "shared" standard, this would imply that the average standard for *interoperability* would not meet FAIR standards (Wilkinson *et al.*, 2016).

The average institutional standard for *reusability* partially fulfilled FAIR standards. While licenses were normally expected, respondents were lacking in providing "a plurality" of descriptive information and metadata that included "detailed provenance" (Wilkinson *et al*., 2016).

As a summation of the current state of data curation, fulfillment of the FAIR principles is also a work in progress, and hopefully with time and experience, the expectations for "average" will advance.

5.3. RQ2: Do existing policies make provisions for standard practices?

As a whole, provisions for data curation are generally lacking. While institutions have largely claimed responsibility for providing the necessary infrastructure, many do not offer a clear indication of how to perform the necessary duties. Provisions, when made, are centered around institutional strategy and legal insurance rather than facilitating standard practices. A clear disparity exists between public, institutional policies and internal practices and departmental policies.

The estrangement between policy and practice is especially apparent when contrasting the differing coverage DOIs, security measures, and preservation practices. DOIs are a basic standard for data records, however, they receive little mention in institutional policies. More weight is given to security by a wide margin. To begin to understand the disparate relationship between policy and practice, it is necessary to consider the motivations behind each. As an issue with legal implications, security may be considered more within the domain of administrative concerns and is therefore more pronounced within policies. However, preservation, an area normally of more concern to data curation practices, is prioritized more highly in institutional policies. Preservation is an instance where the motivations of an institution are misaligned with the available infrastructure. What is being targeted as a priority may not be realistically achievable within current bounds. This can be seen in the issues with accomplishing preservation measures and in missing guidance documentation.

5.3.1. Preservation

Preservation is well-represented in institutional policies, yet the execution of preservation procedures is poor, suggesting either no or inadequate provisions are made for supporting proper preservation actions. Preservation is a prime example of an area where institutional goals are apparent, but direction and appropriate infrastructure have not followed or been addressed.

Aside from the disconnect between institutional motivations and achievable standards is a disconnect between institutional definitions and definitions in practice. This would provide another explanation for the lack of appropriately supported preservation measures. “Preservation” as defined in policy terms may refer to continued access to data within a certain timeframe, while “preservation” as performed by data curators would be the preservation of data for future generations. Recognizing preservation as a long-term action would vastly change the support institutions think is necessary.

Fortunately, this gap is recognized as an issue and hopefully, the current commitment to producing preservation strategies will result in an alignment of institutional values and well-supported data curation practices.

5.3.2. Lack of clear guidance

The absence of appropriate provisions is also evident in the literal lack of guidance documentation. Even when existing guidelines were available, the documentation was often not promoted. It seemed that without clear guidance as support, staff encountered difficulties enforcing policies, and therefore were unable to ensure standard practices. Many institutions also had no requirement to deposit in institutionally supplied repositories, resulting in an inability to consistently track standards for data that were not mediated by repository staff. The lack of institutional support, both in the form of textual guidance and in-built in deposit policies, seemed to undercut the ability of repository staff to require more of researchers.

In addition, there were few concessions for providing additional support to researchers. Data curation requires the cooperation and expertise of researchers, but there was no clear agenda to supplement support for researchers. Enforcing open data standards especially suffered as a result of this oversight. While there are clear motivations and an agenda for furthering data sharing, the lack of provisions for change has resulted in an inability to fully engage researchers in the conversation about open data.

5.4. RQ3: What are the connections between policy and practice?

Arguably, the impetus to invest in research data curation has only truly begun to emerge as a concern for UK institutions within the past four or less years, with the establishment of new Research Excellence Framework (REF) expectations and funder-related obligations. In particular, the growth in institutional policies suggests increasing recognition of the importance and relevance of research data over a short period of time. Therefore, it can be expected that developments in data curation will be continuous during this new period of growth. The relative recentness of institutional policies explains the general lack of infrastructure for practice. Institutional support for data curation has only begun to arise within recent years. However, the rush to develop programs for data curation has resulted in an unsustainable burden being placed on researchers and an emphasis on relatively short-term storage over long-term preservation. Now more than ever, investing in departmental infrastructure to support researchers and establishing a long-term strategy is paramount to the future of data curation.

5.4.1. Burden on researchers

As previously discussed, there is a disparity between policy and practice that can be seen in the imbalance of responsibilities placed on researchers. Policies have assigned primary responsibility for data curation to researchers, and therefore, data curation practices have been reliant on researcher compliance. However, there is no expectation beyond minimum compliance either specified by policies or required in practice.

Policies were instead focused on ensuring that data is managed according to funder and legal requirements, therefore priority is given to providing infrastructure for the facilitation of these requirements. This would explain the existence of repositories and services but a distinct lack of guidance documentation or enforcement of principles. The infrastructure is available if researchers choose to deposit with the institution, but there are no further requirements other than access.

Much of current practice centers around attempting to facilitate productive exchanges with researchers and encouraging them to perform more than the minimum, and continued concentration in this area is recommended. Further analysis between internal documentation and standard practice could also be conducted to ascertain where researchers' responsibilities could be supplemented or redirected completely to more knowledgeable staff.

5.4.2. Storage

The burden on researchers was especially evident in conversations about storage, and storage provides a good example of when responsibilities could be mitigated. As distinguished by the survey respondents, storage is involved in multiple points in the research lifecycle in the form of either active storage or long-term storage. Any issues related to storage, then, may be more nuanced depending on the specific point in the research lifecycle. Within the data curation lifecycle, storage specifically refers to long-term storage, however, researchers have to manage data before this point, during active storage of data. This implies even more responsibility for researchers, as they have to be concerned with storage solutions for both their active research and their published research. However, long-term storage does not need to be the sole responsibility of researchers. In answer, storage was an area where responsibilities were more reasonably distributed in policies, and probably partially as a result, storage activities involved more cooperation. With a line of communication already open, the level of cooperation and engagement could be followed up to develop key areas related to storage deposits, such as metadata.

5.5. Recommendations for future research

Several possibilities for future research can be recommended from the findings in this dissertation, and as the field of data curation develops, further areas of interest will become more appart. In the first instance, building on the dissertation findings with case studies would be the most useful in the short-term for establishing standard practices. Although Perrier, *et al*. (2017) critiqued the overabundance of existing case studies, new case studies supported by *empirical evidence* would be especially helpful during this period of growth. Potential case studies could involve:

- Tracking the growth of a new service and seeing whether *enforcing* a policy would contribute to better practice
- Retroactively discussing the process of setting up a service and accompanying policy and what factors influenced the two (or whether there was any influence)

Additional possibilities for future research include:

- A survey questionnaire or interview that collected information specifically about institutional infrastructure to explore in-depth the connection between institutional support and current practices
- A comprehensive document analysis on internal documentation to compare between public and internal policies
- An investigation into use of storage platforms, including comparing rates of use between different types of repositories and the motivations behind choosing where to deposit

In addition, the question of responsibility became very relevant as the dissertation progressed, and it would be interesting to examine whether research data should actually be the responsibility of an institution rather than a specifically-equipped subject repository and whether a shared institutional repository would not be more sustainable.

With maintained interest in the improvement of data curation practices, best practices will hopefully become more than a “wish list” in the future.

References

Akers, K., *et al.* (2014) ‘Building support for research data management: biographies of eight research universities’. *International Journal of Digital Curation*, 9(2), pp.171-191. doi:10.2218/ijdc.v9i2.327.

Aiman-Smith, L. and Markham, S.K. (2004) ‘What you should know about using surveys’. *Research Technology Management*, 47(3), pp. 12-15. doi:10.1080/08956308.2004.11671625.

Amorim, R., *et al.* (2017) ’A comparison of research data management platforms: architecture, flexible metadata and interoperability’. *Universal Access in the Information Society* 16(4), pp.851-862. doi:10.1007/s10209-016-0475-y.

Austin, C., *et al.* (2015) ‘Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements’. *IASSIST Quarterly*, 39(4), pp.24-38. Available at: http://www.iassistdata.org/sites/default/files/vol_39-4.pdf (Accessed: 08 May 2018).

Bailey, Jr., C. (2018) *Research Data Curation Bibliography*. Available at: http://digital-scholarship.org/rdcb/rdcb.htm.

Berman, F. (2008) 'Got data? A guide to data preservation in the information age'. *Communications of the ACM*, 51(12), pp.50-56. doi:10.1145/1409360.1409376.

Briney, K., Goben, A. and Zilinski, L. (2015) ‘Do you have an institutional data policy? A review of the current landscape of library data services and institutional data policies’. *Journal of Librarianship and Scholarly Communication*, 3(2), pp. eP1232. doi:10.7710/2162-3309.1232.

Buys, C.M. and Shaw, P.L. (2015) ‘Data management practices across an institution: survey and report’. *Journal of Librarianship and Scholarly Communication*, 3(2), eP1225. doi:10.7710/2162-3309.1225.

Casrai (2015) *Data curation*. Available at: http://dictionary.casrai.org/Data_curation (Accessed: 21 June 2018).

Chao, T.C., Cragin, M.H. and Palmer, C.L. (2015) ‘Data Practices and Curation Vocabulary (DPCVocab): an empirically derived framework of scientific data practices and curatorial processes’. *Journal of the Association for Information Science and Technology*, 66(3), pp. 616-633. doi:10.1002/asi.23184.

Cox, A.M., Pinfield, S. and Smith, J. (2016) ‘Moving a brick building: UK libraries coping with research data management as a 'wicked' problem’. *Journal of Librarianship and Information Science*, 48(1), pp. 3-17. doi:10.1177/0961000614533717.

Data Asset Framework (DAF) (2009) *Implementation Guide*. Available at: https://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf (Accessed: 26 June 2018).

Data Seal of Approval (DSA) (2018a) *Requirements*. Available at: https://www.datasealofapproval.org/en/information/requirements/ (Accessed: 26 June 2018).

Data Seal of Approval (DSA) (2018b) *Assessment*. Available at: https://www.datasealofapproval.org/en/assessment/ (Accessed: 26 June 2018).

DCC (2018a) *Curation Lifecycle Model*. Available at: http://www.dcc.ac.uk/resources/curation-lifecycle-model (Accessed: 21 June 2018).

DCC (2018b) *Funders’ data plan requirements*. Available at: http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements (Accessed: 21 June 2018).

DCC (2018c) *Glossary*. Available at: http://www.dcc.ac.uk/digital-curation/glossary (Accessed: 21 June 2018).

DCC (2018d) *Curation Reference Manual*. Available at: http://www.dcc.ac.uk/resources/curation-reference-manual (Accessed: 26 June 2018).

DCC (2018e) *UK institutional data policies*. Available at: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies (Accessed: 28 June 2018).

Dressler, V.A. (2017) ‘The state of affairs with digital preservation at ARL member libraries’. *Digital Library Perspectives*, 33(2), pp. 137-155. doi:10.1108/DLP-08-2016-0030.

Dürr, E., *et al.* (2008) 'Long-time preservation of data sets, results of the DareLux project'. *Information Services & Use*, 28(3/4), pp.281-294. doi:10.3233/ISU-2008-0571.

Evans, J.R. and Mathur, A. (2005) 'The value of online surveys'. *Internet Research: Electronic Networking Applications and Policy*, 15(2), pp. 195-219. doi:10.1108/10662240510590360.

Finnegan, R. (2006) 'Using documents', in Sapsford, R. and Jupp, V. (eds.) *Data collection and analysis second edition*. London: Sage, pp. 138-151.

Fox, R. (2013) 'The art and science of data curation'. *OCLC Systems & Services*, 29(4), pp. 195-199. doi:10.1108/OCLC-07-2013-0021.

Friddell, J., LeDrew, E. and Vincent, W. (2014) 'The Polar Data Catalogue: best practices for sharing and archiving canada's polar data'. *Data Science Journal*, 13, pp. PDA1-PDA7. doi:10.2481/dsj.ifpda-01.

gov.uk (2018) *Recognised bodies*. Available at: https://www.gov.uk/check-a-university-is-officially-recognised/recognised-bodies (Accessed: 21 June 2018).

Helbig, K., Hausstein, B. and Toepfer, R. (2015) 'Supporting data citation: experiences and best practices of a doi allocation agency for social sciences'. *Journal of Librarianship and Scholarly Communication*, 3(2), pp. eP1220. doi:10.7710/2162-3309.1220.

Higgins, S. (2008) 'The DCC Curation Lifecycle Model'. *International Journal of Digital Curation*, 3(1), pp. 134-140. doi:10.2218/ijdc.v3i1.48.

Higman, R. and Pinfield, S. (2015) 'Research data management and openness'. *Program*, 49(4), pp. 364-381. doi:10.1108/PROG-01-2015-0005.

Horton, L and DCC (2016) 'Overview of UK institution RDM policies' Version 6 August 2016, Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies (Accessed: 21 June 2018).

Knight, G. (2012) 'A digital curate's egg: a risk management approach to enhancing data management practices'. *Journal of Web Librarianship*, 6(4), pp. 228-250. doi:10.1080/19322909.2012.729992

Koltay, T. (2016) ‘Are you ready? Tasks and roles for academic libraries in supporting research 2.0’. *New Library World*, 117(1), pp. 94-104. doi:10.1108/NLW-09-2015-0062.

Laughton, P. and du Plessis, T. (2013) ‘Data curation in the World Data System: proposed framework’. *Data Science Journal*, 12, pp. 56-70. doi:10.2481/dsj.13-029.

Lavoie, B. (2014) ‘The Open Archival Information System (OAIS) reference model: introductory guide (2nd Edition)’. *DPC Technology Watch Report*. doi:0.7207/twr14-02DPC.

Lee, C. and Tibbo, H. (2007) ‘Digital curation and trusted repositories: steps toward success’. *Journal of Digital Information*, 8(2). Available at: https://journals.tdl.org/jodi/index.php/jodi/article/view/229/183 (Accessed: 21 June 2018).

Lee, D.J. and Stvilia, B. (2017) ‘Practices of research data curation in institutional repositories: a qualitative view from repository staff’. *PLOS ONE*, 12(3), pp. e0173987.

Lee *et al.* (2017) ‘Open data meets digital curation: an investigation of practices and needs’. International Journal of Digital Curation, 11(2), pp. 115-125. doi:10.2218/ijdc.v11i2.403.

Locher, A.E. (2016) 'Starting points for lowering the barrier to spatial data preservation'. *Journal of Map & Geography Libraries*, 12(1), pp.28-51. doi:10.1080/15420353.2015.1080781.

MacMillan, D. (2014) ‘Data sharing and discovery: what librarians need to know’.
The Journal of Academic Librarianship, 40(5), pp.541-549. doi:10.1016/j.acalib.2014.06.011.

Mohr, A. *et al.* (2015) 'When data is a dirty word: a survey to understand data management needs across diverse research disciplines'. *Bulletin of the Association for Information Science & Technology*, 42(1), pp.51-53. doi:10.1002/bul2.2015.1720420114.

Oczkowski *et al.* (2018) ‘Organ donation in the ICU: A document analysis of institutional policies, protocols, and order sets’. *Intensive and Critical Care Nursing*, 45, pp. 58-65. doi:10.1016/j.iccn.2017.12.005.

Olendorf, R. and Koch, S. (2012) ‘Beyond the low hanging fruit: data services and archiving at the university of new mexico’. *Journal of Digital Information*, 13(1). Available at: https://journals.tdl.org/jodi/index.php/jodi/article/view/5878/5882 (Accessed: 08 May 2018).

Onwuegbuzie, A.J., Frels, R.K., Hwang, E. (2016) ‘Mapping Saldaňa’s coding methods onto the literature review process’. *Journal of Educational Issues*, 2(1), pp. 130-150. doi:10.5296/jei.v2i1.8931.

OpenRefine (2018) *OpenRefine*. Available at: http://openrefine.org/ (Accessed: 26 July 2018).

Open Research Data Taskforce (ORDT) (2017) *Research data infrastructure in the UK*. United Kingdom: Universities UK. Available at: https://www.universitiesuk.ac.uk/policy-and-analysis/research-policy/open-science/Pages/open-research-data-task-force.aspx (Accessed: 28 June 2018).

Perrier *et al.* (2017) ‘Research data management in academic institutions: a scoping review’. *PLOS ONE*, 12(5), e0178261. doi:10.1371/journal.pone.0178261.

Pinnick, J. (2017) 'Exploring digital preservation requirements'. *Records Management Journal*, 27(2), pp.175-191. doi:10.1108/RMJ-04-2017-0009.

Rice, R. and Southall, J. (2016) *The Data Librarian’s Handbook*. London: Facet Publishing.

Rosenthal, D. (2017) 'The medium-term prospects for long-term storage systems'. *Library Hi Tech*, 35(1), pp.11-31. doi:10.1108/LHT-11-2016-0128.

Sallans, A. and Donnelly, M. (2012) ‘DMP Online and DMPTool: different strategies towards a shared goal’. *International Journal of Digital Curation*, 7(2), pp.123-129. doi:10.2218/ijdc.v7i2.235.

Shen, Y. and Varvel, V. (2013) 'Developing data management services at the Johns Hopkins University'. *Journal Of Academic Librarianship*, 39(6), pp.552-557. doi:10.1016/j.acalib.2013.06.002.

UK Data Service (2018) *Data management planning*. Available at: https://www.ukdataservice.ac.uk/manage-data/plan/planning (Accessed: 21 June 2018).

UK Research and Innovation (URKI) (2016) *Concordat on open research data*. Available at: https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (Accessed: 14 August 2018).

University of Cambridge (2018a) *For EPSRC-funded researchers*. Available at: https://www.data.cam.ac.uk/funders/epsrc-funded-researchers (Accessed: 21 June 2018).

University of Cambridge (2018b) *Funders' policies*. Available at: https://www.data.cam.ac.uk/funders (Accessed: 11 July 2018).

Universities UK (2018) *Higher education in numbers*. Available at: https://www.universitiesuk.ac.uk/facts-and-stats/Pages/higher-education-data.aspx (Accessed: 21 June 2018).

Van Zeeland, H. and Ringersma, J. (2017) 'The development of a research data policy at Wageningen University & Research: best practices as a framework'. *LIBER Quarterly*, 27(1), pp. 153-170. doi:10.18352/lq.10215.

Wilkinson *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data*, 3. doi:10.1038/sdata.2016.18.

Appendix 1 – Full list of texts included in literature scan

1. Chao, T.C., Cragin, M.H. and Palmer, C.L. (2015) ‘Data Practices and Curation Vocabulary (DPCVocab): an empirically derived framework of scientific data practices and curatorial processes’. *Journal of the Association for Information Science and Technology*, 66(3), pp. 616-633. doi:10.1002/asi.23184.

2. Friddell, J., LeDrew, E. and Vincent, W. (2014) ‘The Polar Data Catalogue: best practices for sharing and archiving canada's polar data’. *Data Science Journal*, 13, pp. PDA1-PDA7. doi:10.2481/dsj.ifpda-01.

3. Helbig, K., Hausstein, B. and Toepfer, R. (2015) ‘Supporting data citation: experiences and best practices of a doi allocation agency for social sciences’. *Journal of Librarianship and Scholarly Communication*, 3(2), pp. eP1220. doi:10.7710/2162-3309.1220.

4. Laughton, P. and du Plessis, T. (2013) ‘Data curation in the World Data System: proposed framework’. *Data Science Journal*, 12, pp. 56-70. doi:10.2481/dsj.13-029.

5. Lee, D.J. and Stvilia, B. (2017) ‘Practices of research data curation in institutional repositories: a qualitative view from repository staff’. *PLOS ONE*, 12(3), pp. e0173987.

6. Lee *et al.* (2017) ‘Open data meets digital curation: an investigation of practices and needs’. International Journal of Digital Curation, 11(2), pp. 115-125. doi:10.2218/ijdc.v11i2.403.

7. Van Zeeland, H. and Ringersma, J. (2017) ‘The development of a research data policy at Wageningen University & Research: best practices as a framework’. *LIBER Quarterly*, 27(1), pp. 153-170. doi:10.18352/lq.10215.

Appendix 2 – Survey questionnaire

Start of Block: Participant Information

Information about this survey
Participant information for potential respondents

Surveying the current state of data curation: a review of policy and practice at UK HEIs

Introduction
My name is Amy Pham, and I am a postgraduate student currently undertaking dissertation research for an MSc in Library and Information Studies at the University of Strathclyde. My research seeks to review policies and practices related to data curation at universities in the United Kingdom.

Purpose
The purpose of this investigation is to quantify the implementation of standard policies and practices related to data curation programs in the UK. While existing literature extensively covers both topics separately, little has been written about the relationship between policy and practice. The resulting research could contribute to a future gap analysis or inform "best practice" procedures.

Participation
Participation is completely voluntary. You may refuse to participate or exit the survey at any point, up until the "Submit" option.

If you choose to participate, you will be directed to a survey questionnaire, which should take approximately 10-15 minutes to complete. The questions are primarily closed-ended, with some being rating scales or open-ended. Questions will focus on reviewing specific tasks or workflows related to data curation.

You may choose to break and return to the survey at any time. Your answers will be saved for up to two weeks, at which point, answers will be recorded as final. Incomplete surveys will not be included in final data analysis. **The questionnaire will be available for the duration of 3 weeks, from 13 June 2018 to 4 July 2018.**

You have been invited to participate due to your affiliation with a UK higher education institution or due to your relevant professional experience.

There are no known risks associated with participating in this survey.

Confidentiality
Data collected through this questionnaire will be anonymized, and no identifying information will be asked. Data will be stored securely online and require password protection to access.

The University of Strathclyde is registered with the Information Commissioner's Office who implements the Data Protection Act 1998. All personal data on participants will be processed in accordance with the provisions of the Data Protection Act 1998.

Consent

If you choose to participate in this survey, you will be directed to a consent form on the next page. Thank you for your consideration.

Contact details
Please feel free to contact the researcher at amy.pham.2017@uni.strath.ac.uk. All comments are appreciated.

If you would like to contact the supervisor overseeing this dissertation project, please contact Dr. Diane Pennington at diane.pennington@strath.ac.uk.

This investigation was granted ethical approval by the Department of Computer & Information Sciences Ethics Committee.

If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Secretary to the Departmental Ethics Committee
Department of Computer and Information Sciences
Livingstone Tower
Richmond Street
Glasgow
G1 1XH
email:ethics@cis.strath.ac.uk

End of Block: Participant Information

Start of Block: Consent Form

- I confirm that I have read and understood the information sheet for the above project and the researcher has answered any queries to my satisfaction.
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences. If I exercise my right to withdraw and I don’t want my data to be used, any data which have been collected from me will be destroyed.
- I understand that I can withdraw from the study any personal data (i.e. data which identify me personally) at any time.
- I understand that anonymised data (i.e. .data which do not identify me personally) cannot be withdrawn once they have been included in the study.
- I understand that any information recorded in the investigation will remain confidential and no information that identifies me will be made publicly available.
- I consent to being a participant in the project.

○ I consent to being a participant in the project.

○ I do not consent to being a participant in the project.

Display This Question:

If Department of Computer and Information Science, University of Strathclyde Surveying the current s... = I do not consent to being a participant in the project.

Thank you for your time and consideration.

End of Block: Consent Form

Start of Block: Background

Are you affiliated with a Higher Education institution in the United Kingdom?

○ Yes

○ No

Is your position or department directly involved with data curation or data management?

○ Yes

○ No

Does your institution have a data policy?

○ Yes

○ No

Does your institution have documented procedures or standards for the following processes (please select *all* that apply):

□ Data Acquisition (i.e. receipt, selection, ingest, etc.)

□ Metadata

□ Storage

□ Preservation

□ Administration (i.e. daily operations, licensing, etc.)

□ Access

End of Block: Background

Start of Block: Data Acquisition

Which of the following criteria are required for submission of data files (please select *all* that apply):

□ Standardized file formats

□ Standardized file names

□ Standardized metadata

□ Accompanying auxiliary information (e.g. README files, etc.)

□ Submission agreements

□ Licensing agreements

□ Accompanying Data Management Plan (DMP)

□ Not applicable

□ Other (please specify) ____

Is there an appraisal process for data (i.e. to determine the length of time to retain a submission)?

○ Yes

○ No

Do you perform validation and content checks during data processing? (Please select *all* that apply)

□ Validation

□ Content Checks

□ Not applicable

□ Not sure

Please describe how data files are prepared for storage at your institution:

____

____

____

____

____

End of Block: Data Acquisition

Start of Block: Metadata

At what stage of ingestion does quality control first occur?

○ Before

○ During

○ After

○ Not applicable

Please choose the option that is *most* applicable:

○ Researcher is expected to fulfill strict metadata requirements and provide comprehensive metadata

○ Researcher is expected to fulfill core metadata requirements. Additional metadata is supplemented at a later date, either by researcher or a member of staff

○ Metadata is accepted as submitted

○ Core metadata is collected and entered by a member of staff

○ Other (please explain) ____

Are datasets linked to any of the following related records (please select *all* that apply)?

□ Associated publication(s)

□ Author institutional profile(s)

□ Author ID(s) (e.g. ORCID, Scopus, etc.)

□ Institution(s)/organization(s)

□ Subject collection

□ Not applicable

Which of the following information related to provenance is included in metadata records (please select *all* that apply):

□ Creation details

□ Alterations to content or format

□ Ownership

□ Preservation history (e.g. format migration)

□ Not applicable

Is information related to access rights included in metadata records?

○ Yes

○ No

Are data sets assigned unique, persistent identifiers (e.g. DOIs)?

○ Yes

○ No

Please list auxiliary information required upon receipt of data files (i.e. content of data files, software information, etc.):

____

End of Block: Metadata

Start of Block: Storage

Where are data sets hosted at your institution?

○ Institutional repository

○ Data repository (separate from institutional repository)

○ Other (please specify) ____

Does your institution have a storage policy that covers the following (please select *all* that apply):

□ Preservation

□ Security

□ Format

□ Migration

□ Recovery

□ None of the above

Which of the following actions does your institution take to secure data files upon receipt (please select *all* that apply):

□ Encrypt files

□ Convert file to stable file formats

□ Duplicate files (minimum two)

□ Store duplicate(s) off site

□ Not applicable

End of Block: Storage

Start of Block: Preservation

Does your institution currently have a strategy for long-term preservation of data?

○ Yes

○ No

Display This Question:

If Does your institution currently have a strategy for long-term preservation of data? = No

If not, does your institution have a commitment to developing a strategy for long-term preservation?

○ Yes

○ No

Which of the following preservation solutions does your institution currently adopt (please select *all* that apply):

□ Keep original data

□ Use non-proprietary or open data formats

□ Data reappraisal

□ Bit rot repair

□ Format migration

□ Emulation

□ Establish partnerships with external organizations

□ Not applicable

□ Other (please specify) ____

Which of the following actions does your institution take to perform risk assessment (please select *all* that apply):

□ Monitor storage environment

□ Technology watch

□ Self-audit (e.g. DRAMBORA, etc.)

□ Not applicable

□ Other (please specify) ____

End of Block: Preservation

Start of Block: Administration

On a scale of 1-5, how adequate would you rank the level of software support you receive for data curation or data management tasks?

○ Extremely adequate

○ Somewhat adequate

○ Neither adequate nor inadequate

○ Somewhat inadequate

○ Extremely inadequate

End of Block: Administration

Start of Block: Access

On a scale of 1-5, how would you rank your institution's fulfillment of the FAIR Data Principles (Findability, Accessibility, Interoperability, Reusability)?

○ Excellent

○ Good

○ Average

○ Poor

○ Terrible

End of Block: Access

Start of Block: Additional comments

Please use this space for additional comments or to expand on survey answers:

____

____

____

____

____

End of Block: Additional comments

Start of Block: End of survey

You have reached the end of this survey. Your contribution is greatly appreciated!

If you would like to discuss the research topic or request a results report, please email Amy Pham at amy.pham.2017@uni.strath.ac.uk.

Please press the submit button below to record your answers.

End of Block: End of survey