A NOVEL FRAMEWORK FOR USER-CENTERED RESEARCH DATA MANAGEMENT

M I BUGAJE

PhD

2019

A NOVEL FRAMEWORK FOR USER-CENTERED RESEARCH DATA MANAGEMENT

MARYAM IDRIS BUGAJE

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of
Doctor of Philosophy

Research undertaken in the Faculty of Engineering and Environment

February 2019

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee.

I declare that the Word Count of this Thesis is 42,253 words

Name: Maryam Idris Bugaje

Signature

Date: 28/02/2019

Abstract

Recent Open Data policies have led to a large-scale demand for research data repositories. Research data repositories are expected to function as an important instrument for research data preservation as well as for research collaboration and dissemination, helping to realize the advantages that motivated those policies. Existing research data management (RDM) systems and infrastructure, of which research data repositories form an important component, are currently inadequate to support and further this vision. Research data are complex-compound objects, and their use, and also the mode of interacting with them, differs considerably from those of manuscript documents (e.g. research publications). This research proposes a holistic framework for RDM system design that expressly takes into account the needs of system users as well as the peculiar requirements of research data, to develop well-functioning systems. It demonstrates the development process of a simple prototype of a user-centered, data-conscious RDM system, called DataFinder, from the earliest stages of requirements gathering to requirements analyses, design, development, and evaluation. Importance is given as much to the final deliverable (i.e. DataFinder) as to the process of attaining it, since a substantial part of the research preceded the former, and the findings garnered along the way will likely prove useful for purposes of which system design forms only one.

List of Publications

Bugaje, M., Chowdhury, G. (2017). Towards a More User-Centered Design of Research Data Management (RDM) Systems [abstract]. In: Information: Interactions and Impact (i3).; 27-30 June 2017; Aberdeen; 2017:53-55.

Bugaje, M., & Chowdhury, G. (2017). Is Data Retrieval Different from Text Retrieval? An Exploratory Study. In Digital Libraries: Data, Information, and Knowledge for Digital Lives (pp. 97–103). Springer International Publishing. https://doi.org/10.1007/978-3-319-70232-2_8

Bugaje, M., Chowdhury, G. (2018a). Data Retrieval = Text Retrieval? In: Chowdhury, G., McLeod, J., Willett, P., Gillet, V. (eds.) iConference 2018. LNCS, vol. 10766, pp. 253–262. Springer https://doi.org/10.1007/978-3-319-78105-1_29

Bugaje, M., Chowdhury, G. (2018). The Sixth European Conference on Information Literacy, September 24th–27th, 2018, Oulu, Finland : (ECIL) : abstracts / editors Sonja Špiranec, Serap Kurbanoğlu, Maija-Leena Huotari, Esther Grassian, Diane Mizrachi, Loriene Roy, Denis Kos. Oulu : University of Oulu, Department of Information and Communication Studies, 2018.

Bugaje, M., Chowdhury, G. (2018). Identifying Design Requirements of a User-Centered Research Data Management System. In Lecture Notes in Computer Science (pp. 335–347). Springer International Publishing. https://doi.org/10.1007/978-3-030-04257-8_35

Chowdhury, G., Walton, G., Bugaje, M. (2017) The Fifth European Conference on Information Literacy, September 18th-21st, 2017, Saint-Malo, France.

Acknowledgements

First, I would like to give sincerest thanks to my principal supervisor Professor Gobinda Chowdhury for his continuous guidance throughout my research programme. I am deeply indebted to him for his encouragement, kindness, patience, and support throughout the course of my study.

I would also like to thank the staff of the Faculty of Engineering and Environment for the supportive environment they have created.

And I thank my parents and sisters and brothers for their love, kindness, and understanding always.

Table of Contents

List of Figures

List of Tables

1.0 INTRODUCTION

Development of infrastructures for long-term preservation of research data has until recent date been slow, despite an ever present, if rather dormant, demand (Weber & Piesche, 2016). The change, which in the last decade or so took a more accelerated turn, arose from a growing international movement in favor of providing free and open access to research data. This turn of events brought Research Data Management (RDM) into increased prominence. Although not necessarily in a glaringly imperfect state of affairs, RDM at present still stands in need of certain technical, infrastructural, as well as socio-cultural improvements (Nelson, 2009; Hartter et al., 2013; Curdt & Hoffmeister, 2015). This is perhaps not very remarkable considering its relatively recent emergence as a distinct research field worthy of commanding research attention in its own right. Allied fields such as Information Management and Database Management, for example, have been around for a comparatively much longer period. Nonetheless, RDM is at present a trending topic in academic scholarship, frequently to be encountered in relation with “e-science” and “e-research”, which concepts are rapidly gaining a stronghold in the vision for scholarship in the 21st century. The quest for research data solutions to manage an ever-increasing accumulation of research datasets has given rise to a steady demand for RDM products and services. This demand becomes more and more pressing in proportion as both the applicability and the advisability of RDM are recognized across nearly all domains of scientific inquiry (Borgman, 2015). There is, in addition, a general desire to hasten the fulfilment of the widely-acknowledged promises of open data; the chief among which are discussed in the next chapter (see Section 2.3). By way of accounting for these circumstances, the literature commonly cites the increased (and still increasing) uptake of formal Open Data requirements by governments and research funding bodies. Open Data requirements stipulate, especially for research projects financed with public funds, that research data resulting from such projects be maintained in a repository where the same can be freely and openly accessed (Arzberger et al., 2004; Murray-Rust, 2008). This explanation, however, though true, and certainly a sound one, hardly represents the full circumstances of the case which, I believe, are traceable to a phenomenon of an earlier and even more momentous occurrence. It is as follows: modern advances in technology and scientific innovation have naturally led to a corresponding

enhancement, quality and quantity-wise, in the capability of instruments for measuring, recording, and storing data. In addition, the ease and cost of acquisition has reduced proportionately. The result is that, almost across all walks of life, data are now being produced at a rate never before known in history; hence the term “Big Data”, which describes data that combines in its essential characteristics tremendous volume, velocity, and variety (Hey et al., 2009). Moreover, new kinds of data, such as data from social media and wearable technology, now exist for the first time, and often in a more or less minute level of detail. This sheer abundance of data has been vividly described by Borgman (2007) as the “data deluge”, and has in fact opened up a new era of scholarship termed ‘‘the fourth paradigm”, that signifies “data-intensive scientific discovery’’ where ‘‘all of the science literature is online, all of the science data is online, and they interoperate with each other’’ (Hey et al., 2009). Although the latter clause may not apply fully as yet, there has been decided progress in the direction of the former two.

Data creation or collection is so much a core area of research activity in any field that Borgman (2012) called it the “lifeblood of research”. Over the course of a typical research project, between its beginning and completion, researchers often amass a considerable amount of data, which they tend to store on university servers, external storage devices, and local or cloud storage (Weller & Monroe-Gulick, 2014; Chowdhury et al., 2018). Research teams and communities, as well as universities and research institutions, also, at some point, find themselves in possession of valuable datasets that have potential reuse value, and the question then naturally arises how best to preserve those data both for the present and the future, and prevent possible loss that might adversely affect the reuse potential of the data.

The next section provides the basic background to intelligibly set the foreground of the work as presented towards the end of the chapter. A broader and more detailed background account of RDM follows in Chapter 2. For the present, discourse is confined to definition of important terms and to highlighting the key existing issues in RDM. In subsequent sections within this chapter the scope of the work is clearly delineated, its objectives described, and the research questions

it seeks to address distinctly stated. The final subsection gives a brief outline and summary of remaining chapters.

1.1 Research Background

The notion of “data” is a complex one, as Borgman (2015) observes, and a plethora of associated questions and issues will necessarily come into play in considering a subject of such far-reaching magnitude and almost universal scholarly application. I shall begin with a basic review of the meaning of data itself, and of research data, metadata, Open Access, Open Data, Research Data Management, and Research Data Management Systems. All of these terms represent ideas or concepts that are of essential importance in this work. The following sections distinguish between these terms and establish their different connections with one another.

1.1.1 Data

The foundational element in Research Data Management being data, it becomes important at the outset to set forth an appropriate definition of the word as it is meant to be understood in the context of this work. Data is defined in the Oxford dictionary as “facts and statistics collected together for reference or analysis” and as “quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.” This definition, however, by is relative narrowness is inadequate for the purpose of this work; and I am better inclined towards the more comprehensive definition by Uhlir and Cohen (2011), describing data as “digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations)”, which additionally encompasses other “forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines”. This last definition accords better to the sense in which the word is used in the present work, seeming generally to be more inclusive of the kinds of data to be encountered in the different branches of knowledge. And besides, while it fixes

on the above quoted definition, it still leaves room for flexibility of meaning, and does not outright rule out the non-descript. In so doing it seems to endorse the notion that data may after all be or not be, depending on the eye with which it is looked upon, the angle from which it is regarded, or the purpose for which it is considered. Indeed, according to Borgman (2015), data have no “essence” of their own, but “exist in a context, taking on meaning from that context and from the perspective of the beholder”; and their value may or may not be immediately apparent, or may be transient or sustained. Notwithstanding this latter point, however, data have lately been declared by The Economist (2017) as being “the world’s most valuable resource” and, as previously mentioned, have been called “the lifeblood of research” by Borgman (2012). The word “data” may be used in a singular or plural sense with equal correctness (Borgman, 2015). The synonymous term “data set”, also written “dataset”, is in the singular form which can be pluralized, and represents a particular instance of data or a unit in a collection.

1.1.2 Metadata

Metadata is “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource” (NISO, 2004). Metadata are not part of data itself, but are useful, often necessary, for the understanding and potential reuse of the data (Wiley, 2014). It enables users to find data and allows them to decide whether data meets their particular need. Beyond facilitating data discovery and assessability, metadata supplies the needful information for processing or reusing the data; and it advertises an institution’s research efforts, thus being instrumental in creating partnerships and collaborations through data sharing (Wiley, 2014). The concept, use, and application of metadata recurs frequently throughout this and the coming chapters.

1.1.3 Research Data

Research data, as distinguished from simply “data”, are data, but with the difference that they originate from or aim for research or scholarship. Rice (2009) defines research data as data “collected, observed or created for the purposes of analyzing to produce original research results.” But a more comprehensive definition may be the one by Borgman (2015), describing them as “entities used

as evidence of phenomena for the purposes of research or scholarship", which may range in form from digital records (e.g. text, audio, video, spreadsheets, etc.) to physical objects (e.g. laboratory specimens, historical artefacts, soil samples, etc.). A stricter definition by Weber & Piesche (2016) stipulates, in addition, that research data must be associated with useful metadata, or "information describing its creation, transformation, and/or usage context". Open Data is a term now frequently to be encountered in relation to research data and RDM. It defines how research data may be "published and reused without price or permission barriers" (Murray-Rust, 2008), and differs slightly from Open Access in that the latter concerns not only research data, but publications as well. Throughout this work the single word "data" may be occasionally interchanged for the longer phrase "research data", but always hereafter to be understood as meaning "research data".

1.1.4 Research Data Management

The term "research data management" has a broad significance, and entails "all activities that are associated with the processing, storage, archiving and publication of research data" (Simukovic et al., 2015). It is "the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results" (Whyte & Tedds, 2011), consisting of ''a number of different activities and processes associated with the data lifecycle, involving the design and creation of data, storage, security, preservation, retrieval, sharing, and reuse, all taking into account technical capabilities, ethical considerations, legal issues and governance frameworks'' (Cox & Pinfield, 2014). Proper management of data throughout the research process is crucial for making them openly accessible, intelligible, assessable and usable (UK Research and Innovation, 2016). The stakeholders of RDM include:

i. Researchers: including individual researchers as well as teams, and larger research communities;
ii. Service experts: including librarians, archivists, repository or database administrators and managers, RDM training staff, and professionals in some aspect of RDM;
iii. Research funding and governing bodies;

iv. Policy makers, both institutional and national, as well as regional and international; and
v. The public.

1.1.5 Research Data Management Systems (RDMSs)

Research Data Management Systems are "the technical framework to collect, describe, and provide research data" (Curdt & Hoffmeister, 2015). They are interchangeably and more commonly called RDM systems, data repositories or research data repositories; or, occasionally, data archives, project databases or databanks. The RDM ecosystem is formed of many components, research data repositories being the foremost, and uniting in themselves all the essential functions of data management, including storage of research datasets and making them discoverable for potential reuse throughout the data lifecycle (Arend et al., 2014; Cox & Pinfield, 2014; Curdt & Hoffmeister, 2015; Amorim et al., 2016). As to the exact set of functions or features that an RDMS ought to support, there does not seem to be any clearly-defined limit or uniform consensus so far. Razum (2011) opines that RDMSs should provide services for data storage, search, and user right management; while Lotz et al. (2012) suggests more elaborate features including data linkage with metadata, data version control, support of multiple file formats, persistent identifiers, and access authorization. However, notwithstanding the differences of opinion, the following list compiled by Cambridge Concord Associates for the Interuniversity Consortium for Political and Social Research (ICPSR, 2013), although originally meant only for a certain class of research data repositories (domain repositories), gives a general idea of the functions that data repositories potentially can perform or facilitate and hence the features they might support. Data repositories can:

i. *Manage data* in a way that maintains its understandability and usability for the scientific community
ii. *Facilitate data discovery and reuse* through the development and standardization of metadata
iii. *Provide access* while ensuring necessary protections related to confidentiality and intellectual property
iv. *Create systems that facilitate future archiving* (active data curation) while research is undertaken

v. *Respond to the unique and evolving needs* of scientific communities and other stakeholders
vi. *Partner with each community* to create guidelines for data stewardship throughout the data life cycle
vii. *Advocate* for transparency, data access, and data sharing
viii. *Innovate* in the realm of data curation to address new and evolving forms of data
ix. *Add value* through the creation of data products that align with best practices and new technologies
x. *Collaborate* with related disciplines to achieve interoperability across scientific communities
xi. *Mediate* between scientific communities and digital libraries and archives to implement the latest developments in information science

The above long list by the ICPSR (2013) seems to overlook (or, at least, omit explicitly to state) an important function of RDMSs, i.e. a search or browse facility for finding or exploring repository contents. This, I believe to be implicitly implied in the second point, as being a necessary precedent of "data discovery and reuse". Overall, while the list enumerates a set of roles or functions that RDMSs *could* perform, shorter and more instructional lists of what they *should* perform have been set forth by certain research communities and authoritative bodies, most notable among which are the "guidelines" of the UK Research & Innovation Council (UKRI) and the "FAIR principles" of the Future of Research Communication and e-Scholarship (FORCE11). A summary of the recommendations of the UKRI guidelines and FORCE11's FAIR principles are summarized in Table 1.1 below.

Table 1.1. RDM guidelines by UKRI and FORCE11.

Principle	**UKRI Guidelines**	**FAIR Principles**
Findability (or Discoverability)	✓	✓
Accessibility	✓	✓
Intelligibility	✓	
Assessability	✓	
Usability (or Reusability)	✓	✓
Interoperability		✓

1.1.6 Key Issues and Challenges in RDM

These will be expounded in Section 2.3 in the next chapter, but are listed here as they contribute considerably to the motivation of this research:

i. Insufficient metadata;
ii. Researchers' lack of RDM skills;
iii. Lack of standards;
iv. Inadequate infrastructural support for RDM; and
v. Considerable demands on researchers' time.

Having provided a brief, and for the present, sufficient sketch of the key elements that form the background of this work, many of which will recur or be revisited in further throughout these pages, I now turn more particularly to the work itself. In the sections immediately following I discuss respectively the factors that motivated it; the outline of the research framework; the research objectives; research scope; and the specific questions that the research will investigate and seek to answer. This last is followed by an outline summary of the contents of remaining chapters.

1.2 Research Motivation

Three distinct factors motivated this research; they are as follows:

1. The existence of a clear mismatch between the capabilities of current RDM systems and the special requirements of research data;
2. The advantageous potentialities of linked data to RDM; and
3. The promise of the user-centered design approach as being better suited to solve existing design-related issues of RDM systems (see Section 6.1).

Each is now examined by turns. The first occurs largely in consequence of the current predominant usage of Information Retrieval (IR) systems as RDM systems, although it must be admitted that this is done often with evident, if not adequately successful, efforts at making suitable adjustments. Undoubtedly, since the early days of IR dating back nearly 70 years, there have been continual developments and advances in the area beyond the traditional, and, for data, rather simplistic, TF-IDF and language modelling approaches in which the count of query terms is used as the sole indicator of resource relevance (Mitra &

Craswell, 2018). There now exist new IR models that may demonstrably perform well in data retrieval tasks, despite the additional complexities (see Chapter 4, section) involved (Fuhr & Grossjohann, 2001; Fuhr et al., 2002; Gustafson & Ng, 2008; Kim et al., 2009; Park & Yi, 2016). It is unclear, however, how often or to what extent they may be employed in RDM systems. Also, as many of them are, relatively speaking, new, and as more are yet still being developed, it may be somewhat premature to conjecture on their widespread adoption. On the other hand, to proceed on the assumption that RDM systems are generally built upon traditional IR models is equally to conjecture, without grounds enough for certainty. It therefore remains true that traditional IR models, being designed for the unidimensional and simpler nature of text or string-based objects, are rather an ill-suited solution for such complex and multidimensional objects as data. Suffices to say that at present, RDM systems, design-wise and retrieval-wise, leave much to be desired for the potentialities of data. Data, by reason of their greater variability in respect of (1) manner of user-interaction, (2) requisite software for handling and manipulation, (3) metadata use, and (c) file size range, among others, seem necessarily to require a more particular set of system features than could presumably well be met by ordinary information retrieval systems. This unsatisfactory state of the case, although largely on account of the relative novelty of the problem, is, in view of current expectations, disadvantageous especially for the long-run.

For the second motivating factor, a large-scale study by PARSE.Insight involving nearly 2000 researchers and published by the Association of European Research Libraries (PARSE.Insight, 2010) showed that researchers positively welcome the idea of linking research publications with underlying research data: 85% find this useful. Indeed, research publications have been credited with being “the primary means by which most datasets are discovered” and probably “their only public documentation” (Borgman, 2015). As such, research publications can be an important means to research data discovery, and vice versa; and in this way each is more useful and mutually adds value to the other (Borgman, 2015). A further recommendation by Burton & Koers (2016), highlighting how linked data can variously benefit data repositories as well as other RDM stakeholders, is given below. The points have been slightly reconstructed for the particular purpose of the present research, but their original substance is unaltered; as follows:

i. *For data repositories and journal publishers***.** Linking data and scholarly literature will increase the visibility and usage of both; and can support additional services to improve user experience. E.g. for research datasets, providing links to their associated literature can help to place data in context.
ii. *For research institutes, bibliographic service providers, and funding bodies***.** It will enable advanced bibliographic services and productivity assessment tools that track datasets and journal publications within a common and comprehensive framework.
iii. *For researchers.* It will make the processes of finding, accessing, and, importantly, assessing relevant articles and data sets easier and more effective.

The third motivating factor constitutes one of the key issues of RDM, as highlighted in the preceding section. A user-centered approach, as distinguished from other design approaches such as, for instance, the system-centered or activity-centered design approaches, is one in which marked attentiveness to the user characterizes decision-making throughout the design and development process (Bowler et al., 2011). This is discussed in greater detail in chapters 2 (see Section 2.5.1) and 6 (see Section 6.1), especially as to the why and wherefore of its particularly being regarded by this research as the more suitable design approach for RDM systems. Meanwhile I cite here, as some grounds for the present motivation, some of its general advantages as given by the International Organization for Standardization (ISO, 2010) in ISO 9241-210. They are that user-centered design:–

a. Increases user productivity and operational efficiency;
b. Supports reduction of costs, e.g. of training, by resulting in easier and more understandable products;
c. Widens the range of users that may benefit from the products, e.g. by including accessibility features, resulting in better usability and user experience;
d. Reduces discomfort and stress; and
e. Contributes to sustainable goals.

This commendation seems not to be without reasonable grounds, as abundant literature may be found even beyond the confines of the field of RDM that support in theory, substantiate in practice, or otherwise corroborate one or more of the above named points.

1.3 Research Aims and Objectives

The research presented here aims to explore and enquire into the user-centered design approach as applied to RDM system design, for reasons stated in the last section. This object is carried out practically through the means of gathering requirements for and designing, developing, and evaluating a simple prototype of user-centered RDM system, called DataFinder. However, weight is attached as much to the final deliverable (i.e. DataFinder) as to the process of attaining it, since a substantial part of the research work preceded the actual system design and development phases and the various findings accrued may prove useful for many purposes of which system design forms only one; the other benefits or potential applications of the findings could be in developing, for example, researcher training; RDM policies; research services & support, etc. The overall research aim may be broken down into 5 smaller objectives, as follows:

Ob. 1. *To gain an in-depth understanding of RDM system users and their tasks sufficient to enable the specification of their requirements for system development.* A user-centered design entails an in-depth understanding of the different user groups that have stakes in the system and the roles that they play. This will not only guide decisions about user interface design and features that the system ought to support, but also help to prioritize them. The objective involves:

a. Obtaining a sufficiently thorough, descriptive, and discipline-specific appreciation of researchers' experiences, attitudes, concerns, and habits regarding research data services, with a view to discovering the intra-disciplinary similarities as well as inter-disciplinary dissimilarities and similarities;
b. Gathering practical information about researchers' data-seeking needs, strategies and difficulties;
c. Identifying the different user types or groups that will potentially use the system, and their joint as well as unique requirements; and

Ob. 2. *To compare and contrast between the system requirements of research data and of research publications, with a view particularly to modelling the former in a user-friendly way.* This involves:

a. Designing as well as conducting an experiment to explore the key differences between the potential system requirements of data and text objects;
b. Define a relevant set of criteria to assess the degree to which the repositories cater to and are adapted to the special requirements of research data;
c. Reviewing the commonly as well as the less commonly supported features and functionalities of existing RDM repositories; and
d. Assessing the implications of the above, resource-wise and design-wise;

Ob. 3. *To identify and review key problem areas in the status quo;*

Ob. 4. *To understand the user-centered design approach and process, especially as is relevant or applicable to RDM systems;*

Ob. 5. *To synthesize all of the above into design specifications upon which to develop DataFinder.* This involves collating, synthesizing, and translating all the previous findings of the research into a set of system requirements (functional and non-functional) with priority indications;

Ob. 6. *To build a working prototype of the system and test it with real users.*

1.4.1 Research Questions

Through the objectives enumerated in the preceding section, and tying back to the motivating factors set forth previously, this research seeks to investigate and answer the following questions:

RQ. 1. What do researchers as primary system users expect, require, or want of RDM systems; and what variety of roles do they fulfill with respect to their interactions with RDM systems?

RQ. 2. Do background factors, such as researchers' disciplinary domain or extent of experience have any bearing upon the first clause in the point above?

RQ. 3. What are the general requirements of RDM systems and how do they differ from those of information retrieval systems?

RQ. 4. What are the key design requirements of user-centered RDM systems, as distinguished from those of ordinary RDM systems?

RQ. 5. What problems and challenges are current to RDM generally and RDM systems specifically?

RQ. 6. What is the role of metadata in RDM system design and use, and what basic elements of it are required for developing and implementing RDM systems?

These research questions will be enquired into by means of a combination of research methods including questionnaire surveys, face-to-face interviews, a technical experiment, a systematic appraisal of existing RDM services, and also user-evaluation studies.

1.4 Research framework

The main activities of the research are sectioned into three sequential phases: (I) Information Gathering; (II) Prototype Design & Development; and (III) System Evaluation. The development of the research framework was guided chiefly by the consideration of conforming to and being informed by the accepted principles of user-centered design. Accordingly, therefore, the outline presented in the diagram below (Figure 1.1) was broadly built around the four "base activities" identified by Zimmermann & Grötzbach (2007) as common to user-centered design process models. They are:

To–

i. Understand and specify context of use;
ii. Specify user (and organizational) requirements;
iii. Design solutions; and
iv. Evaluate the designs against requirements.

It will be observed that the four activities listed above are correspondingly matched, if in a loose way, by those outlined in Figure 1.1. A cross-tabulation of this is presented in Table 1.2 following.

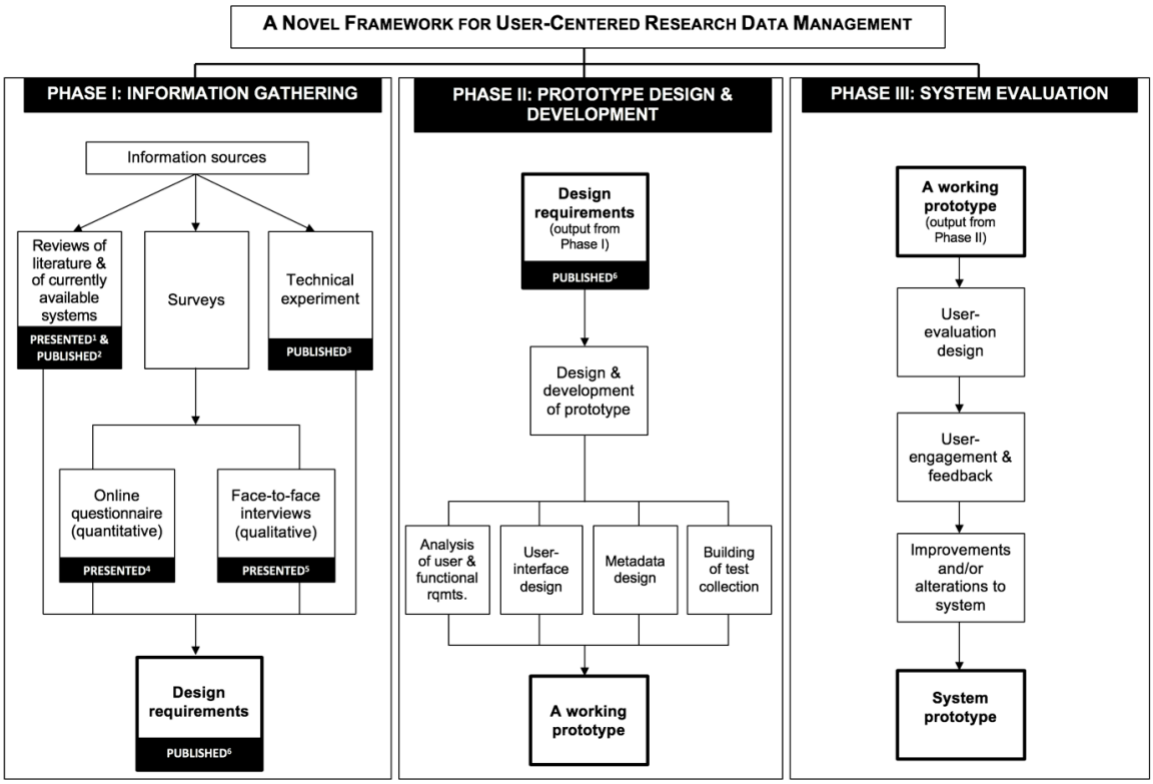

Figure 1.1. Overall outline of the research

Table 1.2. How the research framework fits with user-centered design process models.

User-centered design "base activity"	Research Phase	Chapter
1. Understand and specify context of use	Phase I. Information gathering	2 and 3
2. Specify user requirements	Phase I. Information gathering	4 and 5
3. Design solutions	Phase II. Prototype design and development	6
4. Evaluate the designs against requirements	Phase III. System evaluation	7

1.5 Research Scope

The bounds of this research enclose principally the necessary preliminary work that is needed judiciously to guide and inform the design of a user-centered RDM system, and the development and testing of a simple working prototype of the

same. It is not purposed that the prototype should be a full-fledged and perfectly finished version of the RDM system that the research ultimately proposes, but only a small-scale version with which to demonstrate and evaluate some of the important findings of the research. It does not fall within the purview of this research to probe into the details and technical intricacies of information retrieval, nor into the technicalities of query processing and optimization in databases and database management systems: this research is strictly limited to improvements and enhancements of a less rudimentary order than these entail. Neither is it part of the purpose of this research to examine more closely than is positively relevant to its aims, the innumerable sub-themes concomitant to RDM and RDM systems.

1.6 Thesis Outline

A brief outline of each of the chapters that remain is presented as follows:

- **Chapter 2: Review of Literature.** This chapter critically considers the published literature that theoretically underpins this research, as well as the practical concerns and questions that supply its foundational material. RQ 5 is partially addressed in this chapter;

- **Chapter 3. Methodology.** This chapter details about the various studies conducted in pursuance of the research questions and objectives already set forth. Four different studies detailed therein were conducted each to address specific areas of the same;

- **Chapter 4. Data Analyses.** Here the data collected in the preceding chapter are analyzed using relevant tools and software, and the resulting findings are examined and discussed. The chapter answers RQs 2 and 3, and adds to the answer to RQ 5 partially addressed in Chapter 2. It also supplies preliminary answers to the first part of RQ 1;

- **Chapter 5. Requirements Analyses.** Output from the preceding chapter and also from Chapter 2 are here translated into a list of system requirements. These are then prioritized accordingly for the next step in the process. The chapter fully addresses RQs 1, 4, and 6; and completes the answer to RQ 5 which was partially addressed in Chapters 2 and 4.

- **Chapter 6. System Design.** This chapter ties together all of the previous work into one final deliverable. Using as input the output of the preceding chapter, it develops and also presents the final plan of the system and its actual development. Screenshots of the final product are also shown.
- **Chapter 7. User Evaluation.** Details about the testing of the system with a small number of real potential users are given in this chapter; and
- **Chapter 8. Conclusion and Recommendations.** This makes some pertinent references to, and closing reflections about, the original research questions and objectives. A section is dedicated to considering possible contribution(s) to knowledge made by the research. Observations about future work and wider potential of the research are also made.

2.0 LITERATURE REVIEW

RDM is a multifaceted research area. It unites within itself a wide range of issues of both practical and theoretical concerns, and draws expertise from a variety of disciplinary domains; including among others, Library Scientists for metadata design and digital cataloguing; IT & Computing expertise for software and infrastructure design and development; subject domain expertise for disciplinary advice and support; Data Science knowledge for skills training; etc. This research thus comprehends a multiplicity of topics and interests, RDM systems being a key component of the RDM ecosystem and serving, wholly or partly, many of its important functions. The diagram in Figure 2.1 gives a structured, high-level overview of the field and is accordingly proposed for the general plan of this chapter. The broad outline it presents is not in itself intended as a formal representive model of the RDM ecosystem, but as a temporary structure of convenience to better organize and contextualize the various topics for review. Nonetheless, the outline is the result of a long course of reading and contemplation, and in fact reflecting the actual case as regarded from a certain point of vantage.

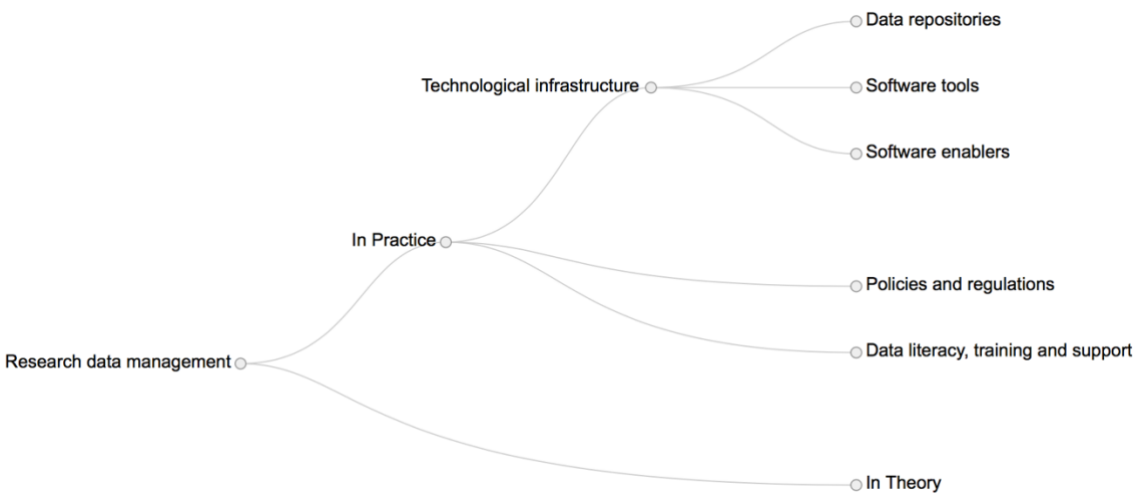

Figure 2.1. An outline overview of the RDM ecosystem

In the sections that follow, each area noted in Figure 2.1 is discussed in accordance with its relevance to the larger work. Accordingly therefore, the section discussing data repositories is the most extensively treated. The chapter concludes by highlighting the key existing problems of RDM as gleaned from the literature, showing a gap which it is hoped that this research will contribute

towards filling. Be it observed, as a side note, that some areas are only briefly covered in this chapter as they seem more indispensably and pertinently to belong to the concerns of one or other of the remaining chapters. Also, by placing everything in so early a chapter I shall greatly be anticipating myself and requiring in later chapters to have recourse to superfluous repetition.

2.1 An Overview of the RDM Ecosystem

The diagram in Figure 2.1 demarcates RDM into areas more practically or theoretically inclined. Although this bipartite demarcation, as previously noted, has no definite recognition in the field, it will be generally agreed, upon consideration, that it nonetheless exists; inasmuch there is one side of RDM more principally occupied with providing a philosophical basis for its proceedings, and another whose occupations produces results of a more materially tangible or applied nature. The former, in other words, leans more towards theory. It comprises formal models and systemized representations of concepts, processes and workflows, including studies of user behavior, among others. The Research Data Lifecycle is discussed under this head in the next section. The latter, of a more practical leaning, largely comprises the development and establishing of those structures and systems (IT and otherwise) for the successful operation of RDM activities and accomplishment of its goals. The present research, being ultimately concerned with the design and development of an RDM system, comes largely under this head. Nevertheless, a substantial part of the work leading up to the design stage involved activities tending towards the establishment of a theoretical underpinning. Models of the research data-seeking or retrieval behavior of users would have been of particular use, therefore, and worthy of a careful review in this chapter, but none, to the best of my knowledge, have yet been developed or proposed in the literature as at time of writing. As Bremer & Gertz (2005) likewise observe, such models “would provide insight into the needs and practices of users that could be applied to both systems design and policy developments for facilitating data discovery.” Nevertheless, some recommendations for the same have resulted from this research and are given in a later chapter (see Section 7.5.1).

Policies & regulations, technological infrastructure, and data literacy, training & support represent the three major areas classed under RDM-in-practice as shown

in Figure 2.1. This is on the grounds that it is primarily the harmonious working together of these three components that moves the machinery of RDM as currently practiced. For, while policies and regulations set requirements and enforce their conformity; technological infrastructure operate in various ways to support or facilitate the activities and core functions of RDM in accordance with the aforementioned policies and regulations; and training and support efforts help to provide or develop the requisite skillset to ensure proper usage of infrastructural products in correct compliance with policies and regulations. Figure 2.2 pictorially illustrates this interplay among the three.

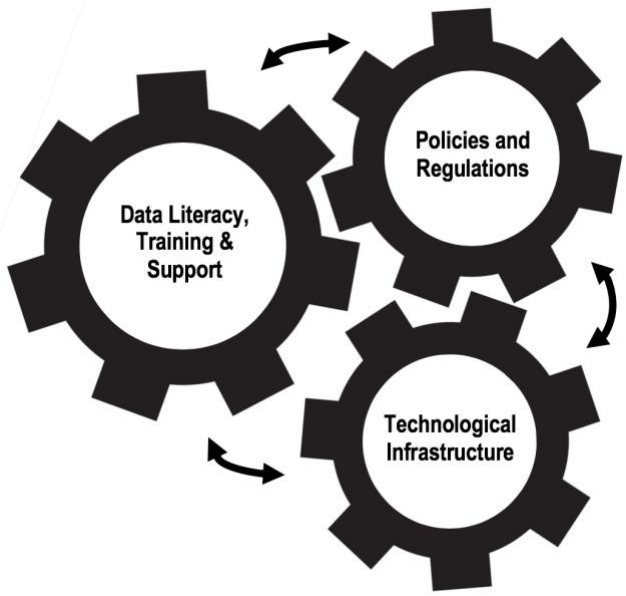

Figure 2.2. The major driving-forces of RDM in Practice

Research data are governed by policies and regulations at various levels of the administrative hierarchy and stages of the research project. At the university level, for example, focus tends to be on how to store data, ensure privacy, and comply with ethical regulations. At the research funder level, the focus is increasingly becoming on ways to promote data sharing. RDM policies, and, specifically, those on data sharing, are covered in Section 2.2.

Technological infrastructure comprises software tools and enablers, and data repositories. Software tools, such as DMPonline[1], assist researchers in accomplishing data management tasks, including the writing of Data Management

[1] http://www.dcc.ac.uk/dmponline

Plans (DMPs) in compliance with funder regulations. As these tools are quite separate entities from RDM repositories, and are scarcely relevant to the main concern of the present research, discourse on them is limited to this brief mention. Software enablers, on the other hand, represent mechanisms that drive the various infrastructural components; e.g. metadata (see Section 5.2.1) and persistent identifiers (see Section 5.2.2). Data repositories are treated in detail in a later section.

Having generally, and with suitable brevity, touched upon those areas of the RDM ecosystem that are of a relatively minor significance to the work at large, I now proceed to discourse at greater length the more important areas, which are as follows:

- Research data policies and regulations;
- Research data sharing and reuse;
- The Research Data Lifecycle;
- RDM system design and development; and
- Current issues and challenges in RDM.

They are considered by turns in the sections immediately following.

2.2 Research Data Policies and Regulations

In the US, the National Institute of Health (NIH, 2003), National Science Foundation (NSF, 2011), Office of Science and Technology Policy (OSTP, 2013), and the Department of Energy (DOE Office of Science, 2014) have all of them issued formal guidelines and requirements for DMPs. In the UK the Natural Environment Research Council (NERC, 2010), Engineering and Physical Sciences Research Council (EPSRC, 2011), Medical Research Council (MRC, 2016), The Wellcome Trust (2017), and the Arts and Humanities Research Council (AHRC, 2018) have likewise done the same. A UK multi-stakeholder group which includes the Higher Education Funding Council for England (HEFCE), Research Councils UK (RCUK) currently called UKRI (UK Research and Innovation), Wellcome Trust, and Universities UK developed in 2016 the “Concordat on Open Research Data”. This “concordat” was a document in which were outlined ten “principles” to “help ensure that the research data gathered and

generated by members of the UK research community is made openly available for use by others wherever possible…" It was moreover stated (Principle #6) that "good data management is fundamental to all stages of the research process and should be established at the outset", and that "the careful management of data throughout the research process is crucial if the data arising from research projects is to be rendered openly discoverable accessible, intelligible, assessable and usable". The statement ended with the assertion that "it is essential therefore that the management of research data is considered from the beginning of the research process and due consideration is given to how research data are to be managed" (UKRI, 2016).

As has been seen, research data sharing mandates, since recent times being issued in ever increasing numbers by research funding and governing bodies, have greatly contributed to the growing demand for RDM and RDM systems especially. Journals and other research publishers, also, now request or require open access to data for submitted research papers (e.g. PLOS since 2014 and Nature since 2016). From the former, the pressure is attributable to a general desire to add value to expensive research and to stimulate cross-disciplinary research efforts for solving grand challenges (Borgman, 2015); whereas the latter has the end in view, of promoting scientific transparency and reproducible research (Borgman, 2007). For these and more reasons besides, grant applications are now generally required to be accompanied with Data Management Plans (DMPs). DMPs are written documents intended to address questions concerning the use and disposal of research data during and after project completion (Strasser, 2015). As Wiley (2014) affirms, DMPs "can identify types of data being collected; use of metadata and data gathering procedures; as well as policies and mechanisms for sharing data". As a large and important portion of the data preservation activities and measures outlined in DMPs fall within the responsibility of RDM systems, it is particularly apt to the purpose of this research that DMPs should be further explored, presently. This, it is believed, will in some way guide system design decisions if not directly to meet these requirements and expectations in some degree, then at least to avoid inadvertent conflict with them. The specific content and structure of a DMP is dependent on the research project and the particular solicitation, as well as the agency awarding the grant (Thoegersen, 2015), but the purpose and the information they contain

are nonetheless generally the same. Strasser (2015) gives the following as representing the main particulars specified in DMPs:

i. A description of the type(s) of data that will be collected or generated during the project;
ii. The standards that will be used for those data and their associated metadata;
iii. A description of the policies pertaining to the data that will be collected or generated;
iv. Plans for archiving and preservation of the data generated; and
v. A description of the resources that will be needed to accomplish data management, including personnel, hardware, software, and budgetary requirements.

The above list, no doubt, is rather condensed. A more comprehensive list is given by the Interagency Working Group of Digital Data (IWGDD, 2009), organized under seven points purposed to be addressed by DMPs, viz. data description, potential data impact, data content and format, data access provision, data protection plans, data preservation plans, and arrangements for transfer of responsibility should it arise. In an analysis of the DMP requirements of 10 federal research funders in the USA, data access was alone found to be the common point addressed by all –a not surprising fact considering that free public availability of publicly funded research factors as a major driving force behind the requisition for DMPs (Thoegersen, 2015).

The next section follows up the thread of the present one by examining certain important aspects of data sharing which might influence the use and consequently the design of RDM systems.

2.3 Research Data Sharing and Reuse

Data do not diminish in value when shared, and hence are a classic example of public good (Vision, 2010, pp 330). As science becomes more data intensive and collaborative, data sharing becomes more important (NSF, 2010). It follows logically that data must be shared to render the possibility of reusing them. The existence of data repositories, where no data will be shared, is at best superfluous

since data sharing is the connecting link between research data held by a certain party and its potential reuse by others. Indeed, it may be argued that nearly all RDM efforts, from the building of repositories to the training of researchers, directly or indirectly tend towards making smoother the task of data sharing or removing existing obstacles thereof, whether this last be infrastructural limitations, researchers' unease about sharing, or anything between the two extremes. According to Borgman (2015), the "fundamental problem", for most researchers, is how to better manage their own data. They require "tools, services, and assistance in archiving their own data in ways they can reuse them, which increases the likelihood that their data will be useful to others later". Data reuse is driven by the notion that data are not only the outputs or by-products of research, but may serve as inputs to new hypotheses (NSF 2008). Indeed, in the opinion of Borgman (2015), this "repurposing" of data for unanticipated questions, as distinguished from simply "reusing" data for the same old questions, is an even "higher goal". The relationship between Open Data and data sharing may be explained by the following analogy: that, considering data reuse to be the end, data sharing may be regarded as the means to it, and Open Data an enabler thereof. Arguments in favor of data sharing abound in the literature, chiefly centering around:

i. *Research transparency***.** By laying it open to validation by others and to public scrutiny, data sharing enhances transparency in the research process, including in data collection methods; thereby minimizing research misconduct (Borgman, 2007; Tsang, 2013; NERC, 2016; Patel, 2016).
ii. *Leveraging public investments in research and scholarship***.** Data are expensive to collect and may be unique or impossible to replicate (e.g. data about a rare event in nature or history) (Borgman, 2015; Murray-Rust, 2008; Henty et al., 2008; RDA, 2014). Also, some data may need to be pooled together from a variety of sources and combined, beyond the compass and resources of one research team, time or location (Borgman, 2007).
iii. *Reproducible research***.** Data sharing enables the integrity of original research to be established, challenged, or reaffirmed by usage (Helbig, 2016; NERC, 2016; Patel, 2016).

iv. *Fostering research and innovation through increased collaboration.* Data sharing may open up new lines of enquiry in old data as well as encourage new hypotheses (NSF, 2008; Kaiser 2013) and methods of investigation. Moreover, new discoveries may potentially arise from fresh or secondary analyses (Markauskaite, 2010; Tsang, 2013; RDA, 2014), and better opportunity is given for cross-disciplinary problems to receive more suitably-qualified research attention (Borgman, 2007; Cragin et al., 2010).

v. *Scholarly recognition***.** Researchers get credit, in the shape of data citation, for sharing data (Tsang, 2013; Patel, 2016). Data citation increases trust in the data (Patel, 2016).

vi. *Better use of time***.** Data sharing may save researchers time that would otherwise have been expended in duplicating research effort through collecting already existing data over again (Patel, 2016). Researcher productivity is thus improved.

Studies by Faniel & Jacobsen (2010), Pienta et al. (2010) and Wallis et al. (2013) indicate that data sharing may be more commonly practised privately than via data repositories, thereby rendering difficult the task of tracking it and impracticable the prospect scaling for use. On the other hand, sharing data on research data repositories may prove ultimately less burdensome to the data-holding party, although control over how the data are reused and by whom, both of major concern to researchers (Chowdhury et al., 2018), is lost. A more in-depth treatment of data sharing is to be found Chapter 4. More pertinent to our discussion for the present, after having presented the rationales principally pleaded in justification of it, is a review of the chief criticisms against it. This is done in hopes that some may prove possible of being in some degree ameliorated by a more judicious design of research data repositories which are its chief instrument.

i. Among the most important criticisms levelled on data sharing is that advocacy for it tends disproportionately to place much greater focus on the mere act of sharing data, and less on the task of making the data possible of being reused when shared. In other words, researchers are encouraged to *share data* rather than to *share reusable data*. As Borgman (2015) observes, making data publicly available and making them reusable are

different issues: once a researcher finds an appropriate dataset, the next important question is whether the data can be reused (Mannheimer et al., 2016). Criticism on this point even argues that mere sharing or publishing of research data might actually be detrimental by contributing to “an increase in noise and opacity” (Günther & Dehnhard, 2015); for, “as more and more data is made available, researchers are finding it increasingly difficult to discover and reuse these data” (Dumontier et al., 2016).

ii. It has also been argued that research transparency is not necessarily increased simply by sharing data; and that what is wanted, rather, is “intelligent openness”, which additionally requires that data be “effectively communicated” (The Royal Society, 2012; Günther & Dehnhard, 2015). This idea of endeavoring for “intelligent openness” rather than simply “openness” of research data becomes of especially paramount importance upon the consideration that data sharing is expected to burgeon cross-disciplinary research, and that this expectation raises new challenges. “Intelligent openness” is particularly crucial given the inevitable differences in meaning and context across disciplines as foreign data is imported for local use; or, to use a term coined by Baker & Yarmey (2009), the “distance-from-data-origin” (Günther & Dehnhard, 2015).

iii. Another major criticism of data sharing relates to the imputation of a possible misplacement of priorities on the parts chiefly of research funders and journals, whom it is supposed misapprehend the greater import of investing in the improvement of existing data sharing systems and infrastructure. As Borgman (2015) observes, data are liabilities as well as assets; and, even while literature has yet to clearly demonstrate whether publicly shared research data are being discovered and reused (Mannheimer et al., 2016), the general preoccupation with increasing the supply of such data has left largely disregarded the scholarly motivations for sharing or reusing them, as well as the required investments in knowledge infrastructure to ensure their use with greater ease and material benefit (Borgman 2015).

Sharing and reuse represent only one activity of RDM over the lifetime of a particular dataset. The next section on the lifecycle of research datasets introduces others.

2.4 The Research Data Lifecycle

The creation and preservation of research data is a process that entails a series of often elaborate steps. Even if not meant for preservation, the disposal of data, especially where sensitive, entails more than mere deletion. Data preservation activities continue for as long as data holds prospect for future use. Many models depicting the successive stages in the existence of research datasets have been proposed, each patterned after a particular idiosyncrasy (e.g. a certain domain of research) or to serve a specific purpose (e.g. for general reference; or to assist in research planning or data management). The models provide a “structure for considering the many operations that will need to be performed on a data record throughout its life” (Ball, 2012), by identifying “the steps to be taken at the different stages of the research cycle to ensure successful data curation and preservation" (NTU, n.d). Some lifecycle models are simpler and can be more generally applied (e.g. the DDI[2], DataOne[3], and UKDA[4] models), while others are more granular and comprehensive (e.g. the I2S2 Idealized Scientific Research Activity[5] and the DCC models[6]). An appreciation of the lifecycle of research datasets is pertinent to the broader purpose of this research, as it points to where in the larger scheme of the existence of research data this work is particularly concerned, as well as its overall significance in relation to the whole. Besides this rationale, Pennock (2007) names the following three excellent factors that necessitate the adoption of a lifecycle approach to RDM. And the changing nature of technology and information systems, coupled with the need to ensure continued accessibility and reusability of stored data, make them particularly worthy to be regarded:

i. Digital materials are fragile and susceptible to change from technological advances throughout their life cycle;

2 http://www.ddialliance.org/Specification/DDI-Lifecycle/
3 https://www.dataone.org/data-life-cycle
4 https://www.ukdataservice.ac.uk/manage-data/lifecycle
5 https://researchportal.bath.ac.uk/en/publications/i2s2-idealised-scientific-research-activity-lifecycle-model
6 http://www.dcc.ac.uk/resources/curation-lifecycle-model

ii. Activities (or lack thereof) at each stage in the life cycle directly influence our ability to manage and preserve digital materials in subsequent stages; and

iii. Reliable re-use of digital materials is only possible if materials are curated in such a way that their authenticity and integrity are retained.

In all three the above considerations RDM systems may have some part to play in helping to mitigate some of the complexities involved.

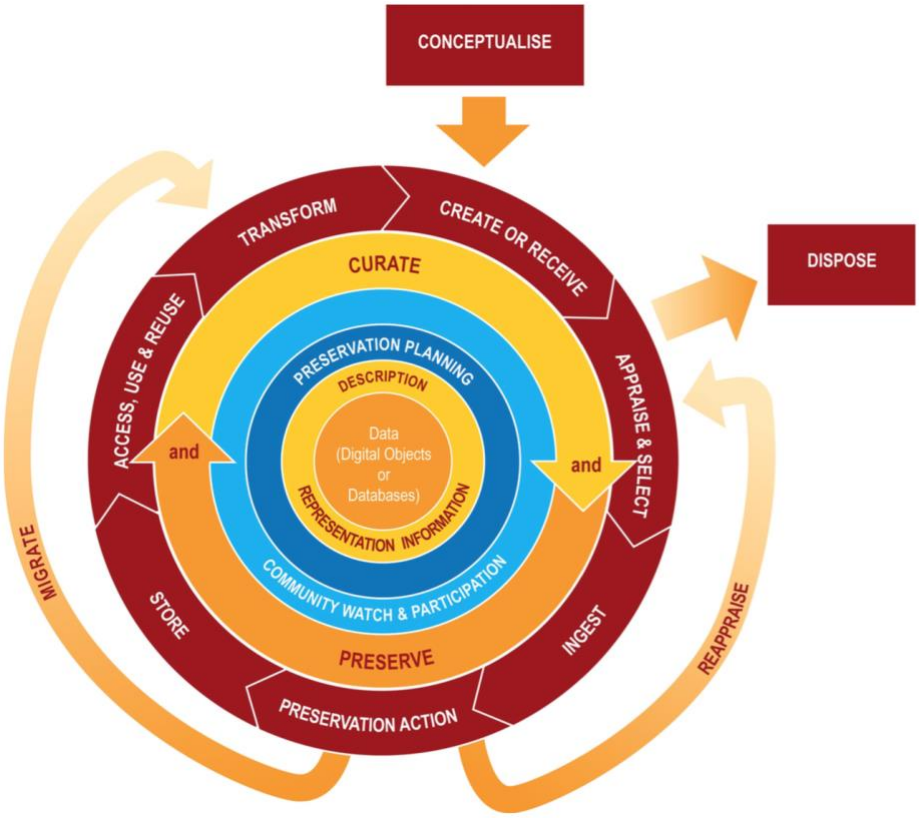

Figure 2.3. The DCC Data Curation Model

Two lifecycle models are here presented for examples: the DCC Curation Lifecycle Model (Figure 2.3) and the UKDA Research Data Lifecycle model (Figure 2.4). Both are quite well-known and more or less typify other models, hence the choice. The former lays emphasis on data curation and preservation, and may be used to plan data management activities. Research data repositories are involved, to a greater or lesser extent, in most of the activities and steps highlighted in the model. On the outermost ring, for example, in which are highlighted data storage, access, use, and reuse, research data repositories, as shown in the preceding section, shoulder much of the burden thereof. Likewise,

as the nexus connecting data creators, consumers, and professionals, research data repositories play an important role in community watch and participation (see inner ring).

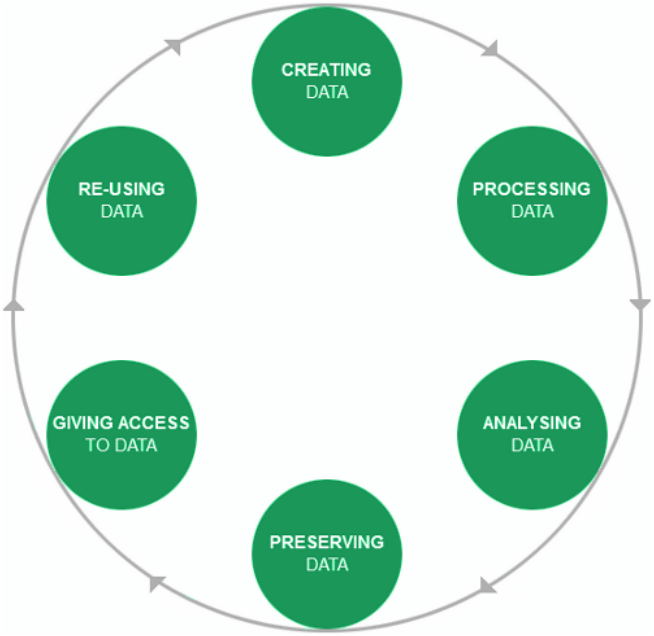

Figure 2.4. The UKDA Data Lifecycle Model

The second model, the UKDA Research Data Lifecycle model, addresses the lifecycle of an actively used dataset, and may aid in comparing how data management activities correspond to the stages of a research project. This particular model resembles other lifecycle models closely enough to make it a fair representation of them. The following are the six main stages in the existence of a research dataset, according to the model:

The–

i. *Data creation* stage, which comprises: designing research; planning data management (formats, storage etc.); planning consent for sharing; locating existing data; collecting data (experiment, observe, measure, simulate); and capturing and creating metadata.
ii. *Data processing* stage, which comprises: entering data; digitizing, transcribing, translating, checking, validating, cleaning, anonymizing where necessary, describing, managing, and storing data.

iii. *Data analysis* stage, which comprises: interpreting data; deriving data; producing research outputs; writing publications; and preparing data for preservation.

iv. *Data preservation* stage, which comprises: migrating data to the best format; migrating data to a suitable medium; creating metadata and other documentation; and backing-up, storing and archiving data.

v. *Data access* stage, which comprises: distributing, sharing, and promoting data; controlling access; and establishing copyright.

vi. *Data reuse* stage, which comprises: follow-up research or new research; undertaking research reviews; scrutinizing findings; teaching and learning.

The present research mainly concerns the final two stages, viz. giving access to data, and facilitating data reuse, in both of which research data repositories play a major if not a primary role. However, the work also touches lightly upon other areas, such as preservation.

2.5 RDM System Design and Development

The primary components forming RDM systems at the basic level, as Figure 2.5 illustrates, are a user interface, a retrieval mechanism, and a database or file storage system. Design and development may begin at the level of any one of these, with the processes and functionalities of the components below it wholly or partially abstracted. For instance, some RDM systems provide only a user interface for search and discovery, relying on third-party retrieval engines for query-processing and external databases for the needful datasets. A distinctive example of this is the Research Data Registry and Discovery Service (RDRDS)[7], developed in the UK by the Joint Information Systems Committee (Jisc)[8], shortly to be presented. According to Witt et al. (2009) research data discovery systems describe the metadata and points of access needed for searching and browsing data repositories; and also ways of helping external users and user agents (such as search engines) to find data. Another set of RDM systems provide search and discovery through their own query processing systems, but are dependent on external databases to provide the needful datasets. The electronic Data Archive

[7] http://researchdiscoveryservice.jisc.ac.uk/dataset
[8] https://www.jisc.ac.uk

Library, e!DAL[9], also to be presented shortly, is an example in this category. Finally, some RDM systems, of which this research is the most concerned, are designed inclusive of all three components. Unfortunately, information and case studies reporting on the various design and development processes of these are, in general, not openly available in the detail and completeness that would have proved of great use for the present research.

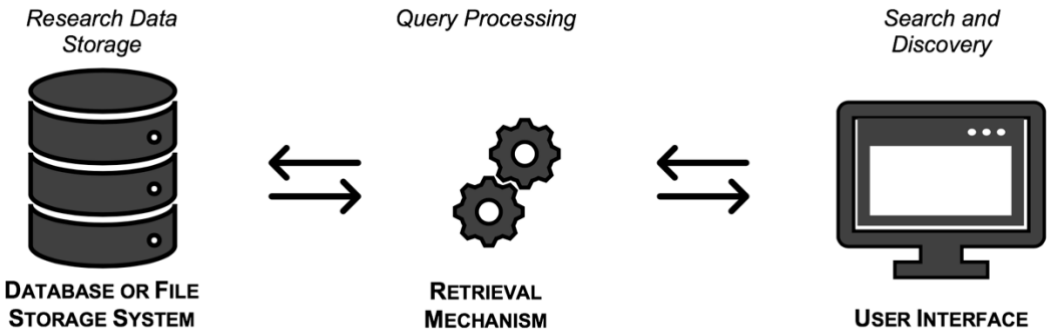

Figure 2.5. Basic components of an RDM system

The RDRDS by Jisc, alluded to in the preceding passage, provides search and discovery services for datasets held in subject-specific data centers and university data repositories across the UK. It holds no datasets of its own, nor assesses the quality of exposed datasets, but works in partnership with a network of UK-based institutions to harvest metadata records and expose them through a national registry of research datasets (Davidson et al., 2014a). The project aimed to build a shared data access facility, and partnered with institutions including the Visual Arts Data Service (VADS), the Natural Environment Research Council (NERC) Data Catalogue, the UK Data Archive, as well as the Universities of Oxford, Edinburgh, Southampton, Glasgow, and Hull (Ball et al., 2014) among others. The RDRDS in the UK is akin to Research Data Australia (RDA)[10] in Australia, and utilizes the experiences of the Australian National Data Service (ANDS) in developing the same (Davidson et al., 2014b). e!DAL, also mentioned previously as an example of an RDM system designed with search and discovery as well as query processing capabilities, comprises several index-based search functions which allow efficient retrieval over metadata. It is an open source system for

[9] https://edal.ipk-gatersleben.de
[10] https://researchdata.ands.org.au

sharing and publishing research data in compliance to open data policies (Arend et al., 2014). It, too, does not own its own data, but "brokes" between in-house file storage and data registries. The scant information available on the design and development processes of these systems limits the possibility of being able sufficiently to compare them with the reasons and processes of making key decisions appertaining to the system that this research seeks to develop. Consequently, it has not been discovered, for instance, that any of the records about the RDM systems mentioned in the preceding passages stated explicitly the design approach that was followed, whether user-centered or otherwise. But, notwithstanding this omission, an attempt is made in the next section to provide with the example furnished by the Biomedical Translational Research Information System (BTRIS)[11], about the design and development of which there is comparatively more information available. Meanwhile, user-centered design approach being an important element in this research, I begin with an overview of it.

2.5.1 User-Centered Design

Guidelines and best practices respecting specifically the design of RDM systems are as yet not certainly established. While at this stage of experimenting with new ideas and methods a danger might reasonably be apprehended, of giving more than due emphasis on the system and the features and capabilities it must support, and less on the system users themselves and their need. User-centered design is a design approach with a standing ethos of showing due consideration to the needs and situation of the individuals that will potentially use a system, in the design of that system. According to Ames (2001) it is a process that entails analysis, design, and user validation; and not just development and implementation. Indeed, in user-centered design user participation seems almost to be regarded as a right, rather than a need, of users. Innumerable formal definitions of the term "user-centered design" exist, such as by Gould & Lewis (1985), Constantine (2004), Zimmermann & Grötzbach (2007), Xia & Li (2009), Martinez-Alcala et al. (2014), and others; but all converge on the one point of the approach being associated with the idea of increased attention to system users. As observed by Bowler et al. (2011), the defining characteristic of user-centered

[11] https://btris.nih.gov

design is user involvement, the goal being “not just to create something that works” but “something that works for the intended user”. A principal rationale and argument for following the user-centered approach in designing RDM systems, besides those born of the intrinsic advantages of the approach in itself, is that research data repositories are targeted to end-users and therefore are end-user systems. User satisfaction becomes in such a case a signal indicator of the system’s ultimate success or failure, and an important factor for or against it.

Next comes a discussion of the design and development process of the RDM system, BTRIS, referred to in the last paragraph but one. The discussion is organized under the basic components of RDM systems as have been identified earlier. BTRIS, considered in the light of the “principles” of user-centered design as given by Satzinger et al. (2016, p. 220), might be judged to have been developed by this approach though it is not explicitly stated in the available documentation. This postulation will be inquired into presently; meanwhile, the “principles” are as follows:

1. Focus early and throughout the project on the user and the user’s work
2. Evaluate all designs to ensure usability
3. Use iterative development

BTRIS was developed by the US National Institutes of Health (NIH) and comprehends a suite of software programs of which a data repository is one, all designed to facilitate researchers’ access to translational and clinical datasets collected at the NIH. The development process followed “good software development practices”, with a focus on four basic requirements (Cimino et al., 2014), namely:

i. The ability to accommodate any type of data that might be encountered,
ii. A database design optimized for the kinds of queries likely to be performed,
iii. Use of a controlled terminology that would include detailed terms encountered in data as well as the high-level concepts that users are likely to include in their queries; and
iv. A user interface that would empower NIH researchers to carry out their own queries (Cimino et al., 2014).

Like the present research, a small-scale prototype of it was developed first, to "better understand the requirements for data representation, storage and retrieval, as well as to elicit use requirements" (Cimino et al., 2014). Table 2.1 below describes the design and development of the system's three basic components, while Table 2.2 evaluates it against the principles of user-centered design that have been quoted above.

Table 2.1. Design and development of the basic composite units of BTRIS

Component	Detail
User Interface (for search and discovery)	BTRIS supports the use of a drag-and-drop implementation of a sophisticated term look-up application called RED Web Search, to conduct searches against specified concept attributes (e.g. names, synonyms, local codes, etc.) within a specified part of the RED hierarchy. The system assembles terms that match each of the user's search term into a corresponding hierarchy on two levels which the user can expand further to reveal more specific terms. The user interface also supports the integration of visualization tools and allows Principal Investigators to view data identified as being associated with their own studies (Cimino et al., 2014).
Retrieval Mechanism (for query processing)	A ''business intelligence'' tool initially served as the user query tool. Data queries were conducted using query templates which had been created for each data type. The query templates required the user to specify one or more research studies of interest, with optional specifications for subsets of research subjects, Boolean relationships, date ranges, value ranges, controlled terms. The query features also allow the user to combine data across multiple domains (Cimino et al., 2014).
Database or File Storage System (for storing research datasets)	The BTRIS database was implemented using a relational database, designed to accommodate the wide variety of NIH datasets. A "convenience sample" of 29 studies were used to create a simple database for the initial dataset, containing approximately 4000 data objects on demographics,

	laboratory test results, medication administration, radiology reports, patient diagnoses, etc. A controlled terminology was then constructed to represent and code the terms found in these datasets, distinguishing between data elements representing relatively stable statements about real-world objects (e.g. date of birth, gender) and facts about those objects that will be added to the database (e.g. body weight, laboratory results) (Cimino et al., 2014).

Table 2.2. Evaluating BTRIS against the "principles" of user-centered design according to Satzinger et al. (2016)

Principle	**Evidence**
Principle 1: Focus early and throughout the project on the user and the user's work	As part of the development process of BTRIS, user groups were assembled to provide feedback on the design features and user interface. A demonstration prototype was made available to the NIH research community for two months, and "invaluable" information was gathered about data and user requirements. Prototype demonstrations also proved useful not only for eliciting feedback from future potential users but for obtaining support from various stakeholders, such as researchers, clinical directors, administrators, and funding committees (Cimino et al., 2014).
Principle 2: Evaluate all designs to ensure usability	User acceptance tests were conducted prior to the release of any new function. Also, responses to user surveys that were carried out indicated areas for improvement, such as the user interface and system response time for large, complex queries which took 10 min or more (Cimino et al., 2014).
Principle 3: Use iterative development	An iterative development process was used, enabling users to provide more precise feedback after preliminary versions of new features were available for use (Cimino et al., 2014).

Having briefly discoursed on user-centered design (more to follow in Chapter 6) and considered some useful examples appertaining to the design and development processes of RDM systems, the next section discusses advances in information retrieval that hold appreciable promise for research data retrieval.

2.5.2 Research Data Retrieval

Textual queries and ranking algorithms are the staple of traditional Information Retrieval systems and not well adapted for retrieving numeric or encoded data (Pallickara et al., 2010). It has been observed in the literature that there is greater variability in the search strategies employed by users when seeking data than literature, and that researchers spend more time evaluating the former than the latter (Kern & Mathiak, 2015). In the opinion of Kunze & Auer (2013), this may be because lists of data cannot be evaluated in the same way and with like efficiency as lists of documents; or it may be because current Information Retrieval models, according to Bremer & Gertz (2005) perhaps do not describe data retrieval practices completely. Information Retrieval systems produce document rankings based on the likelihood of relevance (Gregory et al., 2019), whereas Data Retrieval systems must provide exact matches to user queries (Gustafson & Ng, 2008). Models for retrieval of semi-structured data such as XML commonly enforce relevance ordering by employing query languages like XPath[12] or XQuery[13] which can be extended by a document retrieval operator to rearrange data fragments in order of their relevance to a term subquery (Fuhr & Grossjohann, 2001; Fuhr et al., 2002; Bremer & Gertz, 2005). A conceptually new XML-based approach (also implementable with relational databases) to integrated data and document retrieval, called integrated information retrieval (IIR) was introduced by Bremer & Gertz (2005). The approach of this model is to nest data and document retrieval subqueries into an XML query language, working by degrees on arbitrary, intermediate sequences of document fragments (DFs) in a way that allows for answering new kinds of queries. This approach however is only conceptual and was not demonstrated to give a tolerable performance in the scale that an active, real-life data repository might reasonably be expected to require.

[12] https://www.w3.org/TR/1999/REC-xpath-19991116/
[13] https://www.w3.org/XML/Query/

Another integrated approach to accommodating Information Retrieval models to the requirements of Data Retrieval was proposed by Gustafson & Ng (2008). Their approach works on relational database management systems, and uses a technique of measuring word similarity between queries and data records, and shifting the labor-intensive computational operations imposed on Information Retrieval onto the built-in efficient query processing mechanism of the relational database management system. Among the advantages of this approach are its comparative flexibility, being it works independent of the data to be evaluated; its compatibility with both small or large and information-rich databases, as well as on a wide variety of (unstructured) text data; and its higher precision than an AND-based query search. As Stempfhuber & Zapilko (2009) note, data needs tend to be specific and thus require high precision retrieval.

Kim et al., (2009) presented an interesting probabilistic model for semi-structured data; interesting because of its practicality and the relative ease with which, apparently, it can be plugged into the user interfaces of most of the existing RDM systems. The model is perhaps better described by an example. A supposed user wants to find a qualitative dataset about bird migration in Alaska. Searching with a simple query like “bird migration Alaska qualitative data” is hardly calculated to produce quite the relevant results. The user might fix this by specifying the appropriate fields using the “advanced” options, but it was observed that most users do not use such options. The probabilistic model, given this query, will try to infer the user’s query intent on a per-term, per-element basis to find which document element each query term may be associated with. Thus if an element (e.g., location) is given the highest mapping probability for a given query term (e.g., ‘Alaska’), then the occurrence of the query term in that element is assigned more weight than any other elements, by reason of the inference that the user may have meant the query term ‘Alaska’ as a location. In this way the model can exploit the term-element mapping without loss of information since every element can contribute a score. The mapping that results is then incorporated into the traditional language modeling approach to Information Retrieval proposed by Ponte & Croft (1998), in order to combine element-level scores into a document score. This produces a ranked list of documents. Experimental results for the model in realistic settings show “significant” improvements in retrieval effectiveness over baseline methods.

The next section expands upon the major issues and challenges of RDM at present which were briefly enumerated in the last chapter (see Section 1.1.6)

2.6 Current Issues and Challenges in RDM

As Wilkinson et al. (2016) states, the existing digital ecosystem surrounding scholarly data publication prevents the extraction of maximum benefit from research investments. In spirit, the aim of this work is to help solve at least part of the issues of RDM by developing and demonstrating a new design framework for RDM systems. It is proper therefore to know and as much as possible understand what those issues are. They may be of a technical, socio-cultural, or an ethical nature (Nelson, 2009; Hartter et al., 2013; Curdt & Hoffmeister, 2015); the key ones being:

i. *Insufficient Metadata.* This presents a major barrier to providing rich access and discovery capabilities for research data (Borgman, 2012; Kouper et al., 2013). When data are insufficiently described, their potential re-users are unable to understand to any useful extent the context or content of the data or how they were produced (Dumontier et al., 2016), making reuse 'difficult or impossible' (Koltay, 2015, p. 405). Often, the bulk of the responsibility of metadata tagging and documentation lies with data-holders or researchers themselves; and several studies have shown that that researchers, for reasons digressional to the discussion at present, devote little time for this activity (Carlson et al., 2011; Borgman, 2012; Wallis et al., 2013; MacMillan, 2014; Chowdhury et al., 2017; Chowdhury et al., 2018).
ii. *Researchers' Lack of RDM Skills.* Expertise in research and scholarship does not automatically imply expertise in data management. Studies among academic researchers have shown the existence of a considerable skill gap between what is expected of researchers in their role of data creators and what their current and generally minimal level of skill enables them to fulfil (Borgman, 2011; Cox & Pinfield, 2014; Davidson et al., 2014; Dierkes & Wuttke, 2016; Verbakel & Grootveld, 2016; Chowdhury et al., 2017; Chowdhury et al., 2018). Data discovery is largely dependent upon good metadata (Willis et al., 2012; Borgman, 2015); and data creators, although the primary providers of contextual metadata and other

complementary information about data (Borgman, 2011) are not necessarily skilled in data management or knowledgeable as to its technicalities.

iii. *Lack of Standards.* There are two sides to this problem: on the one hand is the lack of authoritative, well-established, and well-recognized standards, as a result of which there is a proliferation of informal and heterogeneous self-created standards, causing general confusion to both researchers and repository maintainers, besides precluding interoperability (Borgman, 2012; Wallis et al., 2013; MacMillan, 2014; Tenenbaum, 2015; Borgman, 2015; Bourne, 2015; Dumontier et al., 2016; McQuilton et al., 2016). On the other hand is the lack of a single vocabulary providing all key metadata fields required to support basic scientific use cases (Dumontier et al., 2016), due perhaps to the innate complexity and diversity of research data even within one domain, which makes it particularly tricky to develop one all-sufficient standard. Even within the same domain, no one standard is applicable across all individual cases; rather, the specific needs of the case dictate which standard to use, and sometimes only a combination of different parts from multiple standards will fit the case (Tenenbaum, 2015).

iv. *Inadequate infrastructural support for RDM.* Existing RDM infrastructures are unable fully to support researchers in communicating data in a meaningful way (Günther & Dehnhard, 2015). The current inadequacy of RDM systems is among other things chiefly attributable to the fact of their being as yet makeshift adaptations of text or string-based information systems, and not purpose-built solutions specially designed to cater to the particular peculiarities, subtleties, and unique requirements of research datasets (Bugaje & Chowdhury, 2017). This in itself is a source of many drawbacks. Moreover, most university support for research data preservation consists only in the provision of high-availability disk storage and backup solutions; and shared folders are in many cases the sole available instrument for collaboration between researchers (Weber, 2016).

v. *Considerable demands on researchers' time.* While researchers' time is limited, data management processes are time-consuming (Borgman, 2015; Wu et al., 2016) and the benefits or rewards thereof are in many cases not clear or forthcoming. Researchers hence tend to prefer to engage in other

scholarly activities, such as writing research papers, that produce more tangible results for them or are of a better recognized value in academia (e.g. paper citations, h-index, etc.). In fact, as Borgman (2015) observes, many, if not most researchers, view time and resources spent on managing research data as lost to research effort.

2.7 Chapter Summary

This chapter covered some important works, ideas, and concepts, the appreciation of which are invaluable to the proper commencement and subsequent progress of this research. The chapter opened with a high-level overview of the RDM ecosystem and a judicious discussion of its more relevant areas; such as, research data sharing and reuse, research data policies and regulations, the Research Data Lifecycle, etc. The basic components of RDM systems were then identified and discussed, and multiple examples of RDM systems were considered in more or less detail. This was followed by a discussion on the user-centered design approach, and afterwards on research data retrieval. Finally some of the key issues facing RDM at present were discoursed on. Many of the problems owe their existence to the fact of RDM still being in early stages, and the consequent time requirement for developing solutions. It is hoped and intended that this research should be a positive step in that direction. Research question (RQ) 5 was partially addressed in this chapter. Other theoretical underpinnings bearing upon the research work are reported within the text of the relevant chapters and sections appertaining them, as this arrangement seemed to contribute better to the understanding of the overall picture there presented.

3.0 METHODOLOGY

As I have made no specific a priori assumptions concerning the requirements of an RDM system or the behavior of its users, but rely wholly upon what my data gathering and analysis reveal to form any conclusions respecting the same, this research may be said to follow an inductive, rather than a deductive, process. It builds its beliefs and theories in a bottom-up fashion as the research progresses and patterns emerge from the gathered data. In order to choose the appropriate research methodology and data gathering methods it was essential to first re-examine critically the research questions and objectives (see Section 1.3) and note roughly the kind of data most likely to tend to their satisfaction, and also the questions best calculated to produce that data. For this exercise the works of Pickard (2013) and especially Frechtling (2002) were useful as containing detailed descriptions of the kinds of data produced by qualitative and quantitative methodologies, the manner of questions that each was best suited for answering, and also the strengths and weaknesses of each. It thus became evident, from the exercise, that both methodologies would be requisite: in fine, a mixed methods approach. The needful data for the research may be said principally to come under two heads; namely, data relating to RDM systems and data relating to RDM system users. The former, taking a more quantitative leaning and the latter, a more qualitative one, as may be deduced from the detailed statements of research objectives and questions in Section 1.3. A mixed methods approach therefore seems to answer best the purpose of this research. Pickard (2013) defines mixed methods research as “a combination of methodologies to address the same overarching research question”, while Bergman (2008) states more precisely that it is “the combination of at least one qualitative and one quantitative component in a single research project or program”. According to Hammond & Wellington (2013) there are “clear benefits” to be derived from this approach, in that the multiple sources of data it provides may prove useful for contrasting, complementing, or confirming research findings, or as part of a strategy of triangulation, which in fact this research employs. The term ‘triangulation’ takes up different meanings depending on the context in which it is used; however, it is most consistently associated with the use of “more than one method for gathering data, and an explicit concern for comparison of different sets of data” (Hammond & Wellington, 2013). Although an experiment was conducted as part of this triangulation, the research is, in the overall sense, non-experimental, since it has

formulated no hypotheses which it proposes to test. Philosophically, it seems to me to sit more comfortably within the postpositivist paradigm; for, whereas positivism seeks to measure phenomena via quantitative approaches, and interpretivism to find meaning via qualitative ones, postpositivism shows a dualism that inclusively accepts the contributions of both (Hammond & Wellington, 2013; Pickard, 2013).

It has been previously observed that the data required for the present research may be seen as broadly relating either to RDM systems on the one hand, or RDM system users on the other. As to the latter, the selection of data collection techniques is to be made from among the following choices, as given in most standard research methods texts (e.g. Rugg & Petre (2007), Bergman, M. (2008), Hammond & Wellington (2013), and Pickard (2013)):

a. Surveys (commonly questionnaires);
b. Interviews;
c. Observation;
d. Diaries; and
e. Focus groups.

Only the first two, i.e. questionnaire surveys and interviews were used, for reasons that will be explained presently. Triangulating questionnaire surveys with interviews is in fact rather a common practice in mixed methods research (Pickard, 2013), as the two complement each other admirably well, the questionnaires “going wide” and the interviews “going deep”. This ability of the questionnaire to “go wide” by covering a large number of participants with comparative ease, is among its chief advantages. Interviews, on the other hand, are excellent when “qualitative, descriptive, in-depth data” is sought, “that is specific to the individual, or when the nature of data is too complicated to be asked and answered easily” (Pickard, 2013). This almost exactly describes the kind of data about RDM system users which this research seeks. Questionnaires can be conducted manually (e.g. on paper) or electronically (e.g. on the internet), and questions may be closed (e.g. multiple-choice questions) or open-ended (e.g. free-form text questions to be answered in participants’ own words) (Velsen, 2011). They afford a useful means of obtaining quick, bite-size information or

statistics, and were used in this research to identify interesting themes for more productively guiding the direction of the face-to-face interviews which followed. As Hammond & Wellington (2013) remarked, the point of a survey is "to find out *how many* feel, think, or behave in a particular way, and to provide a general picture relatively quickly." The follow-up interviews, in their turn, by adding richness, depth, and dimensionality to the questionnaire data, helped to counteract the quite significant deficiencies inherent in the latter (e.g. see those noted by Labaw (1981), Carter (2007), Hammond & Wellington (2013), and Pickard (2013)). This characteristically reflects the peculiar value of the interview as generally appreciated in the literature, which, in the words of Hammond & Wellington (2013) "allows the researcher to probe into an interviewees account of an event as well as their thoughts, values, feelings, and perspectives… they are interactive, allowing for clarification of questions and identification of unexpected themes".

The techniques of observation, diaries, and focus groups were not resorted to because questionnaires and interviews proved adequate for soliciting and obtaining the needful data for the research. The silent observation of researchers' behavior, for example, seemed, for the purpose of this particular research, not so pertinent as researchers' verbal descriptions and explanations of the whys and wherefores of such behavior on their part. Diaries, also, which are used to log records about specific occurrences over a period of time, were scarcely seen to be at all relevant for the case in hand, where neither time nor individually recurring patterns present factors of any great importance. Focus groups were harder to decide upon. They are held by some to be a type of interview and by others to be a technique in their own right. Irrespective, they are generally described as group discussions consisting of about 3-12 participants and marked by interactive dialogue, questions, answers, and other activities (Velsen, 2011; Pickard 2013; Tracy, 2013). The test for the appropriateness of a focus group for any given study is, according to Tracy (2013), the question whether the topic "could benefit from the group effect". For the present research, no positive grounds could be adduced for an affirmative answer to this question. The advantages of the technique in this case did not sufficiently overcome its drawbacks (see, for example, those mentioned by Kitzinger (1994) and Velsen (2011), for example), and individual interviews seem, in any case, to promise equally well and without those drawbacks. For, an individual setting ensures to each participant the full attention

of the interviewer, and with opportunity to pursue the thread of the dialogue to a satisfactory point, none of which are possible in the competitive setting of a group.

Like the data gathered about RDM system users, that about RDM systems (independent of its users) was also obtained via a dual technique; namely, reviews and an experiment. Indeed the reviews, which were carried out systematically of a number of RDM systems, were inevitable in view of the object of the research itself; since it is necessary to gain a good appreciation of the status quo before setting about devising means for its further improvement. The decision to conduct the controlled experiment was a more deliberate one, and was made in consideration of the obvious additional advantage of obtaining a more operative appreciation of the use RDM systems, aside from the comparatively more superficial appreciation of its features and attributes which were covered in the reviews. Figure 3.1 below illustrates, in summary, the various studies undertaken for data gathering in this research; and Table 4.1 connects each of these to the original research objectives and questions.

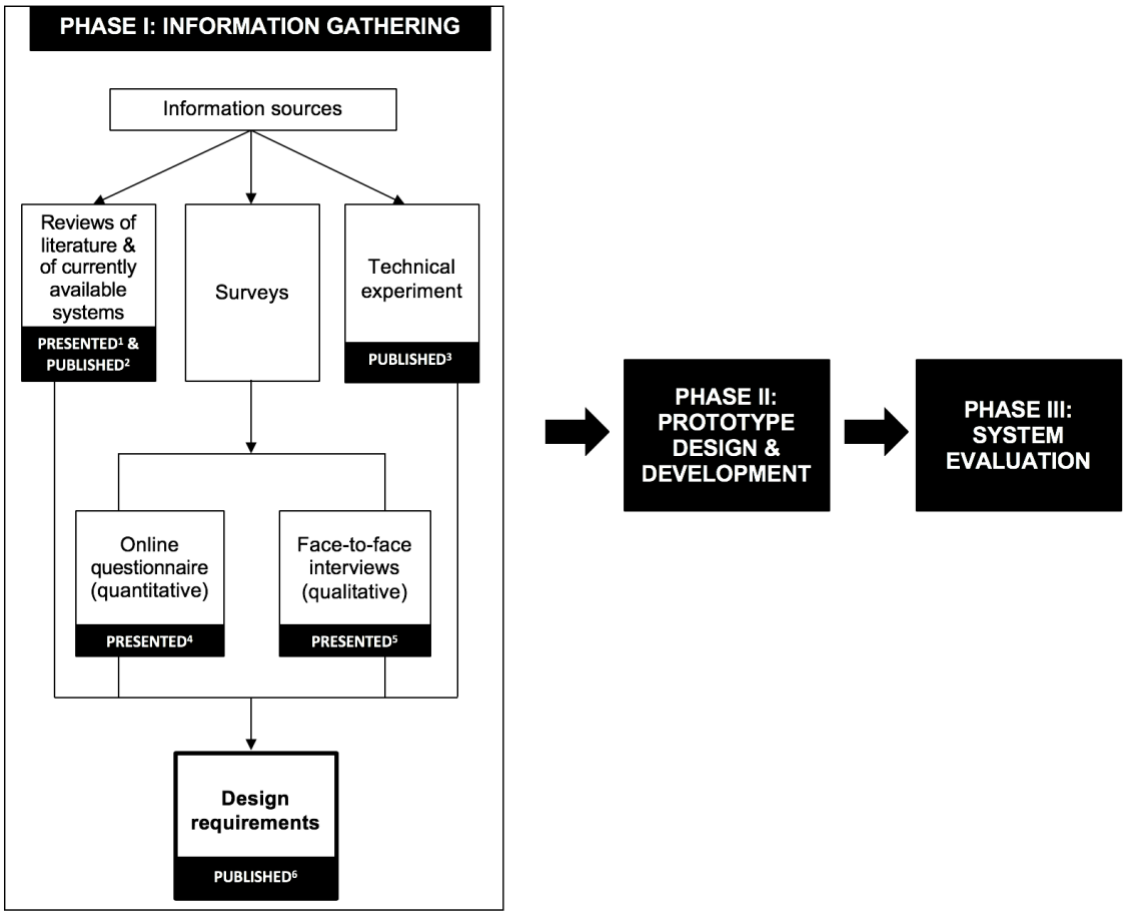

Figure 3.1. Overall outline of the research focusing on methodology (Phase I).

[1] Bugaje & Chowdhury (2017a); [2] Bugaje & Chowdhury (2018a); [3] Bugaje & Chowdhury (2017b); [4] Chowdhury, Walton & Bugaje., (2017); [5] Bugaje & Chowdhury (2018b); [6] Bugaje & Chowdhury (2018c);

Table 3.1. The various research methods employed and their connections to the wider research context.

	Research Method	**Research objective(s) addressed**	**Research question(s) addressed**	**Findings presented in**
1	Market appraisal & review of currently available RDM systems	2b, 2c & 3	5	Section 4.1
2	Online questionnaire survey	1b	1 partially	Section 4.2
3	Face-to-face interviews	1a, 1b & 1c	1, 2, 4 & 6	Section 4.3
4	Technical experiment (comparison between DR and traditional IR)	2a, 2d & 3	3 & 5	Section 4.4

These studies, excepting the technical experiment, were conducted sequentially and in the above order. The sequence emerged naturally; as, the market appraisal and review influenced the design of the questionnaire, and the face-to-face interviews followed the analysis of the questionnaire data, whence it was deemed highly expedient to probe further and gain a deeper insight into some of the findings and hints which, due to the limitations imposed by the comparative rigidity of an online questionnaire survey, were not sufficiently apprehended to form solid groundwork for the design phase. In the remaining sections of this chapter I describe each study as it was carried out, making notes of any limitations in the particular case in hand.

3.1 Market appraisal & review of currently available RDM systems

A systematic review was conducted of RDM systems currently in use to get a sense of the features, capabilities, and services that they offer. This was principally with a view to identifying areas in them that require further improvement, so as to enable them to better fulfil their functions in the RDM ecosystem. The number of RDM systems available as at time of writing is,

according to re3data.org, upwards of 2000. re3data is a global registry of research data repositories, containing perhaps the most comprehensive list of research data repositories available anywhere. Furthermore, its authority is widely recognized by journals and publishers such as Nature, Springer, and Plos; and even by the European commission[1]. Given such a large number of repositories the question which and how many to review requires careful thought. It became necessary therefore to devise a strategy by which as fairly a representative sample as possible might be chosen. Perusing the list with this object in mind, possible groupings began broadly to suggest themselves from the metadata tags (e.g. subject, content type, etc.). Other works (e.g. Kindling et al. (2017) and OpenDOAR[2]) were also consulted in which some of the identifying characteristics of research data repositories were highlighted. This was to help break up the list into smaller and more manageable groups from which samples can be selected that approximate the important characteristics of their respective populations. Six groups emerged at last, as follows:

a. **Disciplinary repositories**, which hold data from particular disciplinary areas or domains;
b. **Institutional repositories**, which serve the staff and student communities of their respective institutions;
c. **Publisher-service repositories**, which are provided by journal publishers for the use of their respective contributors;
d. **Location-based repositories,** which hold research data produced within a certain geographical location or region;
e. **Dedicated content-type**, which hold research data of a certain type or file format; and
f. **Commercial or general-purpose repositories**, housing a wide range of research data with little or no restrictions of any description.

This grouping does not claim formal recognition nor is it considered as being conclusive, but is created simply to facilitate the study. The peculiarity of the study, in terms of its purpose and its application, is also acknowledged. Owing to this circumstance, a standard method of proceeding which exactly appertains to it was

1 https://www.re3data.org/about
2 http://v2.sherpa.ac.uk/opendoar/

not forthcoming. General guidelines for Systematic Literature Review were therefore adopted instead, and suitably adapted; chiefly, those given by Kiteley & Stogdon (2014), due to the detailed descriptions they afforded of the review process. The key steps therein identified were four, viz. search, data extraction, application of appraisal criteria, and information synthesis. The first step, i.e. search, could be said to map onto the perusal of the re3data directory. For the second step, Nature's[3] recommended repositories, besides re3data's statistics, was consulted, and one representative example was carefully handpicked for each of the repository groups formed above. The total number (i.e. six) was deemed adequate for the purpose of the study, which, to articulate them more specifically, are:

To–

1. Distinguish between the different categories of RDM systems and their target users or audience;
2. Identify the relative strengths and weaknesses of each category; and
3. Identify the various design features, as well as other features, of currently available RDM systems, and the degree to which these are common or otherwise.

The third step of the systematic review process, as given in the preceding passage, involves drawing up criteria against which each item for review will be evaluated. With reference to the objectives of the study, given above, five criteria were resolved upon after careful deliberation. Although the exact process by which these were arrived at cannot be easily stated, it was not entirely arbitrary, and a preliminary trying out of a great number of RDM systems as well as other reading all served as input to influence the choice; as follows:

a. **Use of metadata**. The degree to which metadata appears to be exploited to provide features for browsing, searching/querying, filtering and search result presentation;
b. **Querying facility**. The level of expressiveness allowed in searching/querying the repository;

[3] https://www.nature.com/sdata/policies/repositories

c. **Result filtering**. Availability of options for filtering down search results, and the granularity to which this is possible;
d. **Sorting facility**. Availability of options for ordering the arrangement of search results; and
e. **Availability of additional features for data**. This refers to any extra features that improve the overall usability of the repository or that help to comply to a greater degree to the guidelines and principles given in Section 1.1.6, viz. Discoverability, Accessibility, Intelligibility, Assessability, Reusability, Interoperability.

The last step of the review process, i.e. information synthesis, representing the actual result of the study, is presented in the next chapter. The method used was the Qualitative Data Synthesis (QDS) one as described by Kiteley & Stogdon (2014). It was chosen from amongst four others as the best fit for the present scenario. A general limitation of the study is the one inherent in systematic reviews: that of being better at identifying 'what' works than 'why' it works or doesn't work (Kiteley & Stogdon, 2014).

3.2 Online questionnaire survey

Data was collected via online questionnaire surveys conducted at universities UK-wide between the summer and winter terms of the 2016/2017 session. Full ethical clearance was obtained from the University Ethics Committee before the study was conducted. The survey garnered a total of 201 (191 fully complete and 10 nearly complete) responses from researchers from a wide range of academic experience and disciplinary domains; including, Arts & Humanities, Social Sciences, Applied Sciences, Health Sciences, and Natural Sciences among others. A request was sent through the Jisc[4] mailing lists requesting for participation in the survey, with a link to the web-based survey. Both closed and open-ended questions were used, as the latter can capture information that the former cannot (Miles & Huberman, 1994), and all were worded as much as possible in simple, uncomplicated language. The chief aim of the survey is to obtain information as to the following:

[4] https://www.jisc.ac.uk

1. The type, volume, and variety of data used and created by researchers;
2. Researchers' common practices with respect to data storage;
3. Researchers' familiarity with standards, metadata, and their university data policy;
4. Requirements and opportunities for training & support of researchers in RDM;
5. Views, perceptions, and practices pertaining to data sharing and open access; and
6. Researchers' previous experiences of, and impressions about, using research data repositories.

Using the JISC mailing list ensured a UK-wide coverage and helped to mitigate selection biases that might be introduced by confining participation to any particular type or locality of university. Furthermore, and still tending towards the same end, solicitation of participants was extended to all categories of researchers, including Ph.D. students. Despite the steps taken to promote better accuracy of representation, however, it is worth bearing in mind the possibility that responses may be skewed towards those researchers who feel more strongly about RDM. As regards response biases which the design and wording of the questionnaire itself may introduce, care was taken to avoid or mitigate them as far as is possible through providing neutral or otherwise non-committal response options for every question and through steering clear of any ambiguous, leading, double-barreled or loaded questions. The questionnaire can be found in Appendix II.

3.3 Face-to-face interviews

Interviews are useful for obtaining qualitative, descriptive, in-depth data (Pickard, 2013) on the needs and requirements of researchers (Carlson, 2012; Simons & Richardson, 2013). Certain hints and findings that needed further exposition to be fully useful or of value for the greater purpose of this research, emerged from the questionnaire survey previously conducted. These included, but were by no means limited to:

a. Disciplinary patterns in certain tendencies of behavior or attitude of researchers with respect to data sharing and reuse (e.g. Solar Physicists

markedly showed more willingness to do the same, compared to Arts & Humanities researchers), which may have important design implications;

b. Inexplicable inconsistencies, or contradictions between researchers' stated inclinations (e.g. willingness to share data on online repositories) and actual actions (e.g. not sharing data), which may point to design flaws in RDM systems or otherwise indicate opportunities for improving the same;
c. Apparent differences in researchers' conceptions (or misconceptions) of certain key terminologies (e.g. the term "research data"), which may potentially lead to inaccurate responses;
d. Loaded hints from researchers' comments and remarks which were of sufficient importance to warrant closer examination (e.g. what prompts comments such as, "not easy to share data with others either inside the university or outside using our current systems"); etc.

These concerns, among others, occasioned the need for further investigation via face-to-face interviews. Full ethical clearance was obtained from the University Ethics Committee before the interviews were begun. 18 researchers were interviewed: 6 from each of the departments of History, Solar Physics, and Information Science at a British University. The first two disciplines were chosen on the strength of their being fairly representative examples of two polar ends of the disciplinary spectrum (see Table 4.2), while the third provided a middle ground, as regards data sharing practices, use of technology, and the nature or characteristics of the respective research data produced (Borgman, 2015). Table 4.2 below highlights and compares these points across the three disciplines; and it was expected that the broad disciplinary range covered will provide opportunity for learning the unifying similarities of, as well as the contrasting differences between, the disciplines. The 6 researchers interviewed from each discipline comprised: 2 academic staff with varying research experience, 2 postdoctoral researchers and 2 doctoral students in the later stages of their respective researches. Accordingly, in terms of stage of academic career, a total of 6 each of academic staff, postdoctoral researchers, and doctoral students were interviewed, as shown in Table 4.3. The reason for this selection was as much to ensure maximum inclusivity of research experience as to discern the existence of possible peculiarities or differences between the various groups. Participation in

the study was entirely voluntary, and participants were at full liberty to withdraw their consent at any point. An email request, enclosed with a briefing document stating what the interview would entail, and also a declaration of ethical approval from the University, was sent out to a number of eligible participants; and suitable times were arranged for personal meetings with those who agreed to do the interview.

Table 3.2. Disciplinary characteristics which motivated the choices of disciplinary representation for the interviews.

	Solar Physics	**Information Science**	**History**
Nature of data	Mostly quantitative	Both quantitative and qualitative are common	Mostly qualitative
Data creation	Created by natural phenomena	May be created by human or by machine	Usually consists of repurposed ancient artefacts; digital copies may be made
Data collection	Collected by machine; highly automated	May be collected by machine or by human	Usually collected by human
Data format	Digital	May be digital or print	Artefacts
Creative control	Largely regulated by the data collection instruments, policies, and standards	Researcher has creative control over parameters	Limited
Prevalence of standards	Well established standards that are used and shared by a global community of researchers	Limited standards. Usually individual or project-based.	Limited
Size of team	Large	Varies	Small

Table 3.3. Summary of interview participants.

	Level of academic experience	Researchers' disciplinary domains		
		History	Information Science	Solar Physics
1	Academic staff	2 persons	2 persons	2 persons
2	Postdoctoral researchers	2 persons	2 persons	2 persons
3	Doctoral students	2 persons	2 persons	2 persons

Each interview lasted approximately 30 minutes and followed a semi-structured format; beginning on the part of the interviewer with a brief overview about the present research and the objectives of the interview; followed by brief questions to understand the research area/project of the interviewee. The rest of the questions were then slightly modified to suit the research context of the interviewee, but were primarily aimed at, though not limited to, obtaining the following information:

1. Where and how do you obtain data for your research? Do you employ any strategy or have a standard workflow for this?
2. What are some of the problems you've faced before in finding, using, or accessing research data, if any?
3. What data repositories have you used before or do you currently use? What motivates you to use a particular repository rather than another?
4. Have you ever uploaded your own data in an online repository? Why or why not?
5. What are your thoughts on research data sharing and open access?
6. Do you or your research group follow any metadata formats for tagging research data? What are some of the issues you've faced in this regard, if any?

I took hand written notes on each session, and these were transcribed and further elaborated upon immediately within the first few hours after the session. In a few cases the same participant would be interviewed in two different sessions, or otherwise be contacted via email, to further clarify certain points noted from the initial conversation or to confirm that they were no errors in interpreting their words. No voice recordings of the interviews were made, as, according to Pickard (2013), "recording may have a negative impact on the interview; they

[interviewees] may feel inhibited by the fact that their words will be recorded; it makes them conscious of what they are saying and how they say it". This precaution was later justified by the fact that many of the participants, especially from History, showed some degree of hesitancy in venturing positive opinions as to what research data constitutes or data repositories meant.

3.4 Technical experiment (comparison between DR and traditional IR)

This section describes the controlled experiment referred to in the opening passages of this chapter. It was carried out with the aim of demonstrating some fundamental differences between text retrieval and data retrieval, as far as concerns their respective modes of user interaction and resource retrieval. It was deemed as appropriate to approach the study from a disciplinary perspective, as this would enable the possibility of making comparisons with the findings of at least the questionnaire survey and the interviews, to both of which the disciplinary element was also present. Accordingly, after exploring some of the broad disciplinary classifications given by various authorities, the one by Wikipedia[5] was adopted for its simplicity and relevance for the purpose of the study. It organizes the general body of academic disciplines into five broad domains, viz. Arts, Humanities, Social sciences, Natural sciences, and Applied sciences. I made two slight alterations to the original arrangement, by merging Arts with Humanities and choosing Computer & Information Science to represent its parent discipline of Applied Sciences. This was done, in the former case, to make the data more manageable, as it is quite usual to find Arts and Humanities combined together; and in the latter case, that my subject knowledge of Computer & Information Science might be put to better advantage. And, as each of the domains in the original Wikipedia classification are still well-represented the alterations are not very likely to affect the results of the experiment. The next step was to select five keywords or phrases (omitting stop-words) which seem most calculated to represent the respective disciplines. The selections were made from the Wikipedia homepage of each. Upon these carefully chosen terms a search was then conducted for data retrieval as well as for text retrieval. For the latter scenario, the choice of information system fell to Thomson Reuters Web of Science[6] database, being it is considered the most comprehensive database for

[5] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines
[6] https://www.webofknowledge.com/

research publications (Kuncheva, 2014, pp 107); whereas, for the former, it was necessary to make use of more than one system, due to the difficulty of finding any one repository whose data well represents all the required disciplines. Three systems were thus used in combination, viz. UK Data Service[7], DataOne[8], and Dryad[9]. All the three are cited by re3data as among the more well-known of research data repositories, and furthermore are recommended by Nature[10]. The UK Data Service supplied the data for Arts & Humanities and Social Sciences; and DataOne did for Natural Sciences data; while, in the absence of a special Computer & Information Sciences data repository, Dryad, which is generalist, was used.

For both the data retrieval and text retrieval halves of the experiment, only the first 10 items of search results were considered, except in instances when an item so obviously departs from the intended topic, in which case the item is skipped and the next item is considered in its stead. As I have tried to mimic a typical search scenario of a researcher in a real-world situation, the choice of only 10 items was informed by research on user search behavior which shows that well over half of search engine users do not go past the first page of search results (Spink et al., 2001; Jansen & Spink, 2006; Richardson et al., 2007). Also, 10 just happens to be the default minimum number of results on a single page that is common to most search engines (Maley & Baum, 2010; Wu & Marian, 2011), including, in the present case, Thomson Reuters Web of Science and UK Data Service. Each portion of the experiment (i.e. the text retrieval and the data retrieval) yielded therefore 200 observations, calculated thus:

$$4 \text{ disciplinary domains} \times 5 \text{ search terms} \times 10 \text{ items from the search results} = 200 \text{ observations.}$$

The total number of observations drawn from the experiment was thus 400, since the experiment was separately conducted in two different contexts, i.e. text retrieval and data retrieval, each of which yielded 200 observations. This number, as it was drawn from quite a wide range of the disciplinary spectrum, was deemed

7 https://www.ukdataservice.ac.uk/
8 https://www.dataone.org/
9 datadryad.org/
10 https://www.nature.com/sdata/policies/repositories

sufficiently representative to give at least a general indication of the prevailing features and concerns that characterize data and text retrieval. The main information noted down in the experiment was file size, file format, and search hits, although other observations also resulted, which are noted in Section 4.4 in the next chapter. File size, for publications (text retrieval) is represented by the file size of the full research paper, and, for data (data retrieval), by the dataset itself as well as all of its documentation files, if any. The full experimental data will be found in Appendix I. The somewhat idiosyncratic and even arbitrary character of this study is fully acknowledged, and perhaps constitutes its chief limitation. An endeavor was however made as much as possible to base every decision upon sound information from the literature.

3.5 Chapter Summary

This chapter details the research methods involved in Phase I of this research which the research framework (reproduced below) presented in Chapter 1 illustrates. The findings of each is presented in the next chapter. As to the methods associated with the prototype design and evaluation, those are discussed in the respective chapters (Chapters 6 and 7), as they are closely interwoven with the chapters' subject matter.

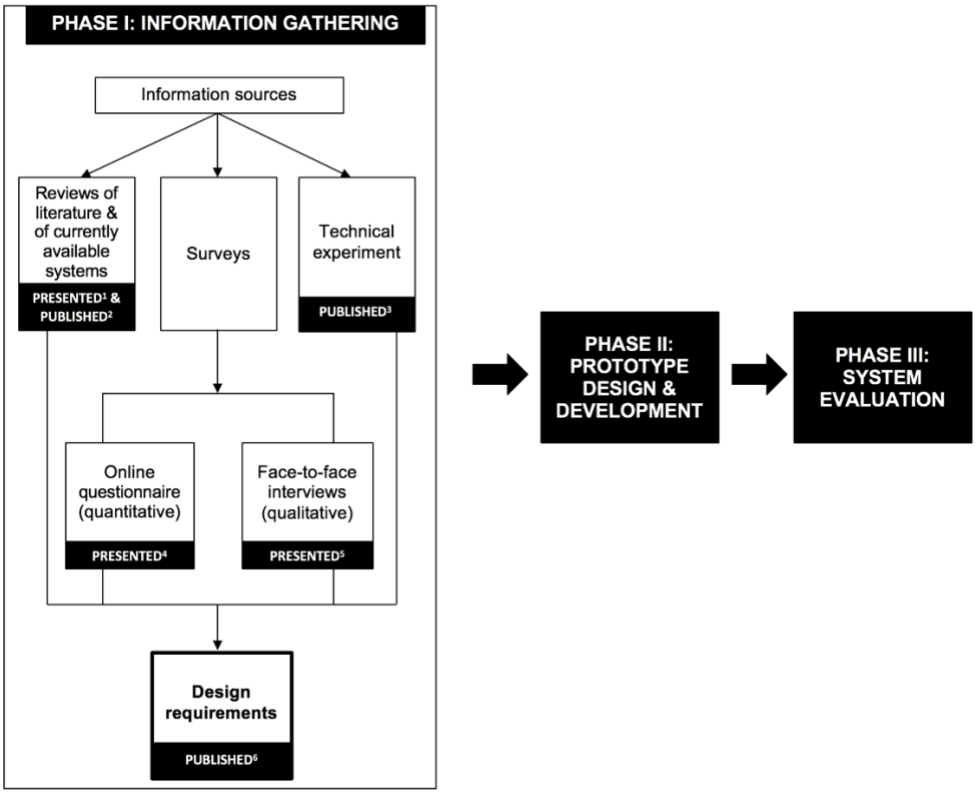

Figure 3.2. Overall outline of the research focusing on methodology (Phase I).

4.0 DATA ANALYSES

This chapter documents the analyses of the data collected in the four studies described in the last chapter, and reports on the raw findings from each. The raw findings are further examined in the next chapter and translated into appropriate design requirements for use in Chapter 6 (System Design). The sections that follow detail the analysis of the data from each of the studies conducted, taken by turns, and followed by discussions of their respective findings.

4.1 Market appraisal & review of currently available RDM systems

This section presents the results of the systematic critique of currently available data repositories. The study provides a useful insight into the usability implications, as well as trade-offs, of various design features for each of the categories of data repositories described in the last chapter (see Section 3.1). They are each discussed by turns in the succeeding sections. The discussions are then summarized at the end in Table 4.1, based on the evaluation criteria proposed in Section 3.1.

4.1.1 Disciplinary repositories

These are dedicated repositories housing research data from a specific disciplinary branch or sub-branch, such as, Dryad[1] for the Biosciences, and the Virtual Solar Observatory (VSO)[2] for Solar Physics data. VSO is a typical example of a disciplinary repository and has therefore been chosen to represent the group. A screenshot of its search interface is shown in Figure 4.1, and the use of rich metadata to enable searching by a number of parameters and variables is evident from the search fields and options supported. This is a typical feature of disciplinary repositories, whose discipline-bounded scope affords opportunity for exploiting metadata that is specific to that discipline, in order to improve, among others, query expressiveness, indexing techniques, retrieval efficiency and search result sorting and filtering to a fine granularity. In our present example, the VSO holds solar data, which is a highly standardized, machine-collected (e.g. with space telescopes) data that is extensively machine-tagged with standard disciplinary metadata (Borgman, 2015). Choosing any one or a combination of

1 datadryad.org/
2 https://sdac.virtualsolar.org/cgi/

the search variables in Figure 4.1 (for example, “spectral range”) leads to another page (Figure 5) where further options respecting that variable may be specified.

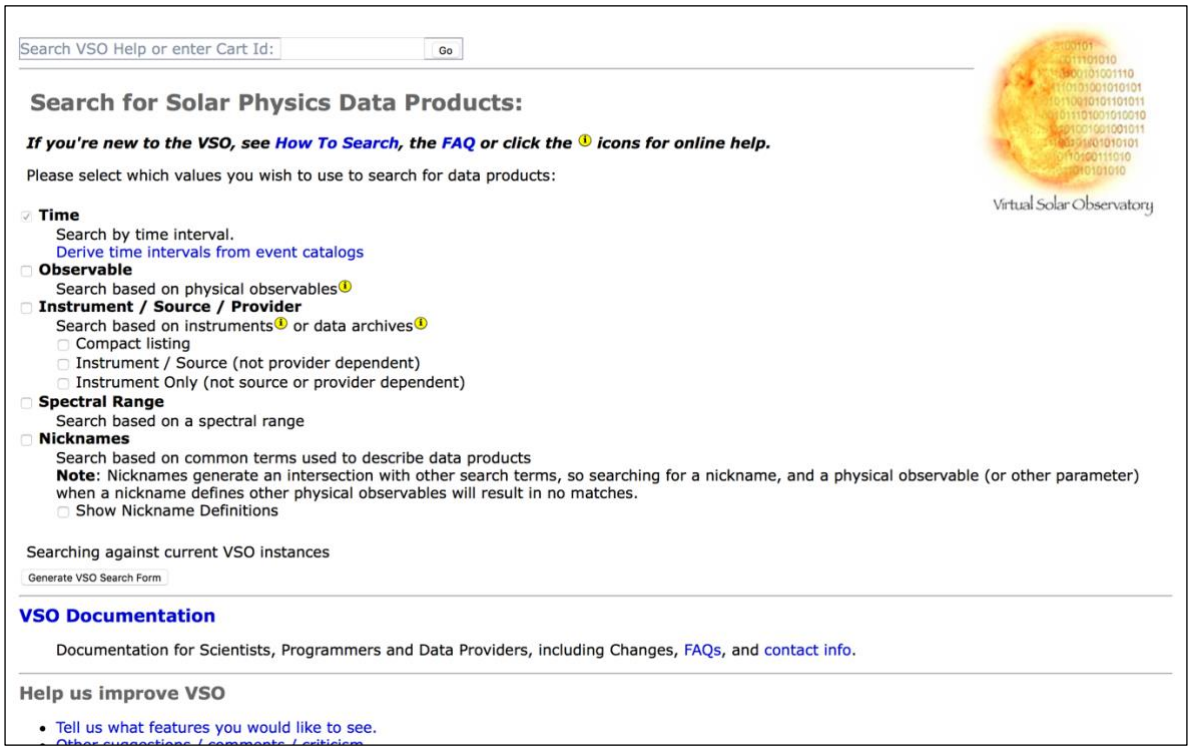

Figure 4.1. The homepage and initial search interface of the Virtual Solar Observatory (VSO)

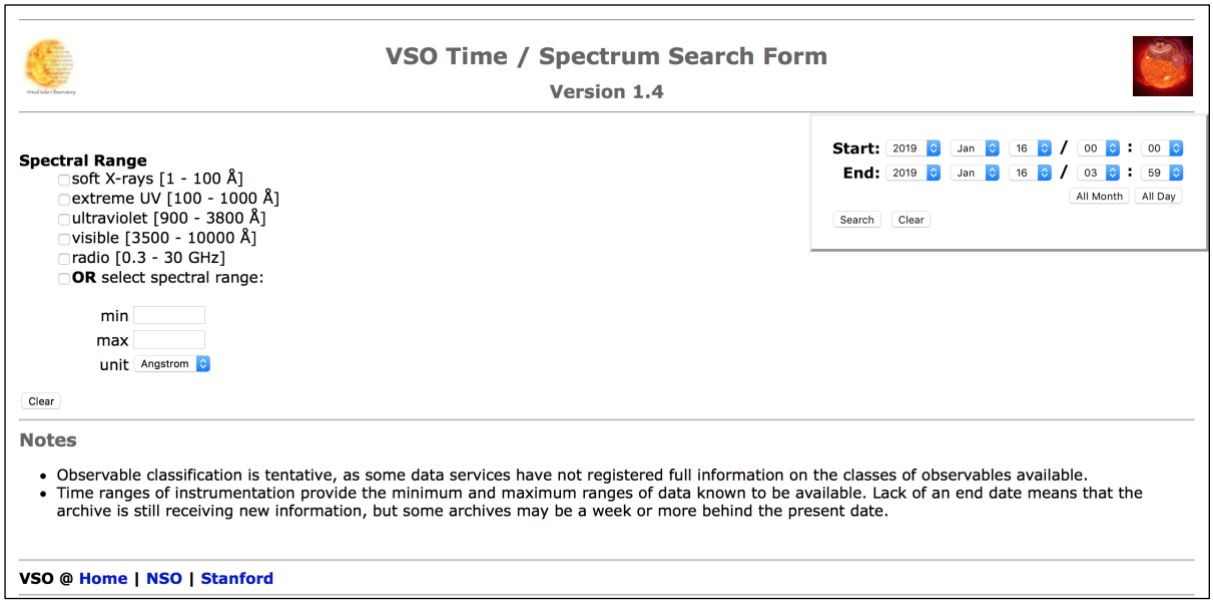

Figure 4.2. Shows how the rich metadata of disciplinary repositories domains allows for minutely specified and fine-tuned search queries.

A possible drawback of this elaborateness is that it may confuse or otherwise overwhelm the user, especially if the user is not familiar with the disciplinary terminology. This may prove a point of concern, since open data, as has been observed in Chapter 1, aims to render research datasets generally accessible and reusable, including to the general public. However, studies have shown that, once acquainted with disciplinary repositories, researchers show an inclination to use them rather than other kinds of repositories (Hayslett, 2015).

4.1.2 Institutional repositories

Institutions of higher learning commonly provide repositories for the exclusive use of their research communities; e.g. Northumbria University's NRL[3] (Northumbria Research Link) and Oxford University's Research Data Oxford[4]. Many universities outsource the provision of this service to third-party vendors (see Section 4.1.6, on commercial & general-purpose repositories). Institutional repositories are rarely designed or meant to hold research datasets alone, but usually function as storehouses for a myriad of research outputs produced by the university, including books, patents, reports, conference presentations, research publications, and doctoral theses among others.

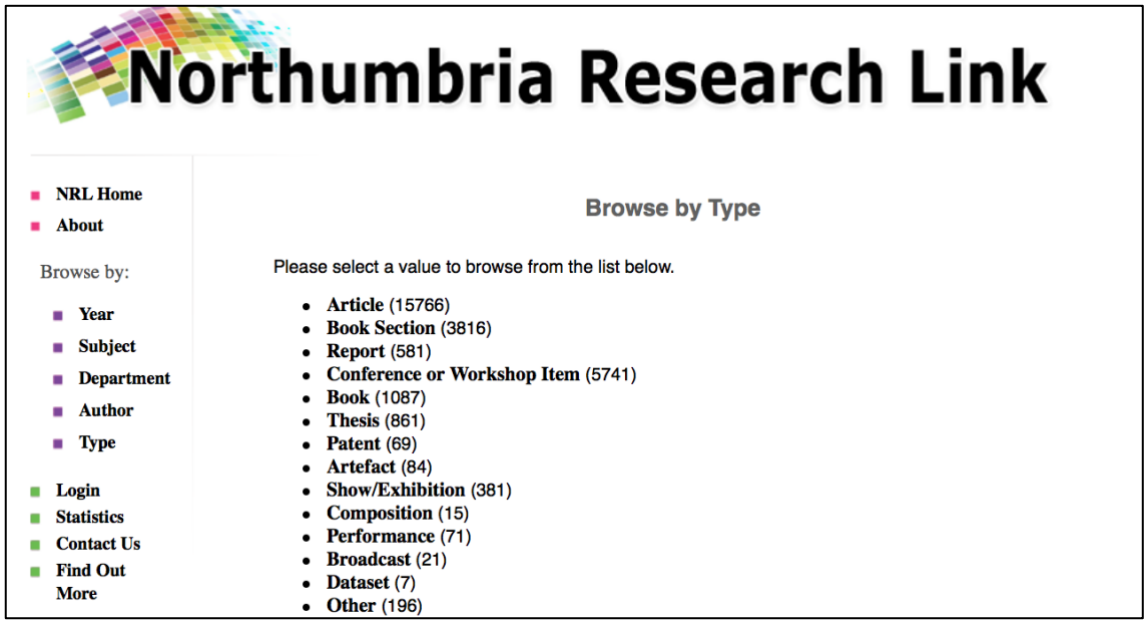

Figure 4.3. Examples of research outputs all held in institutional repositories

[3] http://nrl.northumbria.ac.uk
[4] http://researchdata.ox.ac.uk

As Lynch (2003) observes, institutional repositories are offered to the members of the institution or its affiliated communities for the management and dissemination of digital materials created by them. Figure 4.3 presents a screenshot of NRL, showing the various resource objects typically held in institutional repositories. To accommodate this resource diversity on the high-level, institutional repositories generally provide only very basic and simplified features for searching, sorting, and filtering, as may be seen on the left pane of the same figure. The drawbacks introduced by so doing, however, may not be of a very consequential nature since the number of datasets held in a single institutional repository is usually not very enormous: for example, Figure 4.3 shows the case of NRL holding only 7 research datasets. On another note, the advanced search features of institutional repositories, where provided, are commonly to be found to contain options that are either more specifically relevant to text-based objects or only superficially so to data. An example in the case of NRL is shown in Figure 4.5. The left side of the previous figure, Figure 4.4, shows how the search results of NRL are presented; and it may be noted that even the one additional feature for exporting the results provides options (see right) that are applicable not to the individual result items themselves but to the entire result set that was returned from the search query. In many institutions, the libraries or IT centres, most of whom can play an important role in building up RDM services and solutions, hesitate because policies relating to the treatment of research outputs, such as datasets, are not yet clearly delineated (Weber & Piesche, 2016).

Figure 4.4. An institutional repository showing very basic options for finding research data.

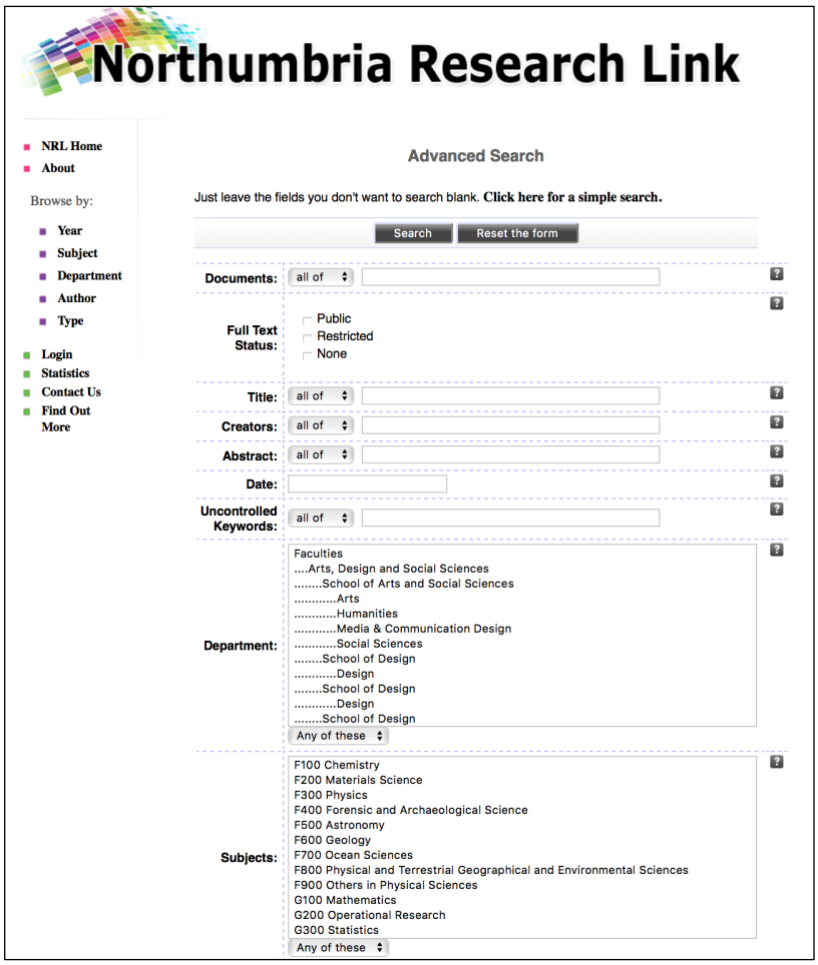

Figure 4.5. Advanced search in an institutional repository

4.1.3 Publisher-service repositories

These are repositories provided by journal publishers, some of whom conduct peer reviews on research data and publish them as standard scholarly outputs, commonly called “data papers”. Nature’s Scientific Data[5], shown in Figure 4.6, is a representative example. Publisher-service repositories are mostly optimized for linking research data with the publications that they underlie; and, as journals generally publish around specific subjects/topics, their repositories may share some of the advantages of disciplinary repositories. However, this type of service is, at present, not widely offered, though growing. A drawback may be introduced by the fact that publishers who publish research datasets also publish journals, and their repositories, which usually hold the two kinds of resources, tend to try to accommodate their differences more or less by approximating down the requirements of research data to those of publications.

[5] https://www.nature.com/sdata/

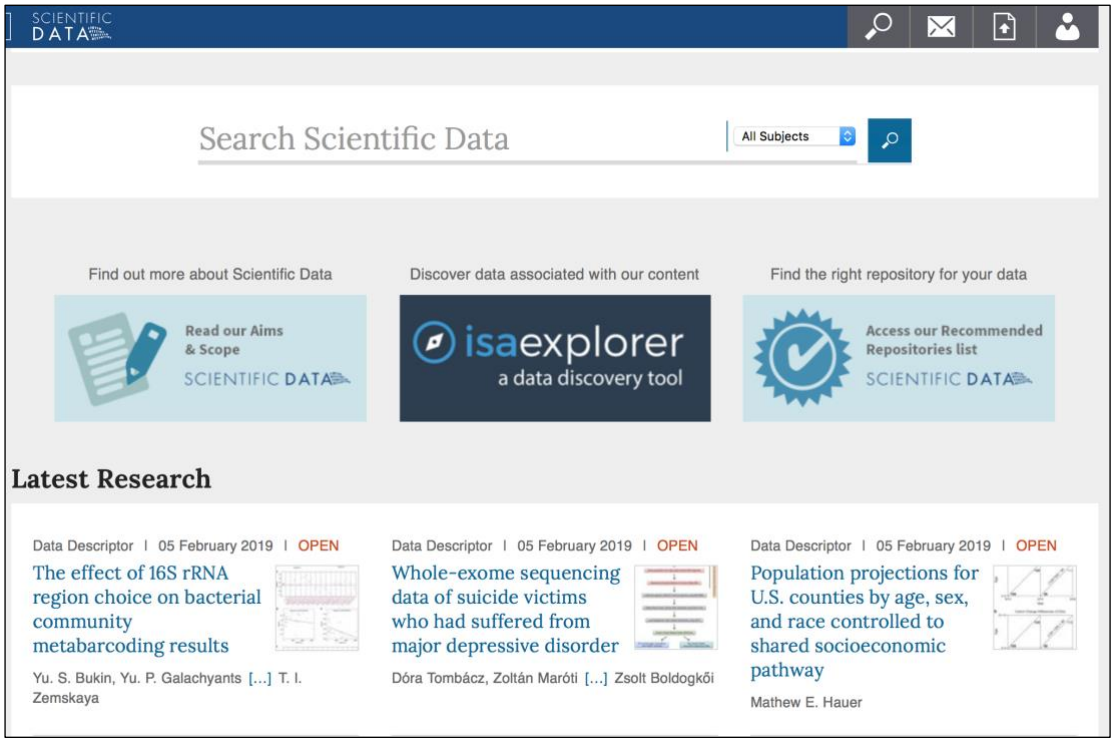

Figure 4.6. Homepage and initial search interface of a publisher-service repository

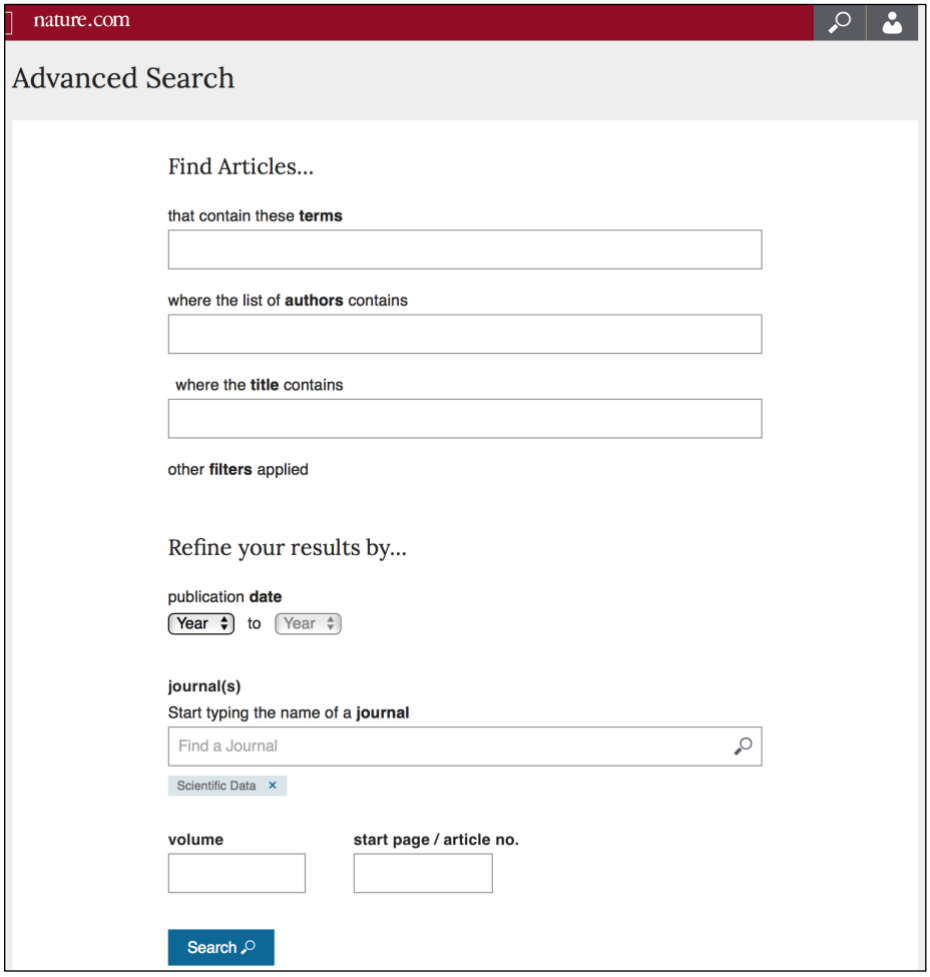

Figure 4.7. Advanced search in a publisher-service repository

As an example, Nature's Scientific Data provides no special features for research data discovery. The search field options in its advanced search page (see Figure 4.7) appears to be more relevant for finding research publications than data (for instance, the first form field shown in the figure accepts only string keywords, consequently precluding use for numerical datasets). Indeed, from the heading of the page itself (i.e. "Find Articles"), it might be concluded expressly have been meant for research publications alone. In fine, therefore, basic keyword search is the only available option for use in searching for data in the repository. The repository also provides a browsing feature which classifies all the resources held in the repository under appropriate subject headings. This may be very useful, as it allows searching where keywords are misleading or unknown. The criteria for sorting (by date or by relevance) and filtering (by article type, journal or date) of the results returned may be perhaps much too generic for narrowing down search results or locating a specific dataset (see Figure 4.8).

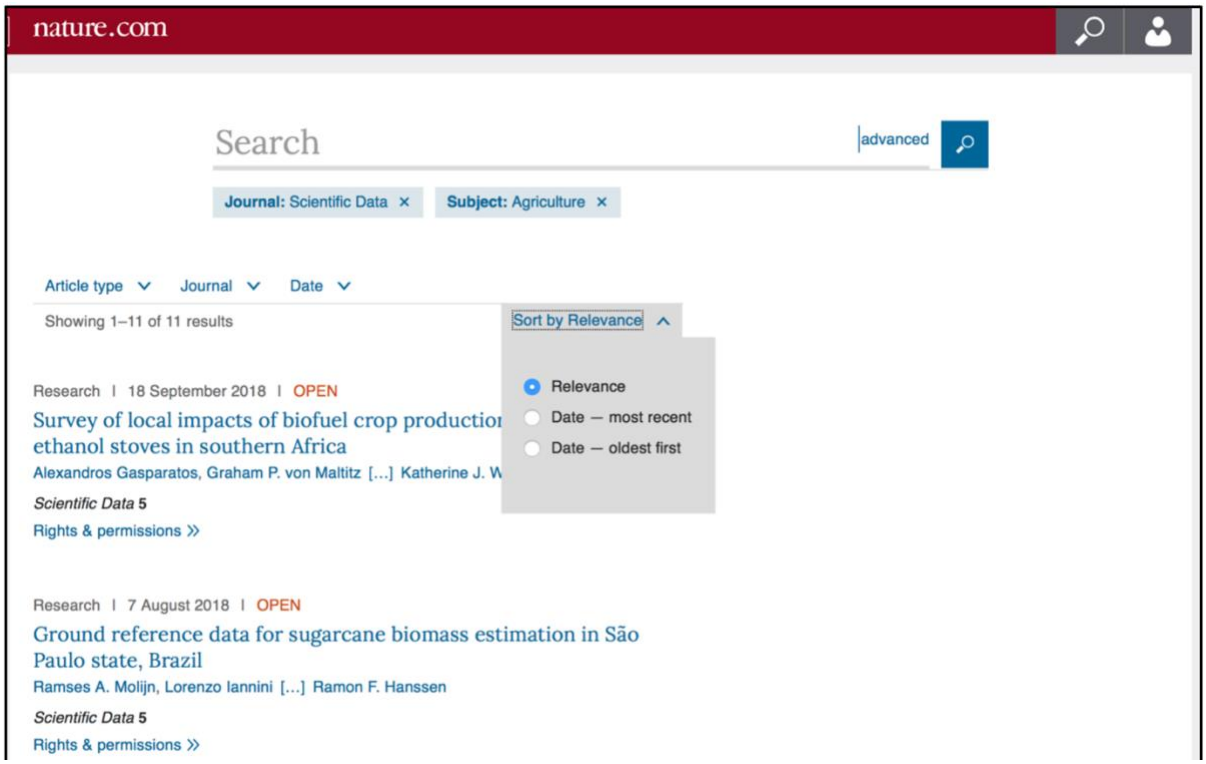

Figure 4.8. A publisher-service repository showing very generic options for sorting and filtering search results.

4.1.4 Location-based repositories

Research data held in these repositories are generally accessible to anyone globally, but data submissions are solicited and accepted only from researchers

within a specified geographical area; e.g. ANDS Research Data Australia[6], and the European Union Open Data Portal (EU ODP)[7]. Location-based data repositories, in part because they hold solely datasets, often show some degree of data-conscious design in the features and functionalities they support. Research Data Australia, for example, besides enabling data discovery through keyword search in various metadata fields (see Figure 4.9) and through browsing by subject, provides a number of advanced search options which are relevant to research data (e.g. relating to the type, subject, geographical origin, access, licence type, etc. of the data. See Figure 4.10). Although faceted advanced search of this kind is not uncommon also in large publication databases such as Web of Science[8] and Scopus[9], it is less commonly to be found in research data repositories. In fact, many institutional, publisher-service, and general-purpose repositories do not provide this feature. The main difference between the faceted search of research data repositories and that of research publication databases is in the choice of the options and their relevance for data discovery.

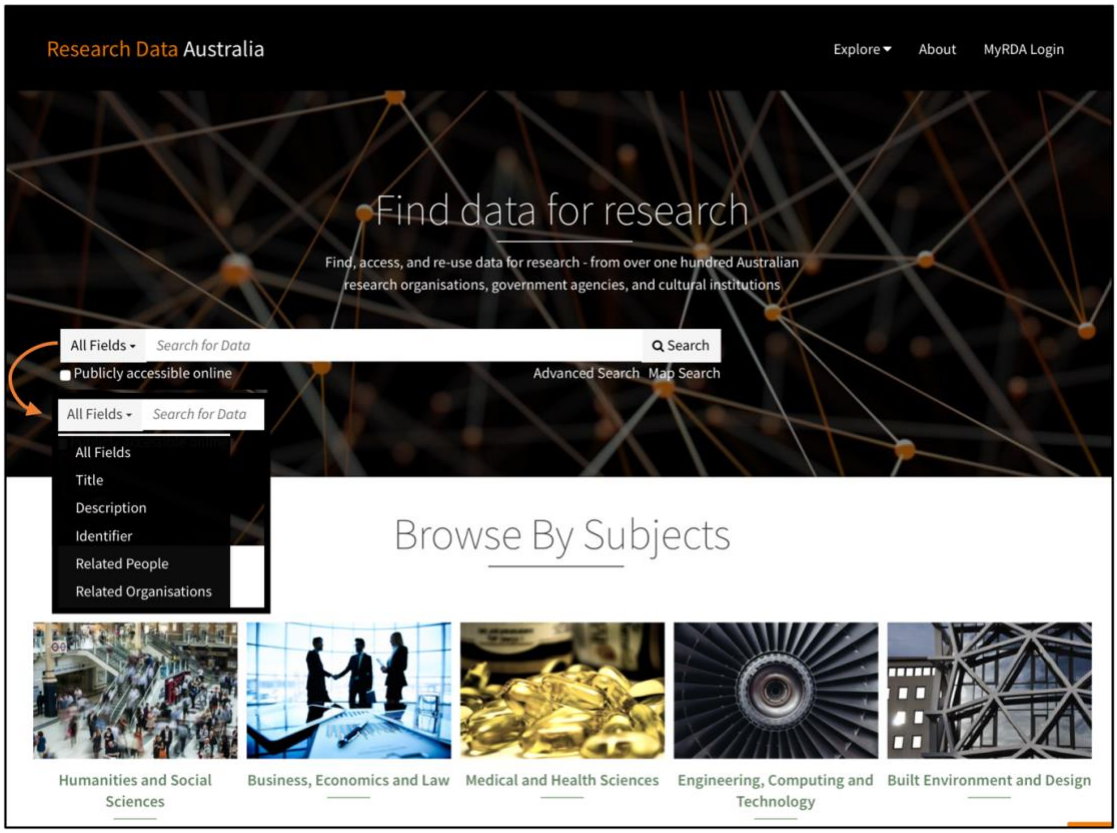

Figure. 4.9. Homepage and initial search interface of a location-based repository

6 https://researchdata.ands.org.au
7 https://data.europa.eu/euodp/en/data
8 https://www.webofknowledge.com
9 https://www.scopus.com

The search parameters usually available for publication search, such as “document type”, “publication”, and “editors” (see Figures 4.11 for Web of Science and 4.12 for Scopus) evidently mean that the resource in question is a manuscript document while, for data search, the options (e.g. data provider, access type, licence) are clearly more geared for use with datasets (see Figure 4.10).

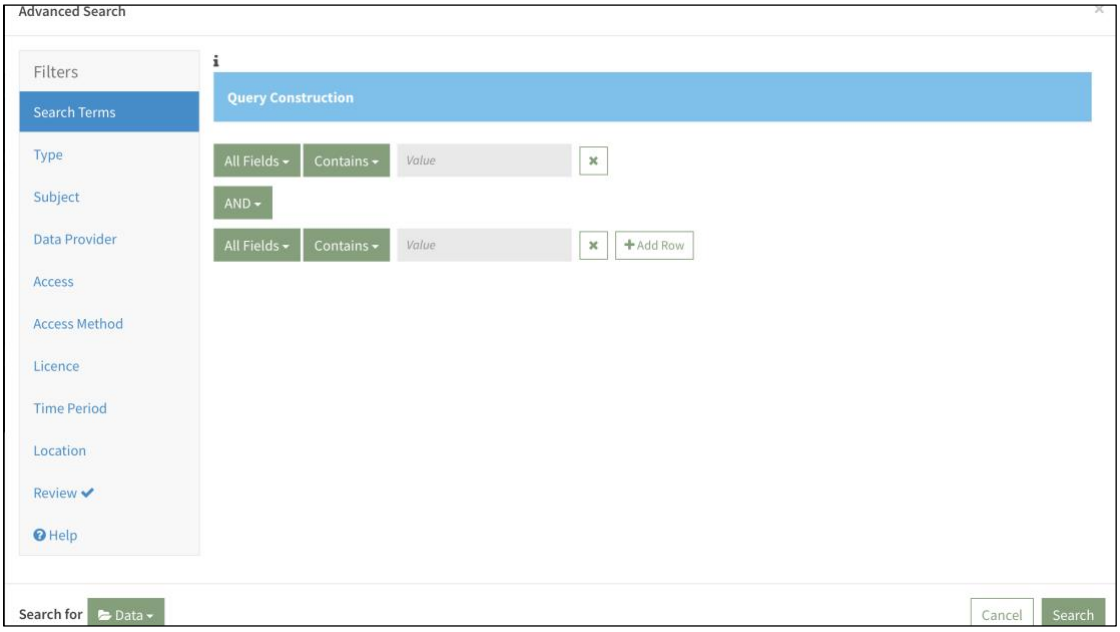

Figure. 4.10. Advanced search options by Research Data Australia

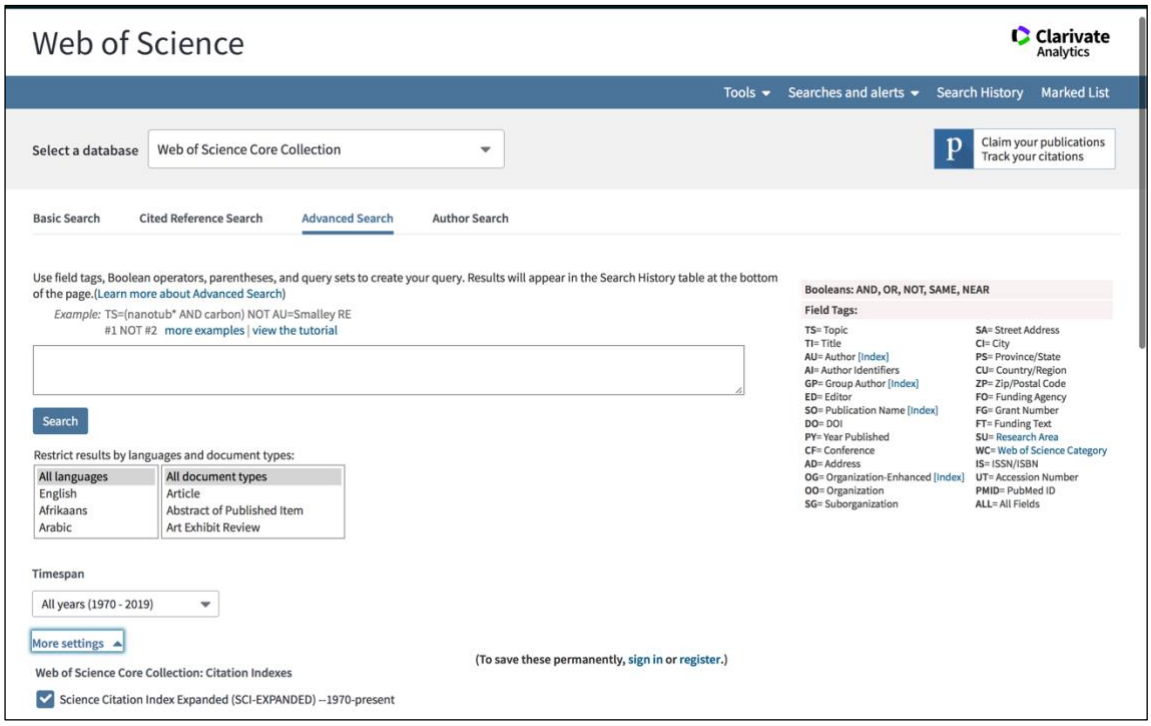

Figure. 4.11. Advanced search options by Web of Science

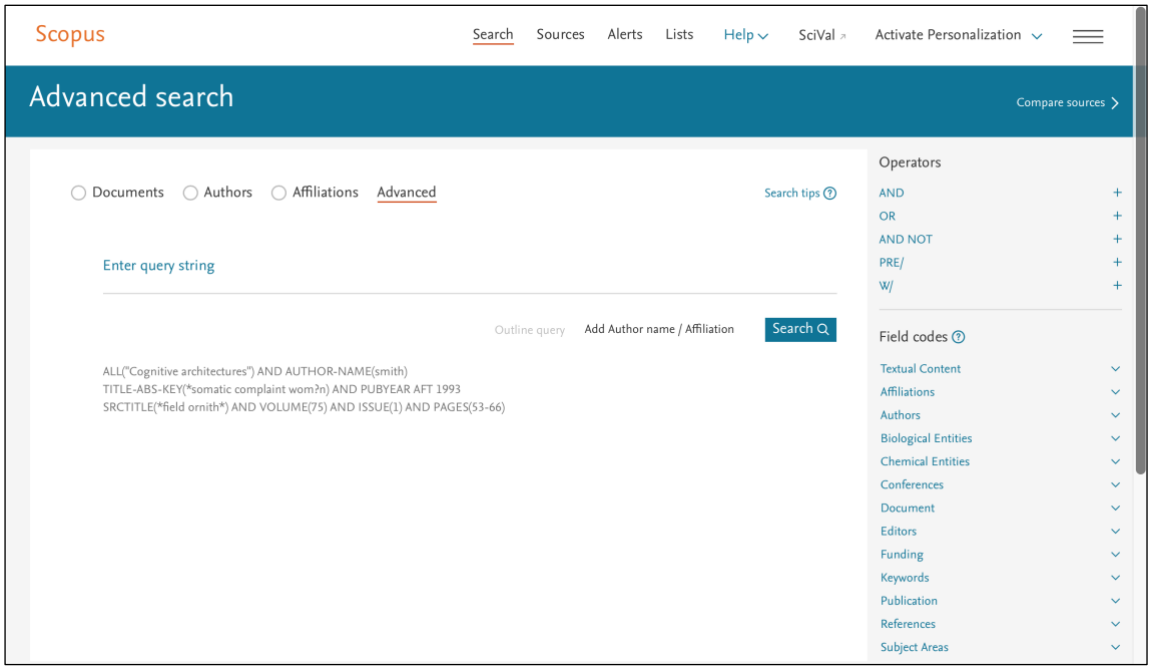

Figure. 4.12. Advanced search options by Scopus

As Figure 4.13 shows, the options for filtering search results are also quite extensive. Despite all this, however, location-based data repositories, in their attempt to accommodate the range of data that fall within their geographical boundaries, sacrifice much of the benefits that come of having a more streamlined content, including the opportunity of using and exploiting less generic and more specific metadata to provide a better service.

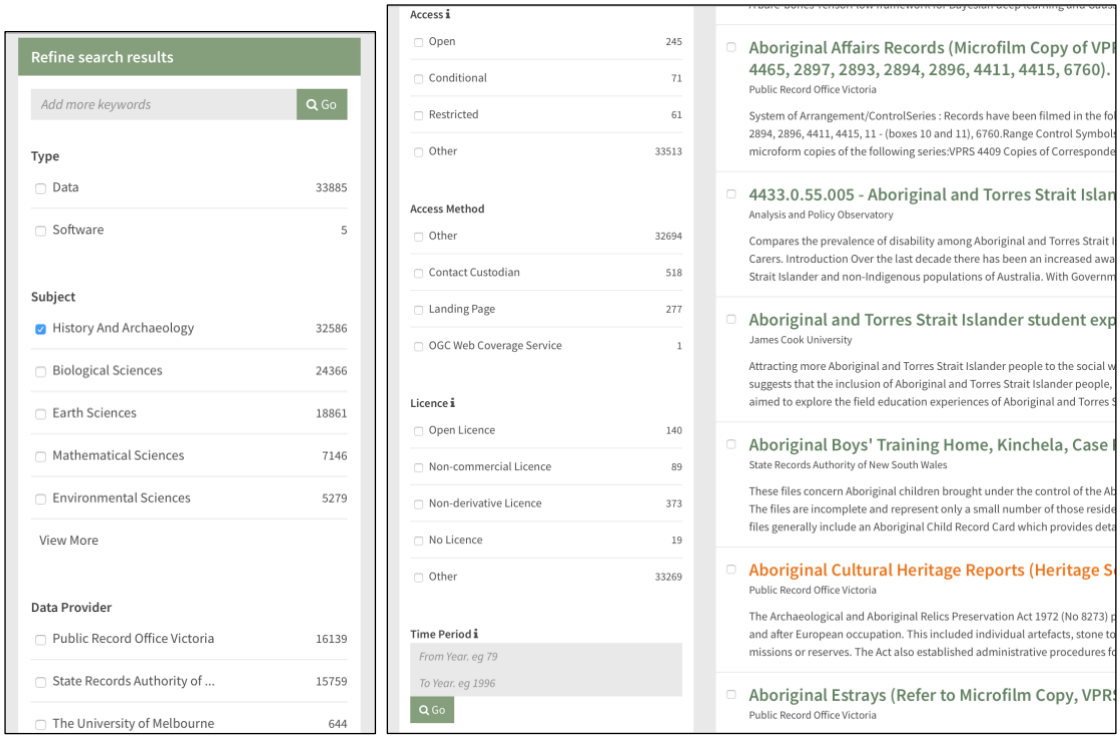

Fig. 4.13. Search result filtering options by Research Data Australia

4.1.5 Dedicated content-type repositories

These exclusively or predominantly house research data of a certain file type or format. The Visual Arts Data Service (VADS)[10], for example, is a repository exclusively for image data. By virtue of this relative homogeneity in their data type or format, dedicated content-type repositories may, potentially and with greater confidence, be designed around facts and concepts that best suit or express the special properties of their content and the possibilities peculiar to it.

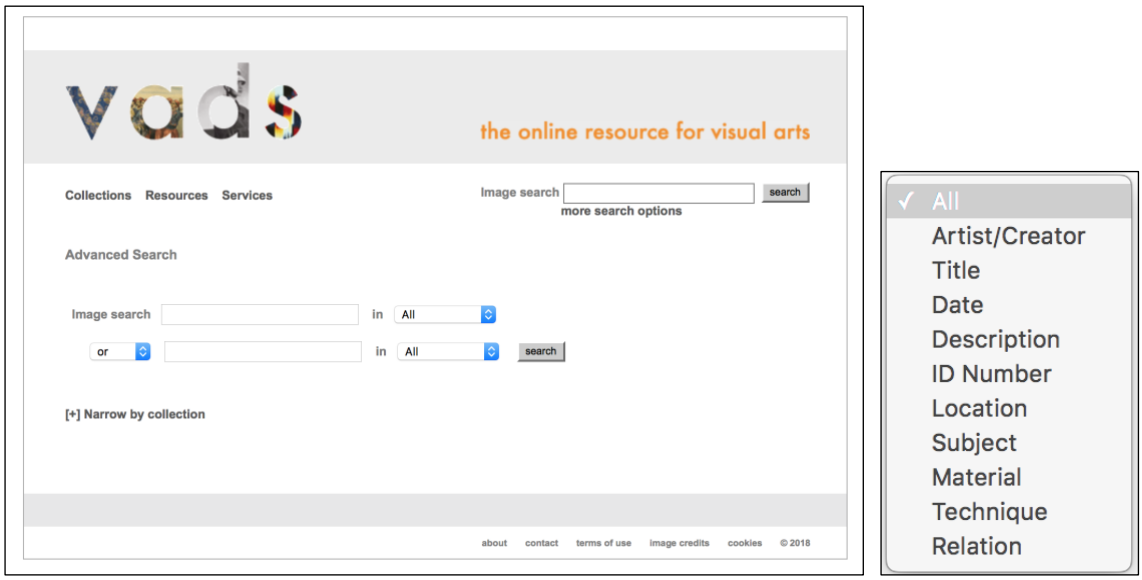

Fig. 4.14. Advanced search options by VADS

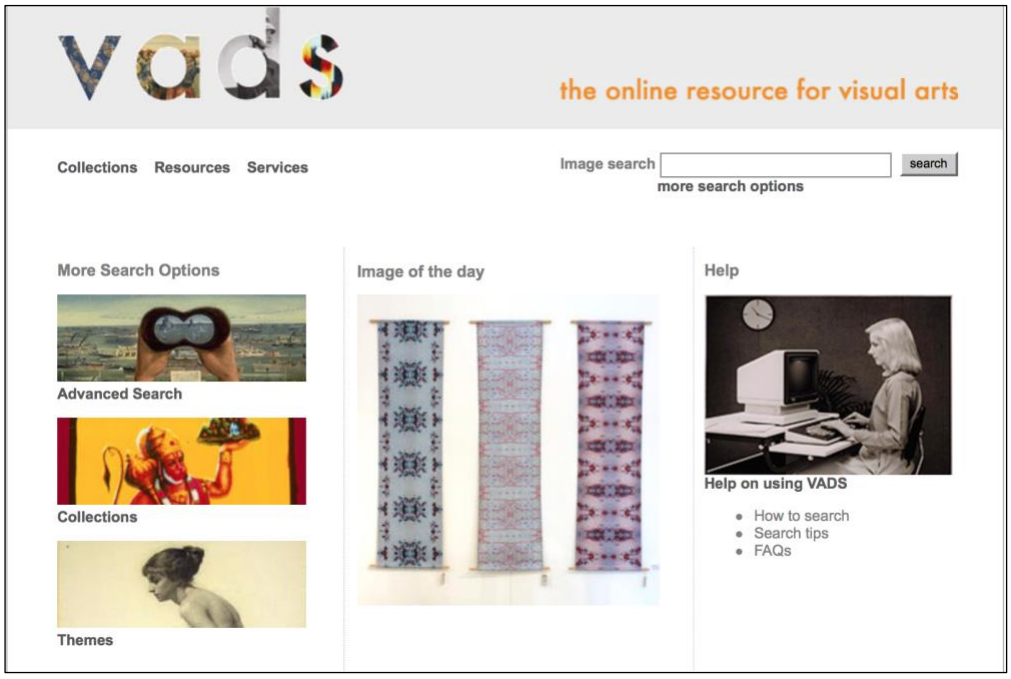

Fig. 4.15. Special browsing options by VADS

[10] https://vads.ac.uk

Figure 4.14 shows, on the left, the advanced search interface of VADS with, on the right, options to search special metadata fields such as “material” and “technique” that uniquely apply to digital or digitized images. Figure 4.15 shows other data discovery options, such as search by image collection or theme. Although this opportunity may, as in the case of VADS, not always be exploited, dedicated content-type repositories give ample scope for providing appropriate options for sorting (by artist, collection, title, or image content type in VADS) and filtering of search results.

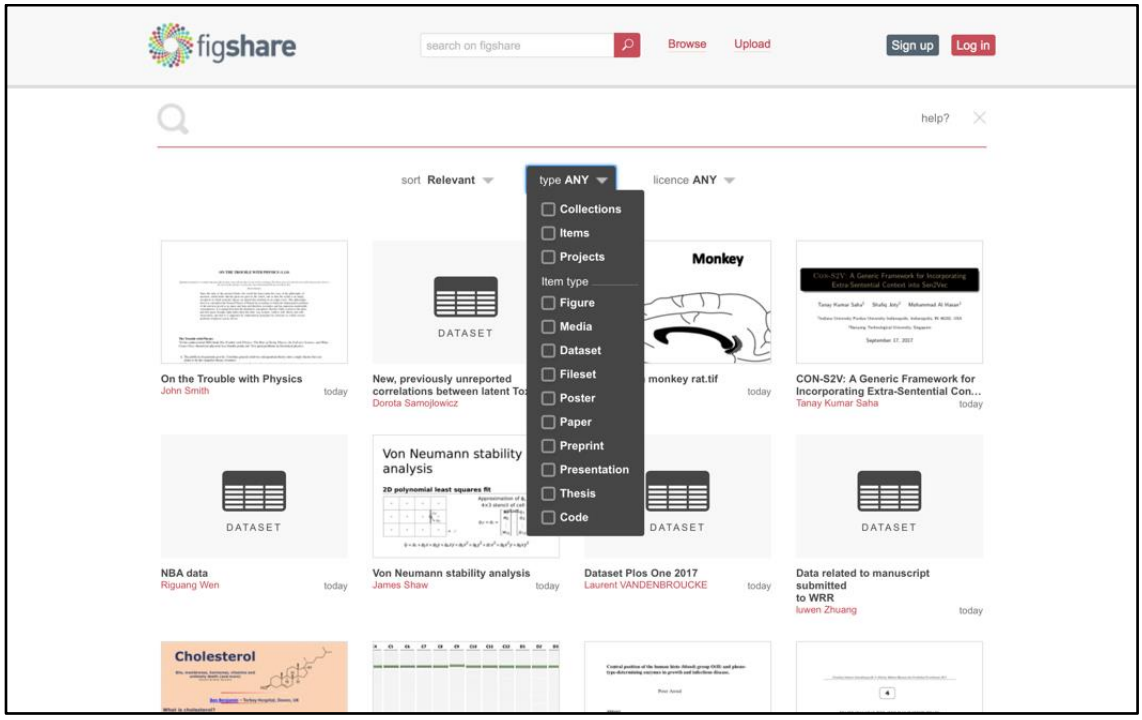

Fig. 1.16. Figshare as an example of general-purpose/commercial data repositories.

4.1.6 Commercial and general-purpose repositories

These repositories accept research data of almost any kind or origin. A popular example is Figshare[11], shown in Figure 4.16. This class of repositories tend to house multidisciplinary data, as well as data from niche disciplines that do not have dedicated repositories. As shown in Fig. 4.16, they also tend to hold other kinds of objects (e.g. in Figshare, posters and theses), by which general inclusivity they acquire many of the drawbacks that have been noted of institutional repositories, such as greater inability to enable fine-grained filtering of search results and support for expressive search queries. This is the case principally

[11] https://figshare.com

because the metadata that is needed to support such functionalities is, in the interest of inclusivity, kept generic at best.

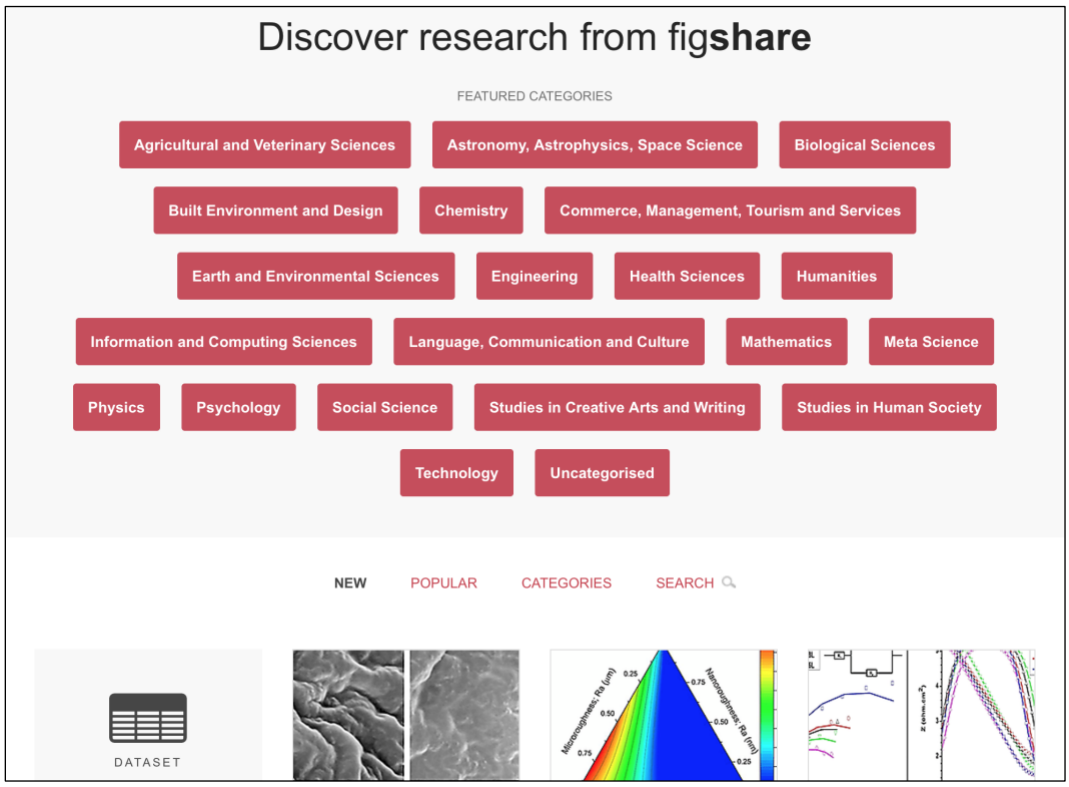

Fig. 2.17. Data browsing features of a commercial/general-purpose repository

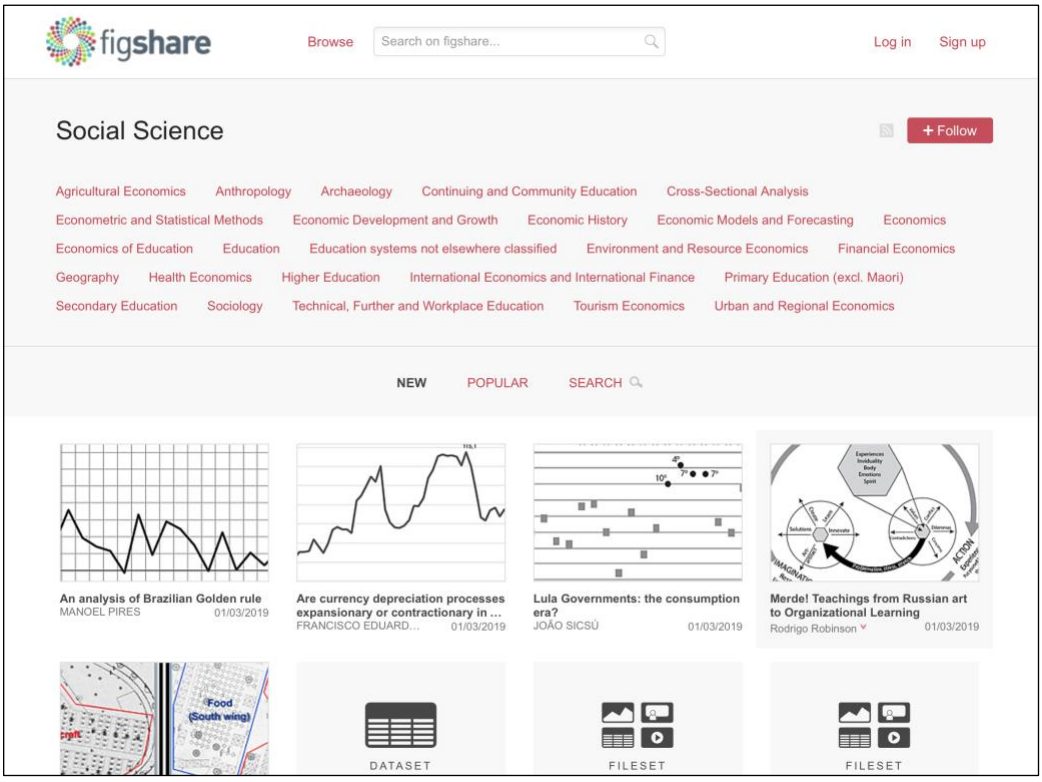

Fig. 3.18. Showing a commercial/general-purpose repositories with very basic result-sorting features

General-purpose repositories, in common with many of the other class of repositories, also provide features for browsing by subject (see Figure 4.17). Beyond this, however, there seldom are provided further filtering options for narrowing down search results (see Figure 4.18 for example). In the case of Figshare, it does not provide advanced search features; but it nevertheless supports the very unique and useful feature of allowing datasets to be previewed prior to download (see Figure 4.19). The potential advantages of providing this feature is discussed further in Section 4.1.7.

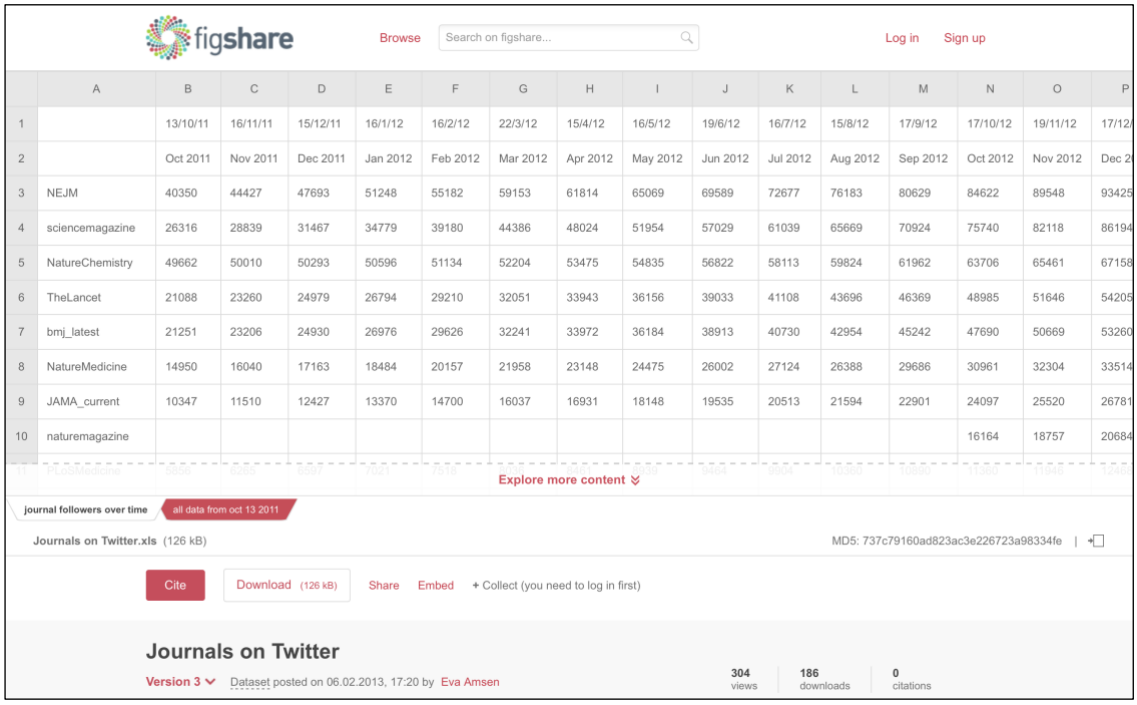

Fig. 4.19. Figshare's data preview feature.

This completes the systematic review of some of the research data repositories currently in use. Table 4.1 presents a succinct summary of the preceding subsections, and the list below highlights the main design issues of RDM systems identified in the study; thus:

i. Limited user interactivity
ii. Insufficient or unavailable metadata
iii. Quality of data questionable or not assured

These issues are annotated in Table 4.2 with comments from the literature and notes on their potential implications on user experience as well as on computing resources.

Table 4.1. Summary of findings with respect to the evaluation criteria of the study (see Section 3.1 in Chapter 3)

Type of repository	**Metadata**	**Querying facility**	**Sorting facility**	**Result filtering**	**Availability of additional features for data**
Disciplinary	Detailed	Expressive, sometimes with browsing features	Extra criteria sometimes provided	Multiple options usually provided	Usually available
Institutional	Very generic	Very basic, with browsing features	Where provided, criteria are generic	Very few options, if any, and usually basic	Unavailable
Publisher-service	Relatively generic	Relatively basic, with browsing features	Usually by generic criteria	Few, basic options	May be available
Location-based	Depends, but usually less generic	Some degree of expressiveness often possible	Extra criteria sometimes provided	Multiple options may be provided	May be available
Dedicated content-type	Detailed	Usually expressive, with browsing features	Extra criteria sometimes provided	Multiple options sometimes provided	May be available
Commercial & general-purpose	Very generic	Usually basic, with browsing features	Where provided, criteria are generic	Very few options, if any, and usually basic	May be available

Table 4.2. Summary of findings from with their corresponding implication(s) on user-experience

Key issues	Comments	Implications
Limited user interactivity	e.g. No feature(s) for previewing dataset content on the web browser before download. This unnecessarily increases the rate of download, making each session highly resource intensive	• Downloading data that ends up unused unduly strains network resources • Poor use of storage space • Renders download count unreliable as a measure of dataset relevance, perceived usefulness, or impact
Insufficient or unavailable metadata	The lack of use of standard metadata to sufficiently contextualize data for discovery (Chowdhury, 2014; Boru et al., 2015) & re-use (Weber & Piesche, 2016) is a major challenge. Deficiency in metadata quality or quantity, along with the fact that using generic metadata for greater inclusivity directly translates into loss of nuanced features, presents a common problem.	• Complex or precise queries cannot be supported • Loosely matching search results • Tedious manual browsing of results • Unproductive use of researchers' time • Threatens the discoverability and, consequently, reuse rate of research datasets
Quality of data questionable or not assured	Researchers tend to reuse the datasets of others whom they trust (RIN, 2008). Many services do not have mechanisms to ensure the quality of user-uploaded datasets; nor are there any standard criteria for measuring the quality of research data.	• Skepticism, which may stunt the rate of data reuse • Time which could be used more productively in active research spent on making inquiries about data.

It is perhaps pertinent at this juncture, before concluding the section, to note the different methods that are used by RDM systems for data upload. These methods are three:

i. Unsupervised upload: whereby the data holder is free to upload their data without express approval of, or checks from, repository staff members. E.g. Figshare. Its advantage is fastness, as it requires less human intervention and monitoring. Its disadvantage is susceptibility to mistakes and errors, deliberate or otherwise.
ii. Semi-supervised upload: whereby the data holder fills out the information in the data upload form, and uploads their data; but before the data is published a repository staff member inspects the submission to ensure it meets quality requirements. E.g. University of Southampton repository. The advantage of this is that errors are likely to be detected and corrected, while its disadvantage lies in its requiring human intervention to inspect and approve each data upload request.
iii. Manual upload by repository staff: in this scenario a repository staff member obtains the required data upload information from the data holder and does everything from filling out the template to uploading and publishing the data. The advantage of this approach is that data are likely to be more correctly documented, due to the manual checks conducted to ensure this; its disadvantages are it does not scale, requires considerable human resources, and is time consuming.

The typical statuses for data records held in repositories, reproduced below in Table 4.3, were suggested by Rumsey & Jefferies (2013).

Table 4.3. The typical statuses for data records held in repositories

Status	**Description**
Draft	Depositor working on record
Submitted	Depositor has submitted record for review
Approved	Reviewed submission approved without modification
Escalated	Reviewed submission to be checked by another member of staff due to issues, such as commercial or legal agreements, or ethics. Note of problem added to admin record

Referred	More/better information needed before submission can be approved. Submission returned to the submitter with a note of the problem and how to rectify it
Rejected	The administrator reviewing the record has decided that there is something fundamentally wrong with it. Reasons for rejection sent to the submitted

4.1.7 Section Summary

It has been noted in Section 1.1.8 that mere publishing of research data offers little benefit, and that effectively communicating it, rather, is what is needed. In this vein, Günther & Dehnhard (2015) further noted that "publishers face considerable challenges when trying to advance from publishing to communicating research data", but that "developing solutions pointing in this direction should be, nevertheless, of primary concern, as publishing without communicating might ultimately be just a waste of resources". Data repositories being almost the sole publishers of research data, it therefore follows that the above quoted statements directly and closely apply to them. The systematic reviews of data repositories presented in the preceding subsections investigate as to the facts concerning the first statement, and partly follow up on the suggestions indicated in the second, to ascertain as to their applicability. To the extent needful for the purpose of this research, I have identified some of the strengths and weaknesses of each class of repository, and the advantages it enjoys or disadvantages it encounters, in consequence of its supported or unsupported features or functionalities. It may be deduced from this study that the narrower the range of resource objects held by a repository, the more the possibility of providing better and more relevant service for data discovery.

4.2 Online questionnaire survey

This section presents the analysis and key findings of the study described in Section 3.2. To facilitate this, and for better coherence, the discussions will be organized under four broad themes, as follows:

a. Research data sourcing and sharing;
b. Research data storage;
c. Research data practices, training & awareness; and
d. Research data attributes;

Of the 201 usable responses, 191 were fully complete and 10 nearly complete. The latter were generally omitted from the analyses except when the question in hand does not require use of the missing columns even for the fully complete rows. The coding and analysis of the data was performed with R, at the 0.05 significance level for the statistical tests (Chi-Square). Table 4.4 below gives a summary of the survey respondents by discipline and years of experience.

Table 4.4: Disciplines and years of experience of respondents (n=199)

		Years of Experience					
Discipline	**n**	**< 5**	**5 - 10**	**11 - 15**	**16 - 20**	**> 20**	**none**
Computer, Library, and Information Science	**23**	8	6	3	0	5	1
Architecture, Design, and Built Environment	**14**	5	2	2	4	1	0
Healthcare, Social Care, Life Sciences	**45**	11	8	7	9	9	1
Sports, Exercise, and Rehabilitation	**4**	2	1	0	1	0	0
Astrophysics and Solar Physics	**4**	0	1	1	1	1	0
Mathematics and Statistics	**15**	3	1	3	5	3	0
Arts & Humanities	**34**	11	5	7	4	7	0
Social Sciences	**43**	13	15	3	6	6	0
Applied Sciences	**6**	3	0	1	0	2	0
Education	**4**	3	0	0	1	0	0
Engineering	**7**	2	0	3	0	2	0
	n=199	***61***	***39***	***30***	***31***	***36***	***2***
	%	***30%***	***20%***	***15%***	***16%***	***18%***	***1%***

4.2.1 Research data sourcing and sharing

Across communities, the avenues for research data discovery commonly include data repositories, journals, websites, and personal networks. Sands et al. (2012) and Faniel & Yakel (2017) attribute this variety to the differing infrastructures available within disciplines. Personal networks are, according to the former, valuable sources of external data, especially in cases of specialized datasets. The present study found that, although few researchers (18%) reported their research data to be openly available, many (41%) further reported that it was available upon request. This supports the statement of Sands et al. (2012) as to the importance of personal networks as sources of research data, and also ties with the discussion on data sharing in Chapter 2 (see Section 2.3). In a similar study by Tenopir et al. (2011), it was found that although as much as three-quarters of the researchers who participated shared their data with others, that only about one-third agreed that their data was easily accessible, and that 46% reported they did not make their data electronically available. It seems therefore as Tenopir et al. (2011) also concluded, that willingness to share data is not lacking, but there appears to be some difficulty attached to putting it to practice; or else there is an apparent preference for sharing only upon request.

My study found research students (60%) to show the most willingness of all other groups to share their data with others; contrasting considerably with the responses of research/postdoctoral staff (0%) and academic staff (21%). Indeed, as Tenopir et al. (2011) observe, scientists do not all share data equally, nor are their perceptions the same about data sharing and reuse. In the present study, among the 199 researchers that responded to the question, highly statistically significant differences were found in 3 instances out of a total of 4, between researcher's post (viz. academic staff, research student, research staff/postdoctoral scientist, and retired academic) and their data sharing behavior. That is, those *not sharing data* ($\chi^2_{(4)} = 41.14$, $p = 0.0003048$); those *sharing with own team* ($\chi^2_{(4)} = 28.996$, $p = 0.0161$); those *sharing with researchers in other institutions* ($\chi^2_{(4)} = 41.41$, $p = 0.0002767$); and (not statistically significant) those *sharing with researchers in the same university* ($\chi^2_{(4)} = 24.496$, $p = 0.05713$). An explanation for the statistical non-significance of the last mentioned is not easily apparent.

Tenopir et al. (2011) found in their study that researchers' discipline, especially, and their age, work focus (research-focused vs teaching-focused), and geographical region (U.S., Europe, and rest of world) significantly influenced their data sharing and data management practices and perceptions. In my study also, discipline made statistically significant differences in all four instances of data sharing behaviour (*N = 199*; *not sharing data* $\chi^2_{(4)} = 31.7$, $p = 0.001537$; *sharing with own team* $\chi^2_{(4)} = 28.189$, $p = 0.005191$; *sharing with researchers in the same university* $\chi^2_{(4)} = 24.916$, $p = 0.01523$; and *sharing with researchers in other institutions* $\chi^2_{(4)} = 34.27$, $p = 0.0006109$). Significant disciplinary correlation was also found to exist among those (*N = 199*) who reported that their data was openly available to everyone ($\chi^2_{(4)} = 24.954$, $p = 0.01504$).

Independent of researchers' habits, preferences, tendencies or concerns regarding data sharing, Sedghi et al. (2011) consider, as a probable barrier to the same, that the existing search functionalities offered by data repositories fail to meet the specific needs or skillset of researchers. Or else, that researchers may simply be unaware that such data or repositories are available, especially if the data falls outside their primary disciplines. The study by Tenopir et al. (2011) reported between 60%-90% of respondents in all disciplines as having agreed that they "would use other researchers' datasets if easily accessible''. A different study by Weller & Monroe-Gulick (2014) showed a disciplinary correlation with data accessibility; quantitative and experimental researchers being less likely than others to "struggle" with obtaining access to data.

The present study found that in general researchers would collect their own data rather than reuse one already existing (see Table 4.5). It is not certain whether this is due to the data needed being unavailable for use or the possible whereabouts of it being unknown to the researcher, or else for other reasons. Researchers with the least years' experience (see Table 4.5) reported more commonly than any other group that they create their own data (i.e. collect primary data), and those with the greatest years' experience create their own data less commonly than any other group. This is readily understandable, because as researchers' years of experience increase they may be assumed to have been accumulating a good store of primary data, or else to have had greater opportunity of forming personal networks for obtaining data from others.

Table 4.5. Researchers' years of experience and mode of sourcing data (n=201)

Experience	Create New Data	From own research team	From own contacts	Other
< 5 years	39 *(57%)*	13 *(19%)*	13 *(19%)*	4 *(5%)*
5 – 10 years	19 *(45%)*	8 *(19%)*	10 *(24%)*	5 *(12%)*
11 – 15 years	15 *(36%)*	11 *(26%)*	8 *(19%)*	8 *(19%)*
16 – 20 years	14 *(45%)*	5 *(16%)*	10 *(32%)*	2 *(7%)*
> 20 years	12 *(31%)*	11 *(28%)*	10 *(26%)*	6 *(15%)*
Total	99 (*44%*)	48 (*22%*)	51 (*23%*)	25 (*11%*)

As may be seen in Tables 4.6–4.8, across almost all disciplines (Table 4.6), years of experience (Table 4.7), and researchers' posts (Table 4.8), legal and ethical concerns represent the chief among researchers' "concerns" about sharing data. This is followed by the fear of others' misinterpreting their data.

Table 4.6. By subject discipline, researchers' concerns about sharing data (N = 201)

Discipline	*n*	**No concerns**	**Losing scientific edge**	**Legal/ ethical concerns**	**Misuse of data**	**Misinterpretation of data**	**Lack of resources**	**Lack of policies & rights**
Computer, Library, and Information Science	*23*	13%	22%	65%	43%	52%	13%	26%
Architecture, Design, and Built Environment	*14*	43%	14%	57%	21%	29%	7%	7%
Healthcare, Social Care, Life Sciences	*45*	9%	20%	80%	40%	51%	18%	18%
Sports, Exercise, and Rehabilitation	*4*	0%	25%	100%	25%	25%	0%	25%
Astrophysics and Solar Physics	*4*	25%	50%	0%	0%	50%	50%	0%
Mathematics and Statistics	*15*	27%	7%	47%	27%	27%	7%	0%
Arts & Humanities	*34*	18%	15%	38%	32%	32%	9%	15%
Social Sciences	*43*	14%	16%	65%	26%	37%	26%	21%
Applied Sciences	*6*	17%	33%	33%	33%	33%	33%	33%

Discipline	*n*	No concerns	Losing scientific edge	Legal/ ethical concerns	Misuse of data	Misinterpretation of data	Lack of resources	Lack of policies & rights
Education	*4*	50%	0%	75%	25%	0%	0%	0%
Engineering	*7*	14%	43%	29%	57%	57%	0%	57%
	total	**17%**	**19%**	**59%**	**33%**	**40%**	**16%**	**18%**

Table 4.7. By years of experience, researchers' concerns about sharing data (N = 199)

Years of Experience	*n*	No concerns	Losing scientific edge	Legal/ ethical concerns	Misuse of data	Misinterpretation of data	Lack of resources	Lack of policies & rights protection
< 5 years	*63*	17%	17%	60%	27%	33%	6%	14%
5 – 10 years	*39*	13%	26%	64%	36%	38%	21%	18%
11 – 15 years	*30*	13%	17%	47%	50%	57%	17%	20%
16 – 20 years	*31*	16%	19%	65%	32%	42%	19%	16%
> 20 years	*36*	25%	14%	58%	25%	36%	22%	28%
	total	**17%**	**19%**	**60%**	**33%**	**40%**	**16%**	**19%**

Table 4.8. By job post, researchers' concerns about sharing data (N = 199)

Researchers' post	*n*	No concerns	Losing scientific edge	Legal/ ethical concerns	Misuse of data	Misinterpretation of data	Lack of resources	Lack of policies & rights
Academic Staff	*127*	17%	18%	63%	35%	42%	22%	19%
Research student	*55*	18%	20%	51%	25%	36%	4%	15%
Research staff/Postdoc	*13*	15%	15%	62%	31%	38%	8%	31%
	total	**17%**	**18%**	**59%**	**32%**	**40%**	**15%**	**18%**

*Other concerns cited are: fear of breaching employment contract, protection of intellectual property, and lack of time.

4.2.2 Research data storage

The present study discerned a common tendency in researchers to store their data on personal devices (e.g. USB sticks and external hard drives, 81%), confirming a similar finding in the study by Weller & Monroe-Gulick (2014), who also found external hard drives/CDs to be the most common method of digital file storage among researchers. Researchers also store data on university central servers (62%) and the cloud (31%). External data repositories (11%) seem rarely to be used. Figure 4.20 shows this graphically. Statistical tests, however, revealed no significant relationships between researchers' data storage choices and their subject discipline (data summarized in Table 4.9).

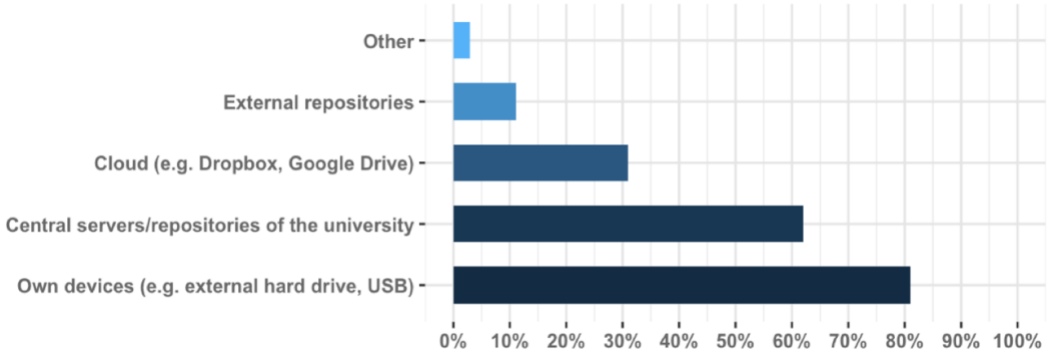

Figure 4.20. Data storage choices of researchers

Table 4.9. Data storage choices of researchers, by discipline (n=199)

	Amount of data produced			Where it is stored				
Discipline	**MB**	**GB**	**TB**	**Own devices (e.g. USB stick, external drives..)**	**Cloud storage (e.g. Dropbox. Google drive)**	**Central servers or university repositories**	**External repositories**	**Other**
Computer, Library, and Information Science	55%	41%	0%	78%	39%	48%	9%	0%
Architecture, Design, and Built Environment	43%	36%	14%	100%	43%	50%	7%	0%

	Amount of data produced			Where it is stored				
Discipline	**MB**	**GB**	**TB**	**Own devices (e.g. USB stick, external drives..)**	**Cloud storage (e.g. Dropbox. Google drive)**	**Central servers or university repositories**	**External repositories**	**Other**
Healthcare, Social Care, Life Sciences	58%	33%	0%	76%	18%	78%	13%	4%
Sports, Exercise, and Rehabilitation	75%	25%	33%	50%	25%	50%	25%	0%
Astrophysics and Solar Physics	0%	75%	25%	75%	25%	75%	50%	0%
Mathematics and Statistics	71%	29%	0%	60%	27%	53%	20%	0%
Arts & Humanities	62%	29%	3%	82%	24%	56%	6%	9%
Social Sciences	53%	40%	7%	88%	30%	65%	7%	2%
Applied Sciences	17%	33%	50%	100%	50%	67%	0%	17%
Education	25%	50%	25%	75%	50%	50%	25%	0%
Engineering	43%	29%	29%	86%	71%	57%	14%	0%

Weller & Monroe-Gulick (2014), in their study, found that researchers' data storage practices vary by research methodology. For example, ease of the storage method (88%) and long-term sustainability (62%) influenced historians more than researchers using other methodologies. Historians also showed less likelihood to be influenced by grant requirements (8%). Privacy and security concerns, on the other hand, were found to motivate statistical (51%), quantitative (50%), experimental (49%) and qualitative researchers (47%) more than historians (34%). Furthermore, concerns over file size and back up needs seemed more pressingly to be felt by quantitative and statistical researchers than by others. On the whole, however, regardless of the research methodology, ease of storage proved the primary influencing factor. 80% of my survey respondents (n = 196) agreed they would like to ensure long-term availability of their datasets

post-project, although more than half (58%, n = 200) also reported that they were unaware of the data policies of their funders or universities, as regards the storage and disposal of their research datasets. Akers & Doty (2013) in a different study found natural scientists as belonging to the discipline most likely to be familiar with funding agency requirements, followed by social scientists. Humanists were observed to be the least likely. Moreover, researchers, data managers, and publishers who participated in the PARSE.Insight (2010) survey opined, in large numbers, that an international infrastructure for data preservation should be built (Tenopir et al., 2011). They named lack of sustainable hardware and software as being the foremost “threat” to “digital preservation”.

4.2.3 Research data practices, training, & awareness

Regardless of research nature or context, observes Qin (2013), data needs to be “stored, organized, documented, preserved (or discarded), and made discoverable and usable” again. Moreover, these processes take up considerable time and labor, and the persons responsible for their discharge require training in technology as well as in subject fields. According to Anderson (2004), “the metadata required to describe data can be more complex than that required to describe written texts”. The great majority (92%, n = 201) of researchers from my study reported that they had had no formal training in metadata, although more than half indicated an interest in acquiring it (see Table 4.10, also graphically presented in Figure 4.21). Over 75% also reported that they had never received other RDM training, such as in version control of data sets, writing of DMPs, consistent file naming or data citation. Another large majority (74%, n = 197) of researchers responded that they never (54%) or only rarely (20%) used standard metadata to tag their datasets. This corroborates a similar finding by Tenopir et al. (2011) in their study, in which about 78% of the respondents either used no metadata (56%) or only “home-grown” metadata (22%) to tag their datasets. Tenopir et al. (2011) also reported that more than half (59%) of their respondents indicated that their organization or the project did not provide training on best practices for data management.

A strong statistical correlation exists between researchers’ subject discipline and their familiarity with metadata ($N = 197$, $\chi^2_{(2)} = 34.29$, $p = 0.0242$), with 100% of Computer & Information Science as well as Education researchers reporting that

they were familiar with it (see Table 4.11). Viewed through the lens of researchers' years of experience, a positive relationship is discernible (see Table 4.12, also graphically presented in Figure 4.22) between that and the number of researchers' who reported that they were familiar with metadata. Statistical tests of correlation between the two produced highly significant results ($N = 198$, $\chi^2_{(2)} = 42.81$, $p = 0.000005378$).

Table 4.10. Researchers' RDM training interests and previous training received (n=201)

Topic	Received training	Not received training	Interested in receiving training
Version control of data sets	5%	95%	47%
Data Management Plan	12%	88%	54%
Consistent file naming	4%	96%	42%
Data citation styles	24%	76%	36%
Metadata	8%	92%	56%
None of the above	64%	36%	18%

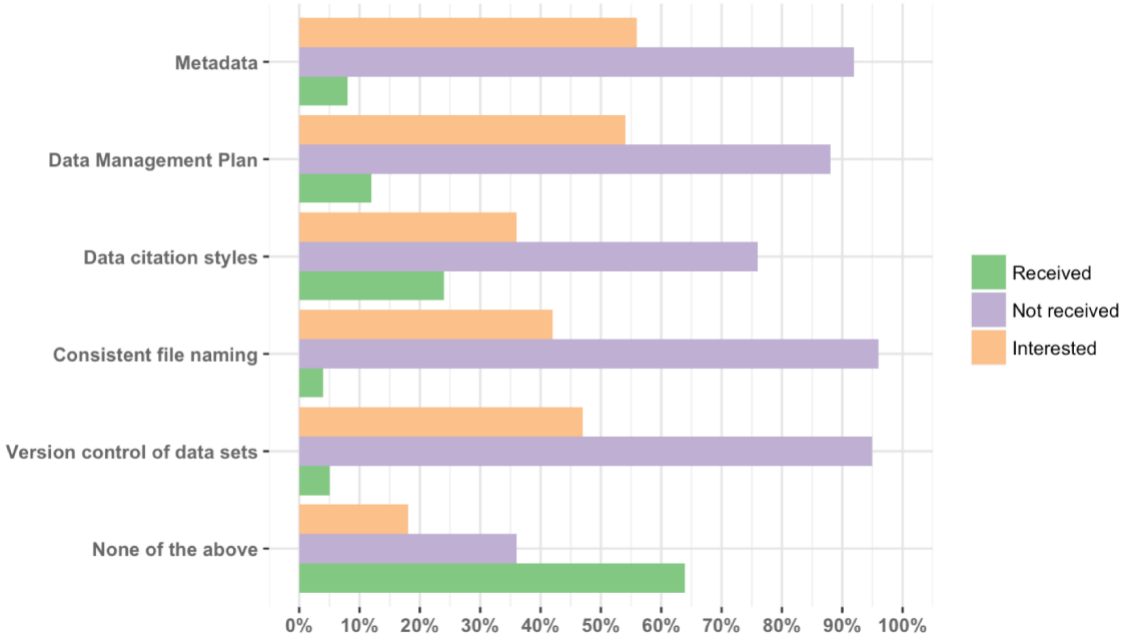

Figure 4.21. State of researchers' RDM training

Table 4.11. Researchers' familiarity with metadata, by discipline (n=197)

Discipline	*n*	Familiar	Not familiar	Uncertain
Computer, Library, and Information Science	*22*	100%	0%	0%
Architecture, Design, and Built Environment	*14*	71%	21%	7%

Discipline	*n*	Familiar	Not familiar	Uncertain
Healthcare, Social Care, Life Sciences	*44*	73%	11%	16%
Sports, Exercise, and Rehabilitation	*4*	25%	75%	0%
Astrophysics and Solar Physics	*4*	75%	0%	25%
Mathematics and Statistics	*15*	87%	13%	0%
Arts & Humanities	*34*	79%	15%	6%
Social Sciences	*43*	77%	9%	14%
Applied Sciences	*6*	67%	33%	0%
Education	*4*	100%	0%	0%
Engineering	*7*	71%	29%	0%

Table 4.12. Researchers' familiarity with metadata, by years of experience (n=198)

Years of Experience	*n*	Familiar	Not familiar	Uncertain
< 5 years	*63*	54%	32%	14%
5 – 10 years	*39*	85%	5%	10%
11 – 15 years	*29*	86%	3%	10%
16 – 20 years	*30*	87%	10%	3%
> 20 years	*36*	97%	0%	3%

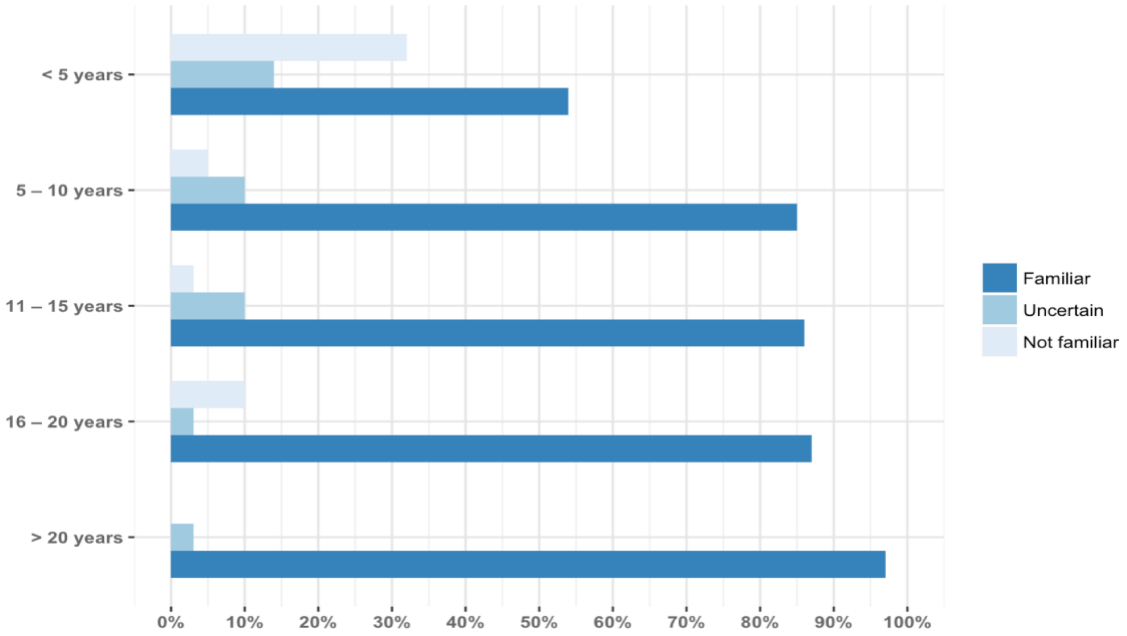

Figure 4.22. Researchers' familiarity with metadata, by years of experience

4.2.4 Research data attributes

From my survey, office documents seemed by far the most common file types used and produced by researchers (88% and 95% respectively, n = 201), shown in Table 4.13, and graphically, in Figure 4.23. Since the sizes of office documents are comparatively small and do not ordinarily exceed that range, this finding supports the one that shows more than half (54%, n = 201) of researchers as producing data in the order of megabytes (see Table 4.15), compared to gigabytes (35%) or terabytes (only 7%). It is worth remarking that the only disciplinary group which reported neither to produce, nor to use data in the order of megabytes is Astrophysics & Solar Physics. This is not unexpected being it is a "big science" discipline (Borgman, 2015). The tabulated summary of responses (see Table 4.14) thus seem to indicate some relationship between researchers' discipline and the volume of data they use or produce; however, statistical tests showed no significant correlation between either.

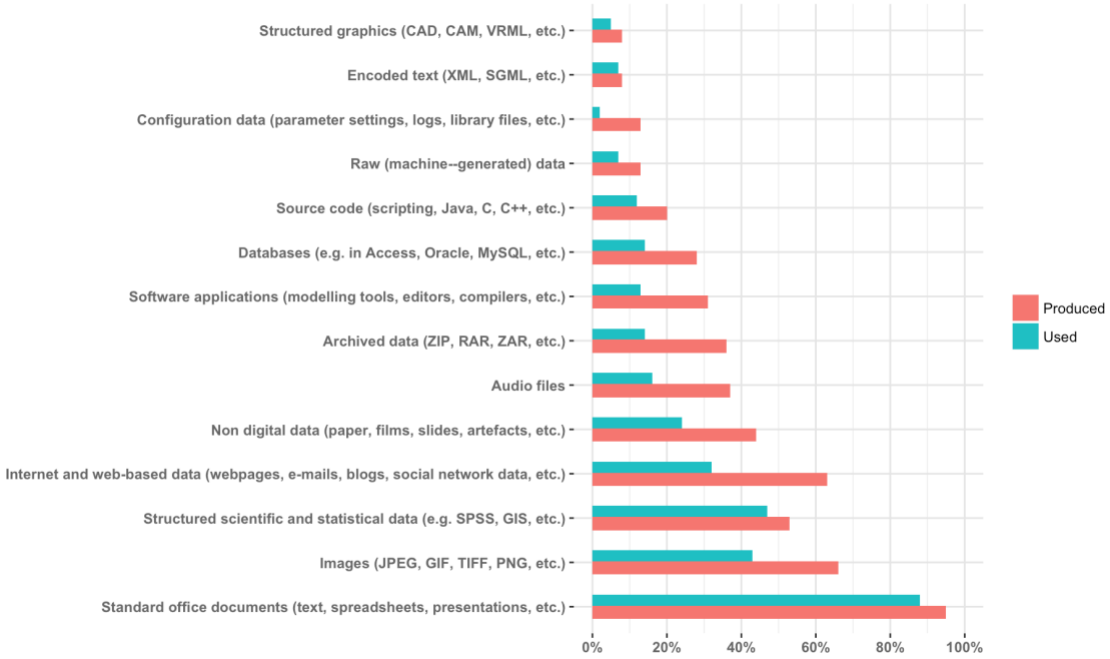

Figure 4.23. Types of data used and produced by researchers

Table 4.13. Types of data used and produced by researchers (N = 201)

Type of Data	Produced	Used
Standard office documents (text, spreadsheets, presentations, etc.)	95%	88%
Images (JPEG, GIF, TIFF, PNG, etc.)	66%	43%

Type of Data	Produced	Used
Structured scientific and statistical data (e.g. SPSS, GIS, etc.)	53%	47%
Internet and web-based data (webpages, e-mails, blogs, social network data, etc.)	63%	32%
Non-digital data (paper, films, slides, artefacts, etc.)	44%	24%
Archived data (ZIP, RAR, ZAR, etc.)	36%	14%
Audio files	37%	16%
Databases (e.g. in Access, Oracle, MySQL, etc.)	28%	14%
Software applications (modelling tools, editors, compilers, etc.)	31%	13%
Source code (scripting, Java, C, C++, etc.)	20%	12%
Raw (machine-generated) data	13%	7%
Configuration data (parameter settings, logs, library files, etc.)	13%	2%
Encoded text (XML, SGML, etc.)	8%	7%
Structured graphics (CAD, CAM, VRML, etc.)	8%	5%

Others noted are:* historic documentary archives, field observation, RAW Image files, physical objects, video.

Table 4.14. By subject discipline, volumes of data used and produced (n=197)

		Produced			*Used*		
Discipline	*n*	**MB**	**GB**	**TB**	**MB**	**GB**	**TB**
Computer, Library, and Information Science	*22*	55%	41%	0%	41%	55%	0%
Architecture, Design, and Built Environment	*14*	43%	36%	14%	50%	29%	14%
Healthcare, Social Care, Life Sciences	*45*	58%	33%	0%	36%	53%	2%
Sports, Exercise, and Rehabilitation	*4*	75%	25%	0%	75%	25%	0%
Astrophysics and Solar Physics	*4*	0%	75%	25%	0%	50%	50%
Mathematics and Statistics	*14*	71%	29%	0%	53%	27%	13%
Arts & Humanities	*34*	62%	29%	3%	41%	44%	6%
Social Sciences	*43*	53%	40%	7%	47%	42%	9%
Applied Sciences	*6*	17%	33%	50%	17%	33%	50%
Education	*4*	25%	50%	25%	25%	25%	50%
Engineering	*7*	43%	29%	29%	0%	71%	29%

Table 4.15. Volumes of data used and produced by researchers (n=201)

Volume of Data	Produced	Used
MB (megabyte)	54%	41%
GB (gigabyte)	35%	44%
TB (terabyte)	7%	10%
Small	<1%	1%
Don't know	2%	3%
Other	2%	2%

4.2.1 Section Summary

The study just reported aimed at understanding RDM system users and their needs, in order to find opportunities for improving and enhancing the usability of RDM systems. Disciplinary patterns appeared in certain tendencies of behavior, attitude, or perception of researchers with respect to the following:

1. Data sharing and reuse; for example, is more common among researchers in Mathematics, Statistics, and Solar Physics than those in Arts & Humanities or the Social Sciences. A possible explanation for this is given by Tenopir et al. (2011), who noted that researchers in the Social Sciences and in Medicine, who often produce sensitive data based on human subjects and consequently must take extra steps (e.g. anonymization) in preparing it for publication, are less likely to share their research data on repositories;
2. Size of data used and produced by researchers; for example, using or producing terabytes of data, except in disciplines such as Applied Sciences, Solar Physics, and some Engineering and Social Science fields, seemed much less common than data in the megabytes and gigabytes range;
3. Familiarity with metadata. 100% of Computer & Information Science researchers (n = 22), compared to only 25% of those in Sports, Exercise, and Rehabilitation (n = 4) reported that they were familiar with metadata. Besides researchers' discipline, years of experience also showed a statistically significant correlation on this point; as, the more the years of experience, the greater the proportion of researchers within that range who report familiarity with metadata;

4. Researchers' conceptions (or misconceptions) of certain key concepts or terminologies (e.g. what are Data Management Plans, or the term "research data"). For example, there were comments, mostly from researchers in Arts and Humanities, about what "data" really constituted.

All the above were enquired further into in the face to face interviews that followed the questionnaire survey. There were also, in addition, some significant remarks by researchers in the comment box, mostly suggesting a need or expressing frustration or discontent regarding certain aspects of RDM. These also warrant a closer examination. Given below are a few of these comments:

1. "Not easy to share data with others either inside the university or outside using our current systems…" (question: why so, and what can be done to ease the process or assist researchers?);
2. "…the data I use…is so specialised that it is hard to see either that other people have anything similar or that my data would be of any immediate use to anyone…" (question: is that what prevents the sharing of the data? does it also affect whether or not the dataset gets tagged with metadata? if the data is so specialized, are there or are there not any metadata standards that the researcher is aware of, to fit such uncommonly specialized data?);
3. "…the University should have its own institutional-open-data repository, linked to NRL [see Section 4.1.2], like Exeter University. External solutions run the risk of appearing cheap and transient to the external community…" (how may such opinions about external data repositories be explained and what may have occasioned them? What did the researcher particularly like about Exeter University's repository and why?); and
4. "…there should also be a dynamic national database that can be queried easily and which references researchers working on specific topics/areas of research, to improve cross-institution collaboration as well as understanding of research gaps." (what would the researcher find most helpful or useful in a service such as the one intimated?)

4.3 Face-to-face interviews

This study built upon experiences and findings from the questionnaire survey just presented. The interview transcripts were coded and analyzed using manual thematic content analysis techniques, as described by Ryan & Bernard (2003). The main themes that emerged were around the following:

1. Resource requirements. This refers to the need for extensive or advanced computing resources (e.g. backup and storage; processing power; required software) for data processing and storage.
2. Data attributes. Refers to details such as data size (in MB, GB, and TB), type (file extensions) and format (digital vs physical data, e.g. manuscripts; qualitative/narrative data vs quantitative/numeric data), all of which vary depending on discipline.
3. Norms and community dynamics. This relates to factors that include, among others, team orientation and data sharing culture. The former, for example, was strongest among Solar Physics researchers, while the latter was weakest among researchers from History.
4. Data sourcing & dissemination. This relates to the origin of the data, the method(s) of collection, and the channels through which data is shared/disseminated. Particulars that fall under this category include: source of data and the steps involved in sourcing it; pecuniary cost of the same to researcher; method of data collection; degree of standardization; and data sharing channels;
5. Other personal habits, practices, and concerns. Such as, data tagging, file naming, and metadata; use of ORCIDs, use of repositories; main cause of frustration with regards to data or repositories.

Researchers' responses were analyzed along and compared across disciplinary lines, to highlight intra-disciplinary similarities and inter-disciplinary differences. A summary of the analysis is presented in the table that follows (Table 4.16), succeeded by a discussion on each theme.

Table 4.16. Summary of the thematic analysis of interview data.

Resource requirements			
Theme	***History***	***Information Science***	***Solar Physics***
Backup and storage	Standard solutions/devices adequate. Personal devices commonly used.	May or may not require more advanced solutions. Personal devices often used.	Institutionally-provided massive storage facilities
Processing power	Standard processors adequate	Standard processors adequate in most but not all cases	Powerful computing resources needed
Software	Standard OS and Office programs	Standard OS, analysis software, and Office programs often suffice. Specialist software seldom necessary.	Usually work in Linux environments, with the command-line interface. MATLAB and/or other specialist software may be necessary
Data attributes			
Theme	***History***	***Information Science***	***Solar Physics***
Data subject/ originator	When human subjects are involved, they are dead humans	When human subjects are involved, they are living humans	Non-human subject (the sun)

Typical data size	MB to GB	GB, occasionally TB	Potentially 100s of TB
File type and format	• Qualitative • Very often physical specimens e.g. manuscripts, buildings • Often images, video, text	• May be qualitative or quantitative • Digital content • Images, audio, video, text, spreadsheets	• Quantitative • Digital content • FITS files, source codes, binary files
Norms and community dynamics			
Theme	***History***	***Information Science***	***Solar Physics***
Team orientation	Solo projects	Solo or team projects	Mostly team projects
Sharing culture	• Most researchers, with time, develop their own personal data archives. • Little or no sharing requests; and little or no willingness to share • Sharing always with caution	• Depends personally upon the researcher's discretion and the sensitivity of the data. • More sharing requests and more willingness to share • Sometimes cautious when sharing	Little to no reservations about sharing; i.e. ground-based data and source code

Data sourcing & dissemination			
Theme	***History***	***Information Science***	***Solar Physics***
Source of data	Archives, libraries	Original data, personal contacts, or the internet (e.g. repositories)	The internet (e.g. central observatories), ground-based instruments, the computer, or personal contacts
Procedure of sourcing	Requesting for access and travelling to multiple physical locations; queuing for access; taking pictures; etc.	Conducting surveys, running computer simulation, downloading from online repositories or colleagues	Running computer simulations or downloading from solar observatories/repositories online. Ground-based data may be sourced from colleagues.
Cost to researcher	The extra financial implications of photocopying, printing, HD cameras, travel, etc. may be great. Additionally, some archives require subscriptions not always offered by the University.	Little to no extra financial implication. Usually comes in the shape of vouchers or coupons for survey participants	Little to no extra financial implication, except costs incurred by ground-based researchers for travel to instrument location for data collection
Method of collection	Manual collection	Manual or machine	Machine collection for space-based telescopes and some degree of manual collection for ground-based.
Degree of standardization	Manual tagging always necessary	Manual tagging often necessary	Pre-tagged data. Space-based data is well documented and standardized, ground-based data less so.

Sharing channels	One-on-one via email or HTTP only personal request. Repository data tends to be uploaded by institutions and not by researchers themselves.	One-on-one or in online repositories. Repository data tends to be uploaded by researchers themselves, usually at request of a journal or the project funder	Space-based data, and a lot of ground-based data openly available to everyone. Ground-based data on personal devices must need be shared via FTP due to size
Other personal habits, practices, and concerns			
Theme	***History***	***Information Science***	***Solar Physics***
Tagging, file naming, and metadata	Researchers tend to follow non-standard, ad hoc systems/methods	Researchers tend to follow non-standard, ad hoc systems or methods	Not necessary for space-based data, which comes pre-tagged and standardized. Ground-based comes with only some degree of tagging and standardization, in which case researchers may add their own tags, typically tending to follow own personal methods developed ad hoc
Use of repositories and/or online services	Rarely, and mostly for locating archives and collections or for trivial/secondary data	Regularly, but satisfaction with services commonly low	Regularly, with satisfaction moderate to high for space-based researchers; and moderate to low for ground-based researchers
Use of or familiarity with personal	Very few know about ORCIDS	All have heard of it, and many use it	Most have heard of it, and some use it

identifiers, such as ORCID			
Main causes of frustration with regards to data or repositories	• Incomplete/censored data in some archives • The tedious process of obtaining access • Variabilities in the keywords needed to search in different archives, due to differences in protocols between archives	• Finding the right data online • Uploading large files onto repositories (time consuming)	Waiting for data to transfer/download, because data are extremely large

4.3.1 Resource requirements

Information about resource requirements is crucial in system development. It helps not only in planning, but also in resource allocation and prioritization, especially where resources are limited, as is often the case. Depending on the predominant data type or format in a discipline, special computing resources may or may not be requisite for data storage, preservation and analyses. The present study shows, for example, that in History, data may be physical manuscripts or other artefacts, and not digital objects. Requirements for computing resources may, in such cases, be minimal. In a 2008 study conducted by the Research Information Network (RIN), few Humanities researchers (including Historians) were engaged in "highly collaborative, highly computationally demanding research". The RIN report further observed that, "although their work was highly complex and varied significantly from project to project, humanists were not as likely to use the state-of-the-art available technology". For other disciplinary domains, however, such as Solar Physics, it is not unusual to require High Performance Computing (HPC) resources.

4.3.2 Data attributes

The attributes of data, such as its size, type, or format determine, in no small measure, the features that the system may or should support for data discovery or presentation. Non-standard, propriety file formats (e.g. SPSS data and Excel spreadsheets) require external plug-ins to open in the browser and special software to be used when downloaded. Figshare, for example (see Section 4.1.6), has been noted as providing a preview feature for datasets, enabling them to be opened and viewed on the web browser. This is however only for datasets with open file formats, such as TXT or CSV. Very large data file sizes may also prove a stumbling block. Of the three disciplines studied, History shows the least use of specialized file formats; followed by Information Science. In Information Science, however, this largely depends upon the specific research area. It is interesting to note that the notion of what constitutes a large dataset varies between the disciplines here studied: History researchers seemed to regard data of a few megabytes to one gigabyte as “large”; whereas, for Information Science researchers, data must be at least a few gigabytes to terabytes to be regarded as such. On a similar note, the majority of the researchers, being questioned on the topic, reported that they found it useful to know the size of a dataset before download. The reason given by some was that they liked to be able to guess how long it might take to download the data, as they might not just then have time for a long wait. Others stated that the file size helped them to decide on the probable relevance of data, depending on whether or not it is likely to fall within that size range; for example, when searching for digitized images one expects that they should be at least a few megabytes in size, and, therefore, may ignore all search results within the kilobyte range.

4.3.3 Norms and community dynamics

Customs and norms exist within disciplines and tend to differentiate it from others. This leads to the formation of the particular disciplinary “culture”. Data sharing is a cultural element and varies between disciplines. Strong cultures of data sharing, for example, exist in molecular biology and ecology (Nelson 2009), while the reverse is the case for Chemistry (Velden & Lagoze, 2009) and History. According to Mannheimer et al. (2016), “if data repositories are established elements of the disciplinary research ecosystem, researchers are more likely to discover and reuse data from those repositories.” Collaboration and teamwork are another

cultural element in disciplines. As at the time of this study, for example, not one of the interviewees from History worked in a team: all were soloists. This contrasted strikingly with the interviewees from Solar Physics, all of whom were part of at least one team within or outside the university. Information Science showed a mix of both soloists and team workers. One of the more important findings of this section of the study, mainly because it confirms as well as justifies a key design feature of DataFinder (i.e. the linking of research data with publications) is that researchers do not always differentiate categorically between content (e.g. tabulated information) in data and in publications. The dividing line between what seems to constitute data and what seems simply to be information in an unusual format appears to be blurred particularly in disciplines such as History, where qualitative data is predominant. Many researchers hence expressed a preference for a repository that would incorporate both these types of resource in a unified manner, rather than enforcing a separation which is not clearly delineated in practice. Weller & Monroe-Gulick (2014), in a similar study to this, found Historians the “least likely to utilize data sets”, more than half of them reporting that they used no data sets. This may be due to the narrower meaning attached to the term “data” by, it appears, many History researchers, to whom the term included only “information that is tabulated or in a spreadsheet”. In reality, however, Humanities researchers (among them, Historians) engage with a “wide range of resources, from paper materials and microfilm to advanced digital resources”. (RIN, 2011).

4.3.4 Data sourcing & dissemination

In addition to the points noted in Table 4.4 pertaining to this, it was observed that most of the researchers from History have preconceived notions that the data they seek would not be available online. A possible explanation to this may be contained in a statement by Weller & Monroe-Gulick (2014), that “Historians find acquiring access to materials the most challenging”. From the present study it may seem to appear that believing this, they therefore make little or no attempt to seek data on research data repositories online. Perhaps, however, it may be because, as many of them commented, their research was unique and had not been done before. This may explain, also, why most of the researchers from Information Science preferred to collect new data for their research (see Section 4.2.1). However, it is true that Humanities scholars often search for “primary

sources, many of which may not be formally published". They use documents such as "diaries, wills, letters, and manuscripts and visual materials such as photographs, portraits, architectural drawings, and films, as well as other types of objects" (Palmer & Cragin, 2009). History researchers, more than researchers from other disciplines, evinced uncertainty as to whether their personal collections may be of much use to others outside of their own very specialized or niche area of research. On the whole, Solar Physics researchers tended the most to reuse datasets and to be open about releasing it on data repositories. But, across all the disciplines, private communication, seems the predominant data sharing method. According to Borgman (2015), this method "can be very effective because scholars can discuss the content, context, strengths, limitations, and applicability of a particular dataset to a phenomenon". Indeed, in view of the current limitations of research data repositories (see, for example, Section 2.6 in Chapter 2 and Table 4.2 in Section 4.1), some of the findings from this study seem to indicate that this might even be preferred by many researchers.

4.3.5 Other personal habits, practices, and concerns

A majority of the History interviewees admitted that they generally preferred to search paper however rather than online ones. The reason given by some was "mistrust" and "certainty" that online catalogues, when searched, do not return all the relevant information or data that they contained. Others gave the entirely different reason that they liked the "feeling" of physically handling ancient manuscripts, and of going through old collections and being pleasantly surprised by finding something unexpected. This finding is unique to History, and supports the remark of Case (1991), that Historians prefer working with original materials. Another peculiarity of the History domain gleaned from the study, and which is especially pertinent to research data repositories, is that many of the interviewees reported that professional help from the archivists in charge of the collections is often needed to search some important online catalogs. The reason given by them was that some of the needful search terms and necessary search procedures were not very simple to know or follow without professional training. Indeed, Weller & Monroe-Gulick (2014), reporting their study, noted a "need" to "help humanists become aware of, adopt and use new tools and digital resources for their research". On the subject of professional training, but quite apart from this, RDM skills training, including in metadata, are not usually part of graduate

courses, and this is especially true outside the information disciplines (see, for example, Section 4.2.3). Expertise in some research domain does not translate into expertise at data management, and few researchers have the requisite skill to document their data to archival standards (Borgman, 2015). The present study, as well as that reported in Section 4.2, confirms this statement; which continues to pose a major problem to RDM (see also point number 2 of Section 2.6 in Chapter 2). Furthermore, it is a double-edged problem in that, (1) unless researchers are trained in RDM skills they will not be able to fulfill their important role as the primary source of contextual metadata needed to make research dataset intelligible and reusable (see the guidelines for open data given in Table 1.1 of Chapter 1); and (2) it is clear, also, that unless there is some clear professional benefit to compensate for the time and the meticulous effort needed to tag data with standard metadata and to otherwise document it up required standards, researchers will continue to follow the ad-hoc methods that they currently use and to document data minimally if at all. Another important finding from this study, confirms a point that has already been previously noted in Table 4.2: that, researchers tend to *reuse* the datasets of other researchers that they trust (RIN, 2008). But, over and beyond this, researchers from both History and Information Science further indicated that they will *share* data only with those whom they trust.

4.3.6 Section Summary

The principal similarity that cuts across all three disciplines is the reluctance of researchers to do manual operations, such as tagging and annotation of datasets, unless these prove absolutely necessary; and even then, to follow intuitive and spontaneous (often arbitrarily developed) personal methods rather than standard schemas and/or conventions. The key findings combined from the questionnaire as well as the interview study are presented in Table 4.17.

Table 4.17. Combined findings from questionnaire survey and interviews.

No.	Point
1	Incomplete documentation or its lack altogether often prevents datasets of interest from being reusable
2	Many users are unskilled information seekers and are unsure as to what search terms to use to find data

3	Researchers commonly follow non-standard, ad hoc methods for tagging or annotating their data with metadata
4	Tools for creating metadata are found to be too hard to use, and very few researchers have received any degree of formal training on metadata or data management
5	Data file sizes in the megabyte range are the most commonly used and produced, closely followed by files in the gigabyte range. File sizes in the terabyte range are rare in most disciplines
6	Google is frequently used for data search, though often with unsatisfactory results
7	Most researchers create new primary data rather than reuse existing data. The main reason(s) given for this is lack of knowledge about or access to existing data
8	The process of obtaining access to data may be particularly tedious in some disciplines (e.g. History)
9	There is a general reluctance among researchers to upload data online before the maximum number of papers have been published on it
10	Many researchers think it useful to know the file sizes of datasets prior to download (see Section 4.3.2 below)
11	Standard office documents (e.g. text, spreadsheets) are the most common file formats used and produced by researchers. Next are images, structured scientific and statistical data, and web-based data (e.g. social media data)
12	Many researchers felt that some way of visualizing datasets would be useful in helping them understand and decide on the usefulness of data
13	Researchers are generally reluctant voluntarily to spend long hours tagging data to upload online, and are more likely to evade this unless it be a requirement

4.4 Technical experiment (comparison between DR and traditional IR)

Unlike publications and other text-based information resources, research data, are complex compound objects (Kouper et al., 2013). This study identifies some significant differences between research data and publications (text) which justify the necessity of developing technological solutions modelled around the peculiar needs and requirements of data. The steps described in Section 3.4 of the previous chapter for the purpose of this study resulted not only in the actual data (see Appendix I) demonstrating the relative differences with respect to file size, between research datasets and research publications (see Table 4.18), but also in a number of observations (see Table 4.19) that will serve in some way to inform system design. As Anderson (2004) remarks, although in the bit level there are no differences between digital information (published papers, reports, proceedings) and digital scientific and technical data, it is nonetheless "important to notice the differences that do matter now, and will continue to matter in the future". Overall, the important findings and observations from the study are as follows:

1. Average file size of retrieved datasets are several times larger than that of retrieved research publication files, and these in turn vary from one discipline to another. This observation is unsurprising and in itself hardly worthy of special remark; however, as other observations or consequences may hang upon or arise from it, a brief mention seemed pertinent. Anderson (2004) described size and volume differences between data and publications as being "first and foremost"; citing as example the United States Geological Survey (USGS) land remote sensing archive which holds 2 petabytes of data, with an additional 1-2 terabytes arriving daily, in contrast with only 17-20 terabytes which is estimated to represent the size of the entire print holdings of the Library of Congress, if digitized. Figure 4.24 shows, for each keyword in each discipline, what proportion of the total file size retrieved constitutes research datasets and research publications. It could be seen, as expected, that on average the file sizes of research datasets generally and significantly exceed those of research publications. In fact, in some cases (i.e. search behavior, face recognition, computer vision, 'renewable energy', and 'ultraviolet light') the whole graph appears to be composed entirely of research datasets, which is however

not really the case. The observation merely demonstrates the overwhelming disparity in average file size of research datasets and research publications (text) in those subjects. Table 4.18 represents this more accurately; it may there be observed in some cases that are by no means exceptional or unusual, that the average file size of a single research dataset exceeds that of a single research publication up to nine hundred-fold.

2. Datasets retrieved in the course of a single search may be of many different file types or formats, and sometimes for the same dataset. Over 20 file types and formats were noted in this experiment, notwithstanding the more or less homogenizing effect of the decision to wherever possible give preference to non-propriety formats (e.g. txt, CSV, tab-delimited) over propriety formats (e.g. STRATA, SPSS, XLS, MATLAB). Whereas, for publications, uniformity was observed in this respect, all of the research papers being in PDF format. As noted by Kennan & Markauskaite (2015) and confirmed in this study, research data are heterogeneous; taking "many forms depending on their origins, the research problem being addressed and the discipline of the researcher".
3. Whereas research publications comprise of only the publication itself, research datasets are almost always accompanied with separate documentation files (up to 22 have been noted in this experiment). Each piece of documentation furnishes further information about the dataset in question and may be necessary for its potential re-use. These documentation files tend be include code snippets, original survey questions, file descriptions, READ MEs, appendices, variable coding information, user guides, instructions, index files, consent forms, ethical clearance certificates, etc.
4. A single dataset item record may constitute several composite files (as many as 524 have been noted in this experiment) comprising fragments of the dataset broken up into smaller file sizes; or versions of the dataset at different stages of processing or under different conditions of observation. This is in contrast with research publications where a single item record comprises only one file representing a whole.
5. Unlike research publications which may be read online in abstract or full text form, research datasets often must be downloaded before they can be

read or used. Consequently, users end up downloading files without full knowledge of the contents or usefulness of the file. This unnecessary downloading of multitudes of datasets, often large in size, besides wastefully consuming valuable storage disk space and network resources, falsely spikes up download count, thereby rendering this metric unreliable as an indicator of data impact, usefulness, or popularity.

The above observations are summarized in Table 4.19 with remarks pertaining to their potential implications.

Table 4.18. Average sizes of files retrieved for research datasets and research publications.

Discipline	Keywords	Data Retrieval*	Text Retrieval*	Approx. ratio of text to data
Arts & Humanities	art museums	6.205 MB	0.820 MB	1:8
	nineteenth century	2.898 MB	1.042 MB	1:3
	“world war”	6.158 MB	0.508 MB	1:12
	medieval	5.158 MB	1.091 MB	1:5
	popular music	9.334 MB	1.000 MB	1:9
Social Sciences	unemployment	4.729 MB	0.455 MB	1:10
	cognition	13.340 MB	1.612 MB	1:8
	“labour law”	2.827 MB	0.410 MB	1:7
	“trade union”	15.939 MB	0.748 MB	1:21
	imprisonment	2.444 MB	0.503 MB	1:5
Computer & Information Science	search behavior	657.707 MB	0.731 MB	1:900
	face recognition	**1.394 GB**	1.535 MB	1:908
	computer vision	**1.339 GB**	2.782 MB	1:481
	research data sharing	1.574 MB	0.521 MB	1:3
	social media data	19.597 MB	1.078 MB	1:18
Natural Sciences	marine life	32.318 MB	1.491 MB	1:22
	“climate change”	2.808 MB	2.497 MB	1:1
	“renewable energy”	766.432 MB	3.606 MB	1:213
	“ultraviolet light”	496.745 MB	1.991 MB	1:250
	“oxidative phosphorlyation”	41.177 MB	1.895 MB	1:22

**Average File Size, inclusive of documentation*

***Average File Size*

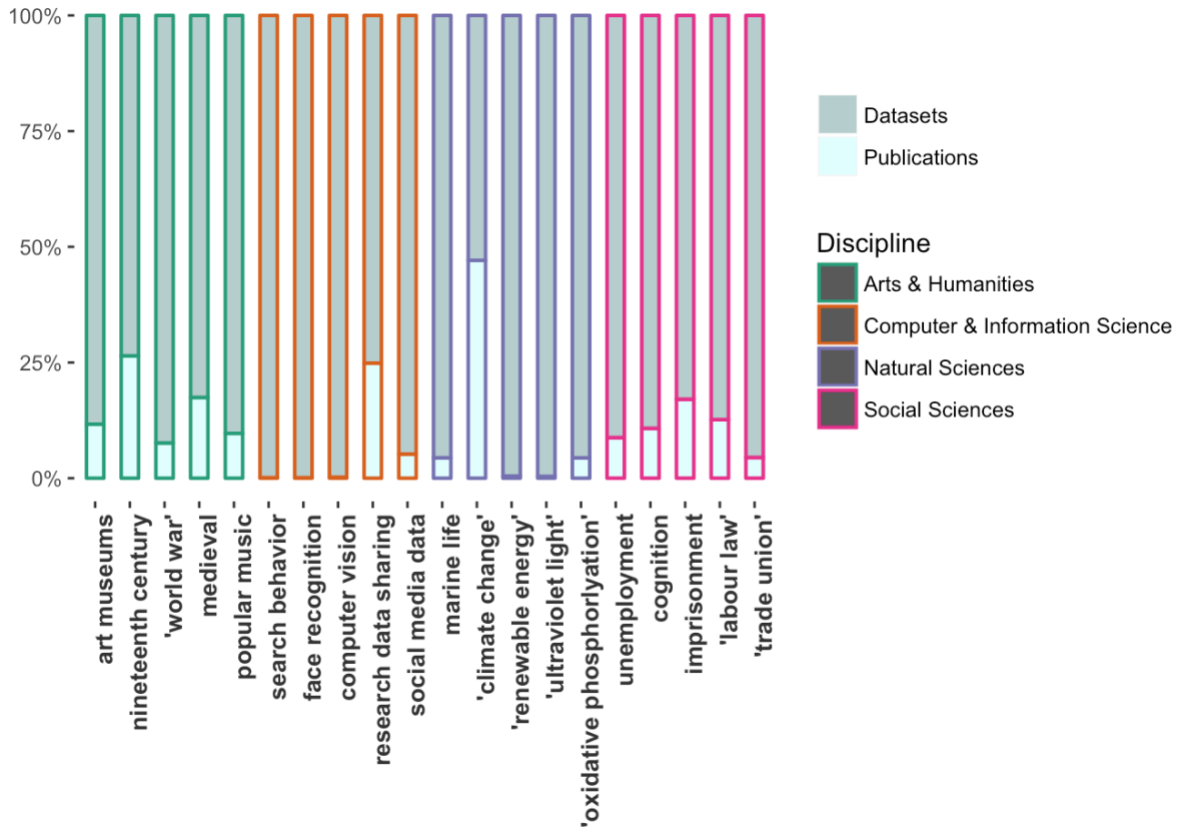

Fig. 4.18. The relative file size proportions for research datasets and research publications out of the overall total file size of all the files retrieved for each keyword.

Table 1.19. Summary of findings from technical experiment comparing DR and IR, with resource implications of each.

No.	Key findings	Implications
1	Text can be read online, while data usually requires downloading prior to being “read” or used. This ties to a previous finding in Section 4.1 (see the first issue highlighted in Table 4.2)	More network (in terms of bandwidth) and storage resources are required for data retrieval.
2	A single data item record may constitute several composite files (as many as 524 have been noted in this experiment)	A system, e.g. metadata schema, for efficiently identifying and linking associated files is imperative
3	Texts (research publications) usually come as a single, self-sufficient file. Data is nearly always accompanied with separate documentation files	A system, e.g. metadata schema, for efficiently identifying and linking associated files is imperative

No.	Key findings	Implications
4	Unlike texts (research publications), the same dataset may come in many different file types or formats	This places additional burden on computing resources (e.g. more storage is required for the same dataset) and also human resources (e.g. in terms of data preservation/curation requirements).
5	The average retrieved file size of datasets is typically several times larger than that of text (research publication).	More network (in terms of bandwidth) and storage resources are required for data retrieval.

4.4.1 Section Summary

The study just reported provides useful insights on data retrieval, and shows, practically, that research data generally have larger file sizes than publications, and are variable and heterogeneous, with file types and formats too innumerable to render feasible the development of a standard or uniform solution. Also, that they commonly come with a host of documentation files and metadata, or may come as multiple broken-up chunks of an originally larger file too large to be easily managed. Also, that whole datasets, unlike full research papers, can rarely be opened on web browsers, usually for reasons of size and software incompatibility.

4.5 Chapter Summary

The various analyses and discussions of findings from the four studies involved in the first phase of this research for the purpose of gathering user and system requirements have been presented in this chapter. The studies are (1) a market appraisal & review of currently available RDM systems; (2) an online questionnaire survey that garnered about 200 multidisciplinary respondents; (3) face-to-face interviews, each lasting about 30 minutes, with 18 researchers from three different disciplines; and (4) a technical experiment comparing between data retrieval and traditional information retrieval. The findings obtained from these served as input for the next stage of this research, i.e. Requirements Analyses, detailed in Chapters 5. This chapter concludes the first phase of the research, concerned with Information Gathering. The next chapter discusses the next

phase, of Prototype Design & Development. The just-concluded phase thus sets the stage for the remaining work. Figure 4.25, reprinted below from Chapter 1, shows the outline of the overall research and places the present chapter in context. The chapter addressed RQs 2 and 3, and adds to the answer to RQ 5 partially addressed in Chapter 2. It also supplies tentative answers to at least one part of RQ 1.

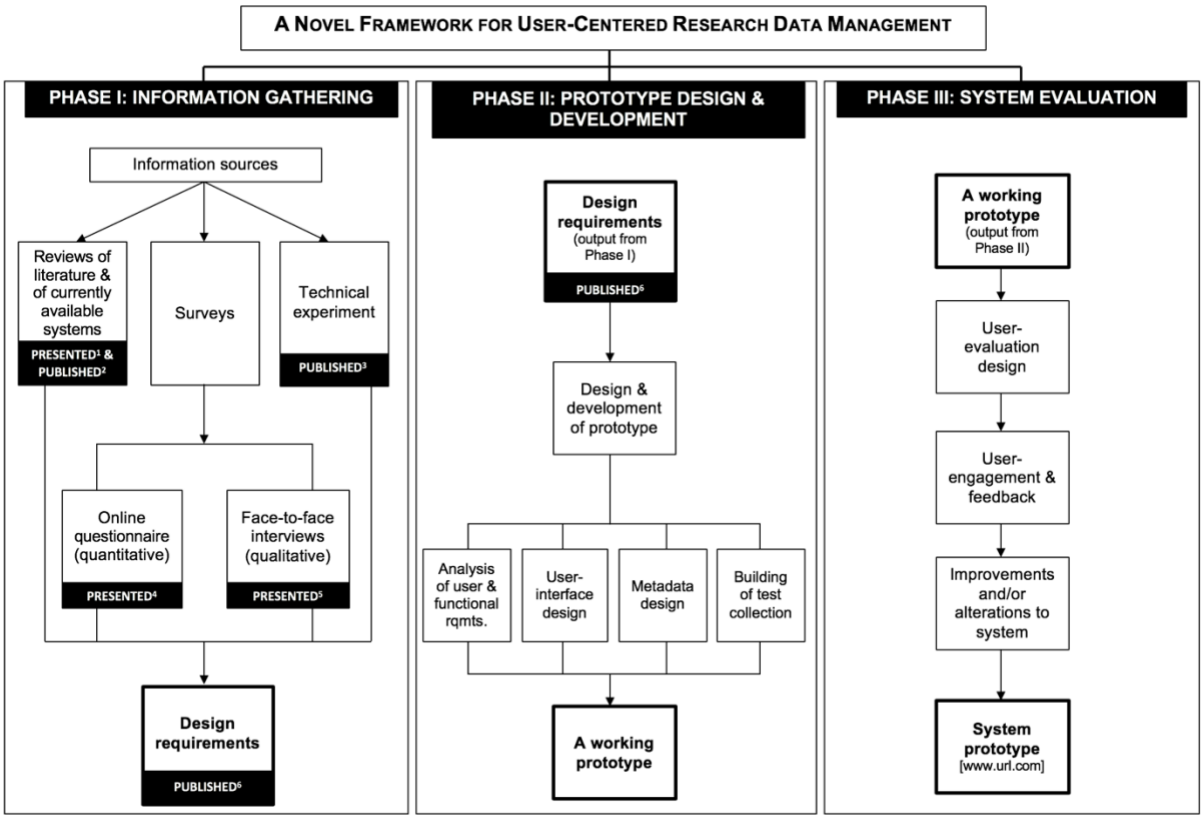

Figure 4.19: Overall outline of the research

5.0 REQUIREMENTS ANALYSES

According to Lazar (2006), the gathering, also called analysis, of user requirements is a “central” activity in user-centered design (p. 98). It forms one of the three key stages of the user-centered design process, the others being design and evaluation (Ames, 2001). Requirements analyses are “those system development activities that enable the understanding and specification of what the new system should accomplish” (Satzinger et al., 2016, p. 4). The present chapter commences Phase II (i.e. Prototype Design and Development –see Figure 5.1 below) of this research, which will be reported in this and the next chapter. The final deliverable of the phase, as of the overall research itself, is a working prototype of DataFinder, developed to meet the objectives of this research as set forth in Chapter 1. The specifications for the prototype will be derived through further analysis of the findings and information obtained in Phase I (i.e. Information Gathering).

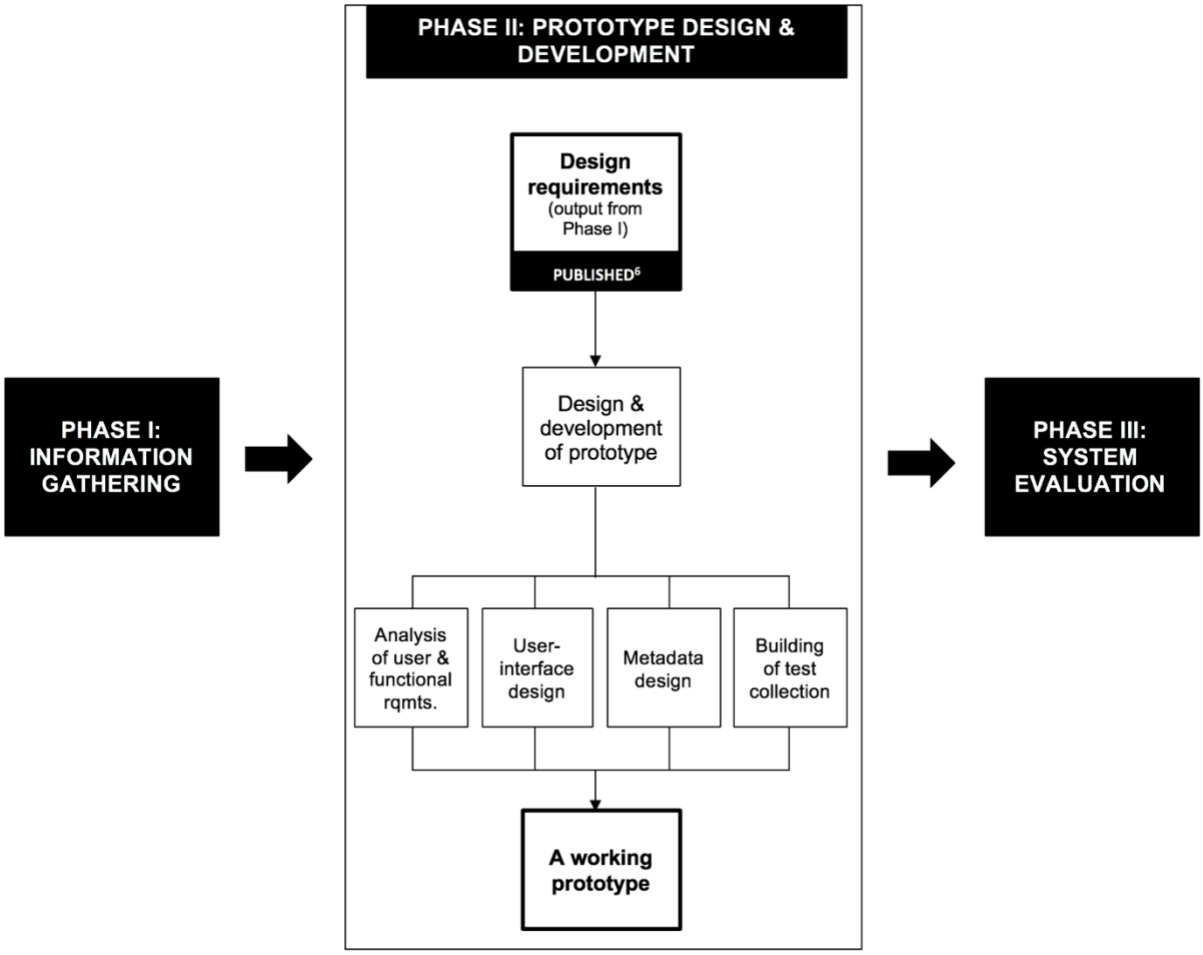

Figure 5.1. Overall outline of the research focusing on prototype design and development (Phase II).

The chapter furthermore reports on other requirements and considerations pertaining to the development of the proposed system; including, among others, metadata and persistent identification (see Figure 5.1 above). All discussions in the chapter will be headed under the two categories of system requirements: functional and non-functional requirements.

5.1 Functional requirements

A requirement is a statement of a service that the system should perform or support, or a constraint it should observe (Kotonya & Sommerville, 1998). Those requirements which relate to core system functions and are obtained from actual or potential system users are known as functional requirements (Maciaszek, 2007). Users of RDM systems may be grouped according to their roles (Bugaje and Chowdhury, 2017a) or their relationship to the system (Alsos & Svanæs, 2011). The latter grouping includes primary users (e.g. researchers), secondary users (e.g. funding bodies), and tertiary users (e.g. search engines). Role-wise, however, they act in the capacity of one or more of the following:

a. The data consumer, whose main activities include searching, browsing, and downloading content from the repository;
b. The data creator or holder, whose main activities include uploading content, setting up access permissions, and furnishing metadata information; and
c. The data administrator, who oversees various activities on the back end of the database, including, in some systems, checking user-uploaded datasets for errors to ensure that quality standards are met.

To achieve maximum system usability, not only is it necessary to identify the different user groups, as above, but also to understand thoroughly the distinctive characteristics of each, and their respective tasks on the system. Research shows that users' information needs, and by inference, data needs, tend to be ambiguous, not definitely articulated (Taylor, 2015), and often recognized only at sight (Morris, 1994). Other important points to note about users, before proceeding to analyze their more specific requirements, are that:

a. Their knowledge of systems may range from very naive to highly skilled and sophisticated (Morris, 1994). Systems must thus accordingly be

designed in such a way as to enable even the less sophisticated user efficiently to search for and find data; and

b. Their data-seeking needs may go beyond a simplistic search for datasets on a single topic (e.g. data on climate change). They may entail more complex conditions such as associative relationships (e.g. climate change data related to ozone depletion) or comparative relationships (e.g. climate change data in which ozone depletion is compared with rise in sea levels) involving multiple topics. The system should provide options to support users' more complex queries.

Before conducting the requirements analysis, it seems well first to consider those core functions which are commonly featured in all RDM systems, and thence to build on other features as required or make amendments as desired. The core features of RDM systems have already been previously encountered in Section 4.1; and include, in varying degrees of sophistication, facilities for:

1. *Uploading data*. Usually by filling out a form with particulars about the dataset in question, e.g. its creator(s), description, keywords, etc. The information collected here constitutes perhaps the most important contextual metadata about the dataset, since, as already mentioned in Section 1.1.9, data creators are its primary source (Borgman, 2011). A delicate tradeoff exists between requiring detailed metadata and more generic, discipline-agnostic metadata at upload. The former, although extra tasking to the user, helps to support more advanced system features. The latter, although considerably less tasking to the user, enables only basic features to be supported. A significant factor in this delicate equation is that researchers are unwilling to bear the burden of metadata creation (Greenberg et al., 2009, Borgman, 2015). Section 5.2.1 presents a review of the most common metadata fields required at data upload;
2. *Conducting searches*. This is generally by the use of keywords; although, occasionally, search by other metadata fields (e.g. creator name or DOI) are enabled. Section 4.1 discusses this in better detail.
3. *Browsing collection or repository contents.* This feature enables data discovery when keywords are unknown or when the data need is vague;
4. *Sorting of search results*. Refers to the order in which search results are arranged. The more common sorting criteria are date (recency) or

relevance or popularity. In general, this feature, although common, appears to be of secondary importance to the rest. The circumstance may, however, vary according to discipline; and

5. *Displaying metadata.* Data discovery is only the first step on the road to data reuse. When data is finally discovered, the information given about it helps the user to decide as to its usefulness for an intended purpose and aids in the actual act of reusing the dataset. Data repositories generally give at least some basic information about datasets, such as its size and file format.

The facility for refining search results, which, from the reviews reported in Section 4.1, may be considered non-core, being it is not commonly supported, is nonetheless highly desirable. This was evident from user responses to the interviews which subsequently followed. The feature was thus in some capacity implemented in DataFinder. The next system features and functionalities to be examined belong to this class: non-core, but generally desired or desirable. For better organization, the discussions are headed under broad themes representing the different issues or difficulties faced by data repository users as suggested in Phase I; namely:

1. Limited interactive features
2. Insufficient or unintelligible metadata
3. Quality of data not assured
4. Disciplinary requirements not met
5. Unacceptable time consumption
6. Legal and ethical concerns
7. Data discovery difficulties

Before proceeding, I venture to note here that, due to limitations of time and lack of a full team of software developers, only a subset of the listed requirements have been implemented in the final prototype. This is certainly an important issue, but not calculated to have significant detrimental effects on the final prototype which this research intends to demonstrate; since, from the outset, the intention was to develop a proof of concept and not a fully-fledged system. The proposed requirements for each of the above points will now be presented by turns, and concisely, since each point has already been considered in greater or lesser detail

elsewhere in this work. A summary is given in Table 5.1, constituting a final list of the functional requirements arranged in order of three levels of priority, viz. high, medium or less. A column in the table also relates each requirement to the UKRI guideline(s) or FAIR principle(s) that it lends towards conforming to (see Section 1.1.6); in addition to which, “System Usability”, being a core concept in user-centered design, has also been considered. The decision of placing a particular requirement in a particular priority level took into account the following factors: a) how many of the issues listed above it solves or contributes towards solving; b) the degree of simplicity or complexity to be expected in implementing it; and c) how many of the aforementioned principles or guidelines it is affiliated with.

5.1.1 Limited interactive features

According to Burgoon et al. (2000), interactivity in systems involves, among other things: a) a bidirectional sending and receiving of verbal and nonverbal messages and feedback, rather than one-way communication or passive participation; b) access to a wide variety of media, e.g. visual, audio, verbal, etc.; and c) synchronous or real time interaction. The present interactive limitations of RDM systems occasion quite serious drawbacks (Bugaje & Chowdhury, 2017b; Bugaje & Chowdhury, 2018a), as noted in Table 4.2 and in Section 4.4. Also, many among the researchers that took part in the qualitative study reported in Section 4.2 indicated an express wish for certain interactive features, such as visualization tools, in data repositories. All things considered therefore, this research proposes:

i. Features that enable the previewing datasets pre-download, such as is offered by Figshare (see Section 4.1.6); and
ii. Plugins for visualizing datasets and their relationships with other objects (e.g. associated publications).

5.1.2 Insufficient or unintelligible metadata

The importance of metadata for data sharing, discovery and reuse (Willis et al., 2012; Noorden, 2013; Arend et al., 2014; Wiley, 2014; Borgman, 2015; Walker & Keenan, 2015), as well as for the overall functioning and performance of RDM systems can hardly be over-emphasized. Problems arising from the lack or insufficiency of metadata distinctly and differently affect at least two among the groups of data repository users identified earlier in this chapter. In the one case,

it renders it difficult for data consumers to find, understand, and reuse datasets; and in the other case, it costs data creators considerable time and effort to adequately supplement the missing information, or to respond to inquiries from parties interested in the data. A practical solution to the conundrum may have been given in form of the following three principles proposed by Qin et al., (2012), as follows:

i. *The least effort principle*, of populating metadata fields automatically or semi-automatically using existing tools and databases; for example, linking the ORCID database to supply necessary information about data creators;
ii. *The infrastructure service principle*, of building new services upon existing infrastructure through metadata modeling of domains of interest and scientific contexts, among others; and
iii. *The portability principle*, of developing flexible, modular metadata schemas in blocks which can be variously merged or assembled together to meet particular needs.

This research further proposes:

iv. Use of associated research publications as supplements to existing metadata by them with data. This will serve the additional purpose of giving researchers the “full picture” of the research work in question. As previously noted in Section 4.3.3, the dividing line between data and information often is blurred; especially in disciplines, such as History, where qualitative data is predominant;
v. Simplifying of the amount of mandatory information required at upload, while also allowing (and even encouraging) data holders to provide as much additional information as they chose;
vi. Ensuring that there are clear statements about the relationship of each associated file(s) provided to the dataset in question (e.g. variable coding information, analysis source code, original survey questions, column header descriptions, etc. See Section 4.4);
vii. Presentation of datasets and their component parts (e.g. their split fragments, versions, and associated files) in a way that is easily comprehensible (see Section 4.4).

5.1.3 Quality of data not assured

In choosing to use others' data, researchers, according to Borgman (2015), put their reputations on the line. It is therefore "crucial" to be able to trust others' data. To satisfy researchers' possible skepticism concerning the quality of datasets found in data repositories, the following are proposed:

i. Adherence to quality standards, guidelines and recommendations as set by the approved authorities, such as CoreTrustSeal[1], and obtaining the certifications thereof. These are to be displayed prominently on repository websites;
ii. Linking of research publications to data, which will show researchers that data are of sufficient quality to produce peer-reviewed publications; and
iii. Establishing of quality control measures to rectify errors and to ensure that user-uploaded datasets meet minimum requirements.

Quality assurance of research data is imperative especially in cross-disciplinary reuse contexts, where the data consumer, to go by the roles earlier identified, may lack the required expertise to properly evaluate data for an intended purpose (Dallmeier-Tiessen et al., 2014).

5.1.4 Disciplinary requirements not met

Data repositories with less disciplinary focus may yet better accommodate disciplinary idiosyncrasies by adding simple options to core features. This research proposes:

i. Search options that allow History researchers to specify multiple date restrictions at once; i.e., to be able, instead of searching within one wide range (e.g. 1893 – 1910), to break it down into more specific dates and date ranges that are relevant (e.g. 1893 – 1898, 1900, 1904 – 1910);
ii. A feature for imposing embargo periods on datasets, particularly for researchers in the Social and Medical Sciences, to enable them to comply with funder requirements of Open Access without losing scientific edge;
iii. Browsing criteria that reflect other attributes of data than its subject domain. This will facilitate data discovery in fields where the nature of the

[1] https://www.coretrustseal.org

content of the data, rather than the content itself, is the deciding factor; for example, Machine Learning researchers looking for categorical data to train models and algorithms.

5.1.5 Unacceptable time consumption

This is partially addressed by the recommendations in point 5.1.2 above, in addition to which the following are proposed:

i. Simplifying, as well as minimizing, of the number of steps entailed for data upload, access, and search;
ii. Keeping of researchers informed as to their progress in the above, by displaying progress messages or bars; such as, “Step 1 of 3” or “45% completed”, etc.;
iii. Clear displaying of the sizes of research datasets, to assist researchers in estimating the time that might be needed for download; and
iv. Phrasing of terms and conditions of access in clear, unambiguous, and intelligible language.

5.1.6 Legal and ethical concerns

This pertains to researchers’ worries about the ethical use of their data by others, and to their compliance with university and funder requirements. The present research opines that by showing data creators what was published with their datasets and by whom, the linking of research datasets to publications will help to reassure them as to the ethical use of their data. Furthermore, data repository developers should understand the requirements above referred to, in order to help researchers to meet them.

5.1.7 Data discovery difficulties

Data discovery is a necessary preliminary step to data reuse. Insight into the practices and perceptions of researchers, acquired from Phase I of this work, has uncovered obstacles as well as opportunities for leveraging the same to facilitate data discovery. The following are proposed:

i. Use of interoperable standards, and optimizing repositories to enable indexing by external and general-purpose search engines such as Google;
ii. Dictionary look-up to assist researchers in choosing the right keywords;

iii. Facility to enable browsing of repository content by other criteria than by subject domain alone (see point 5.1.4 above);
iv. Options for search by various metadata fields or by multiple fields at once; and
v. Integration of a user account creation and management system that supports the saving of user preferences, so that users may optionally be notified about the availability of new data of probable interest; and
vi. Linking of research datasets to research publications.

Table 5.1. Summarized list of user requirements.

No.	Requirement	Section	Guideline/Principle
	Priority level – High		
1	Simple and minimal number of steps for data upload, access, and search	5.1.5	Findability, Accessibility, System Usability
2	Clear display of data file size	5.1.5	Intelligibility, Assessability, System Usability
3	Clear phrasing of access terms and conditions	5.1.5	Accessibility, Assessability, System Usability
4	Additional criteria for browsing the repository, e.g. based on data attributes	5.1.4 and 5.1.7	Findability, System Usability
5	Link research publications to data	5.1.2, 5.1.3, 5.1.6 and 5.1.7	Findability, Intelligibility, Assessability, Reusability
6	Simple metadata template for upload, with option for providing more elaborate information/metadata	5.1.2	Findability, Intelligibility, Assessability, Reusability, System Usability
7	Clear statement(s) of relationship between data and associated file(s)	5.1.2	Intelligibility, Assessability, Reusability, System Usability
	Priority level – Medium		
8	Search options enabling multiple date restrictions at once	5.1.4	Findability, System Usability
9	Quality assurance certification(s)	5.1.3	Assessability

No.	Requirement	Section	Guideline/Principle
10	Quality control measures	5.1.3	Intelligibility, Assessability, Reusability
11	Clear presentation of datasets and their component parts	5.1.2	Intelligibility, Assessability, Reusability
12	The least effort principle (Qin et al., (2012)	5.1.2	Interoperability, System Usability
13	The infrastructure service principle (Qin et al., (2012)	5.1.2	Intelligibility, Assessability, Reusability, Interoperability
14	The portability principle (Qin et al., (2012)	5.1.2	Intelligibility, Assessability, Interoperability
15	Dataset preview feature	5.1.1	Intelligibility, Assessability, Reusability, System Usability
16	Generally being mindful of university and funder requirements in system development	5.1.6	Accessibility, System Usability
17	Use of interoperable standards, and optimizing repositories to enable indexing by external and general-purpose search engines	5.1.7	Findability, Interoperability
18	Options for search by various metadata fields or by multiple fields at once	5.1.7	Findability, System Usability
	Priority level – Less		
19	Display of progress messages or bars, for multistep operations	5.1.5	System Usability
20	An embargo imposition feature	5.1.4	Accessibility, System Usability
21	Data visualization plugins	5.1.1	Findability, Intelligibility, Assessability, Reusability, System Usability
22	Dictionary look-up service	5.1.7	Findability, System Usability

No.	Requirement	Section	Guideline/Principle
23	Integration of a user account creation and management system	5.1.7	Findability, System Usability

5.1.8 Section Summary

Findings and information from preceding chapters were analyzed in this section to produce a practical list of user requirements to solve key user-related issues of RDM systems. The list, along with the non-functional system requirements to be discussed in the coming section, will form the basis for the next stage in the development of the proposed prototype of an RDM system.

5.2 Non-functional requirements

Non-functional requirements relate to system characteristics (Maciaszek, 2007; Satzinger et al., 2016) and comprise, among other things, the technical elements and practical mechanisms that drive or enable the fulfilment of user requirements. For this research, they include the following considerations:

1. The user interface
2. Building of a test collection
3. Metadata
4. Persistent identification
5. Ontological schemas

The above by no means constitutes a comprehensive, nor even an all-essential list. Nonetheless, it represents key requirements for an RDM system such as this research proposes. The user interface and the building of a test collection involve front-end and database designs respectively; both will therefore be deferred for treatment in Chapter 6 (System Design). Only Metadata, Persistent Identification and Ontological Schemas will thus be discussed in this section. As it exceeds the scope of this research to develop a full-blown RDM system, and as the research focus is primary on the user and the user's requirements, the discussions on some of these, particularly the last named, will not be exhaustive. The reason for this, besides the two just given, is that implementing such in the intended prototype will involve a great quantity of elaborate and extraneous research work.

5.2.1 Metadata

An important question that arises in building any data repository is who should create the metadata (White, 2014). Metadata is an essential and vital component not only of RDM systems, but of the entire RDM ecosystem. It is an indispensable, and sometimes sole, driver of all 6 of the principles and guideline of RDM (see Section 1.1.6), namely:

i. *Discoverability.* Metadata includes unique identifiers, such as DOIs or URIs (see Section 5.2.2), using which data may be located directly, as well as indirectly through search engines or linked graphs. Keyword search and data directory services are also made possible through indexing metadata fields;
ii. *Accessibility.* Information about data access rights, terms, and conditions is contained in metadata, which accordingly advises users on such points. The system also utilizes the same metadata in making decisions about granting access to particular users or machine clients;
iii. *Intelligibility*. Contextual metadata helps to decode and generally to understand and make sense of obscure words, symbols and variables that data might contain;
iv. *Assessability*. Before researchers can reuse a dataset, they must first be able to determine whether it is suitable for their intended purpose. To do this, information given in the metadata for that dataset is required;
v. *Reusability*. Represents the ultimate goal of data sharing, preservation, and management. As has been noted throughout this work, contextual, as well as other metadata are the sole supplier of the needful information for repurposing or reusing data;
vi. *Interoperability*. Using standard, widely used metadata formats and schemas in data repository development, and in documenting research data, enables integration and inter-communication with other systems and services.

There are three main types of metadata: descriptive, structural, and administrative metadata (NISO, 2004). Each documents a different of kind information and has different purposes and perhaps subtypes. Researchers prefer descriptive metadata, although they usually are unwilling to create it themselves (Greenberg et al., 2009). Metadata may be created by information professionals, data

creators, or through automatic indexing (White, 2014). For each of these means there are advantages and disadvantages, as well as strengths and weaknesses. A detailed consideration of these, however, is not pertinent to the immediate purpose of this work. For the prototype of the system being designed it is necessary to adopt a metadata standard or set of elements. Dublin Core has 15 elements and is the standard generally used for repositories that hold publications and sometimes even for data repositories (Gómez et al., 2016). However, it was deemed more practical to conduct a small survey to find out what metadata elements were more commonly used by existing data repositories. In pursuit of this, metadata elements were collated from over 18 different data upload templates including those of Figshare, UK Data Archive[2], and the Universities of Edinburgh, Leeds, and Southampton, among others. Two recommendations in the literature, by Rumsey & Jefferies (2013) and Weibel (2005), were also considered. The most common elements from all these sources, numbering a total of 11, formed of 4 mandatory and 7 optional elements, are presented in Table 5.2 below. These elements represent the average number and detail used in the other templates, and are nearly identical, though less in number, to the Dublin Core elements. They were thus judged as being suitable for use in the proposed system.

Table 5.2. Metadata elements to be used for the proposed system.

Mandatory elements	**Optional elements**
Data Title	Funder
Depositor Name	Data License Type (e.g. creative commons)
Data Publisher	Date
Discipline	Keywords
	Data Description
	Related Publication URI (or URL)
	Publication Title

5.2.2 Persistent identification

An identifier is “a sequence of characters that identifies an entity” (McMurry et al., 2017). A persistent identifier identifies digital objects and, according to Hakala (2010) meets the following conditions:

[2] https://data-archive.ac.uk

i. It is assigned only to resources meant for long-term preservation; and
ii. It persists for at least as long as the lifetime of the resource that it identifies.

Further requirements about identifiers, specifically relating to data repositories, as presented by (Grethe, 2015) are that they must be:

iii. Stable;
iv. Unique within repository; and
v. Resolvable, i.e. to a landing page on the repository.

Digital resources may be migrated multiple times or undergo multiple versions and older versions may no longer be accessible or usable. On the internet, persistent identification allows redirection to the latest available version of a resource even when the identifier of an older version of it is used (Hakala, 2010). This prevents “link rot”, or the eventual decay of cited works with time (Pepe et al., 2014; Klein et al., 2014; McMurry et al., 2017). Digital Object Identifiers (DOIs) are persistent identifiers commonly used for purposes of data access and citation of datasets (Simons, 2012). Research data retrieval and discovery is least complicated when the user knows in advance the dataset identifier, such as its DOI. As this is rarely the case, however, developing efficient data discovery techniques continue to be an important area in RDM. For the prototype being developed it was not essential to conform the above listed conditions, since the system would not be a live one and was instead to be used in a very controlled environment. Nevertheless, the real, outer-world DOIs of the test collection datasets and publications were used as primary and foreign keys for unique identification and for linking.

5.2.3 Ontological schemas

An ontology is a “formal model that uses mathematical logic to clarify and define concepts and relationships within a domain of interest” (Madin et al., 2008). It provides a means for building a shared understanding of data, services, relationships, and processes; and can be used to semantically integrate databases (Elzein et al., 2018). Building a linked data application entails mapping between the application model and the underlying ontology of the source dataset. To do this, a comprehensive understanding of the schemas (usually RDF ontologies) underlying the source and target datasets is requisite (Araujo et al.,

2010). RDF is a machine-readable, Semantic Web standard for publishing or exchanging data (essentially metadata) between systems or services on the web. It uses “triples” composed of *subject*, *predicate* (i.e. relationship) and *object* to define facts and relationships between entities, usually in the form: an object *o* has a relationship *p* with subject *s*” (Elzein et al., 2018). In the case of the system prototype being developed, linked data was implemented in a very simplified fashion by adding foreign keys to connect the data and publication tables in the database. Although more advanced relationships can be modelled and elegantly designed using RDF, the kind and amount of work that such would involve is out of research scope. Publishing semantically annotated research metadata helps to improve data quality, information diversity, and knowledge integration (Dimou et al., 2014). Also, linked open data offers new methods of analysis to monitor and evaluate research activity in a larger scale (Dimou et al., 2014).

5.2.4 Section summary

This section contains discussions about key technical requirements, i.e. Metadata, Persistent Identification, and Ontological Schemas, for the design and final construction of the working prototype. Other important system elements, specifically, user interface design and database design (or test collection) were judged more appropriate for the next chapter, which is on design.

5.3 Chapter summary

This chapter has built upon work from preceding chapters to derive a long list of functional (user) requirements. Other requirements of a technical nature have also been considered at lengths sufficient for the purpose of this research. The two sets of output represent, with a few exceptions that will be treated in the next chapter, the full set of system requirements for the prototype described in Chapter 1. The next chapter represents the final stage in the development of the prototype, which will be based on these requirements. Altogether, the analyses and discussions above address to a full extent RQs 1, 4, and 8; and complete the answer to RQ 5 partially addressed in Chapters 2 and 4. RQ 7 has also partially been answered.

6.0 SYSTEM DESIGN

System design signifies activities that enable the detailed description of how the resulting system will actually be implemented (Satzinger et al., 2016, p. 5). It provides foundation for system development and is itself founded upon system requirements, such as have been analyzed and identified in the last chapter. Hence, the present chapter builds on the last and culminates in the development and presentation of the RDM system prototype proposed and described in Chapter 1. The designs of the user interface and the test collection, both of which have been deferred for discussion from the last chapter, are treated. Other important aspects of the system's development, notably user-centered design, is also be considered. The vital significance of the concept of user-centered design to this work, and also its elemental role in the overall composition of the same has been earlier established in Chapter 1. Certainly, user-centered design has influenced much of what has already been reported in preceding chapters, particularly the last two, involving user participation and user requirements analyses. A formal treatment of the subject, however, is also, for the first time, presented here. User-centered design, quite apart from its particular application to any use cases, remains an extensively and currently discussed topic especially in IT, Software Engineering, and allied fields. The discourse to follow will therefore be indicative, rather than exhaustive; and will be curtailed accordingly as is considered pertinent to the overall purpose of the research. The bulk and main substance of it and the last chapter have also been published in Bugaje and Chowdhury (2018c).

6.1 User-centered design

RDM system users, as do system users generally, interact with systems for different intents; with varying degrees of engagement, skill, and ability; and from unique personal and disciplinary backgrounds. These variables, which for each user are peculiarly combined, influence the overall user experience and satisfaction with the system. System usability is a term sometimes used interchangeably with user-centeredness; however, it is of a distinctly different meaning and application. Usability, according to ISO ISO9241 (1998), refers to the "effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments". User-centered design, on the other hand, is a design approach; guided by a philosophy and a set of principles,

techniques, and recommendations (Bowler et al., 2011) to help develop usable systems. Usability is thus only a desired end or outcome of which user-centered design is the means. Another point of distinction is that usability forms a measure of user-centeredness in system design. The techniques above alluded to, of user-centered design, commonly entail a process of iterative prototyping; with user evaluation at the termination of each iteration (Gould & Lewis, 1985; Constantine, 2004; Uflacker & Zeier, 2008; Bowler et al., 2011; Martinez-Alcala et al., 2014). As users thus uncover problems with these intermediary versions, designers correct the problems and test again until a final usable version of the system is attained (Corry et al., 1997). This enables not only that new requirements are discovered in the process, but also that the requirements increasingly match the needs of users under real conditions of use (Bowler et al., 2011). It is important to note, though, that user-centered design does not replace, but complements, traditional system development processes and approaches (ISO, 1999; Zimmermann & Grötzbach, 2007). The user-centered design process has been discussed more extensively in Chapter 2 (see Section 2.5.1), and with examples. This chapter will therefore refer back to that section avoid redundant repetition.

6.2 Prototype design

User-centered design appears, from the above discussion, to be specifically relevant to RDM system development, particularly given the existing issues of RDM systems as has been noted throughout this work (e.g. see Section 5.1). Researchers' data-seeking needs and behavior, unlike those of traditional information-seeking, have yet to be fully understood and formally modeled. Until such a state of the case is attained, data repositories will but imperfectly serve users, the data consumer particularly (see again Section 5.1). However, practical research and exercises in user-centered RDM system design, such as the present work happens to be, can help to uncover useful information to help formalize and develop appropriate models of data-seeking and data-reuse behavior of researchers. Due to scope limitations, only a small subset of the user requirements identified in Chapter 5 are implemented in the prototype. The table summarizing the requirements is reproduced in Table 6.1, with an additional column indicating which of the features have been implemented in the prototype and to what degree. Following this are the discussions on the design of the user interface and test collection.

Table 6.1. Summarized list of user requirements for implementation in prototype.

No.	Requirement	Section	Remark on Prototype
	Priority level – High		
1	Simple and minimal number of steps for data upload, access, and search	5.1.5	Fully implemented (where applicable)
2	Clear display of data file size	5.1.5	Fully implemented
3	Clear phrasing of access terms and conditions	5.1.5	Not applicable (since all are test data contained in local a MySQL database)
4	Additional criteria for browsing the repository, e.g. based on data attributes	5.1.4 and 5.1.7	Not implemented (complexity involved is out of research scope)
5	Link research publications to data	5.1.2, 5.1.3, 5.1.6 and 5.1.7	Simple implementation
6	Simple metadata template for upload, with option for providing more elaborate information/metadata	5.1.2	Partial implementation (all data in local MySQL database manually tagged according to Table 5.2)
7	Clear statement(s) of relationship between data and associated file(s)	5.1.2	Fully implemented
	Priority level – Medium		
8	Search options enabling multiple date restrictions at once	5.1.4	Not implemented (complexity involved is out of research scope)
9	Quality assurance certification(s)	5.1.3	Not applicable
10	Quality control measures	5.1.3	Not applicable
11	Clear presentation of datasets and their component parts	5.1.2	Fully implemented
12	The least effort principle (Qin et al., (2012)	5.1.2	Currently not implementable
13	The infrastructure service principle (Qin et al., (2012)	5.1.2	Currently not implementable
14	The portability principle (Qin et al., (2012)	5.1.2	Currently not implementable

15	Dataset preview feature	5.1.1	Not implemented (complexity involved is out of research scope)
16	Generally being mindful of university and funder requirements in system development	5.1.6	Partial implementation, by using standard uniform identifiers for datasets, researchers, and publications
17	Use of interoperable standards, and optimizing repositories to enable indexing by external and general-purpose search engines	5.1.7	Partial implementation, through using standard uniform identifiers for all entities in the database would enable the functionality in a live implementation
18	Options for search by various metadata fields or by multiple fields at once	5.1.7	Implemented
	Priority level – Less		
19	Display of progress messages or bars, for multistep operations	5.1.5	Not implemented (complexity involved is out of research scope)
20	An embargo imposition feature	5.1.4	Not applicable
21	Data visualization plugins	5.1.1	Not implemented (complexity involved is out of research scope)
22	Dictionary look-up service	5.1.7	Not implemented (complexity involved is out of research scope)
23	Integration of a user account creation and management system	5.1.7	Not implemented (complexity involved is out of research scope)

6.2.1 User-interface design

The “set of inputs and outputs that the user interacts with” to invoke the functions of a system comprise its user interface (Satzinger et al., 2016, p. 219). User-interface design is an important component of user-centered design, as all

interaction between the user and the system takes place through the user-interface. As a matter fact, Satzinger et al. (2016, p. 220) describe user-centered design as embodying “the view that the user-interface appears to be the entire system”, since “to the user of a system, the user interface is the system” (p. 165). User-interface design specifies “the logical model and physical properties of the system” or, simply, its “look and feel”. It includes style guides, system behavior and interactivity and, ideally, states and presets for concrete system screens (Zimmermann & Grötzbach, 2007). The same source highlights the following sub-types of user-interface requirements:

i. *Information architecture and information flow requirements.* These define the overarching logical structure of the user interface.
ii. *Presentation requirements.* These specify where the layout of an entire component or a single element is defined, e.g. widget boxes, screens.
iii. *User-system-interaction requirements.* These define the behavior of the user-interface elements, e.g. status changes.
iv. *Compound requirements.* These specify the interaction between more than one element, such as in the case of drag and drop functionality.
v. *Message requirements.* These define when and how system generates notifications, e.g. errors, alerts.

More concisely, however, Ames (2001) notes the following user-interface components as affecting usability, and therefore user-centeredness: 1) Visual design; 2) Information architecture and design; 3) Interaction design; and 3) Algorithm design. User interface design is closely interconnected with considerations generally involving Human Computer Interaction (HCI). Indeed, HCI is defined by Satzinger et al. (2016, p. 221) as “a field of study concerned with the efficiency and effectiveness of user interface vis-à-vis computer systems, human-oriented input and output technology, and psychological aspects of user interfaces.” In designing and developing the user interface of the prototype presently in question, reference was made to the above as well as other guidelines in the literature, such Satzinger et al. (2012, ch. 7) and Johnson (2007). The former recommends the following, all of which I have considered in designing my interface:

i. Consistency of design across the system. DataFinder consistently uses the same set of fonts, colors, and other design elements across all of its pages (refer to the screen captures in Section 6.3.8);
ii. Shortcuts. DataFinder provides a shortcut for users to correct mistakes in search parameters without going back to the previous page or beginning anew (see left pane of Figure 6.6);
iii. Feedback. DataFinder, for example, shows users clearly the parameters they had used for the current search and provides an option to modify the parameters (see the left pane Figure 6.6). Also, when a required search field is left blank the user is informed of the specific field concerned and asked to rectify the issue;
iv. Dialogues that yield closure. DataFinder seeks to minimize as much as possible the number of screens that the user would need to pass through to complete a search task, and makes it obvious when there are more screens ahead (refer to Section 6.3.2);
v. Error handling. DataFinder has been designed to handle common user input and navigation errors;
vi. Easy reversal of actions. DataFinder preserves session variables to enable the user to undo recent actions or begin anew; and
vii. Reducing short-term memory load. The use of session variables enables DataFinder to remember short-term information about user activity, saving the user the trouble of keeping track of everything.

6.2.2 Test Collection

The test collection simply constitutes the set of research datasets and publications that were used to develop the prototype and test it with real users. The set comprised approximately 150 open research datasets mainly from the IT and Computing fields, but with a few from the Social Sciences too. This was because of the intention to conduct the user evaluation with researchers in these fields. Neither was the choice arbitrary, for it was clear from the findings in Phase I (see especially Sections 4.2 and 4.3) that such would be the most suitable course to adopt, particularly for reasons of availability of test data, skill of test users, and my own better familiarity with the domain. The test datasets, coupled with at least one associated publication, were downloaded from open data repositories, along with any such documentation(s) as might accompany them. The associated research

publications were likewise all open access and were able to be freely downloaded without a legal breach. Each research dataset and publication were carefully tagged with the metadata elements listed in Table 5.2, and the whole were populated into a MySQL database. There, as described in Section 5.2.3, a basic implementation of linked data was contrived by adding foreign keys to connect the data and publication tables in the database. The database schema is represented in Figure 6.1.

6.3 Prototype development

The development of the prototype followed the standard System Development Lifecycle (SDLC) approach, based on the shortlisted version of the user requirements, as indicated in Table 6.1. The SDLC involves a series of sequential steps, viz. Project Initiation, Project Planning, Analyses, Design, Implementation, Deployment (Curtis & Cobham, 2005, p. 412); and it mirrors the natural course taken by the present research. Only one iteration of an alpha version was developed for user evaluation. An alpha version is “a test version that is incomplete but ready for some level of rigorous integration or usability testing” (Satzinger et al., 2016, p. 468). Screen captures of that and the operation of its implemented features are demonstrated in the next section. Figure 6.1 below illustrates the relationships and interplay between the system requirements (functional and nonfunctional requirements) identified in Chapter 5, and how they unify into a whole. Ontological schemas, though part of the functional requirements theoretically identified (refer to Section 5.2), were omitted from the diagram since, as stated with reasons in Section 5.2.3, they have not been implemented in the prototype.

6.3.1 Presentation of prototype

Table 6.2 below contains the shortlist of the system features, the same which will shortly be presented in this section. The system will also be reviewed based on similar criteria as the market appraisal study described in Section 3.1 and presented in Section 4.1. As only a small subset of the full requirements have been implemented, the main limitations of the study have been described in Section 6.3.9.

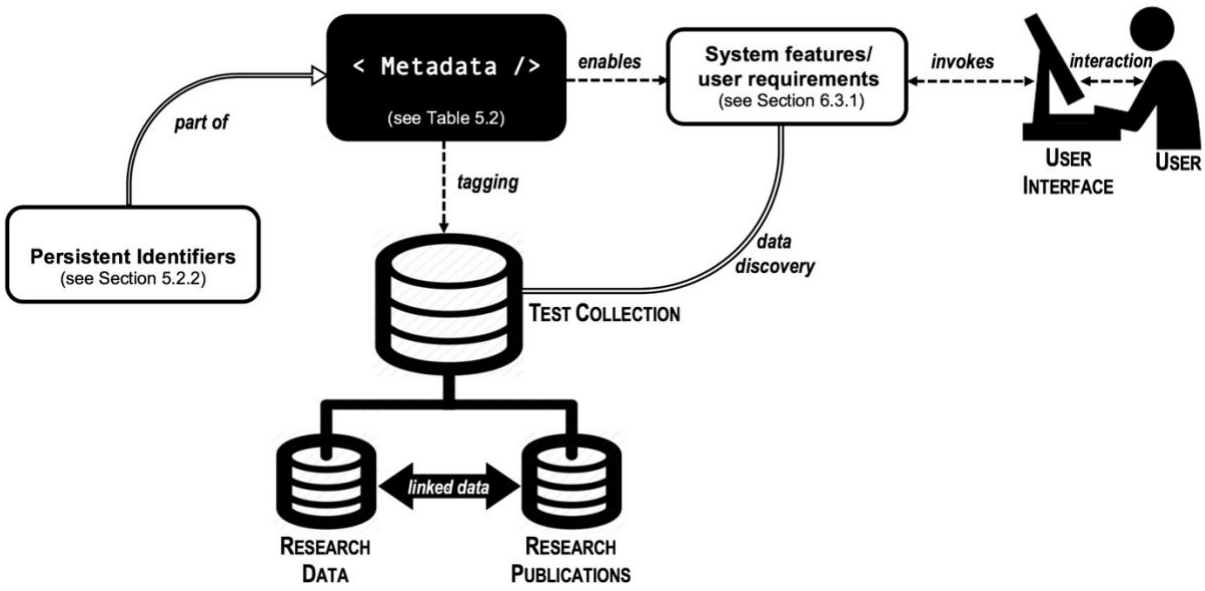

Figure 6.1. Schematic diagram showing the interconnection between the functional (see Section 5.1) and nonfunctional requirements (see Section 5.2) of the system.

Table 6.2. Summarized list of user requirements for implementation in prototype.

No.	Requirement	Presented in	Remark on Prototype
	Priority level – High		
1	Simple and minimal number of steps for data upload, access, and search	6.3.2	Fully implemented (where applicable)
2	Clear display of data file size	6.3.3	Fully implemented
3	Link research publications to data	6.3.4	Simple implementation
4	Clear statement(s) of relationship between data and associated file(s)	6.3.5	Fully implemented
	Priority level – Medium		
5	Clear presentation of datasets and their component parts	6.3.6	Fully implemented
6	Options for search by various metadata fields or by multiple fields at once	6.3.7	Simple mplemented

6.3.2 Simple and minimal number of steps for data upload, access, and search

This feature, described in Section 5.1.5, is applicable in the prototype only to the extent that it concerns data search. In respect of data upload there already exists

a carefully selected and tagged collection of test datasets in the database (see Section 6.2.2); while, in that of data access, the system is not deployed in a live environment to be subject to access rules. The following represent the main particulars of the feature as regards data search:

1. The system involves a maximum number of three different screens from the beginning to the end of any search, whether basic or advanced or for data or for publications. These are:
    a. The search interface page (see Figure 6.4);
    b. The search results display page, with a “quick view” option for a brief further look at the main details about each search result, such as its full description and available documentation (see Figure 6.6); and
    c. The dedicated landing page of the chosen dataset (see Figure 6.7).
2. The basic and advanced modes of searching are, for both data and publications, conducted on the same screen. Search parameters and options are enabled or disabled based on user selection (see Figure 6.5);

6.3.3 Clear display of data file size

Figure 6.3 below shows the structure by which search results are displayed in the system. The following details are easily discernable:

1. The title or name of the dataset;
2. A brief description of the data;
3. The file type of the data;
4. The file size of the data;
5. Number of times the data was downloaded;
6. Link to associated research publication (or link to the latest publication if data has more than one publication); and
7. The date on which the dataset was uploaded. See Section 5.1.5 for further detail.

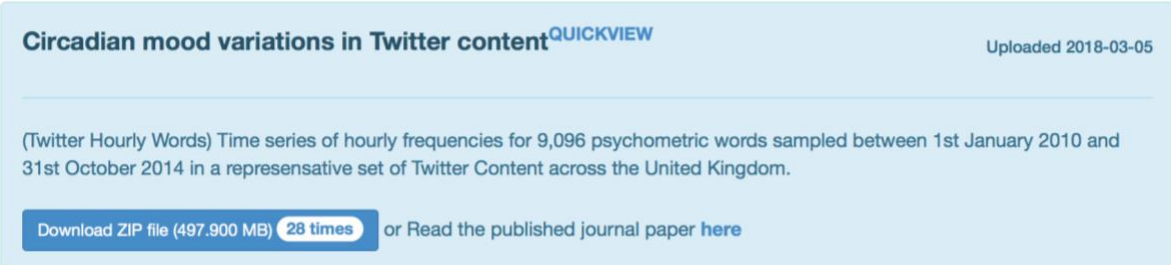

Figure 6.2. Isolated screen capture showing a single search result.

6.3.4 Link research publications to data

The prototyped system adopts a holistic approach to data discovery, in that it allows search on either of the data or publication end, and not only on the same system, but on the same user interface as well. Also, links to associated research data or publications, as the case may be, are conspicuously provided on the search results page itself. See Figure 6.3 above for example, and Sections 5.1.2, 5.1.3, 5.1.6 and 5.1.7 for further information on this requirement.

6.3.5 Clear statement(s) of relationship between data and associated file(s)

As previously mentioned (see Section 4.4), data is often accompanied by one or more files, associated publications being only one example. Others may be source code, original survey questions, file descriptions, READ MEs, appendices, variable coding information, user guides, instructions, index files, consent forms, ethical clearance certificates, etc. All of these require proper labelling, as shown in Figure 6.7, to prevent confusion. See Section 5.1.2 for further details about the requirement.

6.3.6 Clear presentation of datasets and their component parts

This is not only because of the range of documentation that may accompany data, but also because, as mentioned in Section 4.4, larger data files may be broken down into smaller parts. These multiple parts, if not properly labelled and grouped, may be lost or mistaken for documentation. This requirement has been further discussed in Section 5.1.2

6.3.7 Options for search by various metadata fields or by multiple fields at once

Figure 6.4 and 6.5 display the options for data search and for publication search. No change of screen is involved in either search, as search options and parameters are interactively enabled or disabled based on user selection. This was already noted earlier in Section 6.3.2. None of the search fields for either data or publication search are mandatory, therefore users may conduct both narrow and broad searches accordingly as their needs or knowledge of key information suggest. Table 6.3 lists the various search parameters and options for data and publication searches. See Section 5.1.7 for more about this requirement.

Table 6.3. Summary of options and parameters allowed for searching.

Search parameter	Sub-options	Type of feature	Require-ment	Availability
Resource type	• Research datasets • Research publications	Search option	Default	Not applicable
Keyword	Not applicable	Search field	Optional	Data and publications
Resource DOI	Not applicable	Search field	Optional	Data and publications
Author name	Not applicable	Search field	Optional	Publications only
Disciplinary domain	• IT & Computing • Social Sciences • Arts & Humanities	Search field	Optional	Data and publications
Upload date range	Not applicable	Search field	Optional	Data and publications
File size range	Not applicable	Search field	Optional	Data only
Must have associated dataset or publication?	• Yes • No	Search option	Optional	Data and publications, as the case applies

6.3.8 Review of prototype

In section 4.1 a representative selection of research data repositories were evaluated against a set of criteria as described in section 3.1. It seems proper to submit the prototype developed in this section to a similar evaluation using the same criteria, as follows:

1. *Use of metadata.* That is, the degree to which the system exploited metadata to provide features for browsing, searching/querying, filtering and search result presentation. This has been discussed in Chapter 5 (see Section 5.2) and demonstrated in the present chapter (refer to Sections 6.3.3 and 6.3.7). Screen captures of the search interface with different metadata fields being enabled for search are shown in Figures 6.4 and 6.5. Also, the first column of

Table 6.3 reflects the metadata elements given in Table 5.2, and shows how they were used to provide a wider set of parameters for the user to conduct broader or narrower searches to find data or publications. This feature is not commonly provided by some of the categories of data repositories reviewed, mainly institutional and general-purpose repositories, arising due to the heterogeneity of the data that they hold and the consequent necessity to promote general inclusivity by sacrificing particularities;

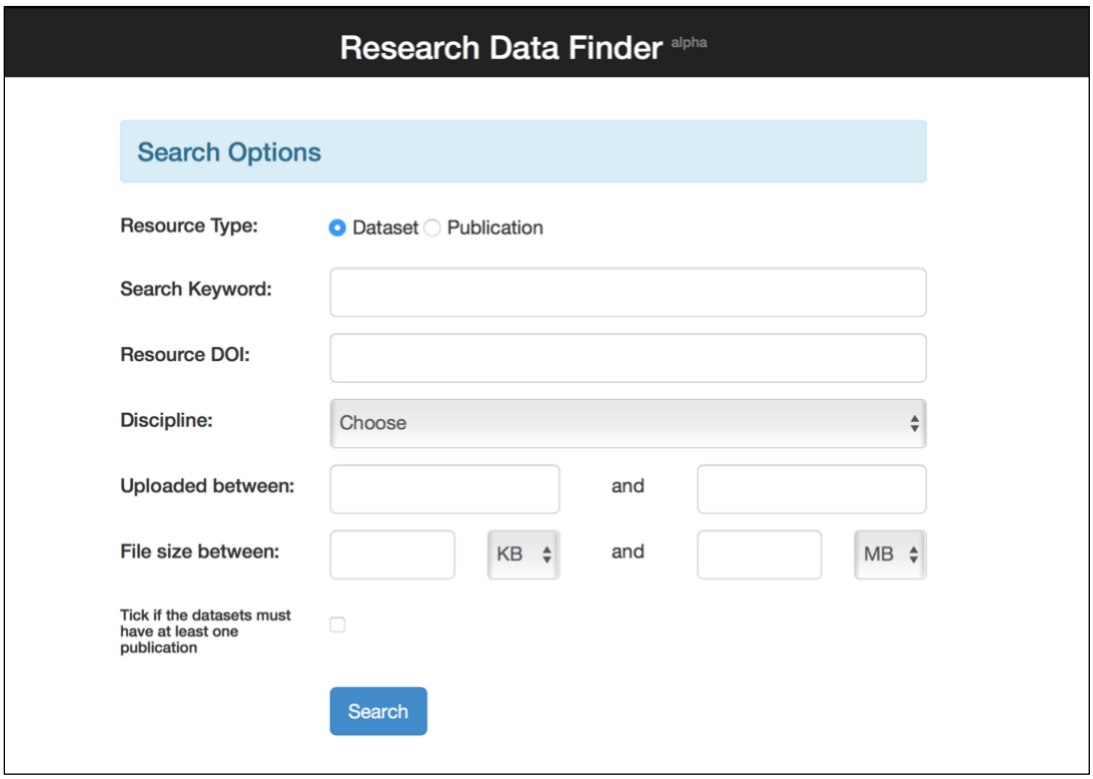

Figure 6.4. Default search screen, showing options for data search.

2. *Querying facility.* Or, the level of expressiveness allowed in searching/querying the repository. The querying facility of the system shown Figures 6.4 and 6.5. Certain data attributes such as have been found by the studies conducted in Phase I to be useful to users were incorporated in designing the query facility. Refer to the preceding point and to Section 6.3.7. DataFinder provides multiple search field options for querying that may be used or left blank if irrelevant. This is an improvement upon not only most institutional and general-purpose repositories, but upon many publisher-service repositories as well, which support advanced search options only for publications and not data;

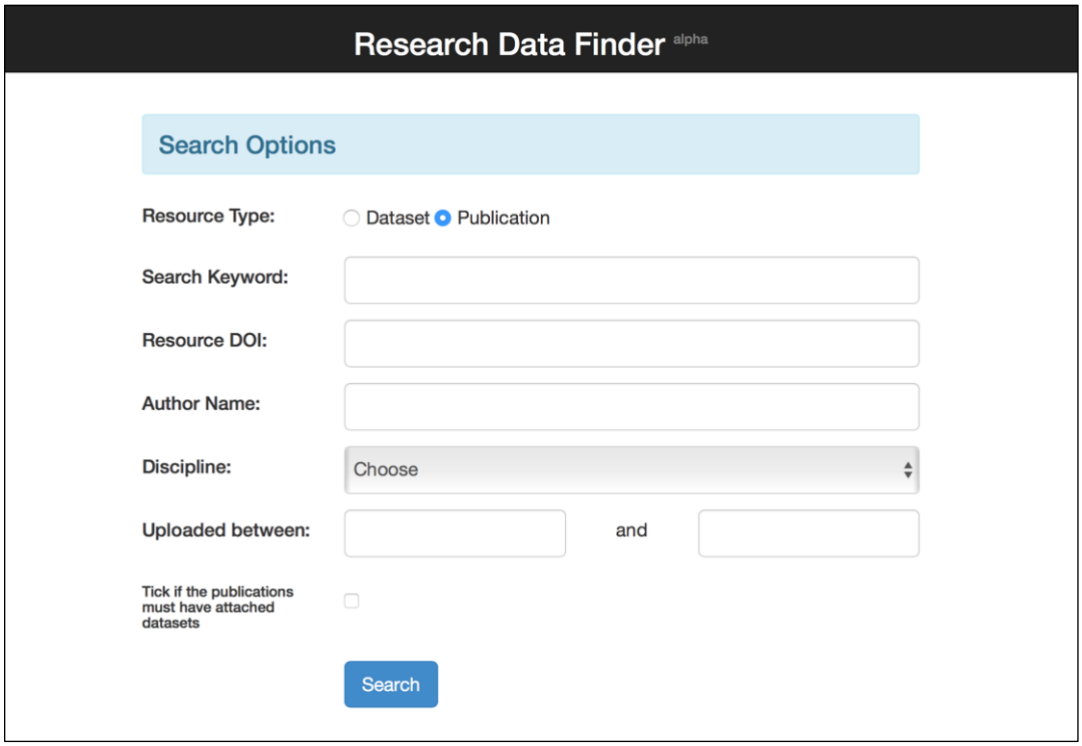

Figure 6.5. Options for publication search.

3. *Result filtering.* Or, availability of options for filtering down search results, and the furthest granularity to which this is possible. Search results may be filtered or narrowed down using the 7 fields allowed for query specification. After search parameters have been specified and the search conducted, the user may make modifications on the left-hand pane shown in Figure 6.6, and enlarged in Figure 6.8;

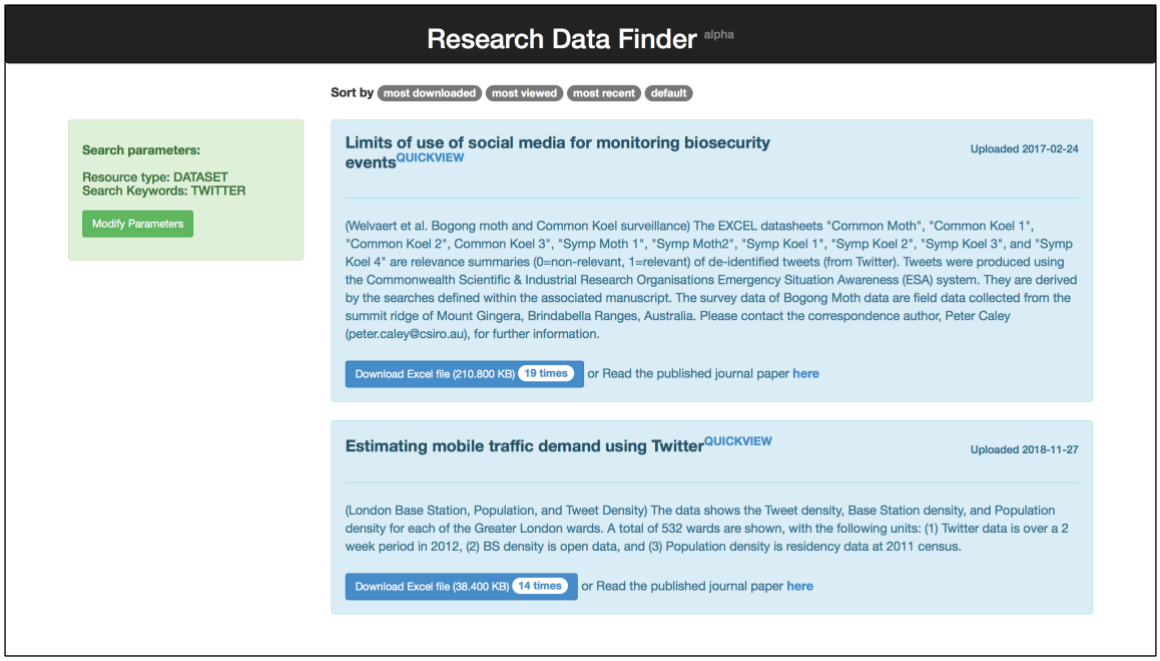

Figure 6.6. Sample search results page

4. *Sorting facility.* That is, the options available for ordering the arrangement of search results. Besides the default arrangement, the system allows three criteria by which to order search results, viz. by the most downloaded, most recent, or most viewed dataset or publication. This feature enlarged from Figure 6.6 and shown below in Figure 6.9;

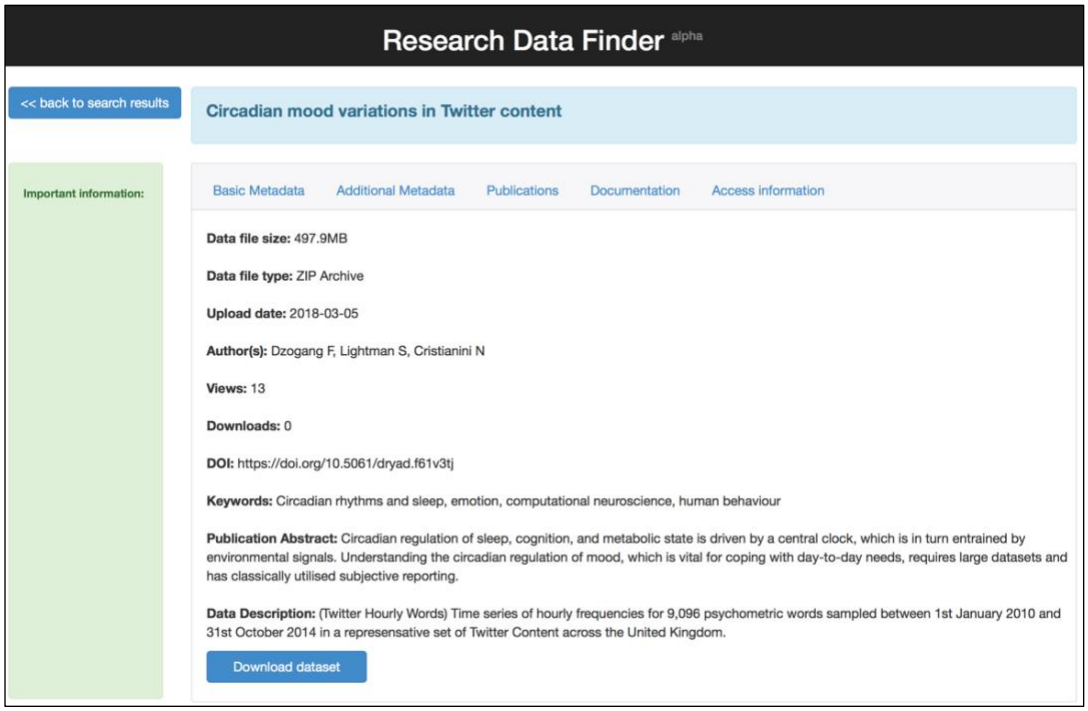

Figure 6.7. Sample dataset landing page.

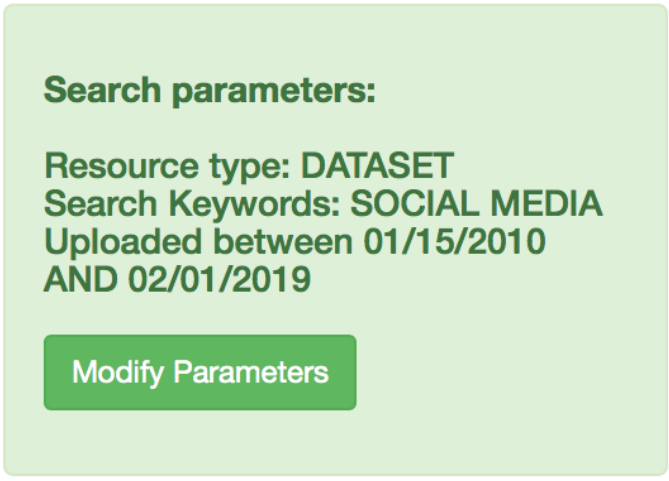

Figure 6.8. Sample search parameters with option to change or modify them.

Figure 6.9. Screen capture showing search result sorting criteria.

5. *Availability of additional features for data.* Aside from those features already mentioned, perhaps the main additional feature of the system, and the most unique, and arguably among the most useful for data discovery, is the ability to search for both datasets and publications on the same interface, and the linking of each to the other in a conspicuous, value-adding manner. It demonstrates a novel method of achieving research data discovery through a linked, hybrid system instead of the predominant separatist method of using different platforms for the two different resources. The advantages of this adopting this method have been noted throughout this work, particularly in Sections 1.2, 5.1.2, 5.1.3, 5.1.6 and 5.1.7.

6.3.9 System Limitations

There are some obvious limitations to DataFinder, arising chiefly from the fact that the full set of system requirements that have been identified were not all implemented. They therefore can be proved neither as positive improvements having the desired effects that have been hoped for or expected, nor as impairments requiring further work. It is thus difficult to measure with any degree of certainty how well or ill the system has achieved its stated objectives (see Section 1.3). Part of this problem is mitigated by conducting a user evaluation (see next chapter), since user feedback, especially for the present research, is an indicator of a system's successes and failures. Also, the review just concluded of the prototype demonstrates certain of its useful features which some categories of repositories that have previously been reviewed (see Section 4.1, Chapter 4) do not support. For example, it had been shown that most institutional and general-purpose repositories provide very scant (if at all) query-formation features beyond simple keyword search, and much too generic (if at all) options for refining and filtering search results, which is not wholly the case with DataFinder. Moreover, although the prototype is an important and even an integral part of this research, much of the work focus is on identifying requirements and sounding their potential usefulness and impact upon user experience and resource efficiency; since these constitute the foundations that, once established, will form the principles of RDM system design and development.

6.4 Chapter Summary

This chapter details the design and development of the prototype as described in Chapter 1. The chapter opened with a discussion about user-centered design in general and in relation to the present work. The long list of user-requirements identified in Chapter 5 were reduced to a shortlist, and a final requirements specification for the prototype was obtained. These were then used to develop as well as to review the prototype. Having obtained a working prototype, this chapter completes Phase II of this research, as Figure 6.10 shows. The next chapter describes the user-evaluation studies conducted with the prototype above presented. The chapter completes the answer to RQ 7 partially addressed in Chapter 6.

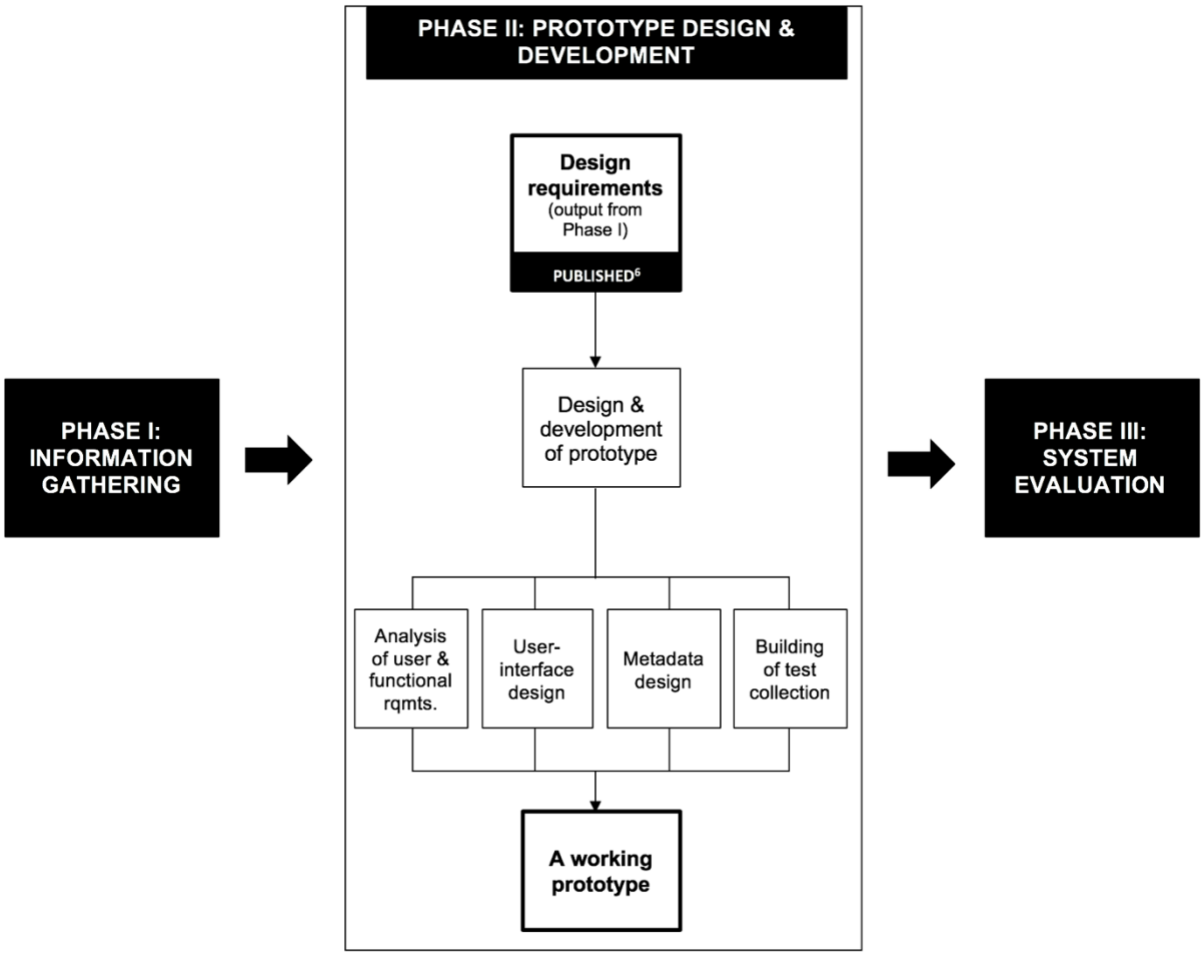

Figure 6.10. Overall outline of the research focusing on prototype design and development (Phase II).

7.0 USER EVALUATION

User evaluation, also called user-centered evaluation, is commonly used for validating user requirements and for improving system design (Zimmermann & Grötzbach, 2007). It helps to test preliminary ideas and to identify system strengths and limitations through activities that focus on gathering the subjective experiences of users. This is in contrast with system evaluation (or system-centered evaluation), which uses performance metrics to objectively measure the effectiveness and efficiency of systems. User-centered evaluation is thus suitable for real users in real-life contexts, while system-centered evaluation is more suitable for developing efficient algorithms (Díaz et al., 2008; Petrelli, 2008). Usability testing and expert reviews are the alternative means of conducting user-centered evaluation (Zimmermann & Grötzbach, 2007); this work will employ the former since it is the method that conforms to user-centered design principles (Ames, 2001; Lazar, 2006, p. 205). A usability test with real users of data repositories has been conducted for the system prototype developed in the last chapter, and is reported in the present chapter. Since iterative prototyping is not feasible within the present research scope, this user evaluation stage represents the final phase (Phase III, refer to Figure 7.1) of research.

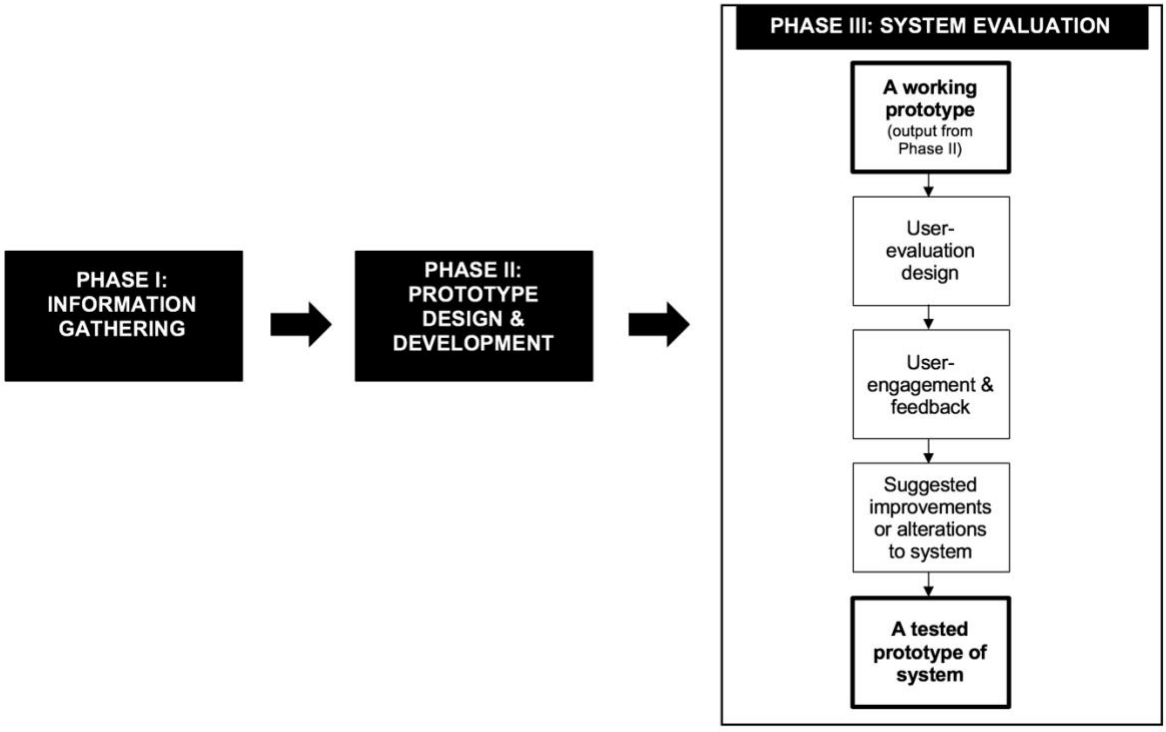

Figure 7.1. The present chapter in the context of the overall research.

7.1 Study objectives

It has been noted above that usability testing is the recommended method of conducting user evaluation for user-centered design. The concept of usability itself has already been explored in Section 6.1. Usability testing describes "a systematic way of observing actual users trying out a product and collecting information about the specific ways in which the product is easy or difficult for them" (Dumas & Redish, 1993, p. 12). According to Ames (2001) it uncovers two types of issues, namely:

1. Issues around the system's look and feel, which accounts for approximately 40% of its usability; and
2. User-system interaction issues, which accounts for approximately 60% of the usability of the product.

These were more precisely stated by Van der Geest (2004) as relating to the following points, each of which will be individually addressed in Section 7.4:

a. *Content and information*, which mostly concerns the content of my test collection (see Section 6.2.2) and the metadata that describes them (see Section 5.2.1);
b. *Navigation and structure*, which concerns partly user interface design (refer to Section 6.2.1, screen captures in Section 6.3.8) and partly user requirements (see Sections 6.3.2 and 6.3.4);
c. *Design and presentation*, also concerns the user interface design (refer to Sections 6.2.1, 6.3.3, and 6.3.6; screen captures in Section 6.3.8); and
d. *Other problems* (which, for the present study, will be any relating to the criteria and the requirements specified in Sections 3.1 and 6.3.1 respectively. Refer to Section 6.3.8 for a review of the DataFinder against the criteria just mentioned).

The specific UI design guidelines that were followed have been discussed in the last chapter (see Section 6.2.1). The aim of usability testing is not to uncover problems relating to the mechanism or technical operation of the system, but relating to its "softer" characteristics. With this view in mind, my specific objectives with regard to the prototype being tested are as follows:

To–

1. Collect feedback from users on the usability and design of the user interface and system features, specifically those highlighted above;
2. Ascertain as to researchers' perceived usefulness of linked research datasets and publications for their day-to-day data and information seeking activities;
3. Compare the overall design and usability of the new system to that of existing repositories; and
4. Uncover problem areas for later improvement.

7.2 Study Population

This study was conducted with 5 users from the IT & Computing discipline who, in response to an email request or by word-of-mouth, freely volunteered to participate. The study was conducted on this relatively small scale, the aim being mainly to obtain in-depth qualitative feedback from a sufficiently representative sample, seeing as the system is only an alpha prototype and the research does not involve a second iteration of system development. As to the choice of the discipline, it was guided by the following considerations, that:

a. My familiarity with the discipline in terms of knowledge of subject area and of established systems, practices, and data characteristics would enable me more deeply to understand users' feedback;
b. The test collection data are predominantly more relevant for researchers in that discipline, because such data had been easier to obtain and tag; and
c. From previous studies conducted in Phase II of this research (see Sections 4.2 and 4.3) users in this discipline showed better acquaintance with RDM concepts.

System usability being the main interest for the evaluation, it was thought better to represent a wide a range of experience. One postdoctoral research fellow, three doctoral research students, and one final-year undergraduate student thus formed the five participants. Admittedly, the development of user-requirement specifications for the system was based on a vastly more representative user group. The samples for those studies (i.e. questionnaire survey and interviews) covered not only various kinds and levels of researchers' academic job post and

years of experience, but also subject disciplines. In comparison, the sample for the present study is much less representative of the entire population of the potential users of the system. It omits entirely an important section of the same, viz. academic and mid to late-career researchers. As this circumstance will consequently skew the data gathered from the study, its findings cannot be generalized onto the larger population. Still, the findings provide useful user-feedback for the further improvement of the system and better definition of its user requirement specifications. This feedback, though partial, is nonetheless relevant.

7.3 Study Design

The study was a 2x5 within-subject design conducted individually with each participant. For the purpose, a meeting lasting about 30 minutes was separately arranged with each participant. The main activity of each session consisted in the participant using DataFinder and one other data repository to search for data on a particular topic, as detailed in the next section. Being there was not a live deployment of DataFinder on the internet, the search tasks were carried out on my local machine by all the participants. The search topic and repository formed the independent variables of the study, the dependent variable being the user's feedback. A combination of three complementary usability methods, viz. Interviews, Observations, and Thinking-aloud were used to conduct the study. The thinking-aloud method has been variously noted in the literature as having the best performance of all the other techniques for usability testing (Henderson et al., 1995; Allwood & Kalén, 1997; Ebling & John, 2000; Donker & Markopoulos, 2002). As its name suggests, it tries to draw out users' thoughts, reflections, and cognitive processes as they engage or interact with a system (Van Oostendorp & De Mul, 1999; Patton, 2002). This method was complemented for this study with post-evaluation interviews based in part on the silent observation of users' expression and manner whilst using the two systems being compared. The advantage of interviews for user engagement has been noted already, in Section 3.3. General guidance for conducting the study was as given by Dumas & Redish (1993, p.22), in the following terms:

1. *The primary goal is to improve the usability of the system: articulate specific goals and concerns when planning the test.* This point has been satisfied in Section 7.1;

2. *Study participants must represent real users.* Those who participated in the study, though representative of a particular segment of real users that may potentially use the system, are non-representative of the entire population. As they consequently cannot give generalizable findings, this condition is only partially satisfied. See Section 7.2 for further discussion on the study sample;
3. *The participants must do real tasks.* This point is addressed in earlier in this and the next section;
4. *Observe and record what participants do and say.* For reasons such as those noted in Section 3.3, only handwritten notes were taken during each session. These were followed by a detailed transcribing immediately after the session. The data obtained was analysed using standard qualitative techniques as described by Patton (2002); and
5. *Analyze the data, diagnose problems, and recommend changes to fix the problems.* See Sections 7.4, 7.5, and 7.6.

7.4 Study Procedures

The steps involved in each of the 5 sessions of the study are detailed below in the order in which they were carried out:

1. In the beginning, the participant was presented with a brief background of the research, followed by a brief overview about the study steps and procedures, and what was expected of him/her during the session. The background was given only and strictly by way of some explanation of the study in which the participant was about to take part, and care was taken to make it as brief and sketchy as possible, so as not to influence the participant's responses. In fact, it was not explicitly stated who the developer of DataFinder was: it was merely intimated that the researcher wanted to compare between it and another system;
2. The participant was asked about his/her own research and means or methods of obtaining research data, and I made a note of these;
3. The participant was presented with a web browser window containing DataFinder's search page, and was asked to open the search page of another data repository of his/her choice; if undecided as to this, Figshare[1]

[1] https://figshare.com

was recommended as a general-purpose repository. Participants were given the first choice of the second repository because I wanted to compare DataFinder with something that they liked or were used to. This, it was hoped, would enable them to give more in-depth and personally relevant feedback, and from longer experience, although it introduced a considerable increase in the variability of the study;

4. The participant was asked to perform a search on each of the two repositories opened above. For DataFinder, since the test collection is not exhaustive, a set of the same 5 available keywords were provided to each participant. From these the participant was asked to choose one for the search task, if desired, since keywords are optional on the system. Participants were allowed the choice of searching from the dataset or publication perspectives. Also, they could use the same or a different keyword for the search on the repository being compared. The suggested keywords for DataFinder were: *open data*, *information networks*, *social groups*, *graph analysis* [data], and *social media* [data]. This, admittedly, is rather a biased comparison that is likely unfairly to favour DataFinder. The focus was therefore shifted more onto evaluating and obtaining feedback for and about DataFinder on its own merit alone, and in a way that does not seem to place it and the other system at mutual variance. This was especially considered since the uncovering of usability, and not of performance issues, motivated the study;
5. Throughout, whilst performing the search tasks, the participant was asked and encouraged to think aloud, i.e. to verbalize his/her thoughts about, for example:
    a. What s/he is trying to do;
    b. How s/he feels about the functions being interacted with;
    c. Impressions respecting the navigation, look, and general design of the system;
    d. If s/he is stuck or confused;
    e. The usefulness or relevance or any system features, or the lack thereof, in his/her particular research context; etc.

6. I took down notes and observations on the above and respond to any questions that might happen to be asked about the search tasks or about DataFinder;
7. After completing the search tasks, the participant was asked to reflect about his/her overall experience, especially where it contrasts between the two systems;
8. A brief post-interview followed in which I asked the participant for feedback particularly on the novel features of DataFinder. Where necessary, I also asked for clarification of doubtful observations or further detail about interesting ones as I might happen to have made in the course of the study. The post-interview typically entailed the following questions:

    a. How much more or less useful would DataFinder be in your particular research context, compared to the usual style of data repository?
    b. Which of DataFinder's search fields or options did you find most useful?
    c. Where, if at all, did you feel stuck, uncertain, frustrated, or confused while using DataFinder?

7.5 Analyses and Results

Feedback, notes, and observations from the previous section were coded and analyzed using manual thematic content analysis techniques as described by Patton (2002). Themes were already pre-decided (refer to Section 7.1 above) from the following sources, as follows:

1. The problem types noted by Van der Geest (2004)–
    a. Content and information;
    b. Navigation and structure;
    c. Design and presentation; and
    d. Other problems;
2. The criteria given in Sections 3.1, by which data repositories are reviewed in this work–
    e. Use of metadata;
    f. Querying facility;
    g. Result filtering facility;

h. Sorting facility; and
i. Availability of additional features for data;

3. The user requirements specifications shortlisted Section 6.3.1–
    j. Simple and minimal number of steps for data upload, access, and search;
    k. Clear display of data file size;
    l. Link research publications to data;
    m. Clear statement(s) of relationship between data and associated file(s);
    n. Clear presentation of datasets and their component parts;
    o. Options for search by various metadata fields or by multiple fields at once;
    p. Generally being mindful of university and funder requirements in system development; and
    q. Use of interoperable standards and optimizing repositories to enable indexing by external and general-purpose search engines.

There is obviously much thematic overlap in the long list above, and it was therefore re-organized into the following four broad themes as follows:

1. Content and information features–
    a. Metadata;
    b. Clear display of data file size;
    c. Clear statement(s) of relationship between data and associated file(s);
    d. Clear presentation of datasets and their component parts;
2. Navigation and structure features–
    e. Simple and minimal number of steps for data upload, access, and search;
    f. Link research publications to data;
3. Design and presentation features–
    g. Sorting facility;
    h. Availability of additional features for data;
4. Operational features–

i. Querying facility;
j. Result filtering facility;
k. Use of interoperable standards and optimizing repositories to enable indexing by external and general-purpose search engines; and
l. Generally being mindful of university and funder requirements in system development.

Using the latter thematic structure each point was categorized accordingly as it belonged the most to a particular theme. What resulted for each theme was then further classified as a "plus" (a usability advantage), a "minus" (a usability problem), or a neutral observation. The final results of the analyses are presented in the series of Tables (7.1 – 7.4) below. The chapter concludes with a section on recommendations for further development and improvement of the prototype.

Table 7.1. Results of user evaluation for the theme of content and information features.

Content and information features	
Usability problem	• The "quick view" link to the search results was generally supposed to mean a quick preview of the dataset itself, rather than of some of its metadata
Usability advantage	• Information about download count was generally found very useful, especially as an indication of how "good" a dataset was • The range of information given for each result item on the search results page was generally appreciated by users (see Figure 6.2) • Key information in DataFinder all clearly labelled/indicated. Two users commented that they felt it would "be hard to make a mistake" about what was what

Table 7.2. Results of user evaluation for the theme of navigation and structure features.

Navigation and structure features	
Usability problem	• Users generally thought that the "Modify Parameters" button (see Figure 6.8) would allow them to perform the said operation on the same page, instead of being taken back to the search options page. Users generally voiced their preference for seeing the effect of their operations as they were being made, rather than to be taken back

	and forth between the search options screen and the search results screen
Usability advantage	• Users commented that DataFinder was easy and straightforward to navigate, most of them saying it would be hard to "get lost" on it • Data Users indicated their preference of the visibility of DataFinder's range of search options, which are conspicuous without having to look for the usual "advance search" link. One user who used Figshare for the comparison voiced frustration at not even finding the said link at all on Figshare, though s/he said s/he was "sure they must have it somewhere" • All the users, including the undergraduate, found the linked data and linked publications feature a "good idea". Two gave the reason that they found it a tedious process to have first to search and find the full associated paper on Google Scholar, be linked to the publisher's website and then scroll down to the end of the page in order to find the location or DOI, before they can at last locate the data • One user mentioned that when reading research papers s/he generally understood it better when s/he followed the analyses presented therein by looking at the actual data. Rather a remarkable finding, this, since popular assumption has been more focused on publications helping to understand data rather than the other way around • The "quick view" popup link received favorable feedback, and its lack on the comparison repository was generally seen as a minus there. Users said they liked to be able quickly to know a few decisive details about search results as they scroll down, without having to open new tabs or navigate away from the page

Table 7.3. Results of user evaluation for the theme of design and presentation features.

Design and presentation features	
Usability problem	• Commenting on the general look of DataFinder, some users said it did not look "real" to them as did the comparison repository, adding that there was "not much" on it. Further questioning revealed this to be because there were hardly any other links to other webpages, e.g. "About", "Terms & Conditions", etc. • Users thought it would be "nice" to add a download link to the "quick view" popup

	• Users thought that the color scheme used in the system "could be improved"
Usability advantage	• Users generally commented favorably on the simple "no clutter" design of the system • Some users said that it looked "fun" • The layout of the search results page and the dataset landing pages were found to be "pleasant" by some users and easy to understand by all the users

Table 7.4. Results of user evaluation for the theme of Operational features.

Operational features	
Usability problem	• Not all the search fields and options were found suitable by all users. Some users felt "at a loss" confronted with "so many options". The least useful search field, according to users, was DOI, commenting that it was easier to remember words than a sequence of numbers • No indications of whether a particular search field was mandatory or optional, and some researchers, thinking them to all be mandatory, were "rather turned off" as they "disliked filling out forms" • The "author" search field which was available only for publications was missed by one user who wanted it for a data search
Usability advantage	• Two users remarked that they were "delighted" at the range of fields for specifying search on DataFinder • The "date range" search field was found very useful for research that used time-series data • The "discipline" filter was generally found very useful. Most users mentioned that its absence in the other repository generated too many irrelevant results for them • The file size filter was found useful by two users who were on an internet data plan that limited their internet use to a quota of a certain number of Gigabytes per month • Another user also found the file size filter "very useful" because s/he predominantly used image classification datasets in his/her research, and such datasets usually fall within a size range. S/he said that DataFinder's option helped him/her to "weed out" irrelevant results

7.5.1 General observations

In addition to the main study observations, the following general points were noted during the course of the study:

That–

1. Users seemed to judge the system accordingly as it satisfies or does not satisfy the particular data need of the research project they were currently engaged in, and not on the grounds of a potential or past need;
2. Users look for data for different reasons which are not always motivated by data content per se, but by data attributes or characteristics. Consequently, data reuse may mean or involve other uses besides analyses for publications or for producing other kinds of primary content. A case in point involves users from the Machine Learning subfield, who often use data mainly for training or validating machine learning algorithms. In such cases, the information needed to decide the suitability of data differs from the typical and usually pertains to quite an entirely different aspect of the data; such as, number of observations in the dataset and their statistical distribution. This remark is a significant one that may help towards modelling researchers' data-seeking needs in different research domains. It also corresponds to findings from Phase II of this research (see Section 4.3.3, for example) pointing to disciplinary idiosyncrasies in various aspects of RDM and the importance of taking these into account when designing RDM systems; and
3. The use of filters seemed for all users to depend on the amount of search results generated by an initial tentative query. All the users at first used only about 2 of the search filters available on DataFinder, which returned only a few search results (due to the relatively small size of the test collection). Many of the search filters provided by DataFinder were however missed or wished for in the comparison repositories which have a narrower range of search options (or sometimes none at all), particularly when a search generates dozens or hundreds of search results.
4. One postdoctoral research fellow, three doctoral research students, and one final-year undergraduate student, as before stated, constituted the five participants of the study. Indeed that is much too small a sample to derive general conclusions from, but it might nonetheless be worth remarking that there does not appear to be any perceptible pattern of responses among

the three doctoral research students to differentiate them from either of the other two participants, nor, similarly, these other two from each other. This seems to indicate that perhaps differences are less to be looked for in researchers' academic post and more in the kind of data that they work with or projects that they work on.

7.5.2 System recommendations

In addition to the recommendations suggested by the usability problems identified in the Tables 7.1 – 7.4, the following come from express user comment:

1. Enhance the "Discipline" filter by adding sub-disciplines and sub-fields;
2. Many users wished for an option for specifying between qualitative and quantitative datasets. This field may be added to the list of metadata elements required at upload (see Section 5.2.1) to help support the feature; and
3. Some users indicated that a user rating feature for datasets, such as that generally found on shopping and other multimedia content websites, would be useful to them in gauging how "good" a dataset possibly was.

7.6 Chapter summary

This chapter completes the overall research reported in this work. It details about the user-evaluation study conducted to test the prototype developed in the preceding chapter, identifying some key usability strengths and weaknesses of the system. The study also confirms some findings from earlier studies, and resulted in some important observations and recommendations for further work on the system and in RDM at large. The scope of this research does not allow for iterative development, but it is hoped that the alpha version of the system thus tested may be taken up for further development in a subsequent work. This next chapter concludes the research with further recommendations based on the overall work.

8.0 CONCLUSION AND RECOMMENDATIONS

The two things that seem alone to compose the soul of RDM are 1) the needs of users and 2) the needs of data. Every problem of RDM seems to boil down to a lack or inadequacy of due attention to either or both of these things. And, likewise, every triumph of RDM seems traceable back to a greater attention to them. RDM is still in early stages and beset with considerable issues all of which require intelligent solving. The key ones have been noted in this work, the first step to problem- solving being problem-identification. And undoubtedly, there is a gradual but progressive closing in of the distance between the state of things as they currently are and the state of things as they are desired to be. The inflow of new ideas and suggestions that tend toward this progress is continual, both from the literature and in practice. The present work itself is an example in point. It is the firm opinion of this research that hope lies in user-centered design and in linked data. User-centered design, because users are indeed, and not only in theory, a central component of the RDM ecosystem and ought therefore to be considered as such in the design of all RDM systems; and linked data, because data and publications in research and scholarship go hand in hand and ought not to be separated. On the one hand, a quick reflection will show that, directly or indirectly, everything in the RDM ecosystem is dependent upon or connected to the user or the user's agency, from the sharing of data on repositories to the finding and potentially reusing of it. And on the other hand, separating research data and research publications engenders an unnatural state of affairs which in itself is a problem. Thus the present research, by producing a solution that combines both user-centered design and linked data, albeit the latter in a very simplified implementation, demonstrates a holistic approach to RDM.

8.1 Contribution to knowledge

This research has resulted in a number of important contributions that add to the depth or breadth of information in the knowledge domain. The chief contribution of the research is a practical one, and constitutes its main deliverable: a simple prototype of an RDM system, by name, DataFinder. DataFinder, among other user-centered features, demonstrates a novel method of achieving research data discovery through a linked, hybrid system in which research datasets and research publications exist on the same platform and are connected together in a mutually value-adding way. Also, the system demonstrates a simple approach to

holistically address user-centredness in RDM design from the inception stage to the implementation stage, and is relatively easy to build. And, finally, it is interoperable with other RDM system services, through its use of standard and universal identifiers such as ORCID and DOI. This practical contribution carries the further advantage of having been evaluated with real users. Sundry other contributions incidental to the process of developing the prototype also resulted. Among these is a longlist of user-requirements for a user-centered RDM system, that was derived mostly from studies carried out expressly for the purpose. The list also happens to be as yet the first of its kind in the literature, and can form the basis for further development. There was also a demonstration of the practical differences between data retrieval and information retrieval, the first of their kind available in the literature. The classification of research data repositories, originally developed for the purpose of this research, may prove useful for other purposes. Finally, this work augments the scant literature on user-centered design in relation to RDM systems. The reported findings of studies conducted by others which relate to the practices, attitudes, and concerns of researchers on the many different aspects and sub-aspects of RDM have also been varied, corroborated, or supplemented by the findings that resulted from the studies conducted in this research. In particular, the findings from the questionnaire survey reported in Chapter 4, although largely similar to other studies already reported in literature, are highly corroborative especially in the areas of researchers' data sharing concerns, their data storage practices, their general want of skill and motivation to tag their data, and their predominant unawareness of institutional data management policies. Overall, the research resulted in 5 contributions to the literature.

8.2 Recommendations

The present work involved a wide range and variety of activities, including mixed-methods studies, experiments, system design and development, and usability testing. This afforded me special scope and opportunity of a gaining a relatively comprehensive awareness of the various windings in the field, and of learning about RDM from diverse perspectives. The recommendations that follow chiefly result from this. Although not all of the features enumerated in the longlist of user requirements were fully implemented in DataFinder (alpha), the fact of their being identified clears the course for further work in the future.

8.2.1 For RDM at large

1. Inter-disciplinary differences and intra-disciplinary idiosyncrasies greatly influence, and often determine, important variables in user and system requirements. These points should be duly considered when developing RDM systems;
2. Perhaps a great part of the difficulty of RDM systems in catering to the needs of researchers is owing to the imperfect understanding of researchers' data-seeking needs, especially with the added complexity of disciplinary variations. Engagement with users can expedite the development of such models, which promises to have multiple uses and applications; and
3. Metadata is a core driver of RDM systems both technically and otherwise. Nonetheless metadata also presents one of its chief problems, since users, whom are the primary sources of the metadata, are prevented through lack of will, or skill, or resources, from supplying the adequate requirement for a better functioning of the RDM ecosystem. It hence becomes important to develop user-friendly software solutions to help users who have no metadata skills to easily tag their data; and to simplify the process for users who are disinclined, in order that the activity may be less time consuming and more effortless.

8.2.1 For DataFinder and RDM systems generally

1. Due to scope limitations, only a subset of the full set of use requirements identified in this work were able to be implemented. It is recommended that a full version of DataFinder be developed with iterative prototyping, addressing the usability problems initially highlighted by the user-evaluation;
2. Dictionary look-up and simple natural language processing functionalities will improve the quality of search results, and also aid in better discovery of data; and;
3. Ontological schemas implemented with RDF triplets will support complex querying and graph analyses for useful insights that might be useful especially to repository proprietors and research funders.

REFERENCES

Akers, K. and Doty, J. (2013). Disciplinary Differences in Faculty Research Data Management Practices and Perspectives. The International Journal of Digital Curation, Vol. 8, No. 2, pp. 5-26.

Allwood, C. M., & Kalén, T. (1997). Evaluating and improving the usability of a user manual. Behavior & information technology, 16(1), 43-57.

Alsos, O., Svanæs, D. (2011). Designing for the Secondary User Experience. Human-Computer Interaction – INTERACT 2011, 84-91. doi: 10.1007/978-3-642-23768-3_7

Ames, A. L. (2001). Users first! An introduction to usability and user-centered design and development for technical information and products. In IPCC 2001. Communication Dimensions. Proceedings IEEE International Professional Communication Conference (Cat. No.01CH37271). IEEE. https://doi.org/10.1109/ipcc.2001.971558

Amorim R, Castro J, Rocha da Silva J, Ribeiro C. A. (2016). Comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society. 16(4):851-862. doi:10.1007/s10209-016-0475-y.

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. Data Science Journal, 3, 191–201. https://doi.org/10.2481/dsj.3.191

Arend D., Lange M., Chen J. et al. (2014). e!DAL - a framework to store, share and publish research data. BMC Bioinformatics. 15(1):214. doi:10.1186/1471-2105-15-214.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., et al. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Science Journal, Vol. 3 November, pp. 135-152.

Baker, K., Yarmey, L. (2009). Data stewardship: environmental data curation and a web-of-repositories. Int J Digit Curation 4(2):1–16. http://www.ijdc.net/index.php/ijdc/article/view/115

Ball, A. (2012). Review of Data Management Lifecycle Models. Bath, UK: University of Bath.

Ball, A., Ashley, K., McCann, P., Molloy, L. and Van Den Eynden, V. (2014). Show me the data: the pilot UK Research Data Registry. Presentation at 9th International Digital Curation Conference (IDCC), 26 February 2014, San Francisco, USA.

Bergman, M. (2008). Advances in Mixed Methods Research: Theories and Applications. Los Angeles Calif. London: SAGE.

Borgman, C. L. (2007). Scholarship in the digital age: Information, infrastructure, and the Internet. Cambridge: The MIT Press.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078 .

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge, MA, The MIT Press.

Borgman, C. L., (2011). The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology, pp. 1-40, 2011. Available at SSRN: https://ssrn.com/abstract=1869155 or http://dx.doi.org/10.2139/ssrn.1869155

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 7, 17-30. doi:10.1007/s00799-007-0022-9

Borgman, C. L., Wallis, Jillian C., Mayernik, Matthew S. (2012). Who's Got the Data? Interdependencies in Science and Technology Collaborations. The Journal of Collaborative Computing 21(6), 485-523

Boru, D., Kliazovich, D., Granelli, F., Bouvry, P., Zomaya, A.Y. (2015). Energy-efficient data replication in cloud computing datacenters. Cluster Computing, 18(1), 385-402

Bourne, P.E., Lorsch,J.R. and Green,E.D. (2015). Perspective: Sustaining the big-data ecosystem. Nature, 527, S16–S17.

Bowler, L. & Koshman, S. & Oh, J. S. & He, D. & Callery, B. G. & Bowker, G. & Cox, R. J. (2011). Issues in User-Centered Design in LIS. Library Trends 59(4), 721-752. Johns Hopkins University Press. Retrieved March 1, 2019, from Project MUSE database.

Bremer, J.-M., & Gertz, M. (2005). Integrating document and data retrieval based on XML. The VLDB Journal, 15(1), 53–83. https://doi.org/10.1007/s00778-004-0150-4

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. (2001). Scholarly work in the humanities and the evolving information environment. Washington, DC: Digital Library Federation

Brown, C. D. (2002). Straddling the humanities and social sciences: The research process of music scholars. Library & Information Science Research, 24(1), 73–94.

Bugaje, M., Chowdhury, G. (2017a). Towards a More User-Centered Design of Research Data Management (RDM) Systems [abstract]. In: Information: Interactions and Impact (i3).; 27-30 June 2017; Aberdeen; 2017:53-55.

Bugaje, M., & Chowdhury, G. (2017b). Is Data Retrieval Different from Text Retrieval? An Exploratory Study. In Digital Libraries: Data, Information, and Knowledge for Digital Lives (pp. 97–103). Springer International Publishing. https://doi.org/10.1007/978-3-319-70232-2_8

Bugaje, M., Chowdhury, G. (2018a). Data Retrieval = Text Retrieval? In: Chowdhury, G., McLeod, J., Willett, P., Gillet, V. (eds.) iConference 2018. LNCS, vol. 10766, pp. 253–262. Springer https://doi.org/10.1007/978-3-319-78105-1_29

Bugaje, M., Chowdhury, G. (2018b). The Sixth European Conference on Information Literacy, September 24th–27th, 2018, Oulu, Finland : (ECIL) : abstracts / editors Sonja Špiranec, Serap Kurbanoğlu, Maija-Leena Huotari, Esther Grassian, Diane Mizrachi, Loriene Roy, Denis Kos. Oulu : University of Oulu, Department of Information and Communication Studies, 2018.

Bugaje, M., Chowdhury, G. (2018c). Identifying Design Requirements of a User-Centered Research Data Management System. In Lecture Notes in Computer Science (pp. 335–347). Springer International Publishing. https://doi.org/10.1007/978-3-030-04257-8_35

Burgoon, J., Bonito, J., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human–computer interaction: a study of credibility, understanding, and influence. Computers in Human Behavior, 16(6), 553–574. https://doi.org/10.1016/s0747-5632(00)00029-7

References

Burton, A., Koers, H. (2016). Interoperability Framework Recommendations. ICSU-WDS & RDA Publishing Data Services Working Group. https://rd-alliance.org/system/files/InteroperabilityFrameworkRecommendations-WDSRDAPDSWG.pdf

Carlson J. (2012). Demystifying the data interview. Reference Services Review.40(1):7-23. doi:10.1108/00907321211203603.

Carlson, J. (2012). Demystifying the data interview. Reference Services Review, 40(1), 7–23. https://doi.org/10.1108/00907321211203603

Carter, P. (2007). Liberating usability testing. Interactions, 14(2), 18-22.

Case, D. O. (1991). The collection and use of information by some American historians: A study of motives and methods. Library Quarterly, 61(1), 61–82.

Cimino, J. J., Ayres, E. J., Remennik, L., Rath, S., Freedman, R., Beri, A., … Huser, V. (2014). The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): Design, contents, functionality and experience to date. Journal of Biomedical Informatics, 52, 11–27. https://doi.org/10.1016/j.jbi.2013.11.004

Chowdhury, G., Walton, G., Bugaje, M. (2017). The Fifth European Conference on Information Literacy, September 18th-21st, 2017, Saint-Malo, France. https://repositorio.ipl.pt/bitstream/10400.21/7706/1/Information%20literacy%20in%20Portuguese%20university%20context_a%20necessary%20intervention.pdf#page=62

Chowdhury, G.G. (2014). Sustainability of scholarly information. Facet Publishing, London.

Chowdhury, G., Ünal, Y., Kurbanoglu, S., Boustany, J., and Walton, G. (2018). Research data management and data sharing behaviour of university researchers. In: ISIC 2018: The Information Behaviour Conference, 9-11 October 2018, Krakow.

Constantine, L. (2004). Beyond User-Centered Design and User Experience: Designing for User Performance, Cutter IT Journal. 17, 2.

Corry, M. D., Frick, T. W., Hansen, L. (1997). User-centered design and usability testing of a web site: An illustrative case study. Educ. Technol. Res. Dev., vol. 45, no. 4, pp. 65–76.

Cox, A. M., Pinfield S. (2014). Research data management and libraries: Current activities and future priorities. J Librariansh Inf Sci. Available: http://lis.sagepub.com/cgi/doi/10.1177/0961000613492542. Accessed 2014 Mar 30.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Curdt C., Hoffmeister D. (2015). Research data management services for a multidisciplinary, collaborative research project. Program. 49(4):494-512. doi:10.1108/prog-02-2015-0016.

Curtis, G., & Cobham, D. (2013). Business information systems: Analysis, design and practice. Harlow, England: Prentice Hall, Financial Times.

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Matthews, B., Mele, S., … Wilson, M. (2014). Enabling Sharing and Reuse of Scientific Data. New Review of Information Networking, 19(1), 16–43. https://doi.org/10.1080/13614576.2014.883936

Davidson, J., Jones, S. and Molloy, L. (2014a). Big data: the potential role of research data management and research data registries. In: IFLA World Library and Information Congress 80th IFLA General Conference and Assembly, Lyon, France, 16-22 August 2014.

Davidson, J., Jones, S., Molloy, L., & Kejser, U. B. (2014b). Emerging Good Practice in Managing Research Data and Research Information within UK Universities. Procedia Computer Science, 33, 215–222. https://doi.org/10.1016/j.procs.2014.06.035

Dierkes, J., & Wuttke, U. (2016). The Göttingen eResearch Alliance: A Case Study of Developing and Establishing Institutional Support for Research Data Management. ISPRS International Journal of Geo-Information, 5(8), 133. https://doi.org/10.3390/ijgi5080133

Dimou, A., De Vocht, L., Van Grootel, G., Van Campe, L., Latour, J., Mannens, E., & Van de Walle, R. (2014). Visualizing the Information of a Linked Open Data Enabled Research Information System. Procedia Computer Science, 33, 245–252. https://doi.org/10.1016/j.procs.2014.06.039

Donker, A., & Markopoulos, P. (2002). A comparison of think-aloud, ques- tionnaires and interviews for testing usability with children. In X. Faulkner, J. Finlay & F. Detienne (Eds.), Proceedings of human computer interaction 2002 (pp. 305-316). London: Springer.

Duff, W. M., & Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behavior of historians in archives. Library Quarterly, 72(4), 472–496.

Dumontier M,Gray A, Marshall Metal. (2016). The healthcare and life sciences community profile for dataset descriptions. PeerJ. 4:e2331. doi:10.7717/peerj.2331.

Díaz, A., García, A., & Gervás, P. (2008). User-centred versus system-centred evaluation of a personalization system. Information processing and management, 44(3), 1293-1307.

Ebling, M. R., & John, B. E. (2000). On the contributions of different em- pirical data in usability testing. Paper presented at the 3rd conference on designing interactive systems: processes, practices, methods, and techniques, August 17-19, Brooklyn, NY, USA

Elzein, N. M., Majid, M. A., Hashem, I. A. T., Yaqoob, I., Alaba, F. A., & Imran, M. (2018). Managing big RDF data in clouds: Challenges, opportunities, and solutions. Sustainable Cities and Society, 39, 375–386. https://doi.org/10.1016/j.scs.2018.02.019

Faniel, I. and Jacobsen, T. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Computer Supported Cooperative Work (CSCW), 19(3-4), pp.355-375.

Frechtling, J. (2002). The 2002 User-Friendly Handbook for Project Evaluation, Chapter 3. National Science Foundation.

Fuhr, N., Govert, N., Kazai, G., Lalmas, M. (2002). INEX: initiative for the evaluation of XML retrieval. In: Proceedings of the ACM SIGIR 2002 Workshop on XML and Information Retrieval

Fuhr, N., Grossjohann, K. (2001). XIRQL: a query language for in- formation retrieval in XML documents. In: Proceedings of 24th Annual International ACM SIGIR Conference on Re- search and Development in Information Retrieval, pp. 172–180

Gould, J.D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. Communications of the ACM 28(3), 300–311.

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata, 9(3–4), 194–212. https://doi.org/10.1080/19386380903405090

Grethe, J. (2015). Data Identifiers Recommendation. https://biocaddie.org/sites/default/files/shared-documents/100_grethe.pdf

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. Journal of the Association for Information Science and Technology, 70(5), 419–432. https://doi.org/10.1002/asi.24165

Gómez, N.-D., Méndez, E., & Hernández-Pérez, T. (2016). Data and metadata research in the social sciences and humanities: An approach from data repositories in these disciplines. El Profesional de La Información, 25(4), 545. https://doi.org/10.3145/epi.2016.jul.04

Günther, A., & Dehnhard, I. (2015). From Publishing to Communicating Research Data. Septentrio Conference Series, (5). doi: 10.7557/5.3663

Gustafson, N., & Ng, Y.-K. (2008). Augmenting Data Retrieval with Information Retrieval Techniques by Using Word Similarity. In Lecture Notes in Computer Science (pp. 163–174). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69858-6_17

Hakala, J. (2010). Persistent identifiers - an overview. http://www.persid.org/downloads/PI-intro-2010-09-22.pdf

Hammond, M. & Wellington, J. (2013). Research methods : the key concepts. London New York: Routledge.

Hartter, J., Ryan, S.J., MacKenzie, C.A., Parker, J.N. and Strasser, C.A. (2013). Spatially explicit data: stewardship and ethical challenges in science, PLoS Biology, Vol. 11 No. 9, p. e1001634, available at: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001634

Hayslett, M. (2015). Data world does not lack standards. Journal of Librarianship & Scholarly Communication, 3(2), 1–5. doi: 10.7710/2162-3309.1245

Helbig, K. (2016). Research data management training for geographers: First impressions. ISPRS International Journal of Geo-Information, 5(4), 40. doi:10.3390/ijgi5040040

Henderson, R. D., Smith, M. C., Podd, J., & Varela-Alvarez, H. (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. Ergonomics, 38(10), 2030-2044.

Henty, M., Weaver, B., Bradbury, S. J., et al. (2008). Investigating Data Management Practices in Australian Universities. Australia: APSR.

Hey, T., Tansley, S., Tole, K. (2009). The fourth paradigm: Data-intensive scientific discovery. s.l. Microsoft Cooperation. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf. Accessed 2010 Sep 20. https://www.icsu-wds.org/files/Domain_Repositories_Call_to_Action_16_Sept_2013.pdf

ICPSR Interuniversity Consortium for Political and Social Research. (2013). Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories

IWGDD Interagency Working Group of Digital Data. (2009). “Harnessing the Power of Digital Data for Science and Society.” Last modified January 14. https://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Jansen, B. J., Spink, A. (2006). How are we searching the world wide web? A comparison of nine search engine transaction logs. Information Processing and Management, 42(1), 248-263.

Kaiser, J. (2013). Rare Cancer Successes Spawn ‘Exceptional’ Research Efforts.

Kay, J. (2001). User modeling for adaptation. In C. Stephanidis (Ed.), User interfaces for all: Concepts, methods, and tools. (pp. 271–294). Mahwah, NJ: Lawrence Erlbaum Associates.

Kennan, M. A., & Markauskaite, L. (2015). Research Data Management Practices: A Snapshot in Time. International Journal of Digital Curation, 10(2), 69–95. https://doi.org/10.2218/ijdc.v10i2.329

Kim, J., Xue, X., & Croft, W. B. (2009). A Probabilistic Retrieval Model for Semistructured Data. In Lecture Notes in Computer Science (pp. 228–239). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00958-7_22

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis. D-Lib Magazine, 23(3/4). https://doi.org/10.1045/march2017-kindling

Kiteley, R. & Stogdon, C. (2014). Literature Reviews In Social Work. Los Angeles, CA: SAGE.

Kitzinger, J. (1994). The methodology of Focus Groups: the importance of interaction between research participants. Sociology of Health and Illness, 16(1), 103–121. https://doi.org/10.1111/1467-9566.ep11347023

Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, et al. (2014). Scholarly context not found: one in five articles suffers from reference rot. PLoS One. 9:e115253.

Koltay, T. (2015). Data literacy for researchers and data librarians. Journal of Librarianship and Information Science. DOI: 10.1177/0961000615616450.

Kotonya, G., & Sommerville, I. (1998). Requeriments Engineering Processes and Techniques. New York: Wiley and Sons.

Kouper, I., Stacy R. Konkiel, Jennifer A. Liss, and Juliet L. Hardesty. (2013). Collaborate, automate, prepare, prioritize: creating metadata for legacy research data. In Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications (DCMI'13), Muriel Foulonneau and Kai Eckert (Eds.). Dublin Core Metadata Initiative 41-46.

Kuncheva, L. (2014). Combining pattern classifiers : methods and algorithms. Hoboken, NJ: Wiley.

Labaw, P. J. (1981). Advanced questionnaire design. Cambridge: Abt books

Lazar, J. (2006). Web usability: A user-centered design approach. Boston: Pearson, Addison Wesley.

Lotz, T., Nieschulze, J., Bendix, J., Dobbermann, M. and König-Ries, B. (2012). Diverse or uniform? – Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. Ecological Informatics, Vol. 8, pp. 10-19, available at: www.sciencedirect.com/science/article/pii/S157495411100094X

Lynch, C. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age. ARL Bimon Rep 226. http://www.arl.org/newsltr/226/ir.html

Maciaszek, L. (2007). Requirements Analysis and Systems Design (3rd ed.). New Jersey: Pearson Education UK.

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

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B. (2008). Advancing ecological research with ontologies. Trends Ecol. Evol. 23, 159e168.

Maley, C., Baum, N. (2010). Getting to the Top of Google: Search Engine Optimization, The Journal of Medical Practice Management, MPM, 25(5), 301-303.

Mannheimer, S., Sterman, L. and Borda, S. (2016). Discovery and reuse of open Datasets: An exploratory study, Journal of eScience Librarianship. doi:10.7191/jeslib.2016.1091.

Markauskaite, L. (2010). Digital media, technologies and scholarship: Some shapes of eResearch in educational inquiry. The Australian Educational Researcher, 37(4), 79–101. doi:10.1007/BF03216938

Markauskaite, L., Kennan, M. A., Richardson, J., Aditomo, A., & Hellmers, L. (2012). Investigating eResearch: collaboration practices and future challenges. In A. Juan, T. Daradoumis, M. Roca, S. Grasman, & J. Faulin (Eds.), Collaborative and distributed e-Research: Innovations in technologies, strategies and applications (pp. 1–33). Hershey, PA: IGI. doi:10.4018/978-1-4666-0125-3.ch001

Martínez-Alcalá, C. I., Muñoz, M., & Monguet-Fierro, J. (2013). Design and Customization of Telemedicine Systems. Computational and Mathematical Methods in Medicine, 2013, 1–16. https://doi.org/10.1155/2013/618025

McMurry, J., Juty, N., Blomberg, N., Burdett, A., Conlin, T., Conte, N., … Parkinson, H. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. Cold Spring Harbor Laboratory. https://doi.org/10.1101/117812

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., & Sansone, S.-A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database, 2016, baw075. https://doi.org/10.1093/database/baw075

Mitra, B., & Craswell, N. (2018). An Introduction to Neural Information Retrieval t. Foundations and Trends® in Information Retrieval, 13(1), 1–126. https://doi.org/10.1561/1500000061

Morris, R. C. T. (1994). Toward a user-centered information service. Journal of the American Society for Information Science, 45(1), 20–30. https://doi.org/10.1002/(sici)1097-4571(199401)45:1<20::aid-asi3>3.0.co;2-n

Murray-Rust, P. (2008). Open Data in Science. Serials Review, Vol. 34 No. 1, pp. 52 64.

Mückschel, C., J. Nieschulze, C. Weist, B. Sloboda & W. Köhler. (2007). Herausforderungen, Probleme und Lösungsansätze im Datenmanagement von Sonderforschungsbereichen. In: eZAI (elektronische Zeitschrift für Agrarinformatik) 2, 1-16.

National Information Standards Organization. (2004). Understanding Metadata. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Nelson, B. (2009). Empty archives, Nature, Vol. 461 No. 7261, pp. 160-163.

Park, J., & Yi, M. Y. (2016). Graph-based retrieval model for semi-structured data. In 2016 International Conference on Big Data and Smart Computing (BigComp). IEEE. https://doi.org/10.1109/bigcomp.2016.7425948

PARSE.Insight. (2010). https://libereurope.eu/parse-insight-survey-report-available/

Patel, Dimple. (2016) "Research data management: a conceptual framework", Library Review, Vol. 65 Issue: 4/5, pp.226-241, https://doi.org/10.1108/LR-01-2016-0001

Palmer, C. L., & Cragin, M. H. (2009). Scholarship and disciplinary practices. Annual Review of Information Science and Technology, 42(1), 163–212. https://doi.org/10.1002/aris.2008.1440420112

Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Glimm, B. (Eds.). (2010). The Semantic Web – ISWC 2010. Lecture Notes in Computer Science. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17746-0

Patton, M. Q. (2002). Qualitative research & evaluation methods (3rd ed.).

Paton, N. W. (2008). Managing and sharing experimental data: standards, tools and pitfalls. Biochemical Society Transactions, 36(1), 33–36. https://doi.org/10.1042/bst0360033

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

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE, 9(8), e104798. https://doi.org/10.1371/journal.pone.0104798

Petrelli, D. (2008). On the role of user-centred evaluation in the advance- ment of interactive information retrieval. Information processing and management, 44(1), 22-38.

Pickard, A. J. (2013). Research methods in information. London: Facet Publishing.

Pienta, A. M., George, C. Alter., Jared A. Lyle. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. http:// deepblue.lib.umich.edu/bitstream/handle/2027.42/78307/pienta_alter_lyle_100331 .pdf?sequence=1.

Ponte, J. M., and Croft, W. B. (1998). A language modeling approach to information retrieval. pages 275–281, New York, NY, ACM.

Qin, J., Ball, A., & Greenberg, J. (2012). Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. International Conference on Dublin Core and Metadata Applications, , 62-71. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3660

Qin, J. (2013). Infrastructure, Standards, and Policies for Research Data Management. In: Sharing of Scienti c and Technical Resources in the Era of Big Data: e Proceedings of COINFO 2013, pp. 214-219. Beijing: Science Press.

Razum, M. (2011). Systeme und Systemarchitekturen für das Datenmanagement; in Büttner, S., Hobohm, H.-C. and Müller, L. (Eds), Handbuch Forschungsdatenmanagement, Bock u. Herchen, Bad Honnef, pp. 123-138.

RDA Research Data Alliance. (2014). The Data harvest: How sharing research data can yield knowledge, jobs and growth. An RDA Europe report (December 2014). https://rd-alliance.org/sites/default/files/attachment/The%20Data%20Harvest%20Final.pdf, last accessed 2017/06/11.

Research Information Network (RIN). (2008). To Share or Not To Share: Publication and Quality Assurance Of Research Data Outputs.; 2008:48. Available at: http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf. Accessed June 25, 2018.

Rice, R. (2009). DISC-UK DataShare project: Final report. Retrieved from Jisc repository: http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf

Richardson, M., Dominowska, E., Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. Proceedings of the 16th international conference on World Wide Web - WWW 07 (2007), 521–530

Rumsey, S. and Jefferies, N. (2013). Challenges in building an institutional research data catalogue', International Journal of Digital Curation, 8(2), pp. 205–214. doi: 10.2218/ijdc.v8i2.284.

Rugg, G. & Petre, M. (2007). A gentle guide to research methods. Maidenhead New York: Open University Press.

Ryan, G. W., & Bernard, H. R. (2003). Techniques to identify themes. Field Methods,15(1), 85–109. doi:10.1177/1525822x02239569

Satzinger, J., Burd, S., & Jackson, R. (2016). Systems analysis and design in a changing world (7th ed.). Australia: Cengage Learning.

Satzinger, J., Jackson, R., & Burd, S. (2012). Introduction to systems analysis and design (6th ed.). Australia: Cengage Learning.

Simons, N. and Richardson, J. (2013). New Content in Digital Repositories: The Changing Research Landscape. Chandos Publishing Oxford, Oxford.

Simons, Natasha. (2012). Implementing DOIs for Research Data. D-Lib Magazine 18(5/6). http://doi.org/10.1045/may2012-simons

Simukovic, E.; Kindling, M.; Schirmbacher, P. (2015). Umfrage zum Umgang mit Digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. Available online: http://nbn-resolving.de/urn:nbn:de:kobv:11-100213001

Spink, A., Wolfram, D., Jansen, B.J., Saracevik, T. (2001). Searching the web: the public and their queries. Journal of the American Society for Information Science, 53(2), 226-234.

Stephanidis, C. (2001). User interfaces for all: Concepts, methods, and tools. Mahwah, NJ: Lawrence Erlbaum Associates.

Strasser, C. (2015). Research data management: a primer publication of the National Information Standards organization. Baltimore, MD: NISO

Tashakkori, A., & Creswell, J. W. (2007). Editorial: Exploring the Nature of Research Questions in Mixed Methods Research. Journal of Mixed Methods Research, 1(3), 207-211. https://doi.org/10.1177/1558689807302814

Taylor, R. S. (2015). Question-Negotiation and Information Seeking in Libraries. College & Research Libraries, 76(3), 251–267. https://doi.org/10.5860/crl.76.3.251

Tenenbaum, J.D., Sansone,S.-A. and Haendel,M. (2014). A sea of standards for omics data: sink or swim? J. Am. Med. Inform. Assoc., 21, 200–203.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tracy, S. (2013). Qualitative research methods : collecting evidence, crafting analysis, communicating impact. Chichester, West Sussex, UK: Wiley-Blackwell.

The Economist. (2017). The world's most valuable resource is no longer oil, but data. Retrieved 4 January 2019, from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

The Royal Society. (2012). Exploiting data revolution is key to scientific and economic progress. https://royalsociety.org/news/2012/science-open-enterprise/

Thoegersen, Jennifer L. (2015). "Examination of Federal Data Management Plan Guidelines." Journal of eScience Librarianship 4(1): Article doi.org/10.7191/jeslib.2015.1072

Tsang, Daniel C. (2013). Research Data Management: Tools & Services. Presentation prepared for a Town Hall for faculty on e-research and research computing, 7 January 2013, University of California, Irvine.

Uhlir, Paul F., and Daniel Cohen. (2011). Internal Document. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences.

UK Research and Innovation. (2016). Concordat on Open Research Data. Available at: https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf. Accessed July 19, 2018.

Van der Geest, T., Jansen, J., Mogulkoç, E., De Vries, P., & De Vries, S. (2008). Segmentation and e-government: A literature review. Enschede: Telematica institute.

Van Noorden, R. (2013). Data-sharing: Everything on display. Nature, 500(7461), 243–245. https://doi.org/10.1038/nj7461-243a

Van Oostendorp, H., & De Mul, S. (1999). Learning by exploration: Think- ing-aloud while exploring an information system. Instructional science, 27(3/4), 269-284.

Velden T., Lagoze C. (2009). Communicating chemistry. Nat Chem 1(9):673–678. doi:10.1038/nchem.448

Velsen, L. V. (2011). User-Centered design for personalization. Unpublished. https://doi.org/10.13140/2.1.3843.0081

Verbakel, E., & Grootveld, M. (2016). Essentials 4 Data Support. IFLA Journal, 42(4), 278–283. https://doi.org/10.1177/0340035216674027

Vision, Todd J. 2010. Open Data and the Social Contract of Scientific Publishing. BioScience 60(5): 330-331. http://doi.org/10.1525/bio.2010.60.5.2

Walker, W., & Keenan, T. (2015). Going beyond availability: Truly accessible research data. Journal of Librarianship & Scholarly Communication, 3(2), 1–8. doi: 10.7710/2162-3309.1223

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8 (7):e67332. doi:10.1371/journal.pone.0067332.

Weber, A., Piesche, C. (2016). Requirements on long-term accessibility and preservation of research results with particular regard to their provenance. ISPRS Int. J. Geo-Inf. 5, 49

Weibel, S. (2005). The Dublin Core: A Simple Content Description Model for Electronic Resources. Bulletin of the American Society for Information Science and Technology, 24(1), 9–11. https://doi.org/10.1002/bult.70

Weller, T. and Monroe-Gulick, A. (2014). Understanding methodological and disciplinary differences in the data practices of academic researchers", Library Hi Tech, Vol. 32 No. 3, pp. 467-482. https://doi.org/10.1108/LHT-02-2014-0021

White, H. C. (2014). Descriptive metadata for scientific data repositories: A comparison of information scientist and scientist organizing behaviors. Journal of Library Metadata, 14(1), 24–51. doi: 10.1080/19386389.2014.891896

Whyte A., Tedds J. (2011). Making the Case for Research Data Management. Digital Curation Centre. dccacuk. Available at: http://www.dcc.ac.uk/resources/briefing-papers/makingcase-rdm. Accessed June 12, 2018.

Wiley, C. (2014). Metadata use in research data management. Bulletin Of The Association For Information Science And Technology, 40(6), 38-40. doi: 10.1002/bult.2014.1720400612

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505–1520. https://doi.org/10.1002/asi.22683

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing Data Curation Profiles. International Journal of Digital Curation, 4(3), 93–103. https://doi.org/10.2218/ijdc.v4i3.117

Wu, M., Marian, A. (2011). A framework for corroborating answers from multiple web sources. Information Systems, 36(2), 431-49

References

Wu, S., Worrall, A., Stvilia, B., et. al. (2006). Exploring Data Practices of the Earthquake Engineering Community. In iConference 2016 Proceedings. iSchools. https://doi.org/10.9776/16187

Xia, Y. & Li, J. (2009). User-Centered & Experience Design trend overview. In 2009 IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design. IEEE. https://doi.org/10.1109/caidcd.2009.5375255

Zhu, X., & Liao, J. (2007). Web usability: A user-centered design approach. Journal of the American Society for Information Science and Technology, 58(7), 1066–1067. https://doi.org/10.1002/asi.20588

Zimmermann, D., & Grötzbach, L. (2007). A Requirement Engineering Approach to User Centered Design. In Human-Computer Interaction. Interaction Design and Usability (pp. 360–369). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73105-4_40

APPENDIX I

Table 1. Summary of keywords and search hits

DISCIPLINE	KEYWORDS	DATA REPOSITORIES SEARCHED IN	DATA NUMBER OF HITS	PUBLICATIONS (WEB OF KNOWLEDGE) NUMBER OF HITS
Arts & Humanities	art museums	UK Data Service	81	9,603
	nineteenth century	UK Data Service	138	33,494
	"world war"	UK Data Service	74	39,335
	medieval	UK Data Service	68	53,494
	popular music	UK Data Service	13	5,296
Social Sciences	unemployment	UK Data Service	1680	30,690
	cognition	UK Data Service	335	110,631
	imprisonment	UK Data Service	22	3,761
	"labour law"	UK Data Service	48	450
	"trade union"	UK Data Service	1221	2,702
Natural Sciences	marine life	UK Data Service & DataOne	63	17,704
	"climate change"	UK Data Service	230	151,303
	"renewable energy"	DataOne	20	43,237
	"ultraviolet light"	DataOne	12	16,872
	"oxidative phosphorlyation"	DataOne	29	15,837
Computer & Information Science	search behavior	Dryad	48	44,439
	face recognition	UK Data Service	76	43,220
	computer vision	Dryad	16	33,590
	research data sharing	Dryad	88	21,611
	social media data	Dryad	17	110,631

Table 2. The main experimental data

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
Discipline	**Arts & Humanities**				
Keyword 1: art museums	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	15.420	2.810	PDF	0.253
File 2	Tab-delimited	4.780	0.435	PDF	0.894
File 3	Tab-delimited	1.980	0.111	PDF	0.378
File 4	Tab-delimited	1.240	0.220	PDF	0.381
File 5	Tab-delimited	13.160	0.033	PDF	0.970
File 6	ZIP	0.595	0.051	PDF	0.843
File 7	Tab-delimited	0.450	0.233	PDF	1.300
File 8	XLS	1.760	0.846	PDF	0.482
File 9	Tab-delimited	16.590	0.070	PDF	1.200
File 10	SQL	1.100	0.162	PDF	1.500
Total		*57.075 MB*	*4.971 MB*		*8.201 MB*
Average		*5.708 MB*	*0.497 MB*		*0.820 MB*
Keyword 2: nineteenth century	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	2.270	0.700	PDF	0.146
File 2	Tab-delimited	3.920	0.824	PDF	0.398
File 3	ZIP	0.537	0.379	PDF	0.464
File 4	Tab-delimited	0.820	0.464	PDF	0.880
File 5	Tab-delimited	0.400	0	PDF	0.130
File 6	Tab-delimited	0.740	0	PDF	3.500
File 7	Tab-delimited	12.400	0.997	PDF	0.542
File 8	Tab-delimited	0.920	0.194	PDF	0.676
File 9	Tab-delimited	1.360	0.023	PDF	0.479
File 10	ZIP	2.000	0.030	PDF	3.200
Total		*25.367 MB*	*3.611 MB*		*10.415 MB*

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
Average		*2.537 MB*	*0.361 MB*		*1.042 MB*
Keyword 3: "world war"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	RTF	2.950	0.850	PDF	0.428
File 2	Tab-delimited	0.110	0	PDF	0.578
File 3	RTF	5.850	0.128	PDF	0.707
File 4	Tab-delimited	0.190	0.054	PDF	0.418
File 5	ZIP	0.535	0.037	PDF	0.110
File 6	Tab-delimited	3.760	0.045	PDF	0.208
File 7	Tab-delimited	0.640	0.157	PDF	0.403
File 8	RTF	7.170	1.027	PDF	0.222
File 9	Tab-delimited	0.850	0.047	PDF	0.902
File 10	Tab-delimited	35.600	1.571	PDF	1.100
Total		*57.655 MB*	*3.916 MB*		*5.076 MB*
Average		*5.766 MB*	*0.392 MB*		*0.508 MB*
Keyword 4: medieval	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	0.900	0	PDF	0.796
File 2	ZIP	18.500	0.696	PDF	0.155
File 3	Tab-delimited	0.100	0.036	PDF	0.199
File 4	Tab-delimited	10.500	0.137	PDF	3.500
File 5	RTF	0.840	0.030	PDF	0.257
File 6	RTF	0.650	0.060	PDF	0.109
File 7	Tab-delimited	0.440	0	PDF	0.159
File 8	Tab-delimited	0.320	0.030	PDF	0.135
File 9	Tab-delimited	2.420	0.022	PDF	1.100
File 10	XLS	15.860	0.036	PDF	4.500
Total		*50.530 MB*	*1.047 MB*		*10.910 MB*
Average		*5.053 MB*	*0.105 MB*		*1.091 MB*

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
Keyword 5: popular music	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	17.300	2.088	PDF	0.521
File 2	Tab-delimited	2.030	0.137	PDF	2.200
File 3	Tab-delimited	1.130	0.105	PDF	0.189
File 4	Tab-delimited	9.920	0.247	PDF	2.700
File 5	Tab-delimited	0.380	0.045	PDF	1.000
File 6	RTF	1.580	0.109	PDF	0.656
File 7	ZIP	35.600	4.651	PDF	0.168
File 8	Tab-delimited	8.820	2.268	PDF	0.368
File 9	Tab-delimited	5.770	0.164	PDF	1.000
File 10	ZIP	1.000	0	PDF	1.200
Total		*83.530 MB*	*9.814 MB*		*10.002 MB*
Average		*8.353 MB*	*0.981 MB*		*1.000 MB*
Discipline	**Social Sciences**				
Keyword 1: unemployment	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	1.290	0.495	PDF	0.509
File 2	Tab-delimited	11.840	6.101	PDF	0.170
File 3	Tab-delimited	2.130	3.526	PDF	0.808
File 4	Tab-delimited	3.060	3.811	PDF	0.657
File 5	RTF	0.250	0.097	PDF	0.556
File 6	XLS	1.540	0.171	PDF	0.412
File 7	Tab-delimited	1.750	0.796	PDF	0.271
File 8	Tab-delimited	3.690	0.916	PDF	0.594
File 9	Tab-delimited	0.780	0.526	PDF	0.309
File 10	RTF	4.250	0.260	PDF	0.263
Total		*30.580 MB*	*16.699 MB*		*4.549 MB*
Average		*3.059 MB*	*1.670 MB*		*0.455 MB*

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
Keyword 2: cognition	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	SAV	0.173	0.290	PDF	0.411
File 2	Tab-delimited	4.070	0.346	PDF	1.600
File 3	SAV	0.250	0	PDF	0.449
File 4	SAV	0.520	0.067	PDF	0.543
File 5	ZIP	4.000	4.066	PDF	5.400
File 6	Tab-delimited	0.210	0.322	PDF	0.294
File 7	Tab-delimited	86.280	7.510	PDF	1.200
File 8	Tab-delimited	7.010	3.026	PDF	0.477
File 9	XLS	1.100	0.280	PDF	5.100
File 10	Tab-delimited	13.200	0.686	PDF	0.700
Total		*116.813 MB*	*16.593 MB*		*16.174 MB*
Average		*11.681 MB*	*1.659 MB*		*1.612 MB*
Keyword 3: imprisonment	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	XLS	1.000	1.807	PDF	0.785
File 2	Tab-delimited	1.820	0.347	PDF	0.181
File 3	Tab-delimited	0.290	0	PDF	0.466
File 4	Tab-delimited	1.650	0.758	PDF	0.689
File 5	Tab-delimited	0.150	0.519	PDF	0.397
File 6	Tab-delimited	0.670	0	PDF	0.702
File 7	ZIP	1.335	0.306	PDF	0.387
File 8	Tab-delimited	7.840	0.782	PDF	0.668
File 9	Tab-delimited	2.040	1.059	PDF	0.618
File 10	Tab-delimited	1.570	0.490	PDF	0.135
Total		*18.365 MB*	*6.068 MB*		*5.028 MB*
Average		*1.837 MB*	*0.607 MB*		*0.503 MB*
Keyword 4: "labour law"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
File 1	XLS	0.846	0.330	PDF	0.364
File 2	Tab-delimited	0.960	0.900	PDF	0.285
File 3	Tab-delimited	1.110	0.899	PDF	0.248
File 4	Tab-delimited	3.990	0.170	PDF	0.469
File 5	XLS	4.070	3.257	PDF	0.667
File 6	Tab-delimited	1.620	1.277	PDF	0.152
File 7	RTF	0.490	0.227	PDF	0.540
File 8	Tab-delimited	1.720	2.240	PDF	0.820
File 9	Tab-delimited	0.380	1.224	PDF	0.345
File 10	XLS	1.480	1.080	PDF	0.207
Total		*16.666 MB*	*11.604 MB*		*4.097 MB*
Average		*1.667 MB*	*1.160 MB*		*0.410 MB*
Keyword 5: "trade union"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	0.660	0.006	PDF	0.398
File 2	Tab-delimited	0.840	0.288	PDF	0.268
File 3	Tab-delimited	0.700	0	PDF	1.400
File 4	Tab-delimited	0.820	0	PDF	0.405
File 5	Tab-delimited	6.100	129.613	PDF	0.392
File 6	Tab-delimited	1.770	3.069	PDF	1.600
File 7	Tab-delimited	0.550	1.050	PDF	1.200
File 8	RTF	3.790	1.672	PDF	0.601
File 9	RTF	1.310	1.115	PDF	0.118
File 10	RTF	4.190	1.847	PDF	1.100
Total		*20.730 MB*	*138.660 MB*		*7.482 MB*
Average		*2.073 MB*	*13.866 MB*		*0.748 MB*
Discipline	**Natural Sciences**				
Keyword 1: marine life	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
File 1	Tab-delimited	1.000	0.370	PDF	0.177
File 2	CSV	76.200	132.794	PDF	0.071
File 3	Tab-delimited	0.720	0.289	PDF	1.900
File 4	Tab-delimited	23.280	11.707	PDF	0.483
File 5	Tab-delimited	9.030	7.346	PDF	3.900
File 6	Tab-delimited	15.080	7.390	PDF	1.200
File 7	Tab-delimited	11.490	2.528	PDF	2.600
File 8	Tab-delimited	12.150	1.578	PDF	1.100
File 9	Tab-delimited	1.260	0.191	PDF	0.479
File 10	Tab-delimited	6.860	1.921	PDF	3.000
Total		*157.070 MB*	*166.114*		*14.910 MB*
Average		*15.707 MB*			*1.491 MB*
Keyword 2: "climate change"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	Tab-delimited	0.730	4.417	PDF	0.570
File 2	Tab-delimited	3.490	1.028	PDF	0.765
File 3	Tab-delimited	1.940	1.094	PDF	0.298
File 4	DOC	0.276	2.726	PDF	1.900
File 5	RTF	1.690	0.685	PDF	0.270
File 6	SAV	0.311	0	PDF	1.200
File 7	ZIP	1.400	0.123	PDF	10.700
File 8	XLS	0.403	0.002	PDF	5.900
File 9	XLS	0.113	0.012	PDF	2.600
File 10	SAV	6.200	1.444	PDF	0.764
Total		*16.553 MB*	*11.531 MB*		*24.967 MB*
Average		*1.655 MB*	*1.153 MB*		*2.497 MB*
Keyword 3: "renewable energy"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	ZIP	62.910	15.430	PDF	1.300
File 2	Tab-delimited	4.560	4.481	PDF	4.000

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
File 3	XLS	139.020	5.816	PDF	3.100
File 4	CSV	1535.030	30.987	PDF	2.600
File 5	ZIP	36.200	7.353	PDF	0.851
File 6	Tab-delimited	1.360	1.237	PDF	0.513
File 7	Tab-delimited	282.390	10.598	PDF	8.200
File 8	Tab-delimited	0.080	0	PDF	8.800
File 9	CSV	5369.900	4.656	PDF	5.000
File 10	Tab-delimited	149.130	3.760	PDF	1.700
Total		*7.580 GB*	*84.318 MB*		*36.064 MB*
Average		*758 MB*	*8.432 MB*		*3.606 MB*
Keyword 4: "ultraviolet light"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	ZIP	101.000	2.147	PDF	0.676
File 2	CSV	16.000	1.939	PDF	2.200
File 3	Octet Stream	2.139	0.020	PDF	1.700
File 4	CSV	0.077	0.070	PDF	2.600
File 5	CSV	0.088	0.07	PDF	2.000
File 6	ZIP	4832.000	3.320	PDF	2.200
File 7	XLS	6.000	0.678	PDF	4.300
File 8	CSV	0.010	0	PDF	2.300
File 9	TXT & XML	0.320	0.05	PDF	1.200
File 10	ZIP & TXT	1.348	0.157	PDF	0.730
Total		*4.959 GB*	*8.451 MB*		*19.906 MB*
Average		*495.900 MB*	*0.845 MB*		*1.991 MB*
Keyword 5: "oxidative phosphorylation"	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	TXT & OOXML	178.068	1.120	PDF	1.600
File 2	TXT & CSV	11.356	2.052	PDF	1.600
File 3	OOXML	0.232	0.247	PDF	1.400

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
File 4	OOXML	0.267	0.203	PDF	4.600
File 5	ZIP	210.000	3.643	PDF	1.200
File 6	OOXML	2.210	1.537	PDF	2.300
File 7	CSV	0.091	0	PDF	0.954
File 8	bitstream	0.020	0	PDF	1.200
File 9	bitstream	0.160	0.545	PDF	2.400
File 10	RAR Compressed	0.014	0	PDF	1.700
Total		*402.418 MB*	*9.347 MB*		*18.954 MB*
Average		*40.242 MB*	*0.935 MB*		*1.895 MB*

Discipline	**Computer & Information Sciences**				
Keyword 1: search behavior	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	TXT	0.041	0.155	PDF	0.413
File 2	XLS	1.377	0.850	PDF	0.293
File 3	ZIP	222.300	3.089	PDF	0.934
File 4	7z	10.672	1.354	PDF	0.269
File 5	TXT	111.070	7.608	PDF	0.341
File 6	XLS	0.072	0.577	PDF	2.100
File 7	XLS	0.176	0.019	PDF	0.839
File 8	ZIP	5589.360	0.977	PDF	0.689
File 9	XLS, ENV, PED, MAP, & ARP	623.712	1.662	PDF	0.329
File 10	XLS	0.843	0.779	PDF	1.100
Total		*6.560 GB*	*17.070 MB*		*7.307 MB*
Average		*656 MB*	*1.707 MB*		*0.731 MB*
Keyword 2: face recognition	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	XLS	0.381	0.781	PDF	0.800

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
File 2	XLS	1.770	0.070	PDF	0.645
File 3	TXT	0.540	1.608	PDF	1.300
File 4	SAV & SPU	1.100	0.777	PDF	2.000
File 5	ZIP	5030.000	4.871	PDF	0.626
File 6	ZIP	22.000	1.763	PDF	3.000
File 7	SAV	0.172	0.021	PDF	1.400
File 8	ZIP	8850.000	18.804	PDF	0.559
File 9	RTF	1.640	0	PDF	0.721
File 10	ZIP	2.800	0.260	PDF	4.300
Total		*13.910 GB*	*28.955 MB*		*15.351 MB*
Average		*1.391 GB*	*2.896 MB*		*1.535 MB*
Keyword 3: computer vision	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	ZIP	2903.100	9.559	PDF	6.900
File 2	CSV & ZIP	519.640	16.997	PDF	3.900
File 3	ZIP	1114.967	21.320	PDF	1.600
File 4	CSV	0.016	0.939	PDF	1.800
File 5	ZIP	228.900	13.394	PDF	1.900
File 6	XLS	0.191	1.299	PDF	0.749
File 7	MATLAB	59.650	16.193	PDF	0.873
File 8	OBO	1.677	2.025	PDF	3.800
File 9	ZIP	8453.200	5.552	PDF	4.100
File 10	TXT & XLS	16.468	4.934	PDF	2.200
Total		*13.298 GB*	*92.212 MB*		*27.822 MB*
Average		*1.330 GB*	*9.221 MB*		*2.782 MB*
Keyword 4: research data sharing	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	CSV	1.268	0.383	PDF	0.474
File 2	XLS	0.023	0.309	PDF	0.366
File 3	SAV	0.317	0.232	PDF	0.448

	DATA RETRIEVAL			TEXT/INFORMATION RETRIEVAL	
File 4	CSV	0.496	0.665	PDF	0.975
File 5	ZIP	1.265	0.380	PDF	0.515
File 6	ZIP	0.412	0.690	PDF	0.330
File 7	XLS	0.171	0.845	PDF	0.267
File 8	CSV	1.843	1.195	PDF	0.887
File 9	CSV	3.672	0.317	PDF	0.358
File 10	CSV	0.674	0.585	PDF	0.585
Total		*10.141 MB*	*5.601 MB*		*5.205 MB*
Average		*1.014 MB*	*0.560 MB*		*0.521 MB*
Keyword 5: social media data	**Data Format**	**Size (MB)**	**Documentation Size (MB)**	**Data Format**	**Size (MB)**
File 1	CSV	0.012	0	PDF	0.726
File 2	XLS	0.211	0.398	PDF	1.100
File 3	XLS	0.091	0.188	PDF	0.469
File 4	TAR	101.800	11.490	PDF	2.400
File 5	ZIP	4.885	4.977	PDF	2.100
File 6	TAR	18.530	2.483	PDF	0.748
File 7	TXT	36.079	11.341	PDF	0.578
File 8	CSV	0.005	0	PDF	0.735
File 9	TXT	1.616	1.807	PDF	0.326
File 10	ZIP	0.052	0	PDF	1.600
Total		*163.281 MB*	*32.684 MB*		*10.782 MB*
Average		*16.329 MB*	*3.268 MB*		*1.078 MB*

Table 3. Data Summary

Discipline	Keywords	Data Retrieval Average File Size (inc. documentation)	Text/Information Retrieval Average File Size	Approx. Ratio (x times as large)
Arts & Humanities	art museums	6.205 MB	0.820 MB	8 times
	nineteenth century	2.898 MB	1.042 MB	3 times
	"world war"	6.158 MB	0.508 MB	12 times
	medieval	5.158 MB	1.091 MB	5 times
Social Sciences	popular music	9.334 MB	1.000 MB	9 times
	unemployment	4.729 MB	0.455 MB	10 times
	cognition	13.340 MB	1.612 MB	8 times
	imprisonment	2.444 MB	0.503 MB	5 times
	"labour law"	2.827 MB	0.410 MB	7 times
	"trade union"	15.939 MB	0.748 MB	21 times
Natural Sciences	marine life	32.318 MB	1.491 MB	22 times
	"climate change"	2.808 MB	2.497 MB	1 time
	"renewable energy"	766.432 MB	3.606 MB	213 times
	"ultraviolet light"	496.745 MB	1.991 MB	250 times
	"oxidative phosphorlyation"	41.177 MB	1.895 MB	22 times
Computer & Information Science	search behavior	657.707 MB	0.731 MB	900 times
	face recognition	**1.394 GB**	1.535 MB	908 times
	computer vision	**1.339 GB**	2.782 MB	481 times
	research data sharing	1.574 MB	0.521 MB	3 times
	social media data	19.597 MB	1.078 MB	18 times

Appendix II

Base data literacy

You are invited to participate in a survey which aims to collect data about the data literacy of academics and research students in higher education institutions. From your responses we will be able to fully understand the current levels of awareness and gaps in knowledge which will help us develop appropriate data literacy training for the higher education community.

Please answer all the questions, and note that this survey is anonymous. It will take approximately 10 minutes to complete the entire survey. By completing this survey you are consenting to the use of your data for research and dissemination purposes. If you have any questions or comments as you are going through the survey, please contact ...

Thank you very much for your cooperation!

There are 26 questions in this survey

PART I: Demographic Information

1 []Your current primary role *

Please choose **only one** of the following:

- ○ Academic staff
- ○ Research student
- ○ Other

2 []Your age *

Please choose **only one** of the following:

- ○ 18-25
- ○ 26-35
- ○ 36-45
- ○ 46-55
- ○ 56-65
- ○ 65+
- ○ Don’t want to disclose

3 []Your discipline *

Please choose **only one** of the following:

- ○ Natural sciences:Mathematics
- ○ Natural sciences: Computer and information sciences
- ○ Natural sciences: Physical sciences
- ○ Natural sciences: Chemical sciences
- ○ Natural sciences: Earth and related environmental sciences
- ○ Natural sciences: Biological sciences
- ○ Engineering and technology: Civil engineering
- ○ Engineering and technology: Electrical engineering, electronic, engineering, information engineering
- ○ Engineering and technology: Mechanical engineering
- ○ Engineering and technology: Chemical engineering
- ○ Engineering and technology: Materials engineering
- ○ Engineering and technology: Medical engineering
- ○ Engineering and technology: Environmental engineering
- ○ Engineering and technology: Environmental biotechnology
- ○ Engineering and technology: Industrial biotechnology
- ○ Engineering and technology: Nano-technology
- ○ Medical and health sciences: Basic medicine
- ○ Medical and health sciences: Clinical medicine
- ○ Medical and health sciences: Health sciences
- ○ Medical and health sciences: Health biotechnology
- ○ Medical and health sciences: Materials engineering
- ○ Agricultural sciences: Agriculture, forestry, and fisheries
- ○ Agricultural sciences: Animal and dairy science
- ○ Agricultural sciences: Veterinary science
- ○ Agricultural sciences: Agricultural biotechnology
- ○ Social sciences: Psychology
- ○ Social sciences: Economics and business

- ○ Social sciences: Educational sciences
- ○ Social sciences: Sociology
- ○ Social sciences: Law
- ○ Social sciences: Political science
- ○ Social sciences: Social and economic geography
- ○ Social sciences: Media and communications
- ○ Humanities: History and archaeology
- ○ Humanities: Languages and literature
- ○ Humanities: Philosophy, ethics and religion
- ○ Humanities: Art (arts, history of arts, performing arts, music)
- ○ Other

In the field "Other", please comply with the classification structure. ie: Social science: political science

4 []Your legal gender *

Please choose **only one** of the following:

- ○ Male
- ○ Female
- ○ Other
- ○ Don't want to disclose

5 []How long have you been involved in research? *

Please choose **only one** of the following:

- ○ < 5 years
- ○ 5-10 years
- ○ 11-15 years
- ○ 16-20 years
- ○ > 20 years
- ○ I have never been involved in research

6 []Your country *

Please write your answer here:

7 []Your institution *

Please write your answer here:

PART II: Awareness of Data Management Issues

8 []Please indicate the file type of data that you normally use for your research *

Please choose **all** that apply:

- ☐ Standard office documents (text, spreadsheets, presentations, etc.)
- ☐ Structured scientific and statistical data (e.g. SPSS, GIS, etc.)
- ☐ Encoded text (XML, SGML, etc.)
- ☐ Internet and web-based data (webpages, e-mails, blogs, social network data, etc.)
- ☐ Databases (e.g. in Access, Oracle, MySQL, etc.)
- ☐ Images (JPEG, GIF, TIFF, PNG, etc.)
- ☐ Audio files
- ☐ Structured graphics (CAD, CAM, VRML, etc.)
- ☐ Raw (machine-generated) data
- ☐ Archived data (ZIP, RAR, ZAR, etc.)
- ☐ Software applications (modelling tools, editors, compilers, etc.)
- ☐ Source code (scripting, Java, C, C++, etc.)
- ☐ Configuration data (parameter settings, logs, library files, etc.)
- ☐ Non digital data (paper, films, slides, artefacts, etc.)
- ☐ Other:

9 []Which of the following better describes the volume of data you use for your research? *

Please choose **only one** of the following:

- ○ MB (megabyte)
- ○ GB (gigabyte)
- ○ TB (terabyte)
- ○ Other

10 []How do you usually get the data for your research? *

Please choose **all** that apply:

- ☐ Create new data
- ☐ From own research team/group at the university
- ☐ From own research network (or personal/professional connections)
- ☐ Always from one known source
- ☐ Always from multiple known sources
- ☐ Search from outside sources (please describe):

11 []How do you usually use data that you get from others/outside sources? *

Please choose **all** that apply:

- ☐ As it is without any problems
- ☐ With a bit of effort for some cleaning and/or modifications
- ☐ After spending a lot of time and efforts to make it usable for the project
- ☐ I do not use data from others/outside sources

12 []What type of data do you produce from your research? *

Please choose **all** that apply:

- ☐ Standard office documents (text, spreadsheets, presentations, etc.)
- ☐ Structured scientific and statistical data (e.g. SPSS, GIS, etc.)
- ☐ Encoded text (XML, SGML, etc.)
- ☐ Internet and web-based data (webpages, e-mails, blogs, social network data, etc.)
- ☐ Databases (e.g. in Access, Oracle, MySQL, etc.)
- ☐ Images (JPEG, GIF, TIFF, PNG, etc.)
- ☐ Audio files
- ☐ Structured graphics (CAD, CAM, VRML, etc.)
- ☐ Raw (machine-generated) data
- ☐ Archived data (ZIP, RAR, ZAR, etc.)
- ☐ Software applications (modelling tools, editors, compilers, etc.)
- ☐ Source code (scripting, Java, C, C++, etc.)
- ☐ Configuration data (parameter settings, logs, library files, etc.)
- ☐ Non digital data (paper, films, slides, artefacts, etc.)
- ☐ Other:

13 []Which of the following better describes the volume of data you produce from your research? *

Please choose **only one** of the following:

- ○ MB (megabyte)
- ○ GB (gigabyte)
- ○ TB (terabyte)
- ○ Other

14 []Where do you usually store the data you produce from your research? *

Please choose **all** that apply:

- ☐ Your own devices (your computer, your tablet, external drive, etc.)
- ☐ Cloud
- ☐ Central servers/repositories of the university
- ☐ Outside repositories
- ☐ Other:

15 []Do you normally assign any additional information to your research data *

Please choose **all** that apply:

- ☐ Administrative information (e.g. creator, date of creation, file name, access terms/restrictions, etc.)
- ☐ Discovery information (e.g. creator, funding body, project title, project ID, keywords, etc.)
- ☐ Technical information (e.g. file format, file size, software/hardware needed to use the data, etc.)
- ☐ Description of the data file (e.g. file/data structure, field tags/descriptions, application rules, etc.)
- ☐ No, I do not assign additional information to my research data

16 []Do you collaborate with other researchers and share data? *

Please choose **all** that apply:

- ☐ No
- ☐ Yes, with researchers in the same team
- ☐ Yes, with researchers in the same university
- ☐ Yes, with researchers in other institutions
- ☐ Any other (Please specify):

17 []Which of the following applies to your research data *

Please choose **all** that apply:

- ☐ My data is openly available to everyone
- ☐ My data is openly available only to my research team
- ☐ My data is available openly upon request
- ☐ My data has restricted access (e.g. only some parts of the dataset is accessible)
- ☐ My data is not available to anyone else

18 []Do you have any concerns for sharing data with others *

Please choose **all** that apply:

- ☐ No concerns
- ☐ Fear of losing the scientific edge
- ☐ Legal and ethical issues
- ☐ Misuse of data
- ☐ Misinterpretation of data
- ☐ Lack of resources (technical, financial, personnel, etc.)
- ☐ Lack of appropriate policies and rights protection
- ☐ Any other (Please specify):

19 []Please answer the following questions *

Please choose the appropriate response for each item:

	Yes	Uncertain	No
Does your institution have a Data Management Plan (DMP)?	○	○	○
Have you ever used a DMP for your research?	○	○	○
Do you have a DMP for your current research project(s)?	○	○	○
Do you think a DMP actually helps researchers in managing research data?	○	○	○
Are you familiar with the term metadata?	○	○	○
Do you think a formal training on metadata would be useful for managing research data?	○	○	○
Does your university have a prescribed metadata set for uploading data to a repository?	○	○	○
Does your research community use/recommend any standard file naming system?	○	○	○
Does your university have a	○	○	○

	Yes	Uncertain	No
standard/consistent file naming system?			
Do you use any standard style for citing research data?	○	○	○
Are you familiar with the concept of Digital Object Identifier (DOI)?	○	○	○
Does your university recommend any specific guideline for citing data (e.g. APA, Harvard, etc.)?	○	○	○
Have you got any unique researcher identification (like ORCID=Open Researcher and Contributor ID)?	○	○	○
Does your university actively encourage you to share data on open access (OA) mode?	○	○	○
Are you familiar with your university and/or funding body's requirements with regard to data storage?	○	○	○

20 []How often do you practice the following? *

Please choose the appropriate response for each item:

	Almost Always	Often	Sometimes	Rarely	Never
Using metadata standard for tagging your data	○	○	○	○	○
Using your own/in-house (your research team) tags and metadata	○	○	○	○	○
Using datasets that are tagged with standard metadata	○	○	○	○	○
Using file naming convention or standard	○	○	○	○	○
Having different versions of the same dataset(s)	○	○	○	○	○
Using systems/techniques for version control to easily recognise a specific version	○	○	○	○	○
Citing research data	○	○	○	○	○
Working with data that are generally in the public domain	○	○	○	○	○
Working with data that have restricted access?	○	○	○	○	○

21 []How strongly do you agree or disagree with the following *

Please choose the appropriate response for each item:

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I am familiar with the open access requirements	○	○	○	○	○
I am comfortable and willing to share my research data with others	○	○	○	○	○
I foresee no problems with sharing my research data?	○	○	○	○	○
I perceive data ethics could be an issue when research data is shared with others	○	○	○	○	○
I would like to store my research datasets beyond the lifetime of the project	○	○	○	○	○
Every university should have a Data Management Plan	○	○	○	○	○
Every university	○	○	○	○	○

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
should have a prescribed metadata set for uploading data into a repository					
Universities should recommend and use a standard file naming system	○	○	○	○	○

22 []In your opinion who should pay for storage and public access to the data set that you created? *

Please choose **all** that apply:

- ☐ Yourself/your team
- ☐ Your university
- ☐ The funding body
- ☐ A national body
- ☐ Other:

23 []Where should the data be stored for long term access? *

Please choose **all** that apply:

- ☐ At your university
- ☐ With the funding body
- ☐ At external storage (unpaid)
- ☐ At external storage (paid)
- ☐ Other (Please specify):

24 []Have you had a formal training on the following *

Please choose **all** that apply:

- ☐ Data Management Plan
- ☐ Metadata
- ☐ Consistent file naming
- ☐ Version control of data sets
- ☐ Data citation styles
- ☐ No, I haven't had training on any of the above

25 []Would you like to have a formal training on the following *

Please choose **all** that apply:

- ☐ Data Management Plan
- ☐ Metadata
- ☐ Consistent file naming
- ☐ Version control of data sets
- ☐ Data citation styles
- ☐ No, I am not interested
- ☐ Other (Please specify):

26 []Any additional information and/or comment you would like to provide related to data management in research

Please write your answer here:

Submit your survey.
Thank you for completing this survey.