Research data sharing, reuse, and metrics: adoption and challenges across disciplines and repositories

Item Type	Thesis or dissertation
Authors	Khan, Nushrat
Publisher	University of Wolverhampton
Rights	Attribution-NonCommercial-NoDerivatives 4.0 International
Download date	2025-05-29 05:19:24
License	http://creativecommons.org/licenses/by-nc-nd/4.0/
Link to Item	http://hdl.handle.net/2436/624837

RESEARCH DATA SHARING, REUSE, AND METRICS: ADOPTION AND CHALLENGES ACROSS DISCIPLINES AND REPOSITORIES

by

Nushrat Khan

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy in Information Science

University of Wolverhampton

2021

Abstract

Data sharing is widely believed to be beneficial to science and is now supported by digitization and new online infrastructures for sharing datasets. Nevertheless, differences in research cultures and the sporadic development of data repositories, support services, guidelines, and policies have resulted in uneven data sharing and reuse practices. An overall understanding of the current situation is therefore needed to identify gaps and next steps. In response, using two case studies and two surveys, this dissertation explores the current landscape and identifies challenges within data sharing and reuse practices. The results demonstrate how present systems and policies could be modified to support and encourage these activities. The researcher survey found that the type and format of data produced, as well as systematic data sharing varied between disciplines, with Physical Sciences and Earth and Planetary Sciences leading and Business and Economics, Engineering, and Medicine lagging in some respects. Surveys and observations were frequently produced in most fields, with samples and simulations being common in science and engineering and qualitative data being more prevalent in the social sciences, business, and humanities. Researchers who had prior data reuse experience shared data more frequently (56.8%, n=1,004) than those who only used their primary data for research (32.6%, n=396). The biodiversity case study and surveys show that secondary data are valuable for many purposes, but most struggle to find datasets to reuse. Data citations can incentivize data sharing, although a lack of appropriate data citations and reliable technologies make it difficult to efficiently track them. In biodiversity, where the sharing and reuse of open data via mature infrastructures is common, citing secondary datasets in references or data access statements has been increasing (48%, n=99). However, users simultaneously exploiting many data subsets in this field complicate the situation. This thesis makes recommendations for handling large numbers of biodiversity data subsets to attribute citations accurately. It also suggests further enhancements for the article-dataset linking service, Scholexplorer, to automatically capture such links. Based on responses from data repository managers, this research further identifies nine objectives for future repository systems. Specifically, 30% (n=34) of the surveyed managers would like integration and interoperability between data and systems, 19% (n=22) want better research data management tools, 16% (n=18) want tools that allow computation without downloading datasets, and 16% (n=18) want automated systems. It also makes 23 recommendations in three categories to support data sharing and promote further data reuse including 1) improved access and usability of data, as well as formal data citations; 2) improved search systems with suggested new features; and 3) cultural and policy-related issues around awareness and acceptance, incentives, collaboration, guidelines, and documentation. Finally, based on researcher feedback, this study proposes an alternative scoring model that combines a dataset quality score and a data reuse indicator that can be incorporated in academic evaluation systems. The outcomes from this research will help funders, policymakers and technology developers prioritize areas of improvement to incentivize data sharing and support data reuse with easily discoverable and usable data.

Table of Contents

Abstract .... *ii*

Acknowledgements .... *vi*

List of Publications .... *vii*

List of Tables .... *viii*

List of Figures .... *ix*

1. Introduction *1*

1.1 Background and problem statement **1**

1.2 Research questions and study overview **4**

1.3 Research design **5**

1.3.1 Case study design 5

1.3.2 Survey design 6

1.4 Key concepts **7**

1.5 Overview of the dissertation **9**

2. Literature Review *10*

2.1 Data sharing – from closed to open science **10**

2.2 Data sharing - determinants and means **12**

2.2.1 To share or not to share – influencing factors 12

2.2.2 Disciplinary differences in data sharing 14

2.3 Role of data repository services **16**

2.3.1 Practices and adoption of data repositories 16

2.3.2 Data repository surveys 18

2.4 Data reuse – why and how **19**

2.5 Data citation and tracking secondary data reuse **21**

2.5.1 Understanding data citation practices 22

2.5.2 Data citation tracking - technical feasibility 23

2.6 Incentives for data sharing and promotion of data reuse **25**

3. Biodiversity Data Case Study *28*

3.1 Data sharing and reuse practices in biodiversity **28**

3.2 Methods **29**

3.2.1 Data collection 30

3.2.2 Data analyses 31

3.3 Results **32**

3.3.1 Dataset quality and citation rate 32

3.3.2 Citation growth rate and distribution of citing articles over the years 34

3.3.3 Types of data use and reuse cases 35

3.3.4 Citation practice for biodiversity datasets 36

3.3.5 Do altmetrics reflect data impact? 37

3.3.6 Is GBIF attributing citations in the best way? 46

3.4 Recommendations **46**

3.5 Limitations **47**

3.6 Discussion 48

3.7 Summary 49

4. Scholexplorer Case Study 51

4.1 Identifying links between datasets and articles 51

4.2 Methods 55

4.3 Results 56
4.3.1 Scholexplorer identified previously unknown dataset-article links 56
4.3.2 Scholexplorer can be used to identify reuse cases of published data 59

4.4 Limitations 62

4.5 Discussion 62

4.5 Summary 64

5. Survey of Repository Managers 66

5.1 Landscape and role of data repositories 66

5.2 Methods 68
5.2.1 Data collection 69
5.2.2 Data analysis 70

5.3 Results 71
5.3.1 Type of data repositories and technical frameworks 71
5.3.2 Data reuse metrics 72
5.3.3 Research data services and quality maintenance 76
5.3.4 Challenges and motivators 78
5.3.5 Priorities 79
5.3.6 Tools needed in the future 81

5.4 Limitations 84

5.5 Recommendations for registries of data repositories 85

5.6 Discussion 86

5.7 Summary 88

6. Survey of Researchers 90

6.1 Disciplinary differences in data sharing and reuse 90

6.2 Methods 93
6.2.1 Questionnaire design 93
6.2.2 Selection of subject categories 94
6.2.3 Data collection 95
6.2.4 Data analysis 96

6.3 Results 97
6.3.1 Type and format of data produced in different disciplines 99
6.3.2 Data sharing across disciplines and research experience 102
6.3.3 Methods of sharing research data 105
6.3.4 Choice of data repositories 109
6.3.5 Data reuse across disciplines and research experience 111
6.3.6 Data sharing vs data reuse 113
6.3.7 Data reuse purposes 115
6.3.8 Finding datasets to reuse 122
6.3.9 Promotion of data and future improvements 126

6.4 Limitations **129**
6.5 Discussion **130**
6.6 Summary **132**
7. Conclusion ***134***
7.1 Key findings **134**
7.2 Contributions and significance of the study **137**
7.3 Summary of findings **138**
7.4 Implications **143**
7.5 Challenges and limitations **147**
7.6 Directions for future research **149**
7.7 Data and code availability **150**
References ***151***
Appendices ***162***

Acknowledgements

This dissertation would not have been possible without the immense support and encouragement from my family, friends, mentors, and colleagues. I am incredibly grateful to my program director, Professor Mike Thelwall for his mentorship and guidance from the very beginning. His kind, encouraging, yet constructive feedback helped me think critically, develop my ideas, improve my writing skill, and most of all grow as a researcher. Thank you for being the most supportive supervisor I will ever know, and for providing your comments and suggestions whenever I asked for it. Also thankful to my second supervisor Dr Kayvan Kousha for his help to shape the intellectual direction of this study through his thoughtful advice, alternative perspectives, and meticulous attention to details. I could not have asked for better mentors.

I could not thank Dr Catherine Pink enough for encouraging me to take up the Scholix project while I worked at the Research Data Services at the University of Bath, and translate this work into a thesis chapter, as well as a publication. My UCL colleagues have been a core support system throughout the final year of my PhD and helped me get through this while I juggled my job and thesis writing. Special thanks to Professor Mario Cortina-Borja for offering me his sincere support. Thank you for taking the time to read over sections of my thesis several times and for providing your valuable feedback. Thanks to Kate Wilson as well for proofreading parts of my thesis.

I am thankful to my close friends for their emotional support and deep conversations when I needed them. I am immensely grateful to my parents for all the sacrifices they made and for encouraging me to pursue an academic career. My father inspired me in early days by never giving up on his ambitions and finishing his PhD when we were young. He always wanted us to follow his path and I know he would have been proud of me for achieving this. Thanks to my sister and brother for always knowing how to lift my spirits.

Finally, I am grateful to my fiancé, Will for believing in me, for reading over my drafts, and for helping me survive during my write-up period when I could not always manage the time to make food. Thank you for your endless encouragement and unwavering support throughout this journey.

List of Publications

Chapter 3, Chapter 4, and Chapter 5 of this dissertation are modified from or informed by the following publications:

Khan, N., Thelwall, M., & Kousha, K. (2019). Data Citation and Reuse Practice in Biodiversity - Challenges of Adopting a Standard Citation Model. *Proceedings of the 17th International Conference on Scientometrics & Informetrics, Rome, Italy,* pp. 1220-1225.

Khan, N., Pink, C. J., & Thelwall, M. (2020), Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations. *Patterns,* Vol. 1 No. 1, 100007.

Khan, N., Thelwall, M., & Kousha, K. (2021a). Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics. Scientometrics, 126(4), 3621-3639.

Khan, N., Thelwall, M., & Kousha, K. (2021b). Are data repositories fettered? A survey of current practices, challenges and future technologies. *Online Information Review*.

List of Tables

Chapter 3

Table 3.1 Type and number of datasets published between 2007-19 and average citations.... 33
Table 3.2 Publication year of all citing articles mentioned on GBIF....................................... 35
Table 3.3 Top 5 blogs mentioning GBIF datasets.. 38
Table 3.4 Distribution of altmetric mentions for different types of datasets 39
Table 3.5 Content from blog mentions of Occurrence datasets 41
Table 3.6 Distribution of content creators for Facebook mentions... 43
Table 3.7 Spearman correlations between citations and altmetric scores for 2016-19............ 45
Table 3.8 Spearman correlations between citations and altmetric scores for different types of datasets 45

Chapter 4

Table 4.1 Links to UKDS datasets associated with University of Bath researchers................ 58

Chapter 5

Table 5.1 Type of data repositories and technical frameworks used 72
Table 5.2 Main priorities for repositories (n=152) 80
Table 5.3 Type of tools and services needed in the future (n=114).. 82

Chapter 6

Table 6.1. Selected disciplines for the survey and numbers of responses 97
Table 6.2 Top 10 types of data produced by discipline with the term frequencies.................. 99
Table 6.3 Terms that frequently appeared before ‘data’ for types of data produced 101
Table 6.4 Logistic regression of data sharing outcomes (n=3,098)....................................... 104
Table 6.5 Data sharing methods in different subject category... 106
Table 6.6 How researchers first found repositories to share data 109
Table 6.7 Factors that influence choice of repositories in different disciplines..................... 110
Table 6.8 Logistic regression of data reuse outcomes (n=3,095) 113
Table 6.9 Comparison between data sharing and reuse across disciplines 113
Table 6.10 How researchers find datasets to reuse in different disciplines 122
Table 6.11 Important factors when searching for existing datasets 123
Table 6.12 Future improvements needed to promote data reuse.. 126

List of Figures

Chapter 3

Figure 3.1 Distribution of different types of datasets published (2007-2019) 33
Figure 3.2. Number of occurrence datasets published, and average number of citations received 34
Figure 3.3 Citation location in randomly selected articles (labels on bars represent the number of articles) 36
Figure 3.4 Recommendation for revised GBIF citation attribution system 47

Chapter 4

Figure 4.1 Dataset links found by Scholexplorer in external repositories associated with University of Bath research outputs, excluding CCDC. 57

Chapter 5

Figure 5.1 Perceived usefulness of different type of metrics (labels on bars represent number of responses) 74
Figure 5.2 Tracking status of different type of metrics (labels on bars represent number of responses) 75
Figure 5.3 Metrics display status of repositories who track these (labels on bars represent number of responses) 76
Figure 5.4 Type of research data services provided by the repositories 77
Figure 5.5 Data quality maintenance types by the repositories 77
Figure 5.6 Type of challenges faced by the repositories 79
Figure 5.7 Motivators for researchers to deposit data in repositories 79

Chapter 6

Figure 6.1 Percentage of responses by research experience and discipline (labels on bars represent number of responses) 99
Figure 6.2 Format of data produced by subject category 102
Figure 6.3 Data sharing across disciplines (labels on bars represent numbers of responses) 103
Figure 6.4 Data sharing in groups with different research experiences across disciplines (labels on bars represent numbers of responses) 104
Figure 6.5 Data reuse across disciplines (labels on bars represent numbers of responses) ... 112
Figure 6.6 Data sharing among those who only use own primary data (labels on bars represent numbers of responses) 114
Figure 6.7 Data reuse among those who do not share data (labels on bars represent numbers of responses) 115
Figure 6.8 Data reuse types across disciplines 116
Figure 6.9 Ease of finding datasets to reuse by discipline 125
Figure 6.10 Ease of finding datasets to reuse by research experience 125

Chapter 7

Figure 7.1 Research data ecosystem 144

1. Introduction

1.1 Background and problem statement

"…[A] one-size-fits-all approach to policy and technology design with regard to openness and sharing of research information is problematic". (Velden, 2013, p. 456)

The primary goal of this study is to understand the current landscape of data sharing, data reuse, and current practices of data citation by researchers in order to support open data sharing by developing technological solutions and reliable impact metrics. For this purpose, I explore the technological challenges faced by different repositories and data systems, as well as disciplinary differences in scientists' data sharing and reuse behavior. The interplay between these factors may suggest how policy and technology design can benefit researchers across disciplines.

The open access (OA) movement in scholarship and publishing has been a major driver for making research outputs publicly available and accessible in recent years, with an increasing number of journals adopting an OA route to publishing. This may have resulted in greater research impact across different disciplines (Antelman, 2004; Hersh, 2017). Over the past decade there has also been a growing interest within the scientific community to share research data in a findable, accessible, and interoperable format that allows the reuse of data by others (Wilkinson et al., 2016). Open research data is a gateway to reproducible science with increased opportunities for collaboration and interdisciplinary research (Borgman et al., 2019). In addition, studies that share research data have a citation advantage (Piwowar et al., 2007; Henneken & Accomazzi, 2011; Colavizza et al., 2020). As a result of the importance of research data, funding bodies are increasingly mandating the sharing of research data at the end of the project funding period (Piwowar, 2013; Kiley et al., 2017). This is reflected in policy and decision making as well, with the Research Excellence Framework (REF) in the United Kingdom including research datasets and databases (category S, p. 112) as standard research outputs in their most recent guideline (REF, 2019).

Sharing research data is not an entirely new concept in the research community, because data has often been informally shared by email upon request from colleagues or bona fide researchers (Federer et al., 2015). However, the norm of publishing journal articles is embedded in the ‘publish or perish’ research culture of scientific community (Van Dalen & Henkens, 2012), whereas a requirement to openly publish data often creates tensions with this due to the extra work involved and the potential to invalidate a researcher’s follow-up publications with the same data (Valden, 2013). Within specific disciplines, the willingness of individual researchers to share data may be guided by the disciplinary research and data sharing culture. For example, Ceci’s (1988) proposal of a new scheme for mandatory data sharing for social scientists was met with opposing arguments from researchers outside of the social sciences. Bioengineering scientists raised concerns, such as the finite nature of their data, losing control of their data for addressing questions at a later time, which were unanticipated at the time the grant application was made, and other scientists benefitting from their hard-earned grants and carefully collected data.

Two decades later, Piwowar (2011) reported that approximately 45% of gene expression studies made their data publicly available, with 25% of the articles in their study sample (*n*=11,603) publishing datasets in best-practice repositories. These repositories are needed for a dataset to be findable and accessible in the long-term. This progress has been made possible by the rapid growth of data repositories over the past decade. The re3data registry of research data repositories, which emerged in 2012, already listed 400 repositories by July 2013 (Pampel et al., 2013). It provides an overview of the landscape of available research data repositories in different disciplines. The sporadic development of discipline-specific, multidisciplinary, and institutional repositories (IR) and disciplinary differences in data sharing mean – a) technological frameworks and solutions can be heterogeneous and b) researchers need to make an informed decision regarding which repository service to use for long-term preservation of their data.

Some data-intensive disciplines, such as astronomy and genomics have progressed more quickly than others (Borgman, 2012). Well-known disciplinary repositories include the NASA Open Data Portal[1], Sloan Digital Sky Survey[2] (SDSS) in astronomy; GenBank[3], National Center for Biotechnology Information (NCBI) in biomedical sciences; Pangaea[4], National Oceanic and Atmospheric Administration[5] (NOAA) repositories for environmental sciences;

---

[1] https://data.nasa.gov/
[2] https://www.sdss.org/
[3] https://www.ncbi.nlm.nih.gov/genbank/
[4] https://www.pangaea.de/
[5] https://www.ncei.noaa.gov/access

Inter-university Consortium for Political and Social Research[6] (ICPSR) in the social sciences. In contrast, the culture of data sharing may be less strong in fields without discipline-specific repositories. These variations confirm that it is important to understand how disciplinary differences shape the data sharing behaviour of researchers.

While sharing research data can benefit scientific progress, it requires a significant amount of time to prepare a good-quality research dataset and share it in a systematic manner. Therefore, researchers may want to know if their shared data will be used by anyone and understand the impact of their research data when it is shared with the broader community (Wallis et al., 2013; Sayogo & Pardo, 2013). The challenges of incentivizing researchers to share their data, developing a universal accreditation system for their efforts and overcoming technological and policy constraints remain.

The ecosystem of research data sharing is complex with multiple stakeholders in play, and this is one of the converging points in scholarship. Researchers, data managers, information professionals, technology developers, policy makers, and journal publishers must come together to develop standardized methods for collecting and implementing metrics in this comparatively new area of scholarship. Two types of research data metrics have been proposed for evaluating the impact of shared research data: 1. Data publication and citation-based metrics and 2. Altmetrics, e.g., social media indicators, readership counts, views and download counts (Lawrence et al., 2011; Costas et al., 2013). Capturing the broader impacts of research data through the implementation of reliable metrics is important as *"it provides tangible evidence of benefit to weigh against the costs of research. For another, it provides an engaging way of comparing peer research programs across the globe, albeit through the lens of proxy indicators, when undertaking strategic decision-making or benchmarking"* (Ball & Duke, 2015, para. 4).

Among different data reuse metrics, data publication and citation-based metrics (citations to datasets or data papers describing a specific dataset) are the most agreed upon and preferred metrics within the research community (Kratz & Strasser, 2015). Consistent data citation is therefore essential to systematically track how existing datasets have been reused by others. Providing datasets with persistent, unique identifiers, such as Digital Object Identifiers (DOI) can promote the practice of data citation and add more value to the data by linking it to a

---

[6] https://www.icpsr.umich.edu/web/pages/

publication (Callaghan, 2014). However, data citation varies from journal article citation in nature (Lawrence et al., 2011), which often results in the inconsistent citation of datasets in publications (Federer et al., 2018; Thelwall et al., 2020). Since measuring the impact of publications based only on citations is often debated, it is unsurprising that methods of assessing the value of research data are yet to be standardized. The social web could be a helpful source to assess the broader impact of research products including datasets (Piwowar, 2013).

Assessing the value of research data based on citations can be influenced by the subject area and data type. Because of this, multiple services, such as Impactstory[7], Plum Analytics[8], and Altmetric[9] have been developed to evaluate the wider impact of research, potentially including datasets with DOIs. However, the application and adoption of such data metrics will depend on their acceptability within the scientific community, technological feasibility, and adaptability for disciplinary differences. Although a content analysis of contextual information of dataset citations and social media mentions might help to reveal how a dataset has proven useful, if at all, this is an area that needs further investigation.

1.2 Research questions and study overview

This is a mixed method and multi-phased study using quantitative and qualitative approaches to examine aspects of current data sharing and reuse in research communication. Two case studies and two surveys were designed to address these goals.

Four overarching research questions drive this investigation, with each study addressing the adoption and challenges of data sharing, data reuse, and metrics from technological and stakeholders' (researchers, data repository managers, funding bodies) perspectives:

- How do researchers reuse and cite secondary datasets in journal articles? Which metrics are informative about the impacts of secondary datasets?
- Is the major data citation tracker (Scholexplorer) able to automatically capture data citation and help identify data reuse?
- How do data repositories vary in data support services, and the technological and operational challenges they face when providing these services?

---

[7] https://profiles.impactstory.org/
[8] https://plumanalytics.com/learn/about-metrics/
[9] https://www.altmetric.com/

- How do data sharing and data reuse practices by researchers vary by research experience and between disciplines?

Together these findings inform the guiding research question: *What can be improved in current systems and policies to support and promote data sharing and data reuse?*

To address the first question, I conducted a case study of biodiversity datasets published in a well-known discipline-specific federated data repository, Global Biodiversity Information Facility (GBIF). I examined the data reuse methods and data citation practices in the articles citing GBIF datasets. Besides dataset citations, altmetric mentions of GBIF datasets in Facebook, Twitter, blogs, and Wikipedia were studied to determine usefulness of such metrics. In the second part of the study, I evaluated the Scholix (Scholarly Link eXchange) framework and Scholexplorer, aggregator of Scholix links (Scholix, 2019), in the context of an institutional repository of an academic institution, to explore its potential of identifying links between datasets and publications and track secondary data reuse. These two case studies identified gaps in current practices and technical solutions and informed areas to investigate in the survey of repository managers – the third part of the study. Finally, drawing from the results of the first questionnaire, I designed a survey for the researchers to understand how they use these data repositories for sharing and reusing data in different disciplines. Each study suggests areas of improvement in current systems and policies to support data sharing in lagging disciplines and to further promote data reuse.

1.3 Research design

1.3.1 Case study design

To explore data sharing and reuse practices for datasets published in a repository, the first step is to identify a repository that publishes such metrics, e.g., download counts or citation counts. This information is available from a limited number of repositories, and even though there are now many research data repositories, the key challenge is to find repositories that allow access to their data through an application programming interface (API) or in a structured format. Having a significant number of datasets or large corpora for analysis is another important factor in choosing the data sources.

For the first case study I choose the field of biodiversity as previous literature suggest that open data sharing is common in this field and there are varied data reuse cases (Magurran et al., 2010). GBIF[10] was used as a data source since this holds a large volume of open biodiversity data and the amount has been growing consistently over the past 11 years. The indexing system held about 39,000 datasets at the time of data collection in 2018 and there was a semi-automated system to provide information on number of citations, including lists of citing articles. Most metadata were accessible through their API, which made it a promising source of information to study.

The second case study explores Scholexplorer by querying the Scholexplorer API to find links between datasets and journal articles in the context of a university and its institutional repository (The OpenAIRE Scholexplorer, 2020). This API allows clients to run queries over the Scholexplorer index in order to fetch links in a given criteria. As stated in the literature review in Chapter 2, a) academic institutions need to comply with funder mandates of data sharing and need to identify other data repositories where their researchers have published data and b) institutional repositories may not be as well-equipped as disciplinary repositories like GBIF to develop their own systems in order to establish links between their datasets and articles that reuse their data. The implementation of the Scholexplorer API for these purposes requires access to both a) publication DOIs and b) dataset DOIs published in their institutional repository. The University of Bath Research Data Archive is one of the earliest institutional data repositories in the UK and the university has been leading data sharing initiatives as a part of the GW4 consortium[11]. Therefore, this was chosen as a data source for the second case study.

1.3.2 Survey design

The first survey explores differences in data repositories and uses openly available metadata from re3data.org to collect contact information provided by repositories. I use this approach to disseminate the survey instead of ad-hoc recruitment of data repository managers to ensure responses from a wide range of data repositories. Since repositories did not always provide an email address as their contact information, a non-personalized version of the survey was created for data repositories that only use web forms as their contact method.

The second survey examines disciplinary differences in data sharing and reuse practices and utilizes publication metadata from Scopus to collect email addresses of the first authors in 20

---

[10] https://www.gbif.org/

[11] https://gw4.ac.uk/

disciplines. When researchers are recruited in an ad-hoc basis, some disciplines tend to have higher response rate than others, i.e., researchers in disciplines where data sharing is less common may not be compelled to respond to this kind of survey. On average 3,500 email invitations per discipline were disseminated to a random sample of researchers, with an aim that in case of very low response rate (e.g., 1%) from certain disciplines, the number of responses could be above the minimum sample threshold of 30. Data collection methods of both the data repository manager and researcher surveys are detailed in Chapter 5 and Chapter 6 respectively.

1.4 Key concepts

Throughout this dissertation I have used the following concepts related to data sharing, data reuse and citation, as defined in existing literature.

Research data: Research data refers to any information that has been collected, observed, generated, or created to answer research questions and validate research findings. The concept of data is difficult to define as it may take many forms, both physical and digital, and hold different meanings depending on the context. In this thesis, the term "data" is used in a broadly inclusive way: "*In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), [data] refers as well to 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*" (Uhlir & Cohen, 2011[12], as reported in Borgman, 2012, p.1061). Therefore, data may include any form of raw data, as well as multimedia files, code and software written to generate, process, analyse and validate research data[13].

Dataset: The term dataset refers to a collection of data produced by an individual or group of data producers or by an institution (Fear, 2013). A dataset may consist of a single file or a collection of files, along with its associated metadata (e.g., an abstract, data collection method, license) that enables understanding and usage of the data in a legal way.

---

[12] Uhlir, P.F., & Cohen, D. (2011). Internal document. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences. 18 March 2011.

[13] https://www.reading.ac.uk/en/research-services/research-data-management/about-research-data-management/research-software-and-code

Data repository: A data repository or data archive is a digital infrastructure that provides storage and long-term access to data and its associated metadata. A repository can be a part of an academic institution (also known as an institutional repository) or can be hosted independently (e.g., Zenodo). The roles of data repositories include standardizing storage and creation of metadata, providing data curation, recommending citations to datasets, tracking data reuse, and promoting good scientific practice (Costas et al., 2013).

Data sharing: Data sharing is the release of research data to be used by others (Borgman, 2012). This general term does not imply a specific method, however. In a standard practice, data should be shared in a findable, accessible, interoperable, and reusable manner (Wilkinson et al., 2016).

Data reuse: Data reuse refers to secondary use of data by users other than the data collectors, including the first use of data collected for a community, e.g., astronomy datasets from sky surveys (Pasquetto et al., 2017). This is also known as secondary analysis in social science (Fear, 2013). Data reuse can be performed for varied reasons, including the following: using an existing dataset in part or as a whole; combining it with another dataset to answer new research questions; teaching and training; or replicating the results of the original study (Zimmerman, 2008; Pasquetto et al., 2017; Bishop & Kuula-Luumi, 2017).

Data citations: Data citations have been defined as inclusion of formal citation to datasets in the reference list of published articles, given that the data were used to support that specific research (Mayernik, 2013). Lawrence et al. (2011) suggests the following information to be included in a data citation: author, publication year (or equivalent), activity or tool that produced the data, and an unambiguous reference to the source of data, e.g., a DOI. However, data are more complex and varied than documents and they introduce new challenges in respect to traditional publications (Silvello, 2018).

Data access statement: Data access statements, also known as data availability statements, are included in publications to describe where the data associated with the paper is available, and under what conditions the data can be accessed. They are required by many funders and scientific journals to ensure that the associated data are accessible for reproducibility and reuse purposes.

Altmetrics: Altmetrics are metrics and qualitative data that are complementary to traditional, citation-based metrics. They can include (but are not limited to) peer reviews on Faculty of 1000, citations on Wikipedia and in public policy documents, discussions on research blogs, mainstream media coverage, bookmarks on reference managers like Mendeley, and mentions on social networks such as Twitter (Altmetric, 2022).

1.5 Overview of the dissertation

The dissertation is divided into multiple chapters, with four based on individual studies and related findings. The next chapter (Chapter 2) provides an overview of the literature relating to data sharing, data reuse and data citation practices across disciplines and how to capture the broader impact of open research data.

Chapter 3 through Chapter 6 contain the findings from each study - Chapter 3 and Chapter 4 consist of two case studies, and Chapter 5 and Chapter 6 consist of two survey studies. These four chapters follow a similar format: 1. the specific methods implemented for that study, with relevant literature; 2. results of that chapter; 3. a discussion of those findings; and 4. chapter summary. The final chapter (Chapter 7) summarizes the findings and key implications from the above sections with a discussion that connects the four chapters to the primary research question, while highlighting the limitations and challenges. I conclude with suggestions for future directions of research in this field.

2. Literature Review

The term “data” encompasses a broad spectrum. For example, to planetary scientists, *“‘The data’ could refer to an image composed of pixels”* or it could *“also be complex graphs generated by spectrometers peering into infrared or ultraviolet light wavelengths, or it could be stellar occultation timings, or measurements of magnetic fields.”* (Vertesi & Dourish, 2011, p. 534-535).

With rapid technological advancement and digitization of data generation methods in most areas, we are inundated with data every day – some of which could be administrative or routine data, and some could be research data, collected for specific purposes. Collecting data to answer new questions is integral to many types of research. Even 40 years ago the federal government in the USA alone spent between $700 to $900 million for the advancement of biomedical and social science (Ceci & Walker, 1983), so it is reasonable to expect that a substantial amount of high-quality research data is created annually that might be usefully shared. Data collected through scholarly work can now be considered as a first-class research output along with publications and should serve two main purposes: verification of findings by reproducing the results and reuse of data for new applications, e.g., research, learning, teaching (Bishop & Kuula-Luumi, 2017). In an ideal world, any data and code generated by research, especially from funded research projects, should be available to all in an accessible and reusable format. However, this ambitious view is yet to be fully realized. This literature review focuses on current practices of data sharing, data reuse and data citation to understand disciplinary differences and technological gaps. The first two sections explore data sharing practices: is data sharing more common in some fields than others? What influences data sharing? In the following section, I explore the evolution of data repositories and the role these infrastructures play: are some data repositories more equipped than others and preferred by researchers when it comes to the discoverability of data? This is followed by how and why researchers reuse openly available data in different disciplines. Finally, I review existing literature on researchers’ data citation practices to understand how it affects current technological solutions.

2.1 Data sharing – from closed to open science

More than three decades ago, Ceci and Walker (1983) emphasized the importance of openly sharing federally funded research data that are not harmful to national security or human subjects and argued that this serves the national welfare by extending the public’s “right to know”. The authors proposed a new scheme for the mandatory sharing of data among social scientists, an open data bank, and outlined its benefits: time and cost efficiency to access data, the application of new forms of statistical and technological methodology, historical data analysis, and transparency.

“It is proposed that in the long term, a national databank be created or extant operations at some agency already versed in the problems of archiving large amounts of data in easily accessible form (e.g., the Census Bureau, the Library of Congress, or the National Archives' Division of Machine-Readable Data) be expanded to accommodate the routine cataloguing and dissemination of raw data from all publicly sponsored investigations” (Ceci & Walker, 1983, p. 418).

This was met with opposing arguments from researchers across many disciplines, especially bioengineering researchers (Ceci, 1988). Ceci mentioned two surveys that were carried out afterwards, finding that most respondents (87%) were willing to share data with their colleagues. However, 59% claimed that their colleagues were not prone to sharing their data, even when the data was obtained with the benefit of federal funds. Since the proposal of an open data bank and publication of these results in 1980s, scientific communities have moved in their suggested direction, supported by technological growth, and use of the Web. For example, ICPSR (Inter-university Consortium for Political and Social Research) is the leading repository for the social sciences. It started as an archive in 1962 and its first website was posted in 1994 (ICPSR, 2021).

Today, most research data are produced in digital format with growing infrastructures and standards in place to support data sharing. While data sharing has increased in many disciplines, cultural impediments and systematic tension between cooperativeness and competitiveness have persisted (Velden, 2013). A timely example is sharing participant-level data from Covid-19 trials, where a recent study found that only 15.7% of 924 clinical trials were willing to do so, even though this could accelerate the identification of effective treatments (Li et al., 2021). Despite increasing willingness to share data with others in some disciplines, it is not clear whether the data will be reused if it is deposited in a repository (Wallis et al., 2013). Hence sharing data directly with other researchers via personal communication remains a common practice (Federer et al., 2015). Thelwall et al. (2020) found that only 13% of 314 primary human genome-wide association studies (GWAS) papers published in 2010 and 2017 reported the complete location of GWAS summary data even though more studies include a data availability statement. Despite the availability increasing from 3% in 2010 to 23% in 2017, this indicates low-level use of standard data repositories in a research field with relatively strong data sharing norms. Therefore, it is important to understand how researchers share data on the web and what influences their data sharing behavior. In the next section, I explore the determinants of systematic data sharing and how they vary between disciplines.

2.2 Data sharing - determinants and means

Publishing research data in a systematic manner opens the door to more complex questions for researchers and policy makers – from how to define a dataset to establishing best practices for citing datasets in a specific field (Borgman, 2012; Kratz & Strasser, 2014; Starr et al., 2015; Silvello, 2018). In a research culture where publications are the primary currency, "*Data citation and publication will ensure that data will be considered as a first class research output that will be available, peer-reviewed, citable, easily discoverable and reusable.*" (Callaghan et al. 2012, p. 113). The practice of data sharing is not homogenous across disciplines, however. Disciplinary culture, size and type of data produced often determine how researchers share their data and what is being shared. For example, researchers in certain data-intensive fields, such as astronomy and bioinformatics, produce a high-volume of data that often needs large-capacity storage systems with computational capabilities. These fields developed standard practices for how data need to be effectively shared by pioneering large-scale database development for open data sharing, such as SDSS, NCBI (Bell et al., 2009). Despite the availability of technological infrastructures, data may not be considered as knowledge commons because the core difference with the model of commons-based peer production lies in the motivation for participation (Fecher et al., 2015).

2.2.1 To share or not to share – influencing factors

Open data initiatives have been growing at different rates within different communities. Borgman (2012) suggests four rationales for sharing data: 1. Reproduce/verify, 2. Serve public interest, 3. Ask new questions, and 4. Advance research. Underlying these rationales are motivation and incentive: "*A motivation is something that causes someone to act, whereas an incentive is an external influence that incites someone to act.*" (Borgman, 2012, p.1066). Previous studies suggest several common factors influencing data sharing: effort required to prepare data and metadata, trust in colleagues and a lack of incentives (Piwowar, 2011; Wallis et al., 2013; Sayogo & Pardo, 2013; Fecher et al., 2015). A survey of biodiversity researchers (n=799) reported that the main obstacles against data sharing in their field included loss of control, possible misinterpretation of one's data by someone else, the time and effort required to prepare a data set for sharing, not being acknowledged for sharing data, missing data standards, missing infrastructure, and unclear legal conditions (Enke et al., 2012).

Anagnostou et al. (2013) points out that even in closely related fields within genomics, data sharing can vary. The authors examined data sharing practices in journal articles published between 2008-2011 and found that researchers in forensic genetics shared data more frequently (86%, *n*=142) than evolutionary (79%, *n*=210) and medical genetics (64%, *n*=72). They suggest strong editorial policies of journals as a proximate cause and the collaborative spirit among investigators in forensic genetics as a remote cause for a higher rate of data sharing in this discipline. Similar results were reported for life sciences data by Thelwall and Kousha (2017): upon signing up for a Joint Data Archiving Policy in 2011, data published in the Dryad repository for articles published in two selected journals had increased rapidly. Journal mandates can play an important role to advance data sharing since perceived ownership of data (reflected in the right to publish first) and a need for control (reflected in the fear of data misuse) impede commons-based exchanges of research data (Fecher et al., 2015).

Kim and Stanton (2016) investigated institutional and individual factors that affect scientists' data sharing by conducting a survey in 43 STEM disciplines (*n*=1,317) and found that the availability of data repositories does not affect data sharing in those disciplines. Their survey included a question on how frequently researchers shared data in different forms in the last two years, but the results were not presented for individual disciplines. Study results from a multiple linear regression model suggested that journal mandates, disciplinary norms, perceived career benefit and scholarly altruism at an individual level have significant positive relationships with data-sharing behaviors, and that perceived effort has a significant negative relationship. Similar results were reported for sociology and political science by Zenk-Möltgen et al. (2018). The authors found that, at the time of their study in 2016, 110 out of 142 journals (77.5%) in sociology and 53 of the 120 selected journals in political science (44.2%) had some sort of data policy for their authors - this was higher than five years ago, when only 18 out of 120 journals (15%) in political science had a data policy in place (Gherghina & Katsanidou, 2013). They selected 1,011 research articles in ten main journals (five journals in each discipline), published between 2012-14 in these two disciplines and surveyed the authors (*n*=446, covering 44.1% of all articles) to examine their motivations, behavioral control, and perceived norms for sharing data. Out of all the empirical articles, only half stated that data were available and for 37% of those the data could be accessed. Similar to Gherghina and Katsanidou (2013), their study found that presence of data policy and data availability were strongly associated with high impact factors and young journals. However, a review of journal data policies reported that journals with a strong data policy were still rare in engineering (four out of 28, 14.2%) and no association was found between data sharing and journal impact factors or OA policies (Wiley, 2018).

2.2.2 Disciplinary differences in data sharing

Previous studies have reported that research data sharing is steadily increasing. From the studies that focused on data sharing in specific disciplines, Piwowar (2011) applied bibliometric methods to identify 11,603 articles that described the creation of gene expression microarray data and found that sharing associated data in best-practice repositories increased from 5% in 2001 to 30%–35% in 2007–2009. Enke et al. (2012) and Huang et al. (2012) conducted separate surveys on biodiversity data sharing (*n*= 799 and 372 respectively). Both studies reported that most respondents were willing to share article-related data but claimed a weak culture of data sharing within their scientific community, although this may have subsequently changed. For example, there has been a substantial shift in attitude in the neuroimaging community since the early 2000s when a journal requirement to deposit fMRI data into the fMRI Data Center was not welcomed. Since then, the community has developed both institutionally coordinated consortia-based data sharing, such as the ADNI and Human Connectome project and long-tail data sharing (Ferguson et al., 2014). Faniel and Yakel (2017) studied three disciplines, social science, archeology, and zoology using a combination of methods for each discipline (interviews, survey, observations, server log entries), and found that disciplinary practices and traditions, along with external forces, such as technological advancements, federal mandates and policies, and repository infrastructure influence data sharing in those disciplines. For example, social science has well-established data sharing standards and infrastructures, such as ICPSR, but researchers in this field only handle a few data formats. Similarly, zoologists have built a strong data sharing standard and infrastructure over the centuries. With Berkley's Museum of Vertebrate Zoology as an example, Star and Griesmer (1989) explained how the community worked together between 1907-39 to handle heterogeneity and collaboration. Darwin Core was ratified as a metadata standard for biodiversity data in October 2009, helping to connect repositories such as VertNet and GBIF (Wieczorek et al., 2012). In contrast, when embarking on digital data sharing, archaeology researchers face challenges from legal and ethical mandates, documenting the variety of data types used, a lack of common data recording practices, and the absence of a standard data repository.

Disciplines that handle sensitive human participant data or qualitative data tend to lag in systematic data sharing, often due to data privacy issues. Federer et al. (2015) surveyed 135 biomedical research staff where 61% (*n*=82) participants responded that they never uploaded data to a repository, even though they had shared data with other researchers via personal communication. In a more recent study, Mozersky et al. (2020) interviewed 30 data repository curators, 30 qualitative researchers and 30 institutional review board (IRB) members to learn from their experience and knowledge of qualitative data sharing. The results suggest that all interviewed professionals in qualitative research fields are unprepared to openly share research data. Compared to other data-intensive fields, such as astronomy or genetics, qualitative researchers are often unfamiliar with the concept of sharing data in a repository. While curators and IRB members encourage qualitative data sharing, they are often not experienced in qualitative data curation and lack standard guidelines to support researchers. Bishop (2009) challenged the lack of archiving qualitative data and pointed out that the UK Data Archive (UKDA) and other major qualitative data archives maintain comprehensive systems for protecting data and provide adequate guidelines. These systems have at least three elements: 1. Consent at the time of data collection for all the key purposes to which the data may be put, 2. Removal of personal or sensitive information, and 3. A rights management framework. The sample sizes for qualitative fields have been too small in most studies so far to compare overall data sharing and reuse attitude against other fields.

Tenopir et al. (2015) conducted a follow-up cross-disciplinary survey (*n*=1,015) to explore changes in data sharing and reuse practices since their baseline study in 2009/2010 (*n*=1,329) and reported that nearly three quarters of the respondents shared some of their data. However, data sharing was not systematic and meaningful in all those cases as researchers were largely reliant on personal data storage options, with only 11.3% using institutional repositories, 9.5% using disciplinary repositories, and 2.4% using a publisher-related repository. There was significant difference within disciplines in terms of disciplinary repository usage, with researchers in Ecology being more likely to store data in such repositories (44.6%) than those in Physical Sciences (7.1%) and Social Sciences (10.8%). These numbers increased slightly in the authors' recent follow-up study, even though the usage of personal storage remained high (Tenopir et al., 2020). Because of limited evidence of cross-disciplinary differences in data sharing practices, it is important to further explore how researchers choose repositories to share data, the type of repositories they use, and which factors influence their choice of repositories.

2.3 Role of data repository services

Data repositories are instrumental in creating a sustainable solution to ad hoc data sharing methods, such as personal storage devices, websites, and email. The number of data repositories has increased over the past two decades, providing a reliable infrastructure to foster adequate scientific data collection, curation, preservation and ensuring long-term access to it. An analysis of 516 studies published between 1991 and 2011 found that the availability of data associated with these articles was strongly affected by article age. The odds ratio from the regression analysis suggests that for every yearly increase in article age, the odds of the dataset being extant decreased by 17%. This highlights the importance of using data repositories for long-term storage and access since the missing datasets were mostly maintained by individual researchers (Vines et al., 2014).

Disciplinary repositories have been developed to meet the specific data format and volume needs of some fields. Around the same time, new open-source repository frameworks have emerged. Eprint was developed in 2000 and DSpace was released in 2002 by the Massachusetts Institution of Technology (MIT) and Hewlett Packard. More academic institutions have adopted institutional repositories in recent years as these allow them to demonstrate the significance of an institution’s research and help preserve and nurture new forms of scholarship (Luther, 2018). Then there are also generalist data repositories, such as Figshare, Dryad, and Zenodo.

2.3.1 Practices and adoption of data repositories

While a six-fold increase in the number of data repositories in the early 20th century is evidence of a perceived increasing need to support data sharing (Pampel, 2013), the large number of diverse data repositories means that determining where to upload data can be confusing for researchers (Federer et al., 2015). Furthermore, data publishing is more complex than literature publishing due to differences in data types and their use cases. Assante et al. (2016) surveyed five generalist data repositories (3TU.Datacentrum, CSIRO DAP, Dryad, Figshare, and Zenodo) that are recommended by data journals on eight key aspects of data publishing: dataset formatting, documentation, licensing, publication costs, validation, availability, discovery and access, and citation. Their study results demonstrate that the services offered by generalist repositories were more similar to those of literature repositories and a lack of a designated community meant an absence of consolidated shared practices. Even though peer review of data can ensure that a dataset has been through a process of scientific assurance (Lawrence et al., 2011), this is not considered in the data publication workflow of generalist repositories. The authors also suggest enabling multiple metadata descriptions and documentation for the same dataset depending on the needs of different target audiences for better reusability. This was supported by a recent study on image sharing in open data repositories (Hansson & Dahlgren, 2021). The authors argued that a lack of image-level granular metadata in the repositories examined would result in the data not being meaningful for reuse to others, and perhaps could be the underlying reason for less use of data repositories by humanities scholars.

These studies suggest that the role of data repositories is crucial to assist users in publishing their data in a meaningful manner, which in turn helps in their research workflow to find relevant data for reuse. Cragin et al. (2010) studied data sharing practices of scientists working in small interdisciplinary fields and stated that their data could be managed well in an institutional repository context. Nevertheless, library practitioners have previously reported limited data repository engagement by academic staff and researchers (Pinfield et al., 2014; Luther, 2018). This lack of engagement was reflected in the unavailability of links to data repositories in published journal articles. Studies that analysed data availability statements found that inclusion of precise data availability information in research outputs is still not common. Many journals now mandate data access statements in articles, but the informativeness of these statements can vary. Colavizza et al. (2020) compared two OA journals and found that the percentage of articles that link to a data repository in their data access statements is only 12.2% (6,656 out of 54,719) for BMC and 20.8% (9,013 out of 43,388) for PLOS. A similar low-rate of data availability for primary human genome-wide association (GWAS) articles was reported by Thelwall et al. (2020). This shows that while the use of data repositories is increasing, it is still a minority activity, even when the data is standardized and has high value for sharing. This may be linked to the absence of data sharing cultures in specific disciplines, but it is important to understand what challenges repository managers are facing in providing these services and whether and how incentives motivate researchers to adopt standard data sharing practices.

With the rapidly evolving role of data repositories, the core requirements of storage and access to data now come with other needs, such as compliance with funder mandates, and proper licensing to ensure reusability. However, the details of non-core practices vary between different types of repositories. In particular, research data services in higher education institutions often lack highly equipped technical services and require proper advisory services for the curation and long-term preservation of active data (Cox et al., 2017). It is therefore essential to understand whether repositories now tend to meet data reusers' needs, support data curation and ensure metadata quality.

2.3.2 Data repository surveys

Several surveys have identified best practices and needs of the data repository community. The Confederation of Open Access Repositories (COAR) received 43 responses to a survey conducted in 2016. Half of the respondents acknowledged using the same platform for publications and research data. The platforms used varied widely, with DSpace and Dataverse being the most common. Engaging researchers in data sharing, lack of institutional policies, and infrastructure for storage and preservation were the top three challenges mentioned (Shearer & Furtado, 2017). A more recent study by LIBER (Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries) focused on implementation of FAIR Data principles (Wilkinson et al., 2016) in 32 repositories with two surveys: one conducted for repository managers (29 responses) and another for the technical staff (14 responses) (Ivanović et al., 2019). Most (81%) repositories were institutional and 41% of the repositories were based on DSpace and 45% had basic data curation support (brief checking, addition of basic metadata or documentation). The study revealed that the understanding and implementation of FAIR principles are often complicated and not fully met by the respondents. In terms of reuse indicators, Kratz and Strasser (2015) surveyed 247 researchers and 71 data mangers, finding that 85% of researchers and 61% of data managers ranked citations to data as the most important reuse metric. Nevertheless, just over 30% of repositories were tracking dataset citations and only 10% were exposing them (i.e., publishing them within their platforms).

The surveys conducted so far have investigated important aspects of research data support. Nevertheless, the sample sizes for repository-oriented surveys have been small, with most respondents being from institutional repositories. This is likely due to participant recruitment methods used in these studies, including mailing lists, personal contacts, social media, and circulation within relevant professional networks. In addition, the rapid evolution of the field can render such surveys obsolete relatively quickly. In response, this thesis reports a more current and larger survey for multiple repository types using openly available metadata from re3data to collect contact information for recruitment purposes that explores the current technical developments, adoption and change in standard practices, challenges, and future needs of data repositories (Chapter 5).

2.4 Data reuse – why and how

Open data unveils new opportunities to explore, aggregate and analyse existing data that would otherwise be difficult to collect by an individual or small group of researchers. Several studies indicate that data reuse is growing with the availability of open data through established data repositories and that researchers would use others' datasets when easily accessible (Wallis et al., 2013; Bishop & Kuula-Luumi, 2017; Tenopir et al., 2020). Sayogo and Pardo (2013) and Curty et al. (2017) reused baseline and follow-up survey results on data sharing from Tenopir et al. (2015) to derive answers to new research questions, which demonstrates data reuse examples within this research area. Meystre et al. (2017) conducted a literature review and content analysis of 324 publications and identified that clinical data are increasingly reused for providing high quality healthcare and developing learning healthcare systems. Primary biodiversity data are fundamental to any systematics and evolution study. Therefore, open data in the field of biodiversity is broadly reused, with primary data uses being ecological studies, taxonomic works, and phylogenetic analyses (Troudet et al., 2018). The authors also identified how different types of data may limit or expand research opportunities and recommended enhancement of observational-based occurrences with ancillary data. Easy accessibility, availability and good quality data with adequate additional information are key to data reuse in biodiversity and ecology research (Zimmerman, 2008; Enke et al., 2012). Similar results were reported in social science, where interviews with 23 social scientists explored the reasons for data reuse failures and found ease of reuse, understanding data through documentation, and lack of support, either from institutions, communities, or individuals (mostly data producers) to be the significant factors (Yoon, 2016). Another study that investigated data reuse in social science, archeology, and zoology reported that trust in repositories played an important role and that data processing, metadata availability, and data selection were important when reusing data (Faniel & Yakel, 2017). The development of standards and the use of repositories varied between the three disciplines studied, with archeology lagging behind the others.

Kim and Yoon (2017) found that the availability of data repositories is one of the main factors influencing data reuse at the disciplinary level. Examples of data reuse cases from specific repositories include the UK Data Service[14] (UKDS) (Bishop & Kuula-Luumi, 2017) and ICPSR (Faniel et al., 2016) for social science data, and the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (NIH) data repository[15] for clinical trials data (Coady et al., 2017). Bishop and Kuula-Luumi (2017) investigated data reuse cases for UKDS by analyzing additional information provided by users for accessing data when they registered with UKDS and found that 64% datasets were used for learning, followed by 15% for research purposes and 13% for teaching. It is noticeable that most of these studies are based on data usage from the key data repositories, perhaps because those are more mature services, well-known within the community and thus attract more users.

Even though there is increasing evidence of data reuse, few studies explored how researchers find datasets to reuse. The social media and web recruitment survey of Kratz and Strasser (2015) suggests a combination of search strategies, with the top three being: checking the literature for references, searching a discipline-specific database, or using a general-purpose search engine. Pasquetto et al. (2019) conducted a meta-analysis of multiple studies on two interdisciplinary consortia, the Centre for Embedded Networking Sensing (CENS) and DataFace Consortium and reported that sources for secondary datasets ranged from established referenced collections to disciplinary repositories. However, these studies did not explore discipline-specific differences in data reuse practices. A recent survey (using multiple online recruitment methods) suggests that even though 52% of 728 respondents self-reported reusing others' research data, the average satisfaction score for obtaining data to reuse was relatively low (Hrynaszkiewicz et al., 2021). Therefore, it is important to understand whether and how data reuse varies by a wide range of research areas and research experience, how researchers find datasets to reuse, and how data reuse practices compare to data sharing practices in those disciplines. Data reuse cases may not be limited to research only and may vary from one discipline to another. A comparative analysis of data reuse purposes across different disciplines will therefore provide an understanding of the importance of data types and sources. At the time of designing this study there was no large-scale investigation of dataset discovery methods, data reuse types and purposes, and factors that influence choice of datasets across disciplines. A recent study on data discoverability and usability of datasets in 31 disciplines that was published later partially fills this gap (Gregory et al., 2020). Nearly half of the respondents in that survey selected more than one discipline. Among those who selected a single discipline, the sample size was small (<30

---

[14] https://ukdataservice.ac.uk/

[15] https://biolincc.nhlbi.nih.gov/about/

responses) for two-thirds of the disciplines except the following: Medicine, Social Science, Engineering and Technology, Computer Science, Biological Science, Arts and Humanities, Physics, Psychology, Health professions, and the group other. Therefore, the authors occasionally reported differences between disciplines where respondents selected only one discipline. Furthermore, this study compares overall data sharing and reuse behaviour without delving into granular discipline-level commonalities and differences. Thus, in less represented fields, such as business and neurology, the findings may not be applicable to the wider population.

To address these areas, this thesis reports the results from a larger multi-disciplinary survey in 20 research areas under nine subject categories that includes previously reported disciplines along with emerging but understudied disciplines (Chapter 6). By recruiting participants using publication metadata from Scopus to collect contact information of the first authors in randomly selected publications, this study systematically targeted responses from multiple selected research areas, which is not possible to control when recruiting participants in an ad-hoc basis, as prior studies have done.

2.5 Data citation and tracking secondary data reuse

Sharing meaningful data can be time consuming, therefore researchers often want to know how their shared data are being reused (Kratz & Strasser, 2015; Wallis et al., 2013). This can be tracked if datasets are cited in a standard manner when reused for research purposes. Therefore, dataset citations can act as a professional reward system and be a major incentive for data sharing (Piwowar, 2011; Edmundus et al., 2012; Enke et al., 2012; Kim & Zhang, 2015; Kratz & Strasser, 2015; Sayogo & Pardo, 2013). The scientific community has worked together to standardize data citation practices and advocate for unique persistent identifiers, such as digital object identifiers (DOI) instead of general URLs for long-term access. Early adopters of this approach include the Natural Environment Research Council (NERC) data centers in Environmental Sciences, the Pangaea data archive in Earth Sciences, the UK Data Archive in Social Sciences, and the Bodleian Library at the University of Oxford (Callaghan et al., 2012). In 2012, NERC assigned DOIs to 14 datasets as a part of DataCite’s pilot project. At present, applying DOIs is considered to be the standard practice for most data repositories, and this also allows versioning when a dataset is updated, by issuing a derivative DOI when a new version is released. Researchers can then include this dataset DOI in the data availability section or cite and refer to a dataset in publications.

Even though it emerged from the concept of traditional bibliographic citation, data citation is more complex than article citation. Whereas text publications have a fixed form that does not change over time, interpretable as standard units, share a common format and representation model, scientific datasets are structured according to diverse data models with varied units (single datum to data subsets or aggregations) (Silvello, 2018). Lawrence et al. (2011) provided detailed recommendations for data citation formats, but a decade later, data citation is still not formalized across all disciplines. For example, whilst the number of publications using and citing GBIF data has rapidly increased since 2007, few datasets are cited in a standard format and the citation style is often determined by the journal editors (Costello et al., 2013). This is similar to life sciences data in Dryad. Mayo et al. (2016) examined data citation practices in 1,125 articles with Dryad data packages published between 2011 and 2014, finding that 74% of articles included an intratextual reference only, 2% included a reference only in the works cited section, and in 4% it was present in both. The number of articles citing data in the works cited (either alone or in conjunction with an intratextual citation) was only 8% as of 2014 (Mayo et al., 2016).

2.5.1 Understanding data citation practices

The examination of data citation practices requires the availability of systems that can automatically identify resources citing datasets. Previous studies have used the Web of Science Data Citation Index (DCI) to analyse data citation practices (Robinson-García et al., 2016; Park & Wolfram, 2017). However, DCI is not free to use and there is evidence that the system is relatively biased towards hard sciences. As of 2016, four repositories represented around 75% of the database (Robinson-García et al., 2016). The current version of DCI indexes more general data repositories, such as Figshare. Nevertheless, the citation information available for datasets indexed by GBIF was not captured by DCI at the time of this study, perhaps because GBIF is a federated data repository that indexes datasets from multiple repositories instead of hosting datasets themselves. This is an important omission, given the importance of this federated repository for biodiversity research and its relatively mature architecture.

To understand data sharing and data-reuse in Genetics and Heredity, Park and Wolfram (2017) used data from DCI, where this was the top subject category with 3.2 million records. Other subject categories with the most published datasets were: Biology, Biochemistry & Molecular Biology, Crystallography, multidisciplinary sciences, Geosciences, Oceanography, Spectroscopy, Plant Sciences, and Chemistry (medical). Data sharing varied widely between subject categories, with some fields having over 1 million shared datasets and others having none. The proportion of data with a DOI in Genetics and Heredity was only 4.6%, which makes it difficult to automatically track data reuse. Among the datasets published between 2010-2015 with a DOI, only 4% were cited. The authors also conducted a manual analysis of 148 citing articles and found that datasets were mostly cited in the main text, followed by the Reference and Supplementary sections. In a similar study, Zhao et al. (2017) performed a content analysis of 600 articles published in PLoS One in 23 research areas, grouped in 12 disciplines (a sample of 50 papers per discipline) and found that formal data attribution methods were used in a limited number of articles: only 9% used a DOI and 59% used a URL. Among the 600 articles in the dataset, 52% (*n*=312) mentioned the use of datasets in their research. Within these 312 articles, only 45% (*n*=140) had data or data-related sections, while the others mentioned data sources in their methods or other sections. Yoon et al. (2019) investigated how scholarly literature cited Health Information National Trends Survey (HINTS) data from the U.S. National Cancer Institute. Their results showed that among 250 articles citing HINTS, only 11.6% cited HINTS data and 50.8% cited HINTS-related documents. These examples demonstrate the need to further investigate data citation practices in specific disciplines to identify gaps and trends in data citation.

2.5.2 Data citation tracking - technical feasibility

While data reuse has been of increasing interest to the research community, most research has focused on scientists' data reuse practices rather than whether data repositories can find links between articles and datasets to track data reuse. Therefore, an important question remains: given the siloed nature of most data repositories and inconsistent data citation practices, how do data repository managers identify links between their datasets and any citing articles? For institutional repositories representing academic institutions, this is of particular importance for two reasons: 1. Compliance with funder mandates and 2. Track reuse of their published datasets. Academic researchers may deposit their data in repositories other than their institutional repositories and being able to track this can help demonstrate compliance with funder mandates. On the other hand, this could help to identify data reuse cases for datasets in institutional repositories and further encourage researchers by developing a reward system that acknowledges their effort to share data. Currently, most institutional repository managers rely on either manual notification of primary publication by the researchers or web search as there is no systematic method to do this automatically. Ghavimi et al. (2016) suggested a semi-automated approach using a quantitative analysis of typical naming patterns in social science datasets. Lafia et al. (2021) developed a machine learning approach using natural language processing technique to detect informal references of datasets within an article text to implement within ICPSR. While promising, these methods are yet to be tested and generalized across different disciplines and repositories. The Scholix (Scholarly Link eXchange) framework -- developed in 2016, aims to fill this gap by connecting links between datasets and articles through metadata published via DataCite and Crossref (Scholix, 2019). However, no empirical studies have been conducted to analyse the output data from the Scholexplorer API to explore its usability.

Technological solutions such as DCI or Scholexplorer are based on data citation in articles. However, previous studies have found that data are frequently used for other purposes as well, such as learning and teaching (Bishop & Kuula-Luumi, 2017; Gregory et al. 2020). Such reuse cases can be identified by requesting information from users when they download a dataset. For example, in the UKDS study by Bishop and Kuula-Luumi (2017), this information was only available when UKDS required user registration for data download purposes. Prior to 2013, none of the data collections were openly available. When datasets are openly accessible and no information about usage purpose is requested from the users, it is difficult to track societal or other impacts. Altmetric sources could be useful in finding such use cases for datasets.

A study on the effectiveness of altmetric scores by Thelwall et al. (2013) compared 11 altmetric sources with Web of Science citations for PubMed articles with at least one altmetric mention in each case. They identified that time from publication, frequency of a particular social media usage, and consideration of user groups and motives would be important for future studies. Costas et al. (2015) expanded their study by focusing on multiple disciplines - social sciences, humanities, and the medical and life sciences. They collected data from Altmetric.com and compared different indicators for those fields. Most of these studies focus on social media mentions of publications, however. For datasets, Konkiel (2013) called for using altmetrics to track various types of engagement that different stakeholders can have with a single dataset, such as discussions, formal references, and recommendations. It is not known, however, whether altmetric scores currently reflect such impact for datasets. Peters et al. (2016) explored citations and altmetric scores for research datasets in three different platforms and found that few cited research data had altmetric mentions, but more in recent years. Importantly, no studies have published content analyses of altmetric mentions to understand whether the scores reflect a particular type of impact, or none. This is particularly difficult because the content of each altmetric source must be accessed individually for such analyses. Therefore, understanding the perceived usefulness of different types of data reuse metrics and any technological barriers to implementing these services is critical for planning purposes.

2.6 Incentives for data sharing and promotion of data reuse

The current academic promotion system relies on the number of original peer-reviewed publications, the impact factors of journals where they are published, and the relative placement of the author in citation rankings (Bierer et al., 2017). Because citation counts can play an important role in research assessment, assurance of attribution motivates some researchers to participate in open science, such as through data sharing (Ali-Khan et al., 2017). To ensure proper attribution to researchers for their efforts to prepare and share data, it is important to develop a standardized method of measuring the impact of shared data. In 2013, over two-thirds of datasets in GBIF were deposited by government organizations. Due to a lack of datasets created by researchers, a more incentivized data publishing model, like that of journals was recommended to encourage scientists to offer datasets to GBIF (Costello et al., 2013).

A citation advantage for associated publications can be another incentive for data sharing. For example, for a sample of 85 cancer microarray clinical trial publications, 48% of trials with publicly shared data received 85% of the aggregated citations (Piwowar et al., 2007). Another study examined the citation rate of 10,555 studies that created gene expression microarray data while controlling for multiple variables known to predict citation rates and found that studies with publicly available data in repositories received 9% more citations than similar studies for which data was not made available (Piwowar & Vision, 2013). Similar findings were reported in Astronomy by Henneken and Accomazzi (2011), where articles with data links accrued more citations after two and four years of publishing, and on average acquired 20% more citations (compared to articles without these links) over a period of 10 years. Citing a publication linked to data does not equate to acknowledging the data, however. In previous studies, appropriate citation to data was mentioned to be the most preferred way to receive credit by the researchers (Kratz & Strasser, 2015). Download counts were regarded as the second most preferred metric for repositories and a number that can be tracked by most repositories. Fear (2013) suggests a combination of metrics for evaluating social science datasets besides citation counts: secondary impact of data reuse publications, diversity of data reuse, and the number of downloads.

Alternative metrics (altmetrics) based on social media mentions can be explored in the context of research data to inform about the usefulness of shared data. However, there does not seem to be a positive correlation for highly cited datasets between citations and altmetric mentions (Peters et al., 2015). With the recent growth of social media use by academics, this landscape may have changed in the past five years and further investigations can help understand whether altmetrics score for research datasets could be used as an incentive for data sharing and the promotion of datasets in a meaningful manner. Cautious steps need to be taken before these metrics are implemented in practice and integrated into decision-making processes in a meaningful way.

Besides compliance-focused and quantitative data-based metrics, other incentives are needed to motivate researchers to willingly engage in the data sharing process. This will in return promote further data reuse. A systematic review of the health and medical research literature revealed that among the studies that explored data sharing rates, most were observational studies and opinion pieces, with only one study that reported testing an incentive (open data badges) (Rowhani-Farid et al., 2017). This study examined the effect of implementing open data badges on data sharing rates for the journal Psychological Science, comparing the articles published in 2012 to those that were published during the first half of 2015. The results showed that data sharing increased from less than 3% in 2012 to 39% in 2015, whereas no changes were seen in four control journals (Kidwell et al., 2016). A similar pilot study was also launched by Springer in 2018 for the journal BMC Microbiology (Pearce, 2018). However, a randomized controlled trial for articles published in BMJ Open did not find any difference between the intervention and control groups, with low data sharing rates (4%) in both groups (Rowhani-Farid et al., 2020). These studies are limited to a single journal in a specific discipline. Further limitations exist in this study design which could be considered when testing such incentives across multiple disciplines.

Bierer et al. (2017) proposed a data authorship model to acknowledge the contribution of dataset creators, adapting to the current academic reward system, such as the Research Excellence Framework (REF) in the UK (REF, 2019). The authors suggest that "*Data authors are responsible for the integrity of the data set but are not responsible for the scientific or clinical conclusions of the analyses drawn from the data unless they were also listed as authors of the original manuscript. This distinction would permit healthy disagreements while acknowledging the use of a data set*" (Bierer et al., 2017, p. 1685). An author can be designated as both data author and manuscript author depending on their contribution to the manuscript. One challenge is that this model must be adopted by journals in all disciplines to ensure an equal attribution system, but this is yet to be implemented.

Hosting competitions for open datasets to find new answers could be another such incentive for data sharing as it generates new opportunities by engaging users from different backgrounds. For example, Project Data Sphere[16] (PDS) is a free digital library-data laboratory that was born from an initiative to reduce the risk of cancer, enable early diagnosis, facilitate access to the best available treatments, and hasten the discovery of new and more effective anticancer therapies. In 2014, a challenge issued by PDS to create a better prognostic model for advanced prostate cancer attracted 549 professionals from 58 teams and 21 countries, which demonstrates that the open data model can empower scientists from diverse backgrounds (Bertagnolli et al., 2017). Other large-scale open data competitions include Kaggle[17], a platform that hosts competitions around a range of open datasets and the Critical Assessment of Protein Structure Prediction[18] (CASP) experiments that have been taking place every two years since 1994. The Economic and Social Research Council (ESRC) of UK Research and Innovation (UKRI) launched the Secondary Data Analysis Initiative[19] (SDAI) open competition in December 2015, which encourages scientists to think of creative ways to use secondary data for research with societal impact. This funding opportunity is still ongoing and offers grant awards of up to £300,000 for projects that seek to use secondary data from one or more existing UK or international data source. The number of innovative research projects supported by the SDAI grant continues to grow. One such example is a public engagement and maternity data linkage study that linked data from the Office of National Statistics (ONS) birth registration and notification data to patient and hospital episode databases for England and Wales. The project was selected as one of the ten ONS Research Excellence Awards 2021 nominees[20]. These examples illustrate new ways to engage with different sources of existing secondary data. While several previous studies have focused on efficient data reuse from the data quality and access point of view, it is still important to understand if other areas of improvement are considered important by researchers and data repository managers to promote data reuse across disciplines.

---

[16] https://www.projectdatasphere.org/
[17] https://www.kaggle.com/
[18] https://predictioncenter.org/
[19] https://esrc.ukri.org/research/our-research/secondary-data-analysis-initiative/
[20] https://mailchi.mp/af20ce0fb5ec/submission2

3. Biodiversity Data Case Study

3.1 Data sharing and reuse practices in biodiversity

This study focuses on biodiversity because sharing and reusing globally collected research data is common in this field, with primary data uses being ecological studies, taxonomic works, and phylogenetic analyses (Magurran et al., 2010; Troudet et al., 2018). For instance, a survey of 370 international researchers in biodiversity sciences indicated that most (84%) agreed that *"sharing article-related data is a basic responsibility"* (Huang *et. al*., 2012, p. 401). GBIF was used as a data source because this group has been working towards developing data publishing standards for biodiversity from an early stage (Moritz et al., 2011) and the platform indexes many diverse datasets from different repositories in many countries. Furthermore, it developed an in-house system to capture and display dataset citations using Crossref Event Data (Noesgaard, 2019) and supports an API to collect citation counts for datasets on a large-scale in an automated way.

Researchers have long recognized the need to provide attribution for dataset reuse. Ingwersen and Chavan (2011) suggested a Data Usage Index (DUI) for biodiversity datasets, an indicator based on search events and dataset download instances to demonstrate the impact of data creators and publishers. However, the use of persistent identifiers for datasets was not common at that time. At present, all datasets indexed on GBIF are provided with a DOI. When a data subset is downloaded from GBIF based on a search query, it is provided with its own DOI, a generic title, and an accession date to cite. The citation attribution system assigns citations to all datasets included in subsets reused and cited by research articles.

Citing subsets complicates developing a standard model to estimate disparate and fractional contributions. As indicated by Kratz and Strasser (2014, p. 6),

> *"…to reproduce an analysis performed on a subset of a larger dataset, the reader needs to know exactly what subset was used (e.g., a limited range of dates, only the adult subjects, wind speed but not direction). Datasets vary so widely in structure that there may not be a good general solution for describing subsets."*

It is crucial that the original datasets are recognized in the right manner, and these can be indexed by relevant systems, such as Google Dataset Search (Patel, 2019), so that the data creators can receive credits. Data citation information is not captured by most data publishing platforms due to difficulties with automating the process, caused by a lack of standards in citation styles. In this context, GBIF's efforts to support data citation make it an interesting source of information to study current data citation and reuse practices in this leading field. This may lead to questions about the best practice to make citations machine-readable and how to develop a standard citation model.

Since GBIF is well-known among the scientific community in biodiversity, analysis of altmetric mentions of GBIF dataset could indicate whether altmetric scores could be used to identify any societal impacts of openly shared biodiversity data (Konkiel, 2013). This study looks beyond the numbers of altmetric sources and citation counts to explore how biodiversity data is reused and cited by the researchers and whether altmetric sources can be relied on to capture the impact of research data beyond academia. The following research questions address the lack of knowledge about data citation and reuse practices and mentions of biodiversity datasets in social media.

1. Does the type of dataset or quality of information available affect citation rates?
2. How quickly do dataset citations accrue? Has the number of articles citing GBIF datasets changed over recent years?
3. How do articles listed as citing datasets on GBIF reuse them, if at all?
4. Does the citation count on GBIF result from coherent citation practices? How does the simultaneous use of many subsets impact citation practice?
5. Do altmetric scores for GBIF datasets reflect a similar type of impact to citation counts? Are altmetric scores informative about the societal impacts of open biodiversity data?

3.2 Methods

This research applied an exploratory method to study the citation and reuse practices of biodiversity datasets and assessed the content of altmetric sources that mention those datasets. Quantitative analysis was used for the GBIF metadata and then content analysis was used for each unique citing article to collect information on citation location (Khan & Thelwall, 2019a). Further information on data reuse context in those articles was then collected to understand the reuse cases of open biodiversity data. Quantitative analysis was conducted for the altmetric scores collected for GBIF datasets and samples from four altmetric sources were then used for content analysis (Khan & Thelwall, 2019b).

3.2.1 Data collection

a) Data from the GBIF API

Metadata from 38,878 datasets was initially collected through the GBIF API in May 2018. The metadata fields retrieved included the dataset key, publishing organization key, dataset DOI, dataset type, title, description, language, homepage URL, citation, citation count, creation date, and last modification date.

A random sample of 1,000 datasets was then selected with a random number generator for a content analysis of articles that cited datasets. About 44% (437) of the datasets in the sample had at least one citing article. Between October 2018 and March 2019, a random citing article and its associated metadata was manually collected for each of the 437 datasets for full-text analysis. Download counts were also manually collected since these could not be directly retrieved through the API.

Within the random sample of citing articles for these 437 datasets, some articles appeared more than once, as these cited data subsets from multiple parent datasets on GBIF. This resulted in 102 unique citing articles. The full text of two articles could not be accessed, which allowed using 100 articles for the content analysis to explore data reuse cases. However, one of these two articles had the associated dataset listed in the references. Therefore, 101 articles were used for citation location count. The publication year, publishing journal, citation location, and contextual information of data reuse were collected for each article.

Since the data collection for citing article selection was completed in 2018, an updated dataset was collected on April 6, 2019. It included metadata from 43,971 GBIF datasets and a list of all citing articles for these datasets. This dataset was used to explore trends of dataset publishing, citation counts, and number of unique citing articles over publishing years.

b) Data from the Altmetric API

The Altmetric API[21] allows free access to the entire Altmetric database for university-affiliated scientometrics researchers. Data from the Altmetric API was collected for 43,971 GBIF dataset DOIs on April 21, 2019. Four altmetric sources – blogs, Twitter, Facebook, and Wikipedia - were selected for this study as these can contain the contextual information necessary to

---

[21] https://www.altmetric.com/research-access/

understand whether and how a dataset was useful. The phase 1 of content analysis explored which types of GBIF datasets appear frequently and what is mentioned about those datasets. Only blogs were chosen for this exploratory phase since these gave the most detailed contextual information (Shema et al., 2014). A random sample of 100 dataset records from all blog mentions was created for the phase 1 content analysis using a random number generator. For every dataset record in the blog sample, citation counts were also collected with Google Dataset Search, where available, to understand its coverage and compare citation counts between GBIF and Google. 57 out of 100 datasets in this random sample had citation counts available through Google Dataset Search. Based on the findings of phase 1, phase 2 focused on Occurrence dataset mentions only, and such datasets with altmetric scores were split into four subsets for content analysis - one for each altmetric source, where each record had received one or more mentions. Where a subset had more than 100 mentions, a random sample of 100 was selected, otherwise all records were included. Each dataset mentioned in the phase 2 sample was accessed on the Altmetric Explorer[22] to collect information about the content creator and type of mentions as this tool allows browsing and reporting on all attention data for every piece of scholarly content Altmetric has found attention for.

3.2.2 Data analyses

Preliminary explorations identified four types of datasets available on GBIF[23]. *Checklist* datasets provide a catalogue or list of named organisms or taxa and can be used as a rapid summary or baseline inventory of taxa in a given context. *Occurrence* datasets provide information about the location of individual organisms in time and space. *Sampling-event* datasets contain more granular information than Occurrence datasets, often containing abundant information to assess community composition for broader taxonomic groups. *Metadata-only* datasets describe undigitized resources in natural history and other collections.

After de-duplicating 169 records, 43,802 datasets were used for analysis. Citation counts (as reported by the GBIF API) were analysed for all types of datasets to explore the first research question. The creation dates for each dataset were processed and average citations were calculated for the years between 2007 and 2019 for Occurrence datasets to explore how long it

---

[22] https://www.altmetric.com/products/explorer-for-institutions/

[23] www.gbif.org/dataset-classes

takes to accrue dataset citations. The list of all citing articles was de-duplicated to identify all unique articles and was used to explore the distribution over each publishing year.

A content analysis for 102 unique citing articles was conducted for a random sample of 1000 datasets for exploring research questions 3 and 4. A Spearman correlation between download and citation counts was calculated for the 437 cited datasets to help assess whether these suggest a similar type of impact.

To explore research question 5, content analyses and Spearman correlation tests were performed for citation counts and altmetric scores using the cor.test[24] function from the R package stats (version 3.6.2). This follows a similar method used in Peters et al. (2016) where the authors studied correlations between citation counts gathered from DCI and altmetric scores from PlumX, ImpactStory, and Altmetric.com. Their study found no correlation between the number of citations and the overall altmetric scores and observed that some research data can have high altmetric scores even though not cited.

For the content analysis, altmetric mentions were examined to understand why these datasets were mentioned on the social web, what users talk about when discussing datasets on social media and whether altmetric mentions demonstrate dataset impact. At first the data were collected and analysed for a random blog sample of 100 records. Based on the findings, the overall dataset was then filtered for Occurrence dataset mentions only. In total, five random samples with more than one altmetric mention were analysed: 1) Blog sample for all dataset types, 2) a. Blog sample, b. Twitter sample, c. Facebook sample, and d. Wikipedia sample for Occurrence datasets.

3.3 Results

3.3.1 Dataset quality and citation rate

GBIF datasets in this sample received 176,283 citations in total. Among four types of datasets, Occurrence datasets are the most frequently cited with 9.82 citations per dataset, presumably because these offer direct evidence of the occurrence of a species (or other taxon) at a particular place on a specified date. Even though Sampling-event datasets are less frequently published (1.34% of all datasets published), these tend to receive more citations as well. In contrast, nearly

---

[24] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test

60% of GBIF datasets are Checklist datasets but these are not highly cited - suggestive of low reuse value (Table 3.1).

Table 3.1 Type and number of datasets published between 2007-19 and average citations

Type	*Datasets*	*Citations per dataset type (%)*	*Citations per dataset*
Occurrence	16,712	1,64,172 (93.17%)	9.82
Checklist	26,216	11,237 (6.38%)	0.43
Metadata-only	286	18 (0.01%)	0.06
Sampling-event	588	779 (0.44%)	1.32

Prior to 2011, Occurrence datasets were the main category of datasets available on GBIF except for a few Sampling-event and Metadata-only datasets published in 2007 and 2008. Despite the evidence of higher numbers of citations received by Occurrence datasets, there was a steady growth of Checklist datasets starting 2013 and a drop in Occurrence dataset publishing. It is unclear why (Figure 3.1).

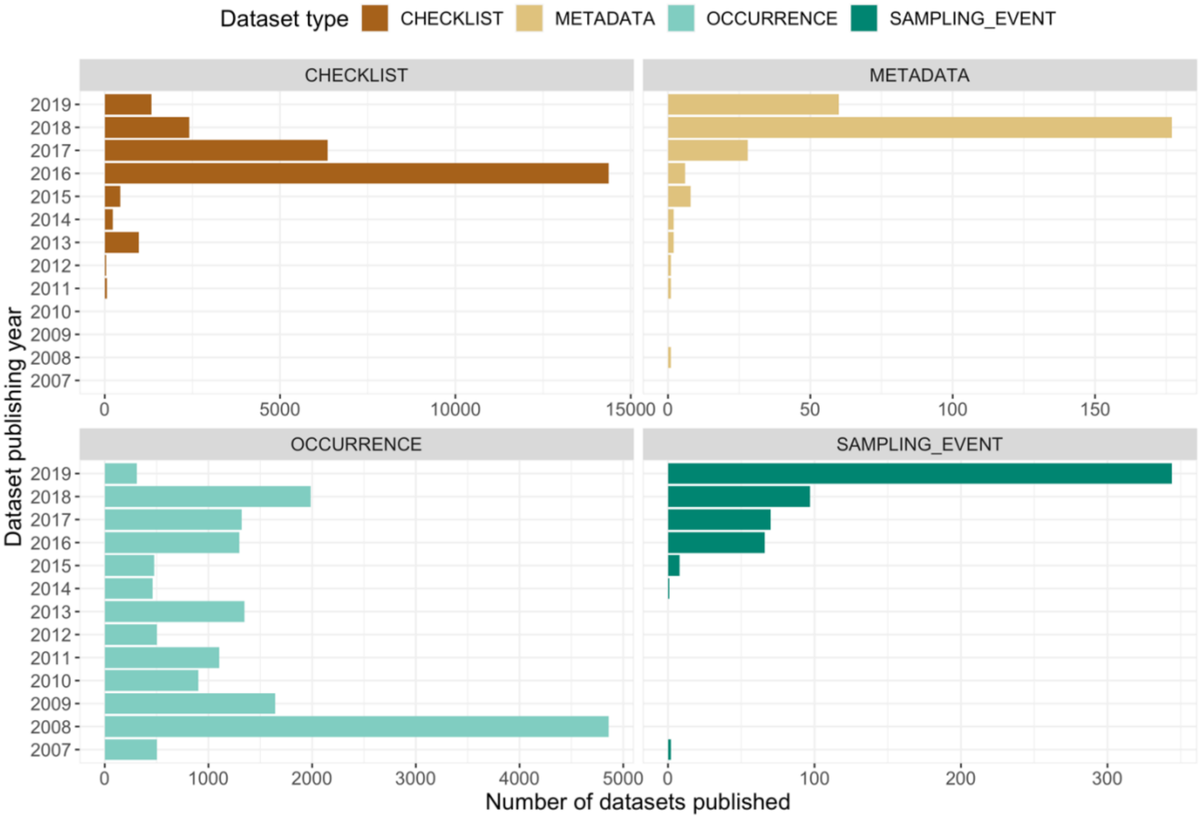

Figure 3.1 Distribution of different types of datasets published (2007-2019)

3.3.2 Citation growth rate and distribution of citing articles over the years

This study examines Occurrence datasets only since these are the type of datasets most frequently reused and cited by articles. Figure 3.2 demonstrates a relatively consistent growth in posting Occurrence datasets. The mean number of citations received per occurrence dataset was 9.8, with the highest of 24.0 for occurrence datasets published in 2015 and a lowest of 0.9 for 2018. The drop in average citations per paper after 2015 indicates that, as for articles, it can take 2-3 years to accrue most dataset citations.

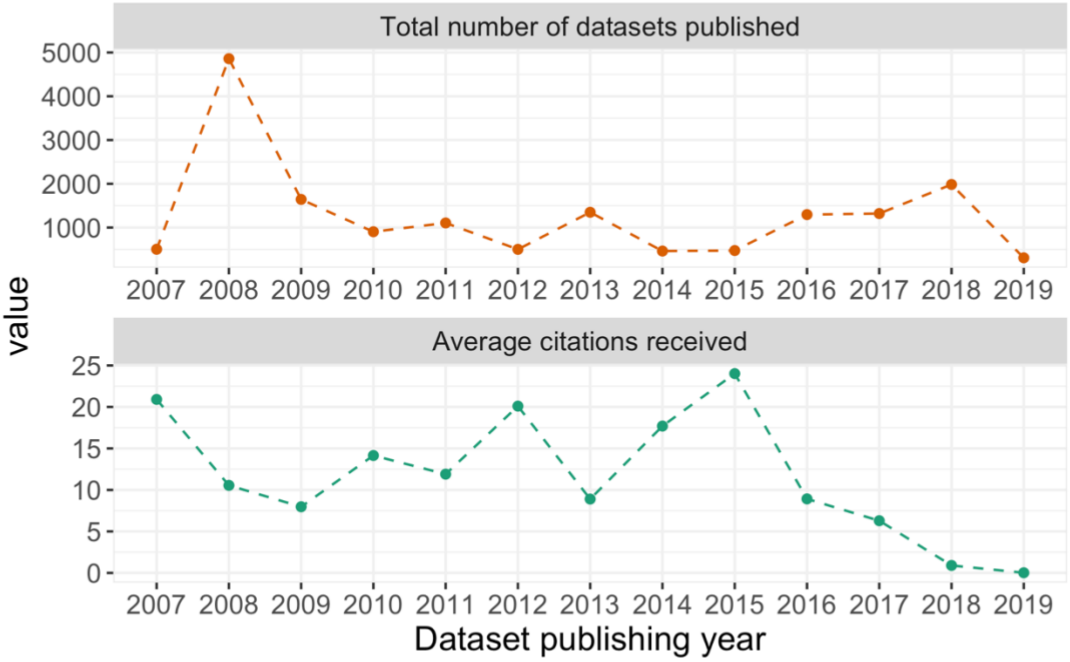

Figure 3.2. Number of occurrence datasets published, and average number of citations received

A Spearman correlation test was performed for download and citation counts for the random sample of 437 cited datasets, finding a very strong positive correlation (rho = 0.787, $p<0.001$). Thus, download counts and citation counts suggest a similar kind of impact. Because of this, early download counts might be a good indicator of longer-term citation counts. Similar to the citation count findings above, Checklist datasets (*n*=92, average downloads=2610) had much lower download counts than Occurrence datasets (*n*=343, average downloads=5211) in general.

As of April 2019, 642 articles have been listed by GBIF as citing a total of 43,802 datasets. From the data in Table 3.2, data reuse in this field (at least from this source) has been increasing since 2013 as the number of unique citing articles has been growing consistently. This is in line with the findings in the conference report from GBIF (Noesgaard, 2019). Data for 2019 only covers articles published before April so the increasing trend may have continued. Growing number of articles citing GBIF datasets indicates the importance of openly available biodiversity data for researchers.

Table 3.2 Publication year of all citing articles mentioned on GBIF

Article Publication Year	*Articles*	*Percentage (%)*
2013	4	0.6%
2014	5	0.8%
2015	23	3.6%
2016	70	10.9%
2017	178	27.7%
2018	260	40.5%
2019	102	15.9%

3.3.3 Types of data use and reuse cases

Out of 102 unique citing articles, excluding the two articles for which full text could not be accessed, the full texts of 100 articles were analysed to identify whether and how these articles used biodiversity data from GBIF. Based on the type of usage, these were categorized as 1) foreground and 2) background data (Wallis et al., 2013). Foreground data are those needed to answer the particular research questions posed in a study and background data are the type of contextual information needed to establish research questions but not to answer the research questions.

GBIF was not cited as a data source within an article published in 2017, even though it was listed as a citing article on GBIF. Among the remaining 99 articles, most used GBIF as a source of data to answer part of their research question. Two coders categorized the articles as ‘foreground’ or ‘background’ and identified in total 81 foreground data use cases and 18 background data use cases. This categorization was based on the information provided in those articles, especially the methods sections (kappa value for interrater reliability = 0.61, equating to “substantial” agreement (McHugh, 2012)). These articles were published in miscellaneous journals with the following appearing at least twice: PLoS One, Phytotaxa, Systematic Biology, Nature, Biodiversity Data Journal, Ecology and Evolution, Ecography, and Journal of Ecology.

Common foreground use cases included the following: using occurrence data to create species distribution model, combining GBIF data with other databases to answer new research questions, to analyse and observe the change in biodiversity data collection practices, to investigate sampling and taxonomic bias, using the GBIF backbone taxonomy to solve taxonomic problems (synonyms), to create a species temperature index (STI), and to generate a conservation network. Data papers were also identified as foreground uses since the accompanying datasets were the basis of those articles. On the other hand, the following background use cases of biodiversity data were found: using test datasets for testing software or tools, to explore and establish the need for further research in that area, simulation of models, data mining, creation of a baseline model, and comparison with previous records of a species' prior occurrence. Such diverse use cases of biodiversity datasets demonstrate the importance of open data in this field and that it creates myriads of opportunities to study biodiversity and related fields.

3.3.4 Citation practice for biodiversity datasets

A content analysis of 102 unique articles was conducted to understand data citation practices in biodiversity articles citing GBIF datasets. Between the two articles for which full text could not be accessed, one included the associated dataset in the references section. This article was included, resulting in 101 articles in total (Figure 3.3).

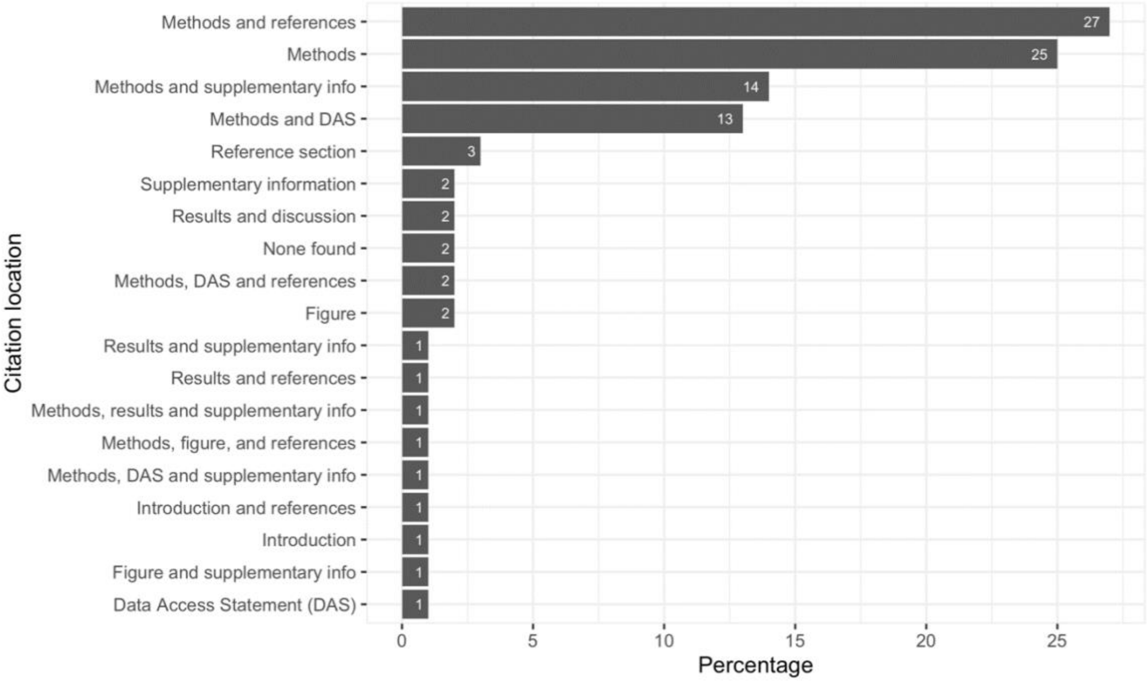

Figure 3.3 Citation location in randomly selected articles (labels on bars represent the number of articles)

Citations to GBIF datasets could not be located in two articles out of 101. For the remaining 99 articles, 48% cited datasets in a standard manner that can help capture dataset DOIs in an automated method. Among these, 32% articles mentioned a dataset in their reference lists, besides citing the dataset within different sections of the article: methods (30%), introduction (1%), and results (1%). 16% of the articles included data access statements (DAS) in addition to citing the dataset in the methods section. One article included a DAS and three included the datasets in the reference section without in-text citations. In addition, 25% mentioned the datasets in the methods section only within the text, which is difficult to find with indexing systems. Mentions in methods and supplementary material sections were also common (14%). 59 articles in the sample were published in 2018 and 23 in 2019, where recent articles are increasingly adopting a standard method of citing data. 16 out of 23 articles published in 2019 cited the datasets in references and data availability sections besides methods, which is encouraging.

Most (53%) articles cited one GBIF subset, but some cited many (9% cited at least 50 subsets). When comparing the number of subsets listed within the articles and on the GBIF records of the corresponding articles, the number of subsets did not match for 5% of the articles. For example, “Species and river specific effects of river fragmentation on European anadromous fish species” (DOI: 10.1002/rra.3386) cited one GBIF subset but the record on GBIF for that article listed 16 subsets. Non-standard citation methods were especially used by articles that utilized large numbers of datasets (12~50), potentially making it difficult to include them all in the reference section.

3.3.5 Do altmetrics reflect data impact?

21.46% (n=9,398) dataset records received a non-zero altmetric or attention score, which indicates how many people have been exposed to or engaged with a scholarly output. Such examples include: mentions in the news, blogs, and on Twitter; article pageviews and downloads; GitHub repository watchers. Out of 43,802 datasets, 4.8% (*n*=2,111) were mentioned in blogs that were included in the manually curated list of 9,000 blogs by Altmetric, 16.6% (*n*=7,271) were mentioned in tweets, 7.9% (*n*=3,459) were mentioned in Facebook posts (curated list of public pages only), and 4.4% (*n*=1,913) were mentioned on Wikipedia.

Content Analysis - Phase 1

Random sample 1 – Blog mentions (All datasets)

Blogs were used as the first sample for content analysis as blog posts do not have limited word counts, can be combined with different media, and bloggers who write about scientific topics are often domain experts (Shema et al., 2014). Science blogging is used for explaining science to the public, and thus can bridge the gap between research and other parts of the society (Bornmann, 2015).

In the random sample of 100 GBIF dataset records that were mentioned by one or more blogs, 99 were Checklist datasets and only one was an Occurrence dataset. In 98% of these cases, the altmetric mentions were for articles associated with datasets instead of the dataset itself. In addition, nearly all blogs mentioned the discovery of a new species. Perhaps this could be due to Checklist datasets cataloguing named organisms or taxa, and the associated articles providing the most in-depth information for learning about these newly discovered species. The most frequently appearing blog was Species New to Science (*n*=52) and Earthling Nature and its Portuguese version, Natureza Terráquea (*n*=18) (Table 3.3). Others included blogs by citizen scientists, academic blogs that list faculty publications and blogs from news outlets. The only Occurrence dataset that was mentioned in a GBIF blog described a study that introduced a novel dynamic occupancy model for handling known sources of bias. While learning about new discoveries in biodiversity is important, the following stages of this study focused on Occurrence dataset mentions only in phase 2. The goal was to understand what is mentioned about these heavily used datasets, with an anticipation that altmetric mentions will be about the datasets instead of associated articles.

Table 3.3 Top 5 blogs mentioning GBIF datasets

Blog name	**Dataset mentions**
Species New to Science	52
Natureza Terráquea	11
Earthling Nature	7
DNA Barcoding	6
Pensoft Blog	5

Google Dataset Search was queried for these 100 datasets in the random sample. This did not return any results for 43 GBIF datasets. Among the 57 datasets that were found, the citation counts varied greatly from what was listed on GBIF. The most striking difference was noticed for the dataset "Artportalen (Swedish Species Observation System)" (DOI: 10.15468/kllkyl) where GBIF listed 109 citations at the time of this search, whereas Google listed only 3 citations. Another example was the Checklist dataset, "Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny" (DOI: 10.1046/j.1439-0469.2001.00155.x), for which Google listed 28 citations and GBIF did not list any. Even though Google Dataset Search was still in its beta form, these differences evidence that GBIF's semi-automated citation attribution system may not be consistent, and the citation counts are not being captured by the general indexing systems at present.

Content Analysis – Phase 2

Checklist datasets received a majority of the altmetric mentions (Table 3.4). This is not surprising given the phase 1 results demonstrated that most of these dataset mentions were linked to their associated articles publishing about new discoveries, and these are likely to be mentioned on social media. Very few of the Occurrence datasets received altmetric mentions, with tweets being the most common and the rest having below 100 mentions. A random sample of 100 was selected for Twitter, and blogs, Facebook, and Wikipedia included all mentions.

Table 3.4 Distribution of altmetric mentions for different types of datasets

Altmetric sources	*Occurrence*	*Checklist*	*Sampling-event*	*Metadata-only*
Blogs	11	2,099	0	1
Twitter	403	6,700	128	40
Facebook	74	3,378	5	2
Wikipedia	5	1,908	1	0

Sample 2a – Blog mentions (Occurrence datasets)

In total 11 datasets had received one or more blog mentions, which includes the two blogs from the random sample in phase 1 (Table 3.5). One post from the Teaching Biology blog was subsequently deleted, so 10 blog posts were analysed. Two of these were from the GBIF Blog and two from the iPhylo blog. The rest of the blog sources were different. All the blogs were written in English except for one blog in Dutch.

Despite being uncommon, the blog posts provided useful insights into the usability and impact of datasets. All the blogged datasets received one or more citations (highest 284), as listed on GBIF, but Google Dataset Search gave very different results again.

While the blog posts are often from the data publishers or creators, these can be useful sources of information for understanding unique use cases of open biodiversity data (Table 3.5). For example, the blog post "App combines computer vision and crowdsourcing to explore Earth's biodiversity, one photo at a time" mentions improving computerized species detection using research-grade observation data published on iNaturalist. Similarly, the blog post "Estimating changes in seasonal site occupancy using opportunistic observations" explains about a study that used the dataset to introduce a novel dynamic occupancy model that attempts to cope with known sources of bias, including lack of data and variation in sampling effort. Other blog posts pointed out why those datasets are important, and how these can be used for research and other purposes in the future.

Sample 2b – Twitter mentions (Occurrence datasets)

For this random sample of 100 Twitter mentions, data was collected on who tweeted and the content of the tweet, except one missing tweet. Most (79) tweets were from the data publisher or the GBIF account, and the rest were from the following: data creators (9), domain experts (6), data management/ research data specialists (2), journal publishers (2), and a museum (1). 70 of these were tweets and retweets about publishing new data, four were on expansions and updates of existing datasets, and three on data paper publishing. A few tweets were less generic. For example, one tweet linked to a news article by the Chicago Tribune on the importance of data published on GBIF, another one indicated that the dataset should be a starting point of many discussions, and another reply to a tweet expressed gratitude for sharing the data.

Table 3.5 Content from blog mentions of Occurrence datasets

Dataset title	*Blog title*	*Blog content*
Global compendium of Aedes albopictus occurrence (10.15468/7apj8n)	GBIF and impact: Crossref, FundRef, and Altmetric	Blog post on understanding the impact of GBIF data.
iNaturalist Research-grade Observations (10.15468/ab3s5x)	App combines computer vision and crowdsourcing to explore Earth's biodiversity, one photo at a time	Blog post on how research-grade observation data published on iNaturalist and GBIF are improving computerized species detection.
EOD – eBird Observation Dataset (10.15468/aomfnb)	eBird 2017: Year in Review	Blog post on eBird data acquisition in 2017 with link to its dataset on GBIF.
International Barcode of Life project (iBOL) (10.15468/inygc6)	iBOL DNA barcodes in GBIF	Blog post by the dataset author on expansion of his dataset to include 2.7 million barcodes.
Artportalen (Swedish Species Observation System) (10.15468/kllkyl)	Estimating changes in seasonal site occupancy using opportunistic observations	Blog listing a study that used data from this dataset to introduce a novel dynamic occupancy model that attempts to cope with known sources of bias, including lack of data and variation in sampling effort.
Published Chenopodium vulvaria observations (10.15468/oyorvb)	Rejuvenating Centuries' Old Botany with Phytogeography	Blog post on biogeography using historical data on plant distribution, with a link to the associated geo-reference in GBIF.

CABI Africa Invasive and Alien Species data (10.15468/pkgevu)	Largest Invasive Alien Plant dataset is now published online!	Blog by the data publisher on the context and impact of this large dataset on invasive series.
Xeno-canto – Bird sounds from around the world (10.15468/qv0ksn)	Vogelgeluiden van tienduizend vogelsoorten online beschikbaar	Blog post with the contextual and content description of the dataset made available on GBIF by the Netherlands Biodiversity Information Facility.
Occurrences of the invasive plant species Heracleum sosnowskyi Manden in the Komi Republic territory (10.15468/zo2svq)	Tracking the invasion of Sosnowsky's hogweed in the Komi Republic	Blog post about a data paper on the collection and use of the associated dataset.
DNA barcoding the fishes of Lizard Island (Great Barrier Reef) (10.3897/bdj.5.e12409)	Tuesday reads	Blog post on a study that conducted short expedition to collect DNA barcodes of the fishes in Lizard Island.

Sample 2c – Facebook mentions (Occurrence datasets)

There were 74 mentions of Occurrence datasets on Facebook, all of which were used for content analysis. Similar to Twitter, most Facebook posts were promotional and contained news on publishing new datasets, data papers, and updating existing datasets. Table 3.6 shows the distribution of content creators, where the Biodiversity Information System of Colombia (SiB Colombia) is the most active promoter.

Table 3.6 Distribution of content creators for Facebook mentions

Content creator	**Number of posts**
Biodiversity Information System of Colombia (SiB Colombia)	41
GBIF	17
Data journal publisher	7
Community	5
Journal publisher	2
Data creator	1

Even though most Facebook posts were for promotional purposes, these posts often contained information on the usefulness of a dataset. Here is an example post from SiB – *"The Selva Association published a set of #OpenData with more than 4,000 biological records from the Serranía del Darién, a key region for the movement of different animals between Central and South America. The objective of the project in which the registries were carried out was to determine the importance that this #biodiversity can have as a source of income and to formulate an ecotourism plan to ensure greater knowledge and better preservation of all the species in this area. These are some of the animals that were registered. Freely consult all the data: http://doi.org/10.15472/7okxxe*."

Another post from GBIF sent out call for application for the 2018 FGVCx Fungi Classification Challenge that used the Danish Mycological Society, fungal records database on GBIF for training and validating images. These types of use cases show creative ways of encouraging open data use outside of academic research.

Sample 2d – Wikipedia mentions (Occurrence datasets)

Wikipedia mentioned only five datasets. Each received an altmetric score of one on Wikipedia. One of the datasets (DOI: 10.3897/bdj.5.e11794) was linked to the Biodiversity Data Journal and the author of the Wikipedia article was the first author of that paper. A second dataset (DOI: 10.3897/zookeys.73.840) was linked to ZooKeys journal and referred to the ZooKeys article for species information. In the Wikipedia article on plant Rheum lhasaense, the dataset (DOI:10.15468/o3pvnh) was used to describe its distribution; an article on Sirgenstein Cave used the dataset (DOI: 10.1594/pangaea.64558) mentioned to describe how the species was organized; and an article on Colpomenia sinuosa used the dataset (DOI: 10.5519/0002965) to refer to a synonym. These demonstrate that use cases of GBIF datasets in Wikipedia articles are similar to their use in academic articles, even though rare.

Correlation tests

Spearman correlations tests were performed by dataset publication years and types of datasets. Few datasets had any altmetric mentions until 2015, ranging between 4 to 17 between 2007 and 2014. In the following years, the number of datasets with altmetric scores were the following: 101 in 2015; 3,136 in 2016; 3,547 in 2017; 1,082 in 2018; and 1,460 in 2019. From the observations above, the number of Checklist datasets published rocketed in 2016 (Figure 3.1) and most of the altmetric mentions were about Checklist datasets (Table 3.4). This explains the rapid growth of altmetric mentions from 2015 to 2016. Due to the lower number of mentions in previous years, correlation tests were conducted for the years 2016 to 2018; excluding 2019 due to fewer datasets (n=7) with any journal article citations to perform correlation tests.

The total number of datasets that received any altmetric mentions for different types of datasets are: 8,773 Checklist datasets, 457 Occurrence datasets, 40 Metadata-only datasets, and 128 Sampling-event datasets. Below are the results for each altmetric sources, in terms of years (Table 3.7) and dataset types (Table 3.8, excluding Metadata-only datasets). The results show no strong correlations between citations and altmetric mentions when compared for different years. Correlation tests for different type of datasets show weak to moderate correlations between the number of tweets and citations for Occurrence and Sampling-event datasets.

Table 3.7 Spearman correlations between citations and altmetric scores for 2016-19

Dataset publishing year	*Blogs*	*Twitter*	*Facebook*	*Wikipedia*
2016	Weak negative, non-significant (*n*=419, r= -0.03)	Weak (*n*=2,410, r= 0.28, p <0.001)	Weak negative, non-significant (*n*=543, r= -0.06)	Weak negative, non-significant (*n*=653, r= -0.02)
2017	Weak negative (*n*=844, r= -0.16, p <0.001)	Weak negative (*n*=2,610, r= -0.039, p=0.04)	Weak negative, non-significant (*n*=2,084, r= -0.02)	Weak negative (*n*=900, r= -0.15, p<0.001)
2018	Weak negative (*n*=285, r = -0.14, p=0.016)	Weak negative (*n*=865, r = -0.13, p<0.001)	Weak (*n*=514, r =0.1, p=0.02)	Weak negative, non-significant (*n*=222, r= -0.06)

Table 3.8 Spearman correlations between citations and altmetric scores for different types of datasets

Dataset type	*Blogs*	*Twitter*	*Facebook*	*Wikipedia*
Checklist	Weak negative (*n*=2,099, r= -0.24, p<0.001)	Weak (*n*=6,700, r= 0.06, p<0.001)	Weak negative, non-significant (*n*=3,378, r= -0.003)	Weak negative (*n*=1,907, r= -0.08, p=0.001)
Occurrence	Moderate, non-significant (*n*=11, r= 0.3)	Moderate (*n*=403, r= 0.46, p<0.001)	Weak, non-significant (*n*=74, r=0.19)	Not enough data (*n*=5)
Sampling-event	No mentions	Moderate (n = 128, r= 0.62, p<0.001)	Moderate, non-significant (*n*=5, r= 0.59)	No mentions

3.3.6 Is GBIF attributing citations in the best way?

The current semi-automated approach of GBIF assigns citations to all original datasets when any subsets downloaded from GBIF are cited by a research article. While theoretically this is the right approach, it does not consider the scenario that after downloading the subsets, researchers often curate the data to meet their purposes. For example, one of the citing articles in this study sample used GBIF data to explore the usefulness of Digital Accessible Knowledge in biodiversity for terrestrial mammals distributed across the Iberian Peninsula. The authors reported that 616,141 out of 796,283 retrieved records were unfit for their use case due to quality issues (Escribano et al., 2019). Even though this might be unavoidable when using open data for specific research purposes, assigning citations to the datasets that contained deleted data would result into erroneous citation counts.

This issue was further explored by examining a dataset that contains a single record of marine mammals (Marine mammals of the Seaflower Biosphere Reserve, DOI: 10.15472/uzo3mq, GBIF with the following UUID: 6f2b8f8d-4e29-40b8-a022-e3a0e642c89e). The only observation the dataset contains is of the Stenella attenuata spotted dolphin (Gray, 1846) from Colombia. However, one of the four articles citing this dataset was, “The shrews (Cryptotis) of Colombia: What do we know about them?” with DOI: 10.12933/therya-19-760. This article used the GBIF subset with DOI: 10.15468/dl.hjv2ad that was derived by searching for “Colombia” and included 5,552,450 downloaded occurrences. However, given shrews are not marine animals, living in forest and cultivated areas[25], the observation in the original dataset was unlikely to be relevant to the article. Therefore, listing it as a citing article would be considered misleading.

3.4 Recommendations

To avoid the issue with attributing citations mentioned above, I propose that GBIF allows data users to deposit the metadata of their cleaned and curated datasets used for research purposes separately and assign DOIs to those subsets. Upon uploading those subsets into GBIF, the platform can continue to use the current semi-automated approach to apply citations to the specific datasets from which occurrences or geo-references were used (Figure 3.4). Furthermore, at present the

---

[25] https://en.wikipedia.org/wiki/Colombian_small-eared_shrew

subsets downloaded from GBIF are assigned a generic "GBIF Occurrence Download" title, which is not informative in terms of understanding the content of that subset or differentiating between multiple subsets when citing those subsets. Therefore, I suggest automatically assigning a title based on the search query or content of that subset. My proposed change will have the following benefits: 1. Assign citations to the correct datasets only, 2. Inform other users about the dataset content by providing meaningful title, and 3. Allow other users to learn about various use cases of GBIF datasets when they explore the list of datasets linked to citing articles, which can lead to generation of newer ideas of research by identifying trends and gaps.

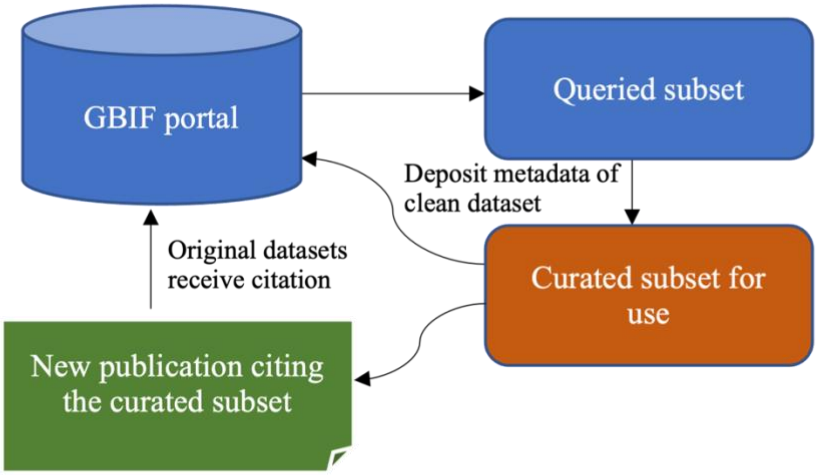
GBIF portal
Queried subset
Deposit metadata of clean dataset
Original datasets receive citation
Curated subset for use
New publication citing the curated subset

Figure 3.4 Recommendation for revised GBIF citation attribution system

3.5 Limitations

The content analyses results presented here are from relatively small samples of citing research articles and altmetric mentions. This is due to multiple reasons: 1. The list of citing articles and altmetric contents in our samples had to be manually curated by accessing individual datasets on GBIF and individual mentions on the Altmetric Explorer, since bulk download is not currently possible; 2. Citing articles were collected from different journals with different access restrictions, which limits downloading all full texts at once; and 3. Due to varied citation practices and reuse cases, manual content analysis was conducted to capture the information that would be needed to develop automated methods in the future. However, this process is time consuming. Easy access to the contents of citing literature and altmetric sources will allow development of a systematic approach using text mining methods that would speed up the process by eliminating manual analyses.

3.6 Discussion

Open biodiversity data creates new scopes of research in miscellaneous ways. Long-term datasets are indispensable as baseline data to compare and judge efforts to reduce the rate of biodiversity loss (Magurran et al., 2010). GBIF as a platform allows access to and analysis of combined data from multiple sources by using Darwin Core - a standard metadata schema. Such data are used by researchers for both foreground and background purposes, as well as to help generate and answer new research questions that would be challenging otherwise. A recent GBIF report supports these findings, claiming that researchers publish studies using GBIF data at a rate of about two papers a day (Noesgaard, 2019). Providing proper attribution is important to recognize the efforts of thousands of data creators and publishers who openly publish their data on GBIF. This case study found that citing data in references or data access statements is becoming more common, but citation practices remain inconsistent across different journals. Noesgaard (2019) found similar results where the number of articles using and acknowledging GBIF-mediated data in a standard manner increased from 15% in 2018 to 60% in 2019. An important finding for biodiversity datasets is that articles using many data subsets (12~50) pose extra challenges for citing them in an appropriate manner. Publishing a data paper for the articles using many subsets and citing the paper itself could be a solution to this issue (Chavan & Penev, 2011; Edmunds et al., 2012). However, a refined and standard model should be adopted to address this problem when a data paper is not available. This study identified that not all downloaded data are reused as data cleaning almost always takes place before using data for specific research and recommends an alternative approach for GBIF to attribute citations to datasets to avoid erroneous mass citations.

This chapter also investigated whether altmetric scores are informative about the impacts of open biodiversity data. The correlation test results of citation counts and altmetric mentions, and content analysis of the first blog sample showed that Checklist datasets tend to receive high altmetric score due to their link to new findings or discovery. However, when researchers download data from GBIF, they predominantly use Occurrence datasets rather than Checklist datasets, as shown by higher citation rates of Occurrence datasets. Citation counts and altmetric mentions are not positively correlated and in some cases are negatively correlated. Occurrence datasets showed moderate correlations for Twitter and blog posts, and a weak correlation for Facebook posts. Even though most of these social media posts are from the data publishers and data creators, perhaps promotion leads to more reuse of such open datasets or popular datasets are promoted more frequently. However, tweets were less rich in content than Facebook and blog posts and did not provide any insight on data reuse cases. This could be because of the previous character count limitation on Twitter and tweeters focusing more on promotion. Among the four platforms compared, blog posts can be informative for those who are keen to learn about usability and use cases of open data. Such blog posts should be encouraged more to advocate creative solutions using open research data and capture societal impact.

3.7 Summary

This chapter has explored data citation and reuse practices in biodiversity. It found evidence that openly available biodiversity data on GBIF is frequently reused by researchers and that the number of articles reusing and citing data retrieved from GBIF has been increasing steadily. Types of data reuse cases are diverse and indicate that open biodiversity data support creative research in the field of biodiversity and beyond. For example, by creating species distribution models with existing data, researchers can identify scopes of new research, determine any changes in biodiversity in a particular area, and compare their findings against a baseline model. This demonstrates the impact of open data, and researchers, data managers, and policy makers should identify how this type of knowledge flow can be encouraged in other fields and for different data formats as well. While this case study produces evidence that standard data citation is increasing among recent studies that cite biodiversity datasets published on GBIF, it is evitable that users of large numbers of subsets make data citation challenging. The recommended approach from this study has been partially adopted by GBIF recently. The platform now allows users to deposit metadata from a cleaned 'derived'[26] dataset that contains a parent dataset relationship. This will assure that citations are attributed to dataset records that were retained after initial data cleaning. GBIF also agrees that using meaningful titles for each subset will be useful in the future as it becomes difficult to refer to a dataset with generic title.

---

[26] https://www.gbif.org/citation-guidelines#derivedDatasets

Finally, this is the first study to conduct a content analysis of altmetric mentions of biodiversity datasets and suggests that mentions of datasets and use cases in blogs, Twitter and Facebook may lead to further data reuse. Future studies can explore whether this is similar for datasets in other disciplines, published via a different data repository.

4. Scholexplorer Case Study

The previous chapter examined data citation and reuse practices in biodiversity by analysing article and dataset links created by GBIF. When investigating article-dataset links, two types of article and dataset relationships are considered important by the data producers (e.g., researchers, doctoral students), institutional repository managers (e.g., data librarians, archive support staff), and research managers (e.g., pro-vice-chancellors, institute directors, and research committees): (a) links between primary datasets and research articles associated with them to prove compliance with funder mandates; and (b) links between a dataset and any articles that reused it, demonstrating the impact of that dataset. Both data repositories and academic institutions can benefit from systematic and automated capture of such links. Therefore, this chapter explores whether and how a current technological solution can automate article and dataset linking process. This case study uses the Scholix (Scholarly Link eXchange) framework, which is based on establishing links between datasets and articles using event data published by DataCite and Crossref (Scholix, 2019). By developing and using a set of Python code to collect openly available Scholexplorer data through the API, this study investigates its usefulness and provide recommendations.

4.1 Identifying links between datasets and articles

4.1.1 Compliance with funder mandates

A major push from funders to make research outputs, including research datasets, openly available (Holdren, 2013) has resulted in an increase in the number of research data infrastructures within higher education institutions and many new policies for research data support. Organizations and committees, such as the Research Data Alliance (RDA)[27], FORCE11[28], and the Committee on Data for Science and Technology (CODATA)[29], are supporting this rapidly changing research environment and tackling emerging issues in the field of research data management. However, due to differences in domain practices and requirements from the funders, fields have been moving at their own pace to adopt and adapt cultures of data sharing.

---

[27] https://www.rd-alliance.org/
[28] https://force11.org/
[29] https://codata.org/

Institutional repository managers and research managers in academic institutions are expected to comply with the funder mandates for research data sharing, thus often needing to identify research datasets published in external data repositories (in cases where an institutional repository is available). Nevertheless, since most research data repositories are designed to act as silos (in the sense of not being connected to other repositories) and searches by author affiliation are often not viable, finding datasets published by institutional researchers and data producers in external archives can be arduous and even unfeasible. The problem is particularly acute for institutional repository managers, who are mostly reliant on web searches or manual notifications from data producers when they publish their data in an external repository, in order to be able to report on policy compliance or impact to research managers.

4.1.2 Identifying data reuse cases

Due to the substantial time and effort required to document and share high quality research data, it is important to know whether shared data will be reused (Kratz and Strasser, 2015; Wallis et al., 2013). Several studies have explored associations between shared research data and the citation rates of articles in different fields, such as cancer microarray data in genomics (Piwowar et al., 2007), astronomy (Henneken & Accomazzi, 2011), astrophysics (Drachen et al., 2016), and for open access journal articles published from PLOS and BMC (Colavizza et al., 2020). All report higher citation impact for articles sharing research data.

Despite the evidence of increased citation impact for articles that share research data, few studies have explored data citation practices and the reuse of shared data. This is largely due to the lack of standards in data citation practices across different fields and journals. Mayo et al. (2016) investigated data citation practices for articles in the Dryad repository and found that the number of articles that cite data in their citation sections increased from 5% to 8% between 2011 to 2014 while intratextual citation had grown from 69% to 83%. As reported in Chapter 3, 27% of articles citing biodiversity datasets indexed by GBIF cited them in their methods and references, 13% in methods and data access statements, only 2% in all three sections, and the rest (58%) intratextually and in supplementary information. Inconsistent data citation was common in articles using a large number of data subsets in biodiversity (12~50). Colavizza et al. (2020) investigated 531,889 journal articles published by PLOS and BMC and report that even though publisher mandates have resulted in data availability statements becoming more common, statements containing a link to a repository are still just a fraction of the total. That could be the reason why an attempt to capture data citation using DCI by Robinson-García et al. (2016) found that 88.1% of records received no citations, since DCI only harvests citations from the references. Mathiak and Boland (2015) and Ghavimi et al. (2016) explored variations in citation practices for social sciences datasets. Both studies proposed automated methods to approach these challenges, an algorithm to find dataset citations in full text documents and a linked data approach by developing an ontology, respectively. However, challenges remain in universal implementation of such methods given the variations of data citation practices.

To resolve these issues, Silvello (2018, p.12-13) suggests that *"…[T]he ideal data citation system should uniquely identify a data set and subsets of it with different levels of coarseness (identification), attribute the ownership and responsibility of the data with variable granularity to the right people/institutions (attribution), guarantee the persistence of the data being cited as well as the citations themselves (fixity), and automatically create complete and consistent citation snippets (completeness and consistency) according to community practices and shared metadata standards.*" While more journal publishers are adopting standardized methods, such as data access statements, to link research datasets within journal articles, there is still no agreement on how data should be cited. This has further complicated the task of finding links between articles and datasets.

Inconsistency in citation practices, as mentioned above, makes it difficult to automatically find evidence of data reuse and citation by secondary data users. As a result, data producers may be unable to demonstrate the impact of their published data or claim full credit for their work. This leaves a big knowledge gap for both researchers and institutions, and especially for institutional repository managers who need to maximize their resources by developing systematic approaches to identify article and dataset links. It is important to fill this gap by investigating whether data producers are complying with funder and publisher requirements and making their data publicly available for both reproducibility and reuse.

4.1.3 Scholexplorer as a technical solution

The Scholix framework could be beneficial to address the needs mentioned above. Data collected using this framework is aggregated by Scholexplorer and made freely available by its REST API (Burton et al., 2017a; The OpenAIRE Scholexplorer, 2020). Multiple articles have discussed the mechanisms and scholarly benefits of this framework (Burton et al., 2017b; Cousijn et al., 2019; Hersh, 2019). Limani et al. (2018) used an alternative approach to establish links between research datasets in the Journal Data Archive (JDA) and publications in economics that were published in the EconBiz portal and reported that the links found using their approach could be valuable for Scholexplorer. Several higher educational institutions, such as University of Manchester (Gibson, 2019), Durham University (Syrotiuk, 2019), and University of Illinois at Urbana-Champaign (Tay, 2018) have explored the Scholexplorer data and developed individual processes to incorporate this into their systems. However, no published empirical studies have analysed output data from the Scholexplorer API to identify: (a) who publishes the linked datasets; (b) whether the data can identify data reuse cases; and (c) how typical links are generated.

To explore the usability and quality of the data derived from Scholexplorer, this chapter built on and extended a set of Python code originally developed by Durham University (Syrotiuk, 2019) to collect data from the Scholexplorer API for approximately 31,890 research outputs published by the University of Bath researchers. This includes all research outputs recorded in the university’s current research information system (CRIS) and publications repository, Pure[30], until April 17, 2019. University of Bath systems were used as the data source for this case study to derive a comprehensive and reliable list of research output DOIs from Pure. This was required as the Scholexplorer API did not yet support affiliation search, and to be able to use the University of Bath’s Research Data Archive (UBRDA) as a benchmark to compare against the Scholexplorer API output. Additionally, dataset DOIs published via UBRDA were also extracted to explore whether these datasets were used in a secondary publication.

By November 2019, there were 332 UBRDA datasets registered on DataCite (University of Bath Research Data Archive, 2019). Staff in the University of Bath Library’s Research Data Service who support the UBRDA have developed methods to locate datasets published by the university researchers in some external archives. However, these methods are too resource intensive for a

---

[30] https://www.bath.ac.uk/services/pure/

small team to routinely use, which is likely to be a common problem internationally that is time-consuming to address. However, as an aggregator of data from many journal and data publishers, Scholexplorer might provide a systematic approach for solving this issue. The following research questions of this case study assess this potential from the perspective of academic institutions and institutional repository managers who need to demonstrate compliance to funder mandates.

1. To what extent can the Scholexplorer API identify previously unknown links between university research outputs and datasets in external archives?
2. Can Scholexplorer identify examples of data reuse cases of published data?

4.2 Methods

To collect data from the Scholexplorer API, a set of Python code was developed, which is available on UBRDA (Khan, 2020). The data analysis was performed using the R statistical programming language (The R Project for Statistical Computing, 2019). The code development was based on prior work from Durham University by Nicholas Syrotiuk, which was then modified to generate output based on study requirements (Syrotiuk, 2019). For example, the results contain many literature-to-literature links in addition to dataset-to-literature links. These could be from citing literature or links to articles in data journals. Therefore, the code was updated to limit this case study to direct dataset-to-literature links only. A reversed version of the code was also created that searches the API for all datasets published via a repository to identify secondary data reuse. This was used to identify all literature linked to datasets published via UBRDA.

The code is designed to search by DOI only, since this is the main standard identifier used by DataCite and has not been tested for other forms of persistent identifiers, such as handle. In its documentation, Scholexplorer mentions that any kind of persistent identifiers can be compatible, including URLs. However, URLs are not resolved in general as the variability of the associated resolvers cannot be handled by one service (The OpenAIRE Scholexplorer, 2020).

API connection failures occurred when testing large numbers of DOIs. To avoid disruption when collecting a large number of metadata, the first program (`get_data.py`) is simpler and searches for any research output DOI that has links to at least one dataset and then creates a set of DOIs for which any dataset-literature links have been found. It also reports the total number of links found, including literature to literature links, the total number of dataset links found including any subsets of datasets, the total number of research outputs for which at least one unique link to a dataset has been found, and the number of research outputs for which no links were found.

The second program (`metadata.py`) then uses the subset of research output DOIs gathered from the first program for which one or more dataset-literature links have been found. In this step it parses the JSON results from the API and gathers metadata for authors of research outputs, DOIs of associated datasets and dataset authors, and dataset publishers. The output is stored in the format of tab separated text (.tsv), which can then be cleaned and analysed. There were some records where the cell alignments were not consistent and required manual cleaning in Excel. A simplified workflow of the python code is shown below:

Step1: All research output DOI → `get_data.py` → Research output DOI linked to at least one dataset

Step 2: Output from Step 1 → `metadata.py` → Parsed JSON in .tsv format

For the first step, 31,890 research output DOIs and 325 dataset DOIs were collected from the University of Bath Research Portal with the assistance from a UBRDA staff member. Then the Scholexplorer API was queried for links between datasets and these article DOIs, generating 1,501 articles and 269 datasets with at least one dataset-literature link. These DOIs were then used for the second program to parse and collect associated metadata. The same data was collected twice – September 27, 2019, and November 17, 2019 – to validate the consistency of the output from the API.

4.3 Results

The Scholexplorer API was tested for University of Bath research outputs and datasets to investigate the research questions.

4.3.1 Scholexplorer identified previously unknown dataset-article links

As of September 2019, UBRDA recorded 48 University of Bath affiliated datasets that were published in external repositories. These had either been reported by the researchers themselves or had been identified by data librarians managing the UBRDA and manually searching for missing connections. The Scholexplorer API identified 1,501 unique research outputs with at least one University of Bath author linked to at least one dataset, a 31-fold increase from the previously known 48 research outputs with at least one dataset. In total 5,002 datasets were associated with these 1,501 research outputs, where one output could be linked to one or more datasets. Most of the datasets were from the Cambridge Crystallographic Data Centre (82.1%). This is in line with the findings by Robinson-García et al. (2016) that crystallography accumulated more than half of all citations to datasets on DCI. However, it can be difficult to find those links using the single search system on the Chemical Crystallographic Data Centre[31] (CCDC) where the search functionality is limited to identifiers, compound names, DOIs, authors, journals, and publication details (year, volume, page). Advanced searching requires registration and a license, creating a barrier to simplified access. Besides CCDC, Scholexplorer also identified 28 other external repositories hosting datasets associated with the University of Bath researchers (Figure 4.1).

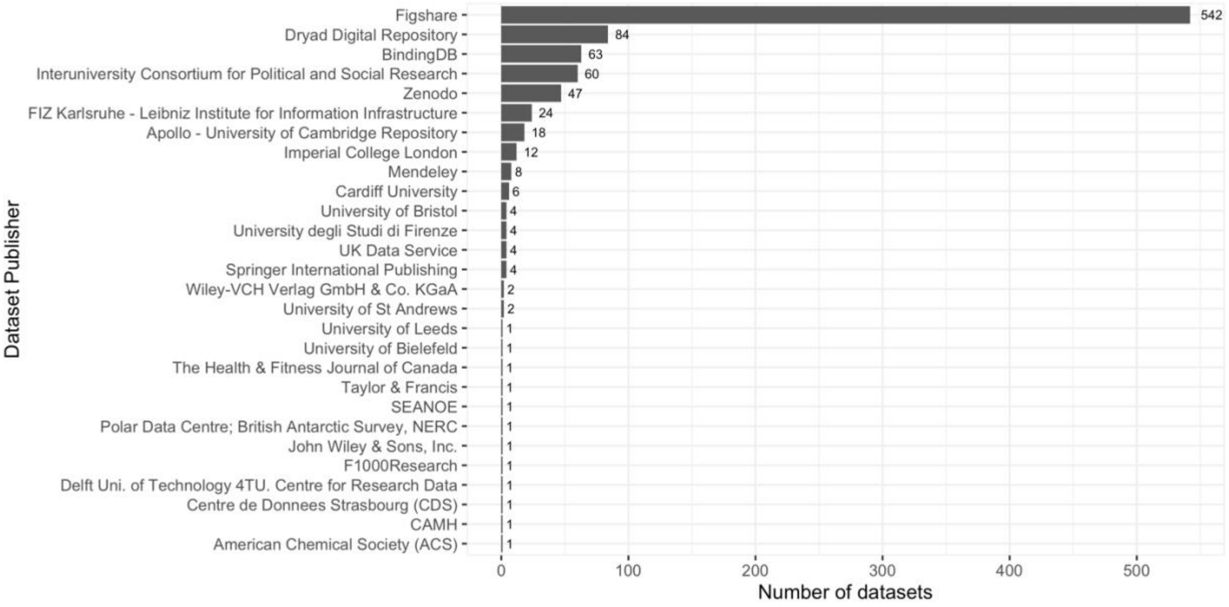

Figure 4.1 Dataset links found by Scholexplorer in external repositories associated with University of Bath research outputs, excluding CCDC.

The Scholexplorer API identified four datasets on UKDS that had linked to two journal articles affiliated with the University of Bath researchers. The API has some gaps in coverage, however. UBRDA had identified 7 dataset records associated with university researchers that were published

---

[31] https://www.ccdc.cam.ac.uk/structures/

on UKDS for Economic and Social Research Council (ESRC)-funded projects, but none were recovered using the Scholexplorer API (Table 4.1). One of the datasets (10.5255/UKDA-SN-852040) had associated journal articles linked on its UKDS record that had not been identified by Scholexplorer. This could be because this dataset was not part of the new beta UKDS website, instead it was on the separate UK Data Service ReShare repository, which is increasingly used by researchers to self-archive datasets from ‘long-tail’ research studies. It is possible that metadata from the ReShare domain may not yet have been consumed by the Scholexplorer API. These datasets in UKDS ReShare have a unique ID ending with UKDA-SN-(6-digit number) for identification.

For the other six missing datasets, UKDS did not link to any related journal articles. However, links to reports, a book chapter, and working papers appear on project websites that are linked to these dataset records (10.5255/UKDA-SN-8303-1, 10.5255/UKDA-SN-8176-1 and 10.5255/UKDA-SN-8397-1). Perhaps because those publications were linked using general URLs instead of DOIs, Scholexplorer had not indexed them. Persistent identifiers for reports and white papers would therefore help with capturing the value of any datasets used, although this is currently not a common practice in many organizations. Similarly, searches of the UK Research and Innovation (UKRI) website for the dataset DOIs 10.5255/UKDA-SN-852527, 10.5255/UKDA-SN-852064, and 10.5255/UKDA-SN-852065 found associated journal articles that were not linked to the datasets on UKDS. As for the dataset record 10.5255/UKDA-SN-852040 mentioned above, these links were not indexed by the Scholexplorer API. Thus, while its coverage is gradually increasing, output from Scolexplorer is not yet comprehensive, and some of the gaps seem to be inevitable.

Table 4.1 Links to UKDS datasets associated with University of Bath researchers

DOIs of UKDS datasets listed on UBRDA	*DOIs of UKDS datasets identified from Scholix*
10.5255/UKDA-SN-8397-1	10.5255/UKDA-SN-5050-16
10.5255/UKDA-SN-8176-1	10.5255/UKDA-SN-7480-1
10.5255/UKDA-SN-8303-1	10.5255/UKDA-SN-7649-1

10.5255/UKDA-SN-852527	10.5255/UKDA-SN-7260-1
10.5255/UKDA-SN-852064, 10.5255/UKDA-SN-852065 (Datasets – part 1 and 2 from the same project)	
10.5255/UKDA-SN-852040	

In the first phase of data collection from the Scholexplorer API in September 2019, only 41 journal articles out of 1,501 (2.7%) included research articles' author names in the API output. These 41 journal articles were further investigated to identify whether the articles are primary research outputs (created by the same authors) or secondary data reuse cases (different authors). For the 121 associated datasets (one publication could be linked to more than one dataset), there were 10 cases of secondary data reuse. Most were from the social sciences and published by ICPSR (*n*=7). The rest were from UKDS (*n*=1), Figshare (*n*=1), and Binding DB (*n*=1). A further analysis of these citing articles was conducted to validate proper data citation practice and is reported below.

4.3.2 Scholexplorer can be used to identify reuse cases of published data

The Scholix schema is based on a simple source and target relationship where the links come from a provider (dataset publisher or journal article publisher). Metadata supported for target and source include identifier, object type, title, creator, publication date, and publisher, where only the identifier and object type are compulsory, and the rest are optional fields (0…N). This limited metadata availability allows simplicity but can leave gaps. For example, author names were only available for 2.7% of the research articles in the September dataset.

As of June 2018, there were more than 870,000 links between Crossref DOIs and DataCite DOIs aggregated by the Scholexplorer API, where most of the links originated from DataCite DOIs and only 22,000 links from Crossref journals. Out of those 22,000 Crossref DOIs, only 16% (3,657) were links between dataset and literature defined by a Crossref type for a scholarly text document and the DataCite metadata resource TypeGeneral of "Dataset". The rest of the links could be literature-literature links (where both source and target objects in the API's JSON output are literature) (Gazra & Fenner, 2018). It is not clear whether the author information was missing mainly from the original publisher information supplied to Crossref. These findings were communicated to the Scholix group members (Adrian Burton, Martin Fenner, Wouter Haak, and Paolo Manghi) by an email on October 07, 2019. A second set of metadata was then collected in November 2019 from the Scholexplorer API, for the same 1,501 research output DOIs extracted in the first phase. In this updated dataset, there was a substantial increase (from 2.7% to 89.2%) in the number of author names available for research articles, perhaps due to updates in the Scholexplorer API software. This suggests that the quality of data is improving, and the Scholix group are responsive to user feedback, which is a positive indicator of the likely continued value of this framework.

Manual comparisons of author and dataset creator names for 319 articles linked to 269 datasets published via UBRDA revealed that all these articles were primary data use cases. The same method was used to compare the author names of articles and datasets of the 41 journal articles and 121 associated datasets from the initial data collection, identifying 10 studies with evidence of data reuse. Citing articles were then examined to understand whether these examples of data reuse were automatically captured due to standard data citation practices. Most of the datasets (7 out of 10) were not included in the reference sections of the articles. Six of these studies reused data from ICPSR and mentioned the datasets in the methods section, and two cited the associated survey websites but not the datasets. Given that Crossref does not provide a text mining feature to analyse full texts of articles for references, the datasets in ICPSR were likely to have been linked manually by a data curator or other staff at ICPSR. This is similar to articles linked to UBRDA datasets since these articles were also manually linked by the staff. While it is useful to learn about such data reuse cases, it does not demonstrate that links are commonly established due to research articles citing datasets directly in the article references.

One article (DOI: 10.1016/j.chembiol.2010.07.018) had links to two BindingDB[32] datasets (DOI: 10.7270/q2jd4v85 and 10.7270/q2n014t9), which were not cited in the article, but the article itself was linked on BindingDB records. By comparing the dataset records and author names and affiliations, the first dataset was identified to be the primary research output and the latter as a case

---

[32] https://www.bindingdb.org/bind/index.jsp

of secondary data reuse. Even though the BindingDB repository links associated articles to the datasets on its platform, dataset creator names were not included in the Scholexplorer records. Both the creator and publisher metadata fields included "BindingDB" only, which may not be useful to automate the identification of data reuse cases. The rest of the articles (*n*=3) reused datasets from Figshare, ICPSR, and UKDS, and had included citation for both primary (*n*=2) and secondary datasets. Among the four UKDS datasets identified by Scholexplorer, three were cited by one journal article (DOI: 10.1186/s12889-017-4665-1) with citations included in the reference section.

Variations in repository names can also cause problems. For example, Dryad and Dryad Digital Repository; FigShare and FigShare Academic Research System; and Zenono and Zenodo Research Shared are all different variations of three repositories that had to be merged when cleaning the dataset, to calculate the total number of links found in each repository for Figure 4.1. Most of the data repositories, including Dryad, Figshare, and Zenodo, are members of DataCite and according to the Scholix website, the method of participation is to feed the data-literature link information to DataCite, which is then aggregated by Scholexplorer. As noted by Aaron Tay in his blog post (Tay, 2018), it is not clear how metadata schemas are implemented in every repository to create dataset-article relationships, how those relationships are mapped by DataCite to feed into Scholix, or whether this affects the API output. Thus, better documentation of metadata mapping and further analysis of the API data would be welcome to develop a benchmark. A controlled vocabulary or use of persistent identifiers could help in the future to aggregate the results efficiently. In case multiple name repository variations need to be captured, either the schema should be extended accordingly, for example, using a metadata field like isSameAs, or integration of persistent identifiers for the host organizations, such as the Research Organization Registry (ROR)[33] or the repository DOI issued by re3data.org would be necessary. Building such PID based relationships is the focus of the EU-funded FREYA project (The FREYA Project: The PID Graph, 2019) and incorporating these developments in the Scholix framework would help with repository name disambiguation.

---

[33] https://ror.org/about/

Another issue was identified in the second iteration of data collection. Fewer article-dataset links were found for the same 1,501 research output DOIs: 5,079 in September 2019 compared to only 5,002 in November 2019. This seems to be a technical issue as no other reason could be established. This should be considered when downloading and using data from the Scholexplorer API for creating metadata records to external datasets in an institutional archive because maintaining consistency is integral to this process.

4.4 Limitations

This case study is limited to research articles and datasets produced by the University of Bath researchers only, since it allowed usage of a benchmark to compare against UBRDA. A comparison of 1. multiple institutions for certain disciplines and 2. different types of repositories can provide further insights into the coverage of Scholexplorer. Furthermore, it uses a small set of article-dataset links to investigate data reuse cases. This is due to unavailability of author names from over 97% article links in the first phase of data collection and analysis. However, increase in author name availability in the second phase means that this could be used for large-scale studies in the future by implementing automated comparison of author names.

4.5 Discussion

This chapter demonstrates that the Scholexplorer API can find links between articles and research datasets published in external archives that data librarians would previously have struggled to find. It also introduced a set of Python code that can be reused by other institutions for this purpose (Khan, 2020). Furthermore, results from this case study show that the information gathered from the Scholexplorer API can be used to provide evidence of compliance to funders’ mandates and validate the impact of research data published by the data producers. In cases where datasets are deposited in a data repository external to the data producer’s host institution, it can help gather information on research collaborators and generate network graphs. For example, several datasets associated with the University of Bath research outputs were deposited in the University of Cambridge, Imperial College London, Cardiff University and University of Bristol’s institutional data repositories (Figure 4.1), demonstrating collaborations between researchers based within these universities.

Besides exposing links to related datasets for articles (e.g., Scopus[34], the bento-box search system from the University of Illinois at Urbana-Champaign (Tay, 2018)), repository managers at higher education institutions can create metadata records using data derived from Scholexplorer to demonstrate the impact of the institutional researchers' datasets (Gibson, 2019; Syrotiuk, 2019). Differentiating between primary research outputs and secondary data reuse cases is therefore important to give an accurate picture of impact, since the latter can indicate the value of data sharing through further reuse. Such connections are currently not straightforward to identify from Scholexplorer because of the unavailability of author affiliations, missing author names in some cases (e.g., records from BindingDB), the lack of a standard naming format (e.g., initials or full first name, order of first and last name), and the lack of author identifiers such as ORCID[35]. Additionally, multiple occurrences of same author names can potentially slow down the processing speed of computer programs designed to compare them. Finally, the API coverage is currently limited to peer-reviewed articles only.

The inclusion of author affiliation information in the Scholexplorer API metadata and the addition of a search function by affiliation would greatly benefit research managers and data librarians who need to find all relevant records from their own institutions. At present (November 2021), it is only possible to search by publisher names on the Scholexplorer API. This will return the datasets published through UBRDA only and does not include any Bath affiliated datasets published in external repositories. Currently there is an identifier field for authors in the Scholix schema that is yet to be implemented in practice. Furthermore, names alone cannot be used accurately to compare and identify a person when different naming formats are used. For example, Fear (2013) took a similar approach to automate data reuse studies by comparing author names of datasets published in ICPSR and the associated articles. However, the author concluded that this may lead to erroneous results due to lack of other contextual information, e.g., author affiliation.

The integration of other persistent identifier (PID) services, such as ORCID would not only help to create more diverse PID relationships, but also encourage researchers to adopt such services. Given that ORCID is not mandatory, this type of added value could be an incentive to promote its

---

[34] http://www.scholix.org/implementors

[35] https://orcid.org/

use. The FREYA project, with similar partners to Scholix (e.g., DataCite and Crossref), planned to integrate PID services to generate meaningful PID graphs (The FREYA Project: The PID Graph, 2019) and the results will hopefully transfer to the Scholix schema as both services mature.

The ideal way of creating article-dataset links would be to ensure the associated dataset is linked to the article when it is published online. However, as found in Chapter 3, data citation is yet to become a standard practice and can vary greatly in different fields. In addition, a substantial portion of article-dataset links that are currently aggregated by Scholexplorer result from the manual linking by data curators. For example, ICPSR had started creating a bibliography of articles citing their datasets and even though a rich source of information, these links are not proof of improving citation practices in journals. Collaborations among journal publishers and repository managers are therefore integral to further improve the data quality of Scholexplorer and ease the process of systematic linking between articles and datasets (Hrynaszkiewicz, 2019). Furthermore, these results show that grey literature, such as reports, book chapters, and working papers are currently not covered by the Scholexplorer API. More studies with larger sample size should be conducted in this area to explore its coverage of grey literature. The research community should identify how the scope of Scholexplorer can be expanded in the future to address this issue since it can be a valuable tool to identify valuable data reuse cases and provide evidence of societal impact.

4.5 Summary

This case study identifies some potentials and shortcomings of the Scholexplorer service. In the absence of a central indexing system for all data repositories that allows searches by author affiliation, Scolexplorer can be a useful tool for academic institutions to automatically identify links between articles published by their researchers and associated datasets. This can be used to demonstrate compliance with funder mandates and to identify areas of improvement. However, at present these article-dataset links captured by Scholexplorer are often manually created by data repository users upon publication of a research article, e.g., repository managers, researchers. As such, this may not be an evidence of standard data citation practice within journal articles and does not allow complete replacement of the manual effort of data repository staff. Better awareness and effort from journal publishers in all disciplines and input from Crossref to automatically create and disseminate these links will be helpful in the future. This can be further enhanced by using a standard vocabulary and persistent identifiers for researchers, institutions, and data repositories, and in turn such features can help systematically identify data reuse cases. Including author affiliations in the Scholix schema will result in more efficient search methods for article-dataset links without having to query the API for a large set of DOIs, which can cause disruptions. A follow-up study in the future will be useful in understanding how current data citation practices and Scholexplorer coverage have changed over time. Finally, the Scholix initiative addresses an important aspect of open science. The best way to promote and support this system is to openly share implementation and integration methods by different institutions for various repository platforms, build community practices, and develop improved guidelines for Scholix and the Scholexplorer API for easy adoption by users.

5. Survey of Repository Managers

The Scholexplorer case study in the previous chapter examined its usefulness to automatically identify article-dataset links and secondary data reuse and demonstrated areas of improvement. At its current stage, data from the Scholexplorer API can supplement the manual linking process by the repository staff rather than replacing it. Chapter 3 introduced GBIF, one of the biggest platforms in biodiversity. Non-standard data citation practices in articles citing GBIF data means that the platform's citation attribution cannot be fully automated yet. Unlike GBIF, repository sizes and their support services can vary depending on discipline and repository type. Therefore, this chapter explores the current landscape of research data repositories, their priorities, and challenges by conducting a survey of data repository managers.

5.1 Landscape and role of data repositories

The development and implementation of repositories has been sporadic and varies between disciplines, with genomics, chemical crystallography, and biodiversity extensively sharing open data more frequently than some other disciplines, such as archeology (Faniel & Yakel, 2017; Robinson-García et al., 2016). Hence the disciplinary data repositories in these areas seem to be relatively mature and robust in terms of technology and policy. For example, GBIF indexes datasets from various biodiversity data repositories and supports both researchers and citizen scientists by making the data easily discoverable and accessible (Chapter 3). On the other hand, there has been an emphasis on developing institutional repositories in higher education institutions in order to ensure compliance with funder mandates and confirm that data published by the institutional researchers meet the necessary standards (Chapter 4). Institutional repositories are also useful for cross-disciplinary data where intellectual property might be complicated due to multiple ownership. This type of data can benefit from planning and negotiation of services for data acquisition and deposition (Cragin et al., 2010). However, differences in types of research data, data repositories and data sharing policies mean that there are no gold standards for different types of data publishing (Assante et al., 2016).

The potential for future reuse by other researchers is a major incentive for openly sharing research data (Wallis et al., 2013). Therefore, being able to track such reuse can be useful to develop an understanding of the value and impact of shared data. Besides, reflecting such impact on repository systems can also act as a reward system for researchers (Costas et al., 2013). While the core responsibility of data repositories has been to publish data, as well as associated metadata, information about how and whether these published data are reused is often not openly accessible from research data repositories. This could be due to a lack of standard and reliable methods or technological barriers to implement data reuse tracking.

Organizations and initiatives, such as the Research Data Alliance[36], European Data Portal[37]E**rror! Hyperlink reference not valid.**, and FORCE11[38] are developing standard practices and technologies for data support services. Examples include the implementation of persistent identifiers, such as DOIs, for research datasets to aid long-term access and standard data citation (Callaghan, 2014). However, the adoption of such services is not consistent across all data repositories. It is also unclear how data services vary across different types of repositories, what are the challenges data repository managers face when offering these services and type of technological solutions they will benefit from in the long-term. This chapter examines the current landscape of data repositories to understand the structure of repository services and types of support offered by them. Furthermore, it investigates the current status of tracking and exposure of data reuse metrics, existing technological barriers and challenges, and type of technologies that could be beneficial in the future.

In order to identify different types of data repositories, this article uses re3data.org, a registry of research data repositories that was established in 2012. It includes a list of data repositories from across the world and publishes information associated with them using the re3data vocabulary. By 2020 the platform listed over 2,500 data repositories, a 6-fold growth in 7 years (Pampel et al., 2013). Availability of these metadata about global data repositories makes it a rich source of information. Using a survey of data repository managers, this chapter explores the following questions:

---

[36] https://www.rd-alliance.org/
[37] https://www.europeandataportal.eu/
[38] https://www.force11.org/

1. How do different types of repositories vary in their adoption of technical frameworks and data support services?
2. What kind of data reuse metrics are currently being tracked and exposed by repositories? Are there any technological barriers to collecting these metrics?
3. How does research data support work for different types of repositories and how do they maintain data quality?
4. What are the current challenges and priorities in supporting research data, and what type of tools would be valuable for the future?

5.2 Methods

A cross-sectional online survey was selected as the instrument to answer the research questions regarding the current landscape of research data repositories and compare with the results of previous studies (Fink, 2003). A questionnaire consisting of 13 questions (Appendix A) was designed where the questions were separated into three main sections: 1) Type of data and technical infrastructure, 2) Research data services and data reuse metrics, and 3) Research data support. These questions were shaped by the existing literature and previous data repository surveys (Kratz & Strasser, 2015; Ivanović et al., 2019; Shearer & Furtado, 2017).

Previous studies have focused on a specific type of data repository (Cox et al., 2017) or reported regional distributions of repositories (Shearer & Furtado, 2017). An overview and comparison of research data services from different types of repositories is therefore needed. In response, this survey was designed to collect information from four main types of repositories based on the type of data they collect: 1. Institutional repositories, 2. Discipline specific repositories, 3. Cross disciplinary repositories, and 4. Repositories supporting specific data formats only. Regional distribution was not taken into account for this survey since discipline specific and cross-disciplinary repositories are often not limited to a specific country or region.

Questions regarding data reuse metrics were adapted from Kratz and Strasser (2015) to understand the current status and interest in tracking different metrics, with an additional question to capture the challenges that repositories are facing. Questions regarding research data services were designed around the size of departments, methods used for data quality checking and engagement with users in order to answer research question 3. A multiple-choice question on current challenges was adapted from the findings of Shearer and Furtado (2017). Finally, two open-ended questions were designed to explore current priorities and future tools to advance functionalities of data repositories. Prior to circulating the survey, a pilot study was conducted with the research data librarians based within the University of Bath to validate the survey questions.

While most of the previous surveys conducted in this area recruited survey participants via professional channels and often focused on specific countries or regions, this study took a different approach. It attempted to reach more varied data repositories by using the Registry of Research Data Repositories, re3data.org. This registry is based in Germany and began in 2012, but it seems to have become a relatively comprehensive source of information about data repositories. It provides a set of relevant metadata about repositories via its API, including contact information. This was selected as the source to collect contact information of repositories where available.

5.2.1 Data collection

a) Metadata and email addresses from the re3data API

Repository metadata was retrieved from re3data.org on February 23, 2019 from its API. In total 2,274 repositories were listed in the registry at the time of data collection. Its metadata fields included a unique identifier for each repository, name, description, contact, type, start date, end date, language, URL, content type, provider type, keywords, subject, entry date, date the record was last updated and remarks.

An initial inspection indicated that some listed repositories had discontinued service and that contact information was not always available since it is not a mandatory metadata field on re3data.org. Furthermore, records had contact information in two different formats: email address or web form to contact. Since web forms are not suitable for the survey platform, all records were manually checked for valid email addresses. Data cleaning based on contact email availability and formatting issues resulted in 1,117 records with an email address. When no email was available, the repository websites were checked instead, finding an additional 70. This resulted in 1,187 curated email addresses for the survey.

b) Survey data

Ethics approval for survey data collection was received from the University of Wolverhampton Life Sciences Ethics Committee (LSEC/201819/MT/85) on March 20, 2019. As recommended and approved by the Ethics Committee, the Jisc Online surveys platform was used to send survey invitations. The survey platform supports two types of invitations: individual email invitations where the link is valid for a specific recipient and an open survey that can be shared via a designated URL. The survey opened on June 27, 2019 and closed on October 4, 2019. 1,187 invitations were sent via email, with 168 responses (response rate 14%). Survey URLs were also sent out to repositories via web forms when no email address could be found, and the open survey URL was forwarded to the previously unknown repositories mentioned by some of the respondents. These methods produced 22 extra responses.

5.2.2 Data analysis

In total 190 responses were received, but 189 were used for analysis after review. One of the responses was excluded from the analysis due to incomplete and premature submission. All responses were anonymized, and any identifiable personal information was removed from the responses.

The survey questions associated with research questions 1, 2 and 3 were either single or multiple-choice questions with an optional ‘Other’ field. These answers were tallied to find the frequency of responses for different groups and content analysis was conducted where open-text answers were included in the ‘Other’ field. To explore research question 4, thematic analyses of answers to open-ended questions were conducted, and two sets of codes were established by the author and another researcher. In total 11 themes were found after reviewing 152 answers for the question regarding current priorities in supporting research data. Similarly, 112 answers for the question regarding future tools and services were reviewed, resulting in 9 themes.

Three coders independently reviewed and coded the responses for each theme: ‘1’ if a response corresponds to a theme or ‘0’ otherwise. This was considered a complicated task due to variations in length and wording of responses. A Fleiss’s kappa test was conducted to calculate interrater reliability (Table 5.2 and 5.3). There were moderate (0.41-0.60) to substantial (0.61-0.80) agreements in most cases (McHugh, 2012). Where kappa values were below 0.4 (Table 5.2), most of the votes fell into one category (code 0 for negative in this case) with a low-level of agreement for the rest. Thus, kappa values were low despite the high level of agreement for a single category. Obtaining high kappa values is difficult for very unbalanced classification tasks, explaining the low agreement rates (Hripcsak & Heitjan, 2002). Disagreements between coders were resolved after a discussion among the coders. These values were reported as the number and percentage of responses in Table 5.2 and 5.3. Lower and upper confidence intervals were calculated using the KappaM (Kappa for m raters) function in DescTools[39] (Tools for Descriptive Statistics) package (version 0.99.44) in R.

5.3 Results

5.3.1 Type of data repositories and technical frameworks

In total 189 responses were received from data repository librarians or data managers, with a majority of responses from institutional (34%) and discipline-specific repositories (47%). A high percentage of repositories (71%) used technical frameworks other than Dspace, Eprint and Fedora (Table 5.1). Other types of frameworks included bespoke systems developed in house, Dataverse, and Figshare for Institutions. Bespoke solutions included custom built systems written with combinations of Comprehensive Knowledge Archive Network (CKAN), Ruby on Rails, Socrata, SQL, Java, XML web application, Solr, Mongo, Dojo, Python, and MySQL. Some repository developers used other software systems, such as Invenio (open-source software to build large-scale information systems - https://inveniosoftware.org/), Islandora (open-source digital asset management system - https://islandora.ca/), Archivematica (open-source digital preservation system - https://www.archivematica.org/en/), LibreCat (institutional repository system – https://github.com/LibreCat/LibreCat), and Adobe Coldfusion (commercial rapid web-application development computing platform).

Institutional repositories were most likely to use DSpace, Fedora, Eprint, Dataverse, and FigShare for Institutions. This is likely because existing repository frameworks are relatively easy to adopt, and this helps academic institutions develop repository services promptly, often without specialist technical support.

---

[39] https://cran.r-project.org/web/packages/DescTools/DescTools.pdf

Table 5.1 Type of data repositories and technical frameworks used

Type of data collected	*Responses*	*Percentage (%)*	*Type of repository frameworks*
Institutional	64	34%	16% Dspace, 12% Eprint, 3% Fedora, 66% other types and 3% did not know.
Discipline-specific	89	47%	8% Dspace, 3% Fedora, 73% other types and 16% did not know.
Any disciplines	22	12%	14% Dspace, 5% Eprint, 77% other types and 5% did not know.
Specific data format	4	2%	100% other types.
Other	10	5%	20% Fedora, 60% other types and 20% did not know.

Most repository services supported the use of PIDs for datasets, with only 8% not supporting any PIDs, and some supporting a combination of PIDs. About three-quarters (76%) of repository services supported DOIs, 22% supported Handle (http://www.handle.net/), and 21% supported other types of identifiers, often specific to a data type or discipline. Handle is more commonly used in repositories for publications and perhaps these 22% of repositories were using the same platform for research data and publications. This was the case for 50% (*n*=20) of the repositories in the study sample of Shearer and Furtado (2017). In addition, just under half (48%) of repositories used DataCite (https://datacite.org/) as a DOI provider, 7% used the Data Citation Index to track data citations and 5% used other data services, such as IRUS-UK (Institutional Repository Usage Statistics UK), Altmetric, and Scholix.

5.3.2 Data reuse metrics

Overall, half of repository managers responded that they currently do not track secondary reuse cases of their published datasets but were interested in doing so. 32% mentioned that they currently track data reuse cases in some format and 19% were not interested in tracking data reuse cases. Among these groups, 25% of institutional, 38% of discipline specific and 18% of cross-disciplinary repository managers currently track data reuse metrics, but a further 64% of institutional, 49% of discipline specific and 41% of cross-disciplinary repository managers were interested in implementing this in the future.

Repository managers were asked which of the following data usage metrics would be helpful, regardless of the repositories' current tracking status of these metrics: citation counts, download counts, landing page views, and links to the data from other websites (e.g., educational use, Wikipedia) (Figure 5.1). Overall, 57% of respondents considered citation counts "extremely useful" and 30% considered them "very useful". Among different types of repositories, 61% of institutional and discipline specific repository managers and 41% of repository managers who collect data from any disciplines considered them "extremely useful". Download counts and links to the data from external websites were considered "very useful" metrics by nearly half: 41% and 44% respectively. Landing page views were less valued, with 28% considering them "very useful" and 29% "moderately useful". This is in line with the findings of Kratz and Strasser (2015) except that there has been a growing interest in links to data from other websites.

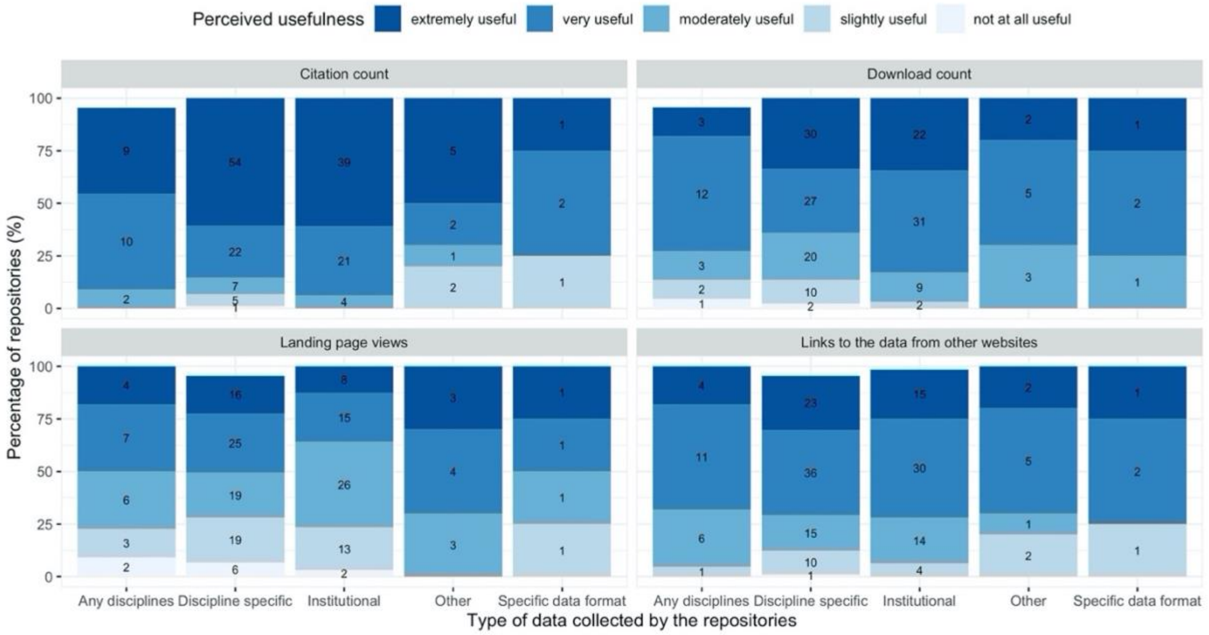

Figure 5.1 Perceived usefulness of different type of metrics (labels on bars represent number of responses)

Concerning secondary use cases of their published research data, 38% of discipline specific, 25% of institutional, 18% of cross-disciplinary, 25% of specific data format supporting repositories and 50% of other repositories tracked this. Figure 5.2 shows the tracking status of four different metrics by different repositories: download counts, citation counts, views, and citations to the repository. Similar to the findings of Kratz and Strasser (2015), downloads and views were more frequently tracked by repository managers, followed by citations to datasets and citations to the repository as a whole. Among those who track these metrics, few were exposing them on their platform (Figure 5.3).

Dataset citations

23% of cross-disciplinary (*n*=5), 46% of discipline specific (*n*=41), 33% of institutional (*n*=21), 25% of specific data format (*n*=1), and 50% of repository managers in other groups (*n*=5) reported that they track citations to datasets (Figure 5.2). Within these groups, all cross-disciplinary repository managers and nearly half of the institutional and other repository managers displayed this metric. The percentage is slightly lower (39%) for discipline specific repositories and the repository manager in the specific data format group did not respond (Figure 5.3).

Repository managers reported the following reasons for not being able to track or expose dataset citations: it is difficult to enforce and track dataset citations because research articles often do not include dataset citations in their main references; DCI does not harvest data related to all repositories; a lack of reliable technologies to automate the process, resulting in users having to manually report any citations. Some repository managers reported using Google Scholar, euroPMC and Dimensions as sources of citation data, but lack of trust and reliability in citation data may have resulted in less repositories exposing the results.

Downloads

Most repository managers (80-100%) in all groups reported tracking downloads (Figure 5.2). However, among these groups, 50% of cross-disciplinary (*n*=9), 73% of discipline specific (*n*=53), 29% of institutional (*n*=15), 50% of specific data format (*n*=2) and 80% of other repository managers (*n*=8) did not expose download counts in their repositories (Figure 5.3). Besides technical difficulties and privacy concerns, lack of interest was mentioned as an important factor in displaying download counts. Some repository managers offer this service only internally. One participant mentioned concerns about the reactions of researchers to these numbers. Another participant raised the technical concern that sections of datasets were often downloaded instead of entire datasets, so a download count for whole datasets was not considered meaningful.

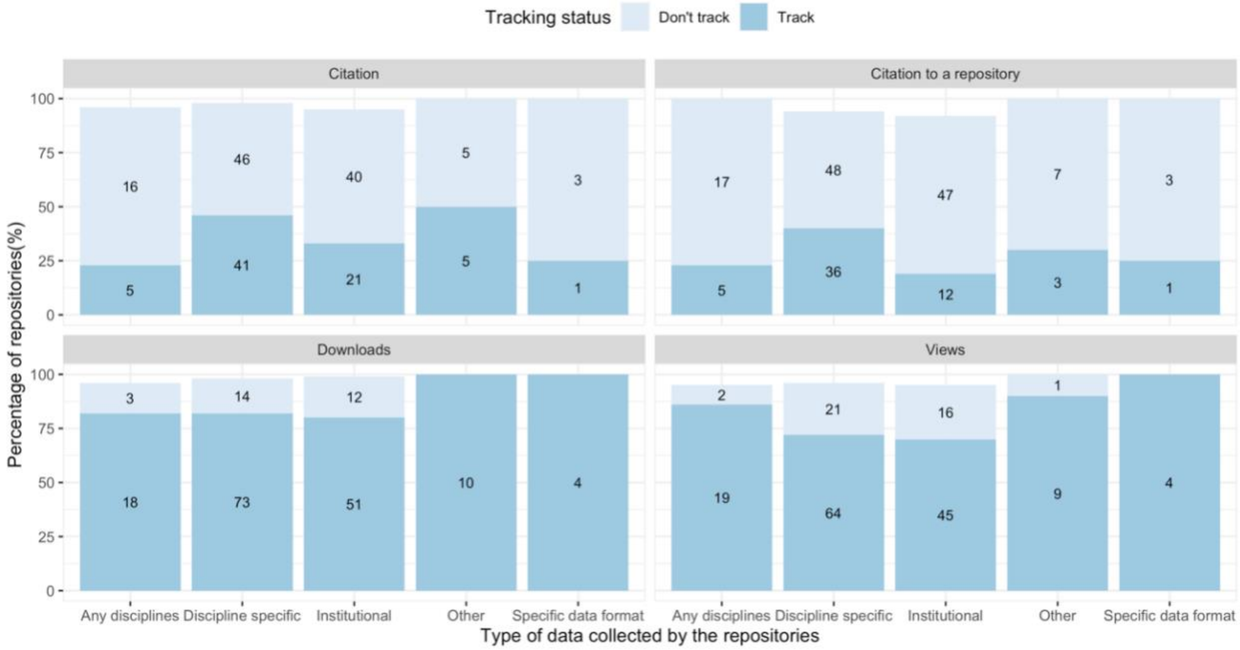

Figure 5.2 Tracking status of different type of metrics (labels on bars represent number of responses)

Views

Similar to download counts, 70-100% of repository managers for different types of repositories mentioned tracking views, even though most of them did not expose these numbers (Figure 5.2 and 5.3). While some repository managers shared view counts internally, many mentioned that page views were of less interest to stakeholders as this metric is not significant and can be manipulated easily. A few participants mentioned privacy issues such as disabling the tracking of metadata discovery and views due to GDPR regulations and distrust in web-trackers.

Citations to the repository

This metric was the least tracked by all types of repository managers. Whilst 40% (*n*=36) of discipline specific repositories mentioned tracking this, 39% (*n*=14) of them exposed this metric. Most of the repository managers did not consider this a valuable metric compared to citations to individual datasets and found it difficult to technologically implement.

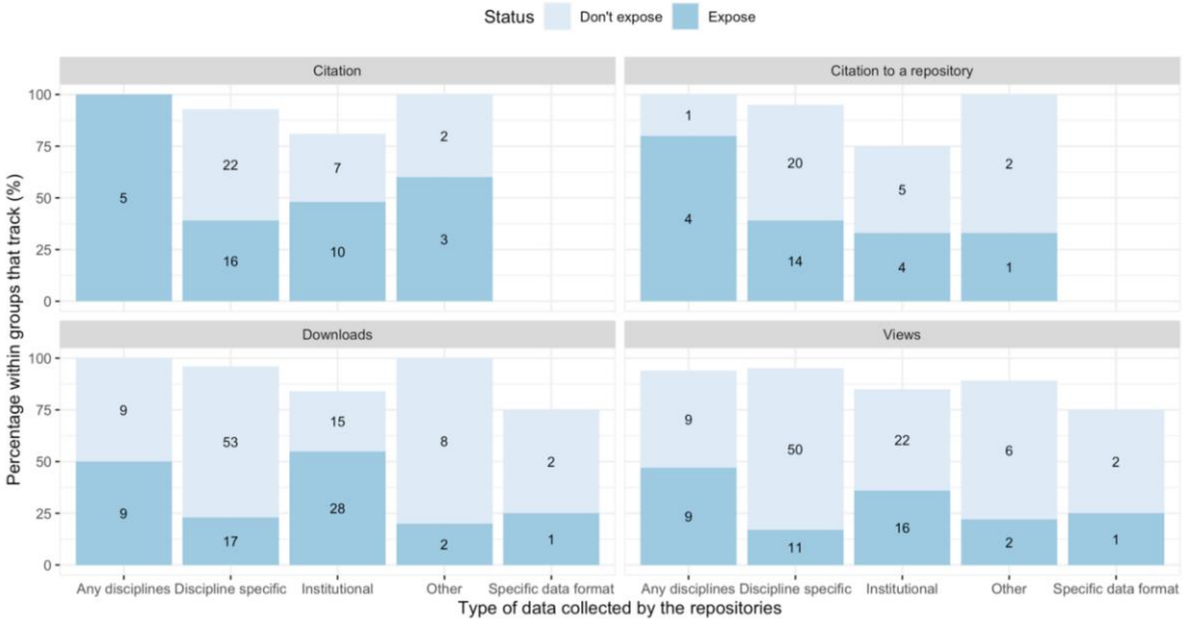

Figure 5.3 Metrics display status of repositories who track these (labels on bars represent number of responses)

5.3.3 Research data services and quality maintenance

Overall, 34% (*n*=64) of research data services operated as small departments of two or three members. Among the rest 66%, 25% (*n*=48) were larger departments with more than three people, 19% (*n*=35) were supported by individual staff, 6% (*n*=12) provided no institutional support, 15% (*n*=28) mentioned other approaches, such as spreading services over multiple departments without a designated research data service department, and 1% (*n*=2) did not respond. Figure 5.4 shows the distribution of types of service for different repositories.

Most repository managers supported dataset and metadata quality checks via librarians, subject specialists, or information professionals (Figure 5.5). This was similar across different types of repositories and research data support services, except where no institutional support was provided

– automated checks were more frequently used in those cases. Where participants mentioned other types of quality maintenance methods, most combined automated checks, (e.g., using scripts to look for errors and duplication) and manual checks by a designated member (e.g., librarian).

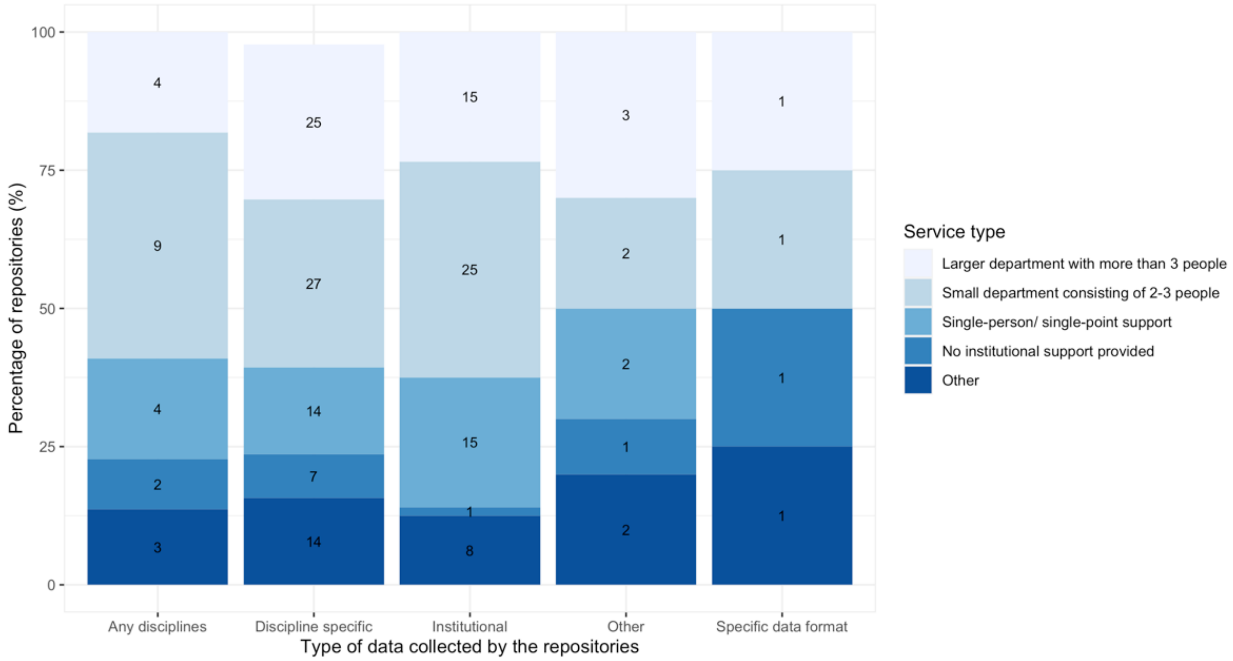

Figure 5.4 Type of research data services provided by the repositories

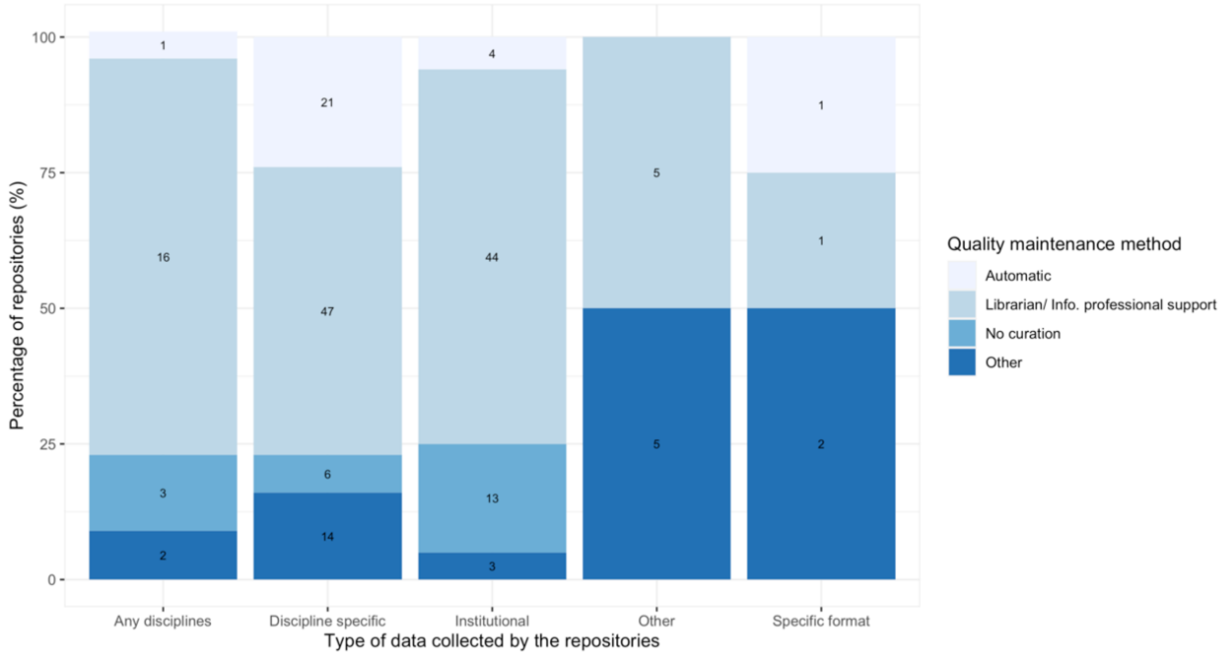

Figure 5.5 Data quality maintenance types by the repositories

5.3.4 Challenges and motivators

Lack of engagement from the users and a shortage of human resources were the top two challenges mentioned by repositories across all groups (Figure 5.6). 73% of institutional, 64% of cross-disciplinary, 50% of specific data format and 40% of discipline specific repository managers mentioned lack of engagement to be a challenge. Long-term maintenance was also a major concern among all repositories, whereas insufficient user need was only mentioned by some. Inadequate funding was seen as a challenge by nearly half of the discipline specific repositories.

Other user challenges include a lack of understanding of standards requirements among researchers, multiple user defined data protocols, trends to put resources in multiple websites, and diverse user needs. Identifying standards, legal or data ownership issues and deciding a long-term solution in an evolving field were some service challenges mentioned by the participants. Adding new functionalities to existing systems, improving metadata quality, and assessing the quality of datasets for publishing without domain expertise are also challenging issues. One participant mentioned demonstrating the value of published data and being able to integrate any technological methods on top of current repository systems:

"*Tracking usage of data to demonstrate value of the repository […]. We have minimal resourcing to better implement solutions that do exist.*"

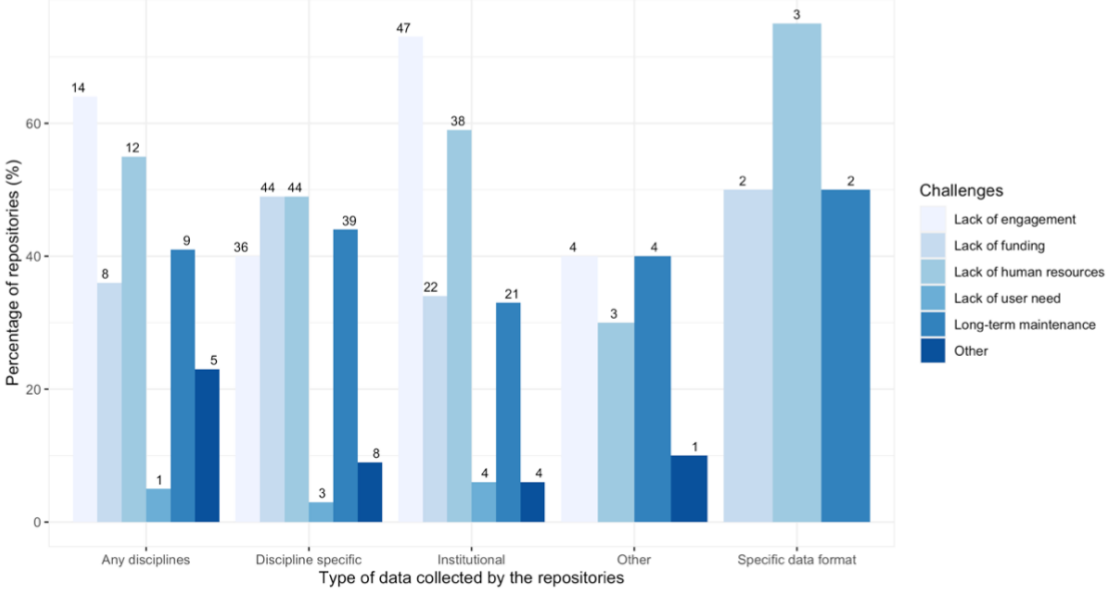

Figure 5.6 Type of challenges faced by the repositories

On average, outreach was mentioned as the most common means to motivate researchers by different repositories (69%). This was followed by funder policies (59%) and training programs (56%). Funder policies was a significant motivating factor for academic researchers who use institutional repositories (77%), compared to 54% discipline specific and 50% cross-disciplinary repositories (Figure 5.7). 23% responses selected other, which included journal mandates, developing innovative programs, such as Research Data Champions, research data management policy, and utilizing different channels of communication. Several respondents in the other group mentioned no active input to motivate researchers as the repositories are well established and used by researchers according to their needs.

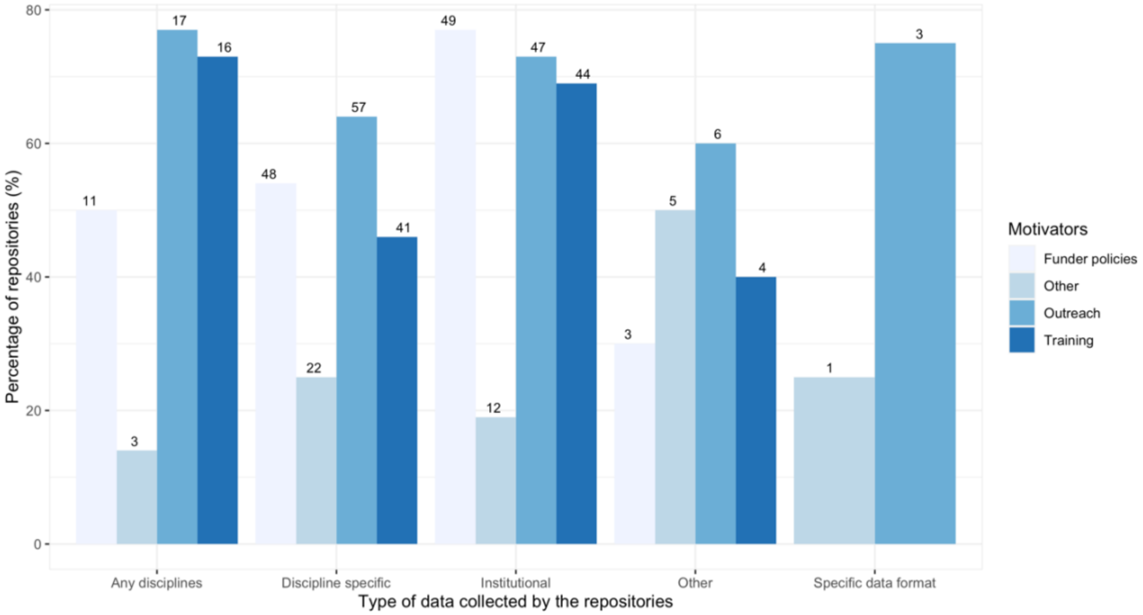

Figure 5.7 Motivators for researchers to deposit data in repositories

5.3.5 Priorities

In total, 152 responses were received for the open-ended question regarding current priorities in research data support. These responses are grouped under 11 categories where each response could be grouped under one or more categories (Table 5.2). 49% of respondents mentioned ensuring data is FAIR (Findable, Accessible, Interoperable and Reusable) as a top priority (Wilkinson et al., 2016). Other high priorities included providing user support for research, e.g., data management plan (DMP) review (36%), as well as building relationships and developing best practices, e.g., provide research data management (RDM) training (29%). As repositories are handling an increased volume of data, ensuring data and metadata quality (15%), simplified data handling (10%), and building robust infrastructures with improved search systems and other data management features (15%) are also considered important. With growing needs from users, some repositories are considering inclusion of other data support services to support data usage and analysis (6%). The following response demonstrates how data repositories are dealing with a multitude of challenges and having to set priorities accordingly:

"*Educating those gathering data about improving data management practices, e.g., FAIR Principles (New Zealand is very behind on this); improving application of data management practices among scientists; providing simple to use DM tools; managing data privacy issues, e.g., for data collected on private land; dealing with the challenges of big data and data science, e.g., data volumes; managing data and metadata where Edge computing and sensors are being used; provenance, repeatability and fine grained metadata*".

Table 5.2 Main priorities for repositories (*n*=152)

Type of priorities	*Number of responses*	*Percentage (%)*	*Agreement rate (%)*	*Kappa value*	*LCL* of kappa (95%)*	*UCL** of kappa (95%)*
FAIR data	75	49%	68.4%	0.58	0.49	0.67
User support (DMP)/ research support)	54	36%	75.7%	0.61	0.52	0.70
Outreach, RDM training, build relationships, and develop best practices	44	29%	91.2%	0.51	0.4	0.58
Data quality check	22	15%	69.7%	0.49	0.52	0.70
Robust infrastructure (improved search	22	15%	84.2%	0.55	0.45	0.64

system, development and inclusion of new data management features)						
Simplified data handling for ease of use	15	10%	82.2%	0.36	0.27	0.45
Need for a systematic approach (norms/ standards/ compliance)	14	9%	83.6%	0.18	0.09	0.28
Support data services (e.g., support data access, use, analysis etc.)	9	6%	90.1%	0.45	0.36	0.54
Secure funding	6	4%	90.1%	0.18	0.08	0.27
Better usage metrics	4	3%	95.4%	0.35	0.26	0.44
Inclusion of data access statement	2	1%	95.4%	0.45	0.35	0.54

* Lower confidence level, ** Upper confidence level

5.3.6 Tools needed in the future

114 participants responded to the open-ended question regarding the type of tools or research data support system they envision for the future. These responses were grouped under nine categories (Table 5.3). Among different recommendations that emerged, integration and interoperability between data and systems was considered important by the most (30%). One participant viewed convergence of data, publications, and research intelligence functions as the ultimate solution to move forward, since current systems are isolated and therefore not interoperable. Other participants mentioned integration of internal institutional systems, such as Current Research Information System (CRIS), DMP tool, repositories to allow reuse of metadata, as well as

integration with specialized services (e.g., visualization, data aggregation) on top of their archived data.

Improved research data management tools, e.g., machine readable DMP, as well as building community of practice across the country and developing and sharing more training material was suggested by 19% of the respondents. For example, a DMP wizard that gives the researchers all features and issues to consider when starting a data production or packaging a project. 16% of participants recommended automated systems for data handling, linkage between publications and datasets, metrics tracking, and tools to support the most frequent types of data analysis and visualization without downloading individual datasets.

In terms of adopting a repository service, a national infrastructure or federated repository was recommended by 15% of respondents, as it would allow a simpler local setup and enable them to shift their emphasis to other new data services and features. Similarly, better data processing, discoverability and storage for repository systems were recommended by some participants, such as tools for long-term preservation of data, streamlined PID based systems, improved search systems, and integration of tools to capture and manage data, e.g., electronic lab notebooks, Open Science Framework (OSF) for research workflow.

Table 5.3 Type of tools and services needed in the future (*n*=114)

Type of tools/ services	*Number of responses*	*Percentage (%)*	*Agreement rate (%)*	*Kappa value*	*LCL* of kappa (95%)*	*UCL** of kappa (95%)*
Integration and interoperability between data and systems (e.g., data exchange between different domains/ journals and repositories	34	30%	87.7%	0.77	0.67	0.88

via national framework, federated systems, ontology tools)						
Better RDM tools (enhanced DMP Wizard, machine readable DMP), promote standardization, develop community practices across countries	22	19%	79.8%	0.49	0.38	0.59
Tools that allow computation (e.g., analysis and visualization of data) without downloading datasets	18	16%	82.5%	0.54	0.43	0.65
Automated systems (data identification, quality check, import/export of data/metadata, linking between publication and data, metrics tracking)	18	16%	87.7%	0.67	0.56	0.77
Repository framework with simpler local setup with emphasis on visualization and analytical service; APIs; new features	17	15%	80.7%	0.44	0.33	0.54

Tools supporting long-term preservation of large volume of data	8	7%	87.7%	0.49	0.38	0.59
Streamlined persistent identifier (PID) based systems with better handling of versions and subsets	6	5%	95.6%	0.69	0.58	0.8
Powerful search engine for better data discoverability	6	5%	91.2%	0.51	0.41	0.62
Integration of tools to capture and manage data (e.g., lab instruments, OSF, Electronic lab notebooks, Sandbox)	5	4%	93.9%	0.67	0.57	0.78

* Lower confidence level, ** Upper confidence level

5.4 Limitations

The sample size analysed here is small given over 2000 research data repositories are indexed in re3data.org. This shortfall is partly due to the limited availability of email addresses provided for repositories in the metadata. Where web forms are used as a contact method, automated distribution of the survey via the platform is impossible. In addition, email addresses provided by repositories tended to be generic email addresses. Therefore, in many cases the email invitation to the survey had to be forwarded to the repository manager to answer the questions. This indirect route slowed down the process and reduced the number of responses. This may also bias the sample against non-English repositories (less likely to forward an email in English), and small repositories (without a dedicated contact email). The sample sizes for certain repository types, such as repositories supporting specific data formats, is relatively small and the results may not accurately represent this group.

5.5 Recommendations for registries of data repositories

This study used openly available metadata from re3data.org as a source of contact information to recruit participants. Registries of data repositories, such as re3data.org, and general repositories, such as OpenDOAR (https://v2.sherpa.ac.uk/opendoar/), will be valuable for similar studies in the future but would benefit from more granular and structured metadata in some data fields. It is recommended that the following changes are undertaken to help with this issue. It should be noted that the following suggestions are secondary to the primary purpose of registries, which is to collect and index the world's data repositories.

5.5.1 Contact types

OpenDOAR and re3data currently record any available contact information in an optional Contact field. Based on the data collection process of this study, it is recommended that re3data defines contact types in this field into two main groups: 1. Email address, and 2. Web address/contact form. The email address field should accept valid email address only, in order to avoid any data loss when these metadata are processed. Additionally, most repository services provide a general email address instead of a specific individual's email, which can lengthen the response time and reduce the response rate. Therefore, where possible, it would be beneficial to include specific metadata fields to accommodate different roles, such as general email address, email address of a repository or data manager, technical services manager, and data librarian. Repositories should also receive a notification to review and update this information on an annual basis.

5.5.2 Repository types

Currently there are only three options on re3data.org to select repository types: disciplinary, institutional, and other. While conducting this survey, it was found that federated infrastructure data portals are also listed as research data repositories, such as GBIF (DOI:10.17616/R3J014). While these systems provide access to data hosted in multiple repositories, they do not offer standard repository services, such as data deposit, maintenance, and curation support. A more granular categorization to accommodate these differences between repository types would be useful. Differentiating between other repository types, such as cross-disciplinary, government data repository, project based, would also help future analyses.

5.5.3 Software

re3data has an optional Software field to record the repository framework used. Currently only 12% of records include this information and 17% declare that it is unknown. Mandating this field when first registering a repository would allow future studies to analyse software-based differences, which may be important to the data repository services.

5.5.4 Quality management

Although re3data supports a metadata field to record whether data repositories manage data and metadata quality, currently only 0.6% complete this field. Furthermore, this field accepts three answers: yes, no, and unknown. Given data quality is one of the main factors that drives future data reuse, it would be beneficial to further extend this field based on existing research in order to provide more information on the type of quality management implemented. This would benefit researchers searching for data repositories to deposit data or to find existing datasets for reuse purposes.

5.6 Discussion

Data repositories are evolving rapidly to accommodate the needs of funding bodies and researchers, as well as to support different types of data. Due to differences in the nature of disciplinary data, most discipline-specific repositories develop bespoke technical frameworks. In contrast, a major incentive for institutional repositories is to support academic researchers following funders' policies, but they often lack the technical expertise available to large-scale discipline specific repositories (Cox et al. 2017) and rely on existing frameworks, such as DSpace, Eprint, and Dataverse. Shearer and Furtado (2017) and Ivanović et al. (2019) found similar results for technical framework adoption by institutional repositories. While differences in institutions and data types mean that a single repository framework may not be a good fit for all purposes, there are opportunities to use community driven approaches for developing research data management policies, training materials and best practices.

As indicated by this study, a lack of user engagement and shortage of human resources are the major challenges faced by all repositories, but these issues were more prominent among institutional repositories than for discipline-specific repositories within the study sample. Shearer and Furtado (2017) also reported an absence of user engagement as the top challenge among their respondents. Comparatively higher percentages of institutional repositories therefore heavily rely on outreach and training to motivate and engage researchers, even though funder mandates were the main motivators for academic researchers who use these repositories. Outreach, RDM training, building relationships, and developing best practices was mentioned among the top three priorities in this study, and institutional repositories can benefit from working together to develop training materials and policies where a lack of human resources is an operational issue.

Most research data repositories started as siloed services to provide storage and access to data. As these services mature, better data discoverability and interoperability are becoming increasingly important to promote data reuse. This is reflected in the findings of this survey, given integration and interoperability between data and systems was the most desired future service. There are two different routes to achieve this: 1. Interoperability between data repositories and 2. Interoperability between journal systems and data repositories. The first route requires a data portal or a global discovery service that would break the silo of individual repositories and allow users to search for data across multiple repositories. Federated infrastructures do this by connecting multiple data repositories and act as an access point for data across those repositories. Some disciplines, such as biodiversity (GBIF.org), and national initiatives, such as the National Research Data Infrastructure (NFDI) for the European Open Science, show that this can be possible (Chamanara et al., 2019; Goldstein, 2017). Another relatively new data discovery service is Google Dataset Search, which allows dataset keyword searches across all supported repositories. This system is based on a linked data model and relies on repository services adopting the Schema.org metadata standard so that dataset metadata from these repositories can be indexed by Google and added to the search system (Patel, 2019).

Interoperability between journal systems and data repositories is necessary to automate tracking of data citations to understand how a published dataset has been reused. Repository managers considered dataset citations to be the standard metric to estimate the scholarly value of data, as well as valued evidence of educational use. Download counts and views are considered less valuable since they do not demonstrate evidence of secondary use and can be easily manipulated. Despite their clear importance, there is no standard method to count dataset citations which can be implemented across different repository systems. Initiatives and technical frameworks, such as Scholix are currently in progress (Chapter 4). The previous chapter suggests further enhancements of the Scholix schema and enrichment of Scholexplorer metadata using controlled vocabularies, as well as the adoption of standardized data citations by journals to establish links between datasets and literature. Google Dataset Search also displays citation counts for datasets, but Chapter 3 found discrepancies between the citation counts displayed by Google and GBIF for biodiversity datasets. These services can be potentially utilized to identify data reuse cases when they mature in the future. In the meantime, repositories should follow and implement the data citation roadmap (Fenner et al., 2019), and carefully consider the guidelines for using indicators to evaluate data outlined by Konkiel (2020).

5.7 Summary

This study identified the key current practices of data repositories and the types of challenges data repository services face. The results show that the sporadic development of different types of data repositories has resulted in the adoption of bespoke technical frameworks by most repositories, especially the ones that are discipline specific. However, developing and implementing new technological solutions for different platforms can be challenging for institutional repository services as it was found that they often had small teams. Whilst it seems logical that disciplinary repositories would often need bespoke services, this makes full interoperability between services difficult to achieve. Nevertheless, integration and interoperability between data and systems was considered important by the respondents. A common language that can be used by all repository systems can help break this silo, such as the Schema.org metadata standard for datasets to be indexed and discovered by the Google Dataset Search, and adherence to standard data citation practices by both researchers and journals. Additionally, this will help repository services track and expose data reuse metrics, such as citation counts for datasets, as suggested by the survey results.

In the long-term, the use of federated systems and simpler local set-ups will allow repository services to shift their focus to build new features, such as the integration of tools to capture and manage data (e.g., lab instruments, Open Science Framework (OSF), Electronic lab notebooks, Sandbox), and the development of new visualization and analytical tools. Whilst most repository services are currently struggling with a lack of user engagement, these new improvements will help demonstrate the value of research data and attract more users.

Given the apparent mismatch between the features desired by repositories and the availability of large enough teams to implement them, current collaborative initiatives appear promising to help develop shared community practices and reduce the burden on individual institutions. Global initiatives, such as the implementation of FAIR data principles, Scholix, and Google Dataset Search will benefit all repository types by promoting standardisation, improving data discoverability, and automating secondary data reuse tracking. While institutional policies and types of outreach activities to engage researchers can differ between academic institutions, shared resources to implement technological solutions (e.g., how to implement the Schema.org metadata standard for a specific repository framework), guidelines and training materials for research data management will be helpful, especially for smaller scale academic institutions. This will also ensure that different data repositories are not developing siloed services but have a common interoperable system in place. Data sharing and data protection rules can vary across different countries and regions. For example, Europe has the General Data Protection Regulation (GDPR) in place and some survey participants mentioned this as a barrier to exposing certain data metrics. Regional collaboration will be valuable in these cases to tackle these issues in a systematic manner.

6. Survey of Researchers

The previous chapter explored how different types of repositories vary in the adoption of technical frameworks, perceived usefulness of data reuse metrics, data support services, as well as in their current priorities and challenges faced when providing these services. In order to ensure that adequate funding and policies are in place to support data repositories of different scales, it is important to examine whether and how researchers in different disciplines use these resources for data sharing and reuse purposes. Therefore, this final empirical chapter focuses on researchers' data sharing and reuse behaviour across disciplines.

6.1 Disciplinary differences in data sharing and reuse

Collecting and producing new data is an integral part of research in many disciplines. Over the past decade there has been a growing interest within the scientific community to share research data in a findable, accessible, and interoperable format that allows the reuse of data by others (Wilkinson et al., 2016). Open research data is a gateway to reproducible science with increased opportunity for collaboration and interdisciplinary research (Borgman et al., 2019). In addition, studies that share research data have a citation advantage (Piwowar et al., 2007; Henneken & Accomazzi, 2011; Colavizza et al., 2020). The growing importance of research data is acknowledged by the Research Excellence Framework (REF) in the United Kingdom, since research datasets have been included as a standard research output in their most recent guideline (REF, 2019).

Despite funding bodies increasingly requiring that all data generated as a part of research to be made openly available (Kiley et al., 2017), these mandates are often not strongly imposed by journals across disciplines. In absence of stringent journal policy or an incentive to share data, standard data sharing practices on the web can vary by discipline and research experience. For example, even though the availability of data access statements increased in PLOS ONE from 2014 to 2016, only 20% of those statements included links to data shared in a repository that is accessible in a meaningful manner (Federer et al., 2018). Similar results were derived from the Engineering field, where 76% of the journals indicated research data sharing to be optional (Wiley, 2018). In cases where researchers use a data repository, it is also important to explore the types of repositories most frequently used by researchers in different disciplines. Tenopir et al. (2015) found that the use of disciplinary repositories was significantly different between disciplines. The sporadic growth of repositories and a lack of a central data discovery system means that more needs to be known about how scientists find data repositories for sharing data, as well as datasets to reuse. This information would help stakeholders and policy makers determine areas which need new interventions to increase data sharing in certain disciplines.

Data sharing and reuse practices are now partially understood (as reviewed in Chapter 2). Openly shared research data helps science progress by answering new research questions from the same data, by combining multiple datasets or through secondary analysis of existing data (Bishop & Kuula-Luumi, 2017; Coady et al., 2017). Multiple studies have explored researchers' data reuse behaviour and the factors that influence the reuse of existing data (Faniel & Yakel, 2017; Kim & Yoon, 2017). However, these studies are difficult to generalise due to their focus on data sharing and reuse in the context of select disciplines (Piwowar, 2011; Wallis et al., 2013; Federer et al., 2015; Faniel & Yakel, 2017; Zenk-Möltgen et al., 2018; Sardanelli et al., 2018) or data repositories (Bishop & Kuula-Luumi, 2017; Coady et al. 2017; Borgman et al., 2019). In cases where samples were collected across multiple disciplines, ad-hoc participant recruitment via email and social media led to fewer responses from disciplines where data sharing is less common, such as Business and Economics, Arts and Humanities (Tenopir et al., 2015), obscuring the general picture. Kim and Stanton (2016) conducted a large-scale survey, but it was limited to STEM disciplines. Therefore, there is limited evidence that types of research data produced, and the way research data is shared and reused vary between disciplines, especially for disciplines with a less established culture of data sharing than others. For example, lack of standards and absence of subject-specific data repository infrastructures lead to difficulty in interoperability in archaeology, whereas biodiversity has a longstanding culture of sharing research data with established standards, which is supported by advanced infrastructures, such as GBIF (Faniel & Yakel, 2017). Understanding these disciplinary differences is essential in developing new national, international, and disciplinary research policies.

Additionally, even though sharing and reuse of data are considered important by different stakeholders, there is a lack of studies that specifically investigate incentives to promote and encourage these practices. Open Science Framework implemented badges to acknowledge open practices and several journals participated in this pilot, which led to substantial increase in data sharing (Kidwell et al., 2016). Although several studies explored what factors affect scientists' data reuse behavior (Yoon, 2016; Faniel et al., 2016; Kim & Yoon, 2017; Curty et al., 2017), it is important to identify a set of factors that the research community agrees upon, which can be used to develop incentives to encourage data reuse.

To address the above, this chapter explores similarities and contrasts in data production, sharing and reuse practices across 20 different disciplines under nine subject categories. It seeks to understand trends in data sharing and reuse by comparing research areas represented in previous studies, as well as explore understudied research areas in the Arts and Humanities, Business and Economics, and Engineering. Furthermore, it aims to explore how researchers share data on the web and how they find datasets to reuse, in order to understand their interaction with the tools that are currently available.

Current practices in data production, sharing and reuse are investigated by addressing the following research questions.

1. How do types and formats of data produced by researchers vary across disciplines?
2. How do researchers share data on the web? Does data sharing vary across disciplines and research experience?
3. How do researchers find repositories to share data and what factors influence their choice of repositories?
4. How frequently do researchers reuse existing data in different disciplines and for which purposes? How does it compare to data sharing in those disciplines?
5. How do researchers find datasets to reuse? Which factors are considered important when searching for existing datasets? How easy is it to find relevant datasets for reuse?
6. How often do researchers promote datasets? What can be improved in current systems to encourage and promote data reuse?

6.2 Methods

Consulting active researchers is the most direct method to gain insights into how research data is currently being shared and reused. A cross-sectional online survey was therefore selected as the instrument to answer the research questions regarding the current landscape of research data sharing and reuse in different disciplines, as well as to compare with the results of previous studies (Fink, 2003). Geographic location was not considered for this survey since discipline-specific and cross-disciplinary repositories are often not limited to a specific country or region, therefore not limiting use of these resources for data sharing and reuse purposes. Furthermore, differences in data sharing and reuse practices were explored across different stages of research experience.

6.2.1 Questionnaire design

A questionnaire consisting of 15 questions (Appendix B) was designed with the questions separated into six main sections: 1) Research area and experience, 2) Data production, 3) Data sharing, 4) Data reuse, 5) Measuring data reuse, and 6) Incentives for data reuse. These questions were shaped by the existing literature and previous surveys on data sharing and reuse (Kratz & Strasser, 2015; Tenopir et al., 2015). When a researcher indicated that they had prior experience of reusing existing data, they were forwarded to the data reuse section - this was designed to understand how researchers find datasets to reuse, for which purposes existing data were reused, and whether those who reuse existing datasets promote their own data as well, to draw the attention of other users. Questions on measuring data reuse and understanding challenges and motivations in finding datasets to reuse were included to develop incentives in the future.

To answer the first research question, an open-ended question was included on the type of data produced by researchers, since specific subject knowledge would be required to design a comprehensive list of options. A multiple-choice question was designed for data formats since these are better known. Questions regarding data sharing methods were adapted from Kratz and Strasser (2015) and Tenopir et al. (2015) with additional questions on how researchers find repositories to share data and the factors that influence their choice of repositories to answer both research questions 2 and 3. Questions about data reuse purposes were designed using the data reuse typology suggested by Pasquetto et al. (2019). A multiple-choice question on how researchers find datasets to reuse was adapted from Kratz & Strasser (2015) to answer research question 5, with additional questions on important factors when searching for existing datasets and ease of finding datasets to reuse. Finally, an open-ended question was designed to explore what can be improved in current systems to encourage and promote data reuse. Prior to circulating the survey, a pilot study was conducted with the researchers at the University of Wolverhampton to validate these questions and identify necessary adjustments. For example, adjustments were made in the range of years to indicate research experience, the inclusion of an ‘Other’ category under each broad subject category to allow researchers indicate their specific field of research if not listed, and the inclusion of an optional question to record respondents interested in being informed about the findings of this survey.

6.2.2 Selection of subject categories

Two sources were selected to compare subject categories: 1. Web of Science (WoS) and 2. Scopus. At first 100 subject categories from WoS, 27 main Scopus categories and 334 subfields in Scopus were reviewed. Both sources were compared for availability of similar subject categories. It was evident that some categories in WoS are too broad, such as Engineering, Chemistry. Additionally, the Social Sciences appear as a single category, but Education appears in a separate category, which may have overlapping results. 27 of the main Scopus categories were also considered to be too broad to understand differences in different subject areas, e.g., Aerospace engineering and Biomedical engineering both fall under the broader category of Engineering but are expected to have differences. Therefore, a subset was selected from the 334 Scopus subfields to identify specific disciplinary differences.

Since previous studies focused on data intensive STEM disciplines (Kim and Stanton, 2016), specific repositories (Pasquetto et al., 2019) or had relatively small samples of less represented disciplines (Tenopir et al., 2015), an overview and comparison of different qualitative and quantitative disciplines was missing. To fill this gap, the survey collected information from 20 disciplines under nine subject categories (Table 6.1). Publications in Scopus from 2018 and 2019 were selected (in 2020) to understand recent trends in data sharing and reuse, where a minimum of 15,000 results was considered as the threshold for each subject area. The selection of disciplines was based on subject classifications in Scopus to control for disciplines addressed in previous studies and to include new disciplines which had not been previously reported, such as Visual and Performing Arts.

6.2.3 Data collection

a) Population and sampling

While most previous surveys often recruited survey participants via professional channels and social media (Kratz & Strasser, 2015; Tenopir et al., 2015), this study took a similar approach to the STEM survey of Kim and Stanton (2016). It selected random samples of published articles in each of the 20 selected disciplines from Scopus (https://www.scopus.com/). To focus on currently active researchers, metadata from 8,000 randomly selected studies were collected for the year 2018 and 2019: 4,000 studies from each year and for each discipline. Email addresses of the first authors were then extracted from Scopus, where available, resulting in 3,500 emails on average per discipline. After de-duplication, a total of 70,060 researchers were identified for the study. However, due to the interdisciplinary nature of some disciplines and articles in Scopus, researchers were allowed to identify their discipline differently than the suggested Scopus subject category and select ‘Other’ or only select a broader subject category (Table 6.1).

b) Survey data

Ethics approval for survey data collection was received from the University of Wolverhampton Life Sciences Ethics Committee (LSEC/201920/MT/125) on June 12, 2020. The Jisc Online surveys platform was used to send individual survey invitations, complying with UK data protection requirements, and as approved by the university ethics committee. The survey opened on July 14, 2020 and closed on August 17, 2020. In total, 70,060 invitations were emailed, and 3,257 responses were received in the nine subject categories (response rate 4.65%). 214 participants only selected a broader subject category and did not report their specific disciplines. The survey platform does not record whether emails have been returned, in case a researcher is no longer affiliated with an organization. Therefore, the underlying response rate may have been slightly higher.

6.2.4 Data analysis

All 3,257 responses were anonymized, and any identifiable personal information was removed from the responses. Originally 402 responses were reported under ‘Other’, outside of the nine categories defined. However, after reviewing the responses, 149 were found to be different variations of the disciplines listed in the survey, e.g., Information Sciences, librarianship (interdisciplinary), besides others. Therefore, these were merged with the main categories, leaving 253 responses in the ‘Other’ category.

Four out of the 20 disciplines received fewer than 30 responses: Organic Chemistry; Radiology, Nuclear Medicine and Imaging; Aerospace Engineering; and Biomedical Engineering (Table 6.1). These disciplines were not included when analysing discipline specific differences due to their small sample sizes. Therefore, the study primarily analyses and reports disciplinary differences using broader subject categories and explores further discipline-specific differences in Social Sciences, Arts and Humanities, Business and Economics, Biomedical Sciences, Environmental Sciences, and Earth and Planetary Sciences, where the sample size was greater than 30 for each discipline. The cut-off 30 was chosen as a common statistical sample size threshold, in the absence of a theoretical reason to pick a given number.

The survey questions were either single or multiple-choice questions with an optional ‘Other’ field and an open-ended question. These answers were analysed to find the frequency of responses for different groups and content analysis was conducted where open-text answers were included in the ‘Other’ field. Free texts from the open-ended question on data types were analysed to find term frequencies in broader subject categories. A manifest content analysis with an inductive approach was used to analyse the final open-ended question (Bengtsson, 2016).

Research experience in years and disciplines were used as independent variables in the analyses to understand their relationship with other outcome variables, such as prior data sharing and reuse experience. Chi-square tests were used to examine the independence between these categorical variables and binomial multiple logistic regression method was used to explore the effect of research experience and disciplinary differences on data sharing and reuse experiences. The assumptions for binary logistic regression were met by the following: 1. The dependent variable is binary, 2. Each observation is independent of each other, i.e., there are no repeated measures, 3. There is no multicollinearity among the independent variables, and 4. Sample size is adequate – minimum of 10 cases for each independent variable. The glm function[40] in stats package (version 3.6.2) in R was used to perform binomial logistic regression.

6.3 Results

Among the 3,257 responses received, most were from researchers with over 10 years of research experience (64.4%), followed by 6-9 years (15.5%), 3-6 years (13.8%) and 0-3 years (6.2%). These proportions were similar for responses within all subject areas (Figure 6.1). Perhaps those with more research experience are more familiar with the concepts of data sharing and data reuse, and therefore were more inclined to respond to this survey. Social Sciences had the highest number of responses (22.5%) within the broader subject categories, and Medicine had the lowest (5.2%). The percentage of responses in specific disciplines ranged from 5% (Organic Chemistry) to 60% (Astronomy and Astrophysics). Many selected 'Other' disciplines under a broader subject category (on average 39%), with the highest in Engineering (63%) and lowest in Environmental Sciences (17%) (Table 6.1). The number of responses in previously underreported disciplines was significantly higher than those reported in previous studies, such as education, linguistics, visual and performing arts, literature, business, and economics. This larger sample size allows the results to be more generalizable.

Table 6.1. Selected disciplines for the survey and numbers of responses

Subject category	*Discipline*	*Scopus subject code*	*Number of responses in each discipline**	*Percentage of responses within broader subject category (%)*
Social Sciences (*n*=733, 22.51%)	Linguistics and Language	3310	72	10%
	Education	3304	114	16%
	Library and Information Sciences	3309	252	34%
	'Other' in Social Sciences		211	29%

---

[40] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

Arts and Humanities (*n*=334, 10.25%)	Visual and Performing Arts	1213	64	19%
	Literature and Literary Theory	1208	103	31%
	‘Other’ in Arts and Humanities		139	42%
Business and Economics (*n*=592, 18.18%)	Business and International Management	1403	193	33%
	Economics and Econometrics	2002	251	42%
	‘Other’ in Business and Economics		123	21%
Physical Sciences (*n*=220, 6.75%)	Astronomy and Astrophysics	3103	133	60%
	Organic Chemistry	1602	11	5%
	‘Other’ in Physical Sciences		63	29%
Biomedical Sciences (*n*=264, 8.11%)	Neurology	2808	39	15%
	Pharmacology	3004	46	17%
	‘Other’ in Biomedical Sciences		145	55%
Medicine (*n*=170, 5.22%)	Radiology, Nuclear Medicine and Imaging	2741	27	16%
	Infectious Diseases	2725	38	22%
	‘Other’ in Medicine		97	57%
Environmental Sciences (*n*=324, 9.95%)	Ecology	2303	147	45%
	Pollution	2310	108	33%
	‘Other’ in Environmental Sciences		56	17%
Earth and Planetary Sciences (*n*=188, 5.77%)	Geology	1907	56	30%
	Oceanography	1910	53	28%
	‘Other’ in Earth and Planetary Sciences		73	39%
Engineering (*n*=179, 5.5%)	Aerospace Engineering	2202	14	8%
	Biomedical Engineering	2204	19	11%
	Environmental Engineering	2305	30	17%
	‘Other’ in Engineering		113	63%
Other			253	8%

* Sample sizes under 30 are highlighted in pink.

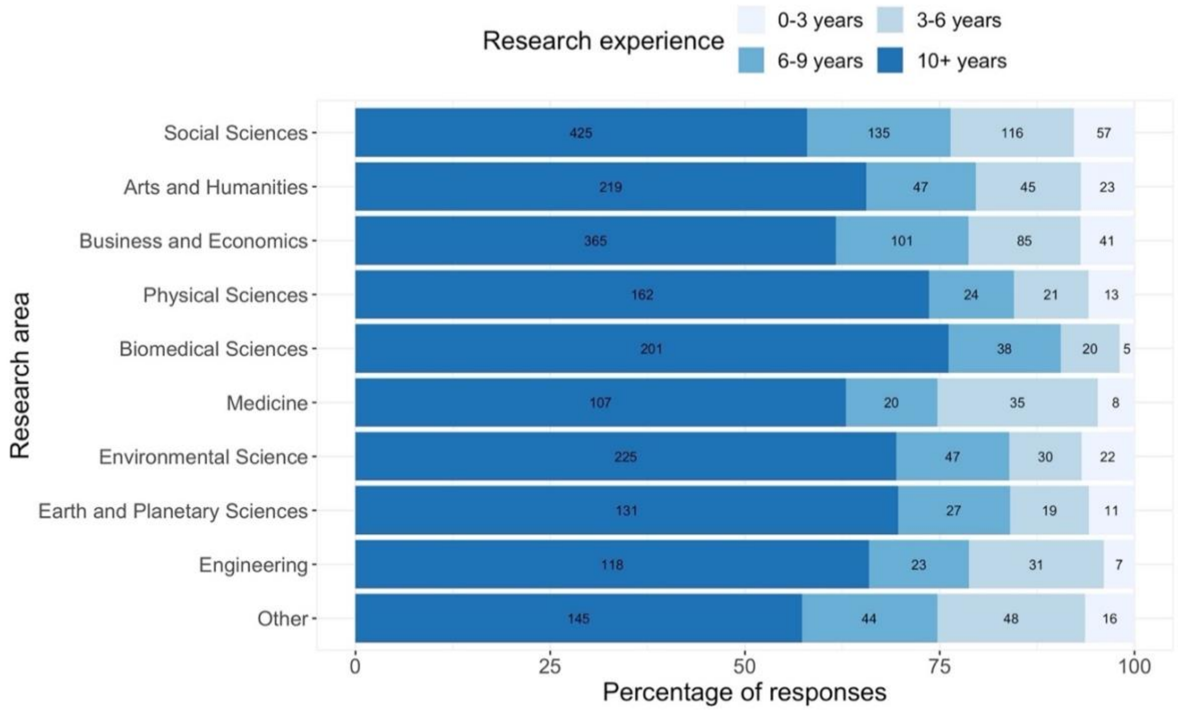

Figure 6.1 Percentage of responses by research experience and discipline (labels on bars represent number of responses)

6.3.1 Type and format of data produced in different disciplines

Researchers were asked to provide examples of the type of data their research generates (open-ended). Surveys and observations were the most common types of data produced across all disciplines (Table 6.2). Qualitative data, audio, and video were common in Social Sciences and Arts and Humanities. In comparison, samples, measurements, simulations, and images were common in Science and Engineering.

Table 6.2 Top 10 types of data produced by discipline with the term frequencies

Subject categories	*Data types*
Social Sciences	Survey (481), interviews (234), observations (113), qualitative (98), transcripts (52), quantitative (44), audio (38), video (36), recordings (34), experimental (34)

Arts and Humanities	Survey (60), observations (43), texts (37), images (33), interviews (33), video (28), audio (26), qualitative (24), literary (23), historical (22)
Business and Economics	Survey (345), secondary (83), interviews (67), observations (31), qualitative (26), experimental (18), financial (16), quantitative (16), economic (15), time series (14)
Environmental Sciences	Survey (105), samples (54), observations (48), water (35), field data (30), experimental (29), measurements (23), images (22), soil (22), species (19)
Earth & Planetary Sciences	Models (40), Observations (32), Samples (25), Survey (25), Measurements (24), Water (17), Field data (16), Numerical (16), Chemical (15), Temperature (12)
Biomedical Sciences	Survey (45), images (32), behavioral (30), samples (26), experimental (21), imaging (19), recordings (13), clinical (12), EEG (12), brain (11)
Medicine	Survey (60), Clinical (35), Observations (26), Images (21), Imaging (11), Qualitative (11), Medical (10), Samples (10), Measures (9), Trials (8)
Physical Sciences	Images (50), observations (48), simulations (45), spectra (36), survey (27), software (24), astronomical (22), numerical (18), catalogues (13), physical objects (13)
Engineering	Survey (33), Experimental (27), Simulations (26), Numerical (16), Images (15), Samples (15), Observations (12), Software (12), Measurements (10), system (8)

Further text analysis of bigrams identified the words that preceded ‘data’ in the responses from researchers in each discipline (Table 6.3). Similar to frequently occurring single terms (Table 6.2), ‘survey’ was the most common term that appeared with data in all disciplines, indicating that this is the most common method to generate new knowledge. Qualitative data and data from interviews were frequently mentioned in social sciences, arts and humanities, business and economics, and medicine. Even though ‘secondary data’ refers to data generated by others, this term appeared 65 times in the business and economics group – perhaps indicative of reliance on secondary data from non-academic sources (e.g., government statistics) by researchers in this field. Discipline specific data collection methods were mentioned in STEM fields.

Table 6.3 Terms that frequently appeared before 'data' for types of data produced

Discipline	*Bigrams (words appearing with 'data')*
Social Sciences	Survey (326), interview (61), qualitative (40), experimental (19), raw (15).
Arts and Humanities	Survey (49), interview (8), qualitative (8), textual (6), raw (5).
Business and Economics	Survey (231), secondary (65), interview (15), raw (11), qualitative (10).
Environmental Sciences	Survey (80), experimental (15), environmental (7), raw (7), genetic (6).
Earth & Planetary Sciences	Survey (20), raw (5), chemical (5), model (4), numerical (5).
Biomedical Sciences	Survey (28), behavioural (17), imaging (13), experimental (12), raw (10), sequencing (6).
Medicine	Survey (38), clinical (9), qualitative (6), imaging (5), interview (4).
Physical Sciences	Survey (20), simulation (12), raw (10), numerical (6), spectroscopic, (5).
Engineering	Survey (21), experimental (15), raw (9), simulation (7), numerical (4).

Data formats also varied between disciplines (Figure 6.2 and Appendix C; one participant could select multiple formats). Numerical data was popular overall across all research areas except Arts and Humanities (25%). Text was the most common format in Social Sciences (73.7%) and Arts and Humanities (88.3%). These two research areas along with Engineering were the top producers of multimedia (audio and video) data; Visual and Performing Arts (33%) and Linguistics and Language (38%) were the highest. Physical Sciences (45%), Engineering (40%), and Oceanography (45%) in Earth and Planetary Sciences (36%) generate many computer programs. Biomedical Sciences (55.3%) produces many images; this category was common overall in all research areas except Social Sciences and Business and Economics.

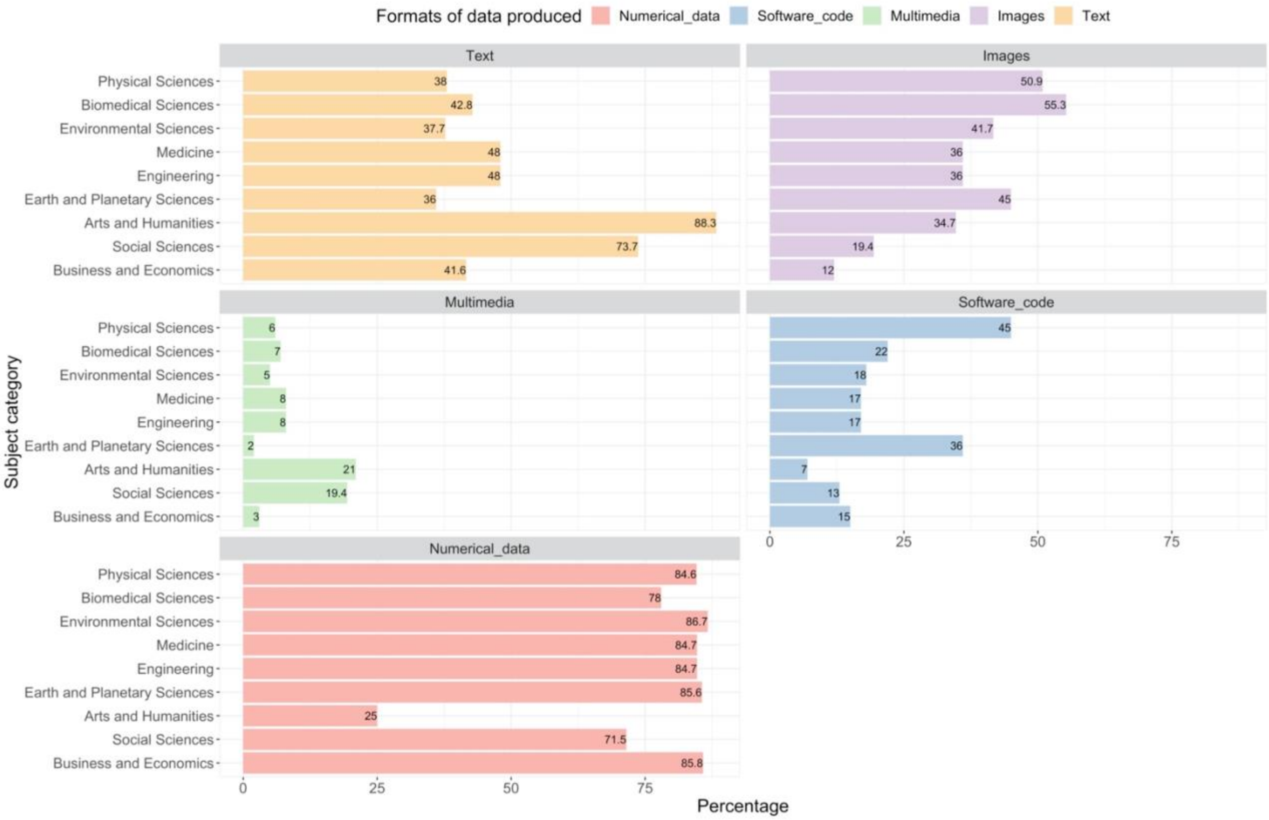

Figure 6.2 Format of data produced by subject category

6.3.2 Data sharing across disciplines and research experience

Nearly half (46.8%, *n*=1,523) of the participants reported sharing data on the web. Self-reported data sharing experience varied among researchers in different stages of their research career (X^2=36.85, p<0.001), and data sharing became more common with more research experience. The percentage of researchers with previous data sharing experience increased from 39% during 0-3years to 50% after 10+ years.

Differences in data sharing were significant between disciplines (X^2=200.17, p<0.001). Data sharing was most common in Physical Sciences (73%), followed by Earth and Planetary Sciences (70%) (Figure 6.3). In comparison, data sharing was less common in Business and Economics (33%) and Medicine (38%).

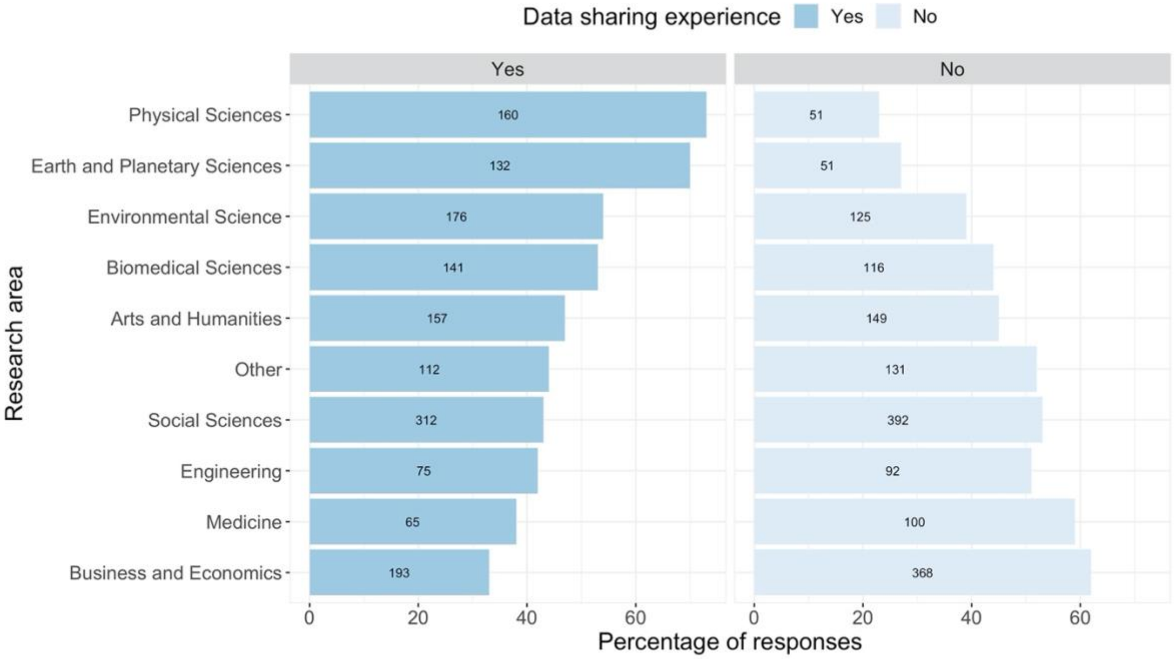

Figure 6.3 Data sharing across disciplines (labels on bars represent numbers of responses)

A binomial multivariable logistic regression was performed to understand the effect of disciplinary difference on data sharing behavior while controlling for research experience. Out of 3,257 responses 159 "I don't know/ not sure" responses were excluded as these represent only 4.8% of the responses on average. A simple model was tested with discipline as the predictor variable. Research experience was added to the model as this was also considered a predictor of data sharing and interaction between research experience and discipline was introduced in the model to explore any relationship between these two factors, but no significant outcome was observed, as shown in Figure 6.4. The forward stepwise variable selection method was applied to select the final set of variables. This was selected as the final model for the lowest AIC value (4122.6) and greatest significance of the likelihood ratio test. Table 6.4 shows the regression results, which indicates that compared to Arts and Humanities, being in Business and Economics, as well as Medicine significantly decreased the chances of data sharing by 0.5 and 0.63 times respectively. In contrast, the chances of data sharing increased by 2.9 times for researchers in Physical Sciences and 2.4 time for those in Earth and Planetary Sciences.

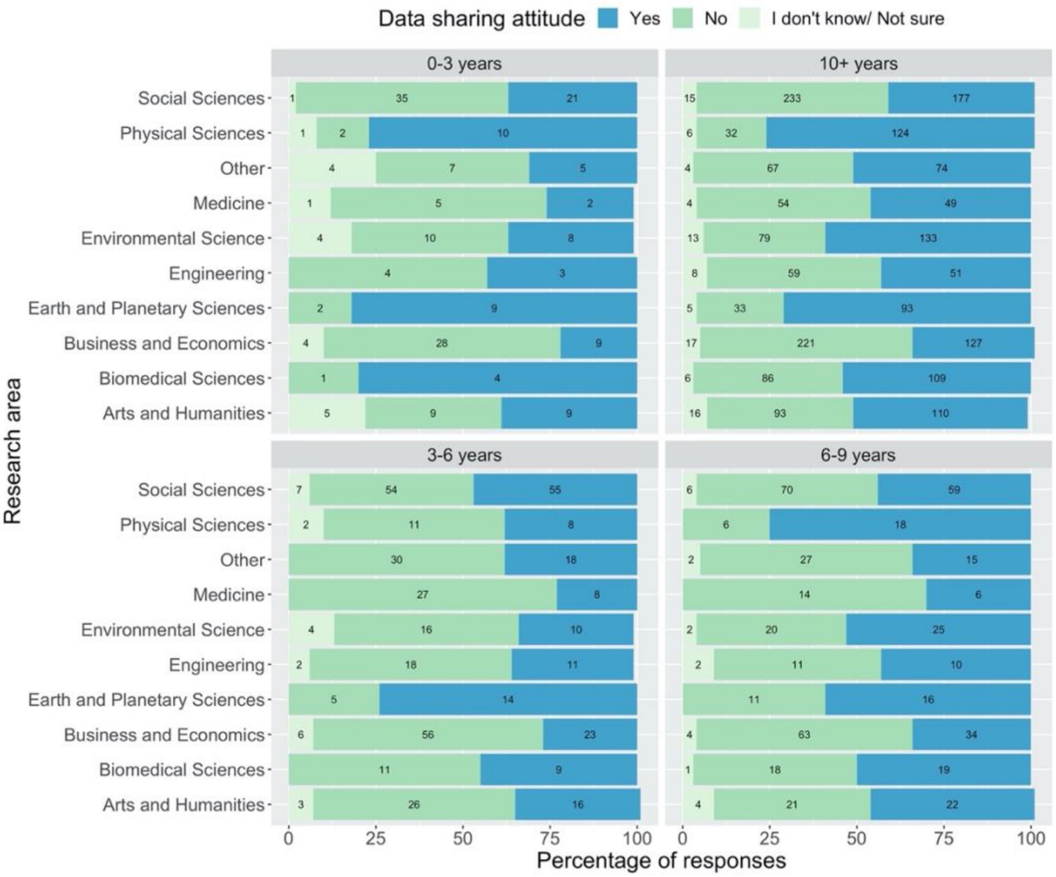

Figure 6.4 Data sharing in groups with different research experiences across disciplines (labels on bars represent numbers of responses)

Table 6.4 Logistic regression of data sharing outcomes (*n*=3,098)

Predictor	*Estimate (β)*	*Std. Error*	*z-value*	*p-value*
Intercept	-0.15238	0.18822	-0.810	0.4182
Biomedical Sciences	0.10910	0.17032	0.641	0.5218
Business and Economics	-0.69173	0.14524	-4.763	1.91e-06 ****
Earth and Planetary Sciences	0.89144	0.20112	4.432	9.32e-06 ****
Engineering	-0.25182	0.19368	-1.300	0.1935
Environmental Sciences	0.27336	0.16406	1.666	0.0957 *
Medicine	-0.46562	0.19683	-2.366	0.0180 **
Other	-0.18376	0.17283	-1.063	0.2877
Physical Sciences	1.07162	0.19778	5.418	6.02e-08 ****
Social Sciences	-0.26114	0.13775	-1.896	0.0580 *

10+ years	0.30144	0.16074	1.875	0.0607 .
3-6 years	-0.08549	0.18425	-0.464	0.6426
6-9 years	0.11700	0.17991	0.650	0.5155
Overall model evaluation				
Likelihood ratio test	$X^2 = 15.225$	df = 13		0.00163**
Significance codes: 0 ‘****’; 0.001 ‘***’; 0.01 ‘**’; 0.05 ‘*’				

Differences exist within specific disciplines as well as between broad subject categories. For example, under the Business and Economics research area, only 26% (*n*=50) of 193 respondents in Business and International Management had previous data sharing experience, compared to 44% (*n*=110) of 251 respondents in Economics and Econometrics. Similarly, within the social sciences, data sharing was less common in Education (27%, *n*=31) compared to Library and Information Sciences (LIS) (46%, *n*=117), and Linguistics and Language (46%, *n*=33). While Earth and Planetary Sciences and Physical Sciences have a strong culture of data sharing, Oceanography (89%, *n*=47); Astronomy and Astrophysics (82%, *n*=109) were the highest within these areas. Similarly, within Biomedical Sciences, Neurology (54%, *n*=21) has a higher rate of data sharing than Pharmacology (39%, *n*=18); and Ecology (63%, *n*=92) is higher than Pollution (43%, *n*=46) in Environmental Sciences. Despite data sharing being less common in Medicine, researchers in Infectious Disease (45%, *n*=17) more commonly share data than do those in Radiology (26%, *n*=7).

6.3.3 Methods of sharing research data

Among different methods of sharing data on the web, over half of the participants mentioned institutional repositories (53.4%, *n*=813), followed by journal-supported repositories (30%, *n*=457) and personal websites (24.5%, *n*=373) (participants could select more than one method). Chi-square tests show that significant disciplinary differences exist in the types of method used for sharing data (Table 6.5, with the highest response rate in each data sharing method marked in bold). Institutional repositories were the most common method of sharing data across all disciplines. The use of disciplinary repositories was more common in most STEM disciplines but less common in Engineering (1%), Arts and Humanities (7%), and Business and Economics (8%). Interdisciplinary repository usage was relatively common in Biomedical Sciences (26%), Earth and Planetary

Sciences (24%) and Social Sciences (23%) and least common in Medicine (6%). Sharing data on personal websites was most common in Physical Sciences (37%) and lowest in Medicine (9%). Examples of commonly used repositories show non-standard data deposit practices in Social Sciences and Arts and Humanities, such as Academia.edu and Google drive, which are not ideal solutions for long-term storage.

Table 6.5 Data sharing methods in different subject category

Subject category (n=previously data shared)	*Institutional repository*	*Disciplinary repository*	*Inter-disciplinary repository*	*Journal supported repository*	*Personal website*	*Commonly used repositories*
Social Sciences (*n*=312, 42.6%)	169 (54.2%)	42 (14%)	71 (23%)	66 (21%)	75 (24%)	Academia.edu, Zenodo, DANS, ICPSR, Figshare
Arts and Humanities (*n*=157, 47%)	105 **(66.9%)**	11 (7%)	21 (13%)	34 (22%)	53 (34%)	Academia.edu, Google drive, Zenodo, Mendeley, OSF
Business and Economics (*n*=193, 32.6%)	89 (46%)	16 (8%)	18 (9%)	75 **(39%)**	62 (32%)	Data in Brief, Figshare, ICPSR, Dataverse, American Economic Association
Physical Sciences (*n*=160, 72.7%)	77 (48%)	44 **(28%)**	31 (19%)	58 (36%)	59 **(37%)**	Zenodo, CADC, GitHub, CCDC, NASA databases, SDSS, SciFinder
Biomedical Sciences	66 (47%)	34 (24%)	37 **(26%)**	50 (36%)	27 (19%)	DDBJ, OSF, Figshare, GenBank, MRI

(*n*=141, 53.4%)						Image Consortium, NCBI, EMBL, PubMed, PubChem, The Cancer Imaging Archive, GitHub, GEO
Medicine (*n*=65, 38%)	35 (54%)	15 (23%)	4 (6%)	24 (37%)	6 (9%)	dbGAP, NCBI, GEO, GenBank, Zenodo, Dryad, EGA, IADR, fMRI database, PLoS ONE
Environmental Sciences (*n*=176, 54.3%)	90 (51%)	40 (23%)	30 (17%)	64 (36%)	22 (13%)	Dryad, GenBank, NCBI, PANGAEA, SeaBass, GitHub, MorphoSource, Barcode of life, ForestPlots.NET, NASA, NSF Arctic Data Centre, data papers and journals
Earth and Planetary Sciences (*n*=132, 70.2%)	86 (65%)	32 (24%)	32 (24%)	33 (25%)	22 (17%)	biorXiv, arXiv, PANGAEA, GitHub, DeepBlue, GIRO, NASA, NOAA, NCAR, NSF Arctic Data Centre, Zenodo

Engineering (*n*=75, 42%)	44 (59%)	1 (1%)	11 (15%)	23 (31%)	19 (25%)	Elsevier, Zenodo, Figshare, OSF, GitHub, Mendeley
Chi-square test result	$X^2 = 30.62$, $p < 0.001$	$X^2 = 66.35$, $p < 0.001$	$X^2 = 40.22$, $p < 0.001$	$X^2 = 36.03$, $p < 0.001$	$X^2 = 55.14$, $p < 0.001$	

Significant differences exist in data sharing methods within disciplines under the broader subject categories discussed above. Within the Social Sciences, institutional repository usage was the highest among LIS researchers (59%, *n*=69 out of 117) and lower in Linguistics and Language (48%, *n*=16 out of 33). However, the latter has a higher usage of disciplinary repositories (21%, *n*=7), interdisciplinary repositories (30%, *n*=10), journal-supported repositories (24%, *n*=8), and personal websites (36%, *n*=12). Researchers in both Literature and Literary Theory, and Visual and Performing Arts within Arts and Humanities rely heavily on institutional data repositories (over 70%). Journal supported repositories were highly used in Literature and Literary Theory (31%, *n*=15 out of 48) but 42% (*n*=13 out of 31) in Visual and Performing Arts relied on personal websites for data storage. Although self-reported data sharing was relatively lower in Business and International Management than Economics and Econometrics, non-standard data sharing using personal websites was much higher in Economics (41%, *n*=45 out of 110) than in Business and International Management (22%, *n*=11 out of 50), and opposite trend in institutional repository usage – 58% (*n*=29) in Business and International Management and 35% (*n*=39) in Economics. Repository usage was in general higher in Neurology than in Pharmacology. Within Environmental Sciences, disciplinary (25%, *n*=23) and interdisciplinary repository (21%, *n*=19) usages were higher in Ecology among those who shared data (*n*=92). In comparison, journal-supported data sharing was more frequent in Pollution (48%, *n*=22 out of 46). Earth and Planetary Sciences had an overall high data sharing rate. Researchers in Oceanography more frequently shared data in discipline-specific repositories (38%, *n*=18 out of 47) than in Geology (7%, *n*=2 out of 30), and the opposite trend was identified for journal-supported repository usage: 13% (*n*=6) and 53% (*n*=16) respectively.

In terms of research experience, there was no significant difference in data sharing methods except disciplinary repositories (X-squared = 9.01, p-value = 0.03): lowest among those with 0-3 years of experience, and personal website (X-squared = 9.24, p-value = 0.03): highest among researchers with over 10 years of research experience.

6.3.4 Choice of data repositories

When asked how they first found repositories to share data, most researchers responded that they were already aware of them, even though this varied between disciplines (Table 6.6). For example, in Physical Sciences and Biomedical Sciences, over 60% of respondents were already aware of relevant data repositories. This was followed by consulting with colleagues and consulting with experts, which was common across all disciplines. General web searches for repositories were more common in Engineering (29%) and Arts and Humanities (20%). Searching re3data, the registry of research data repositories, was not a preferred method by researchers in any discipline (5% or less), so this is potentially more of a professional librarian's tool.

Table 6.6 How researchers first found repositories to share data

Subject category (n=previously data shared)	*Already aware*	*Search re3data.org*	*Web search*	*Consult with colleagues*	*Consult with experts*
Social Sciences (*n*=312)	174 (55.8%)	8 (3%)	57 (18%)	106 (34%)	80 (26%)
Arts and Humanities (*n*=157)	66 (42%)	1 (0.6%)	31 (20%)	52 (33%)	37 (24%)
Business and Economics (*n*=193)	87 (45%)	2 (1%)	31 (16%)	46 (24%)	32 (17%)
Physical Sciences (*n*=160)	108 (67.5%)	4 (3%)	17 (11%)	47 (29%)	25 (16%)
Biomedical Sciences (*n*=141)	86 (61%)	1 (0.7%)	26 (18%)	51 (36%)	24 (17%)
Medicine (*n*=65)	26 (41%)	3 (5%)	11 (17%)	24 (38%)	15 (24%)
Environmental Sciences (*n*=176)	87 (49%)	2 (1%)	27 (15%)	62 (35%)	32 (18%)

Earth and Planetary Sciences (*n*=132)	63 (48%)	2 (2%)	18 (14%)	52 (39%)	28 (21%)
Engineering (*n*=75)	34 (45%)	1 (1%)	22 (29%)	29 (39%)	12 (16%)
Chi-square test result	$X^2 = 38.18$, $p < 0.001$	$X^2 = 8.13$, $p = 0.42$	$X^2 = 15.61$, $p = 0.048$	$X^2 = 13.35$, $p = 0.1$	$X^2 = 13.05$, $p = 0.11$

Ease of use (53.8%, *n*=820), repository reputation (46.9%, *n*=714), disciplinary norms (41.1%, *n*=626), and appropriateness for the data type (40.5%, *n*=617) were the top reasons for choosing a data repository. Other factors that influence researchers' choices are requirement from funding bodies, journals and institutions, accessibility, privacy, security, zero cost, digital object identifier (DOI) assignment, interdisciplinary research support, and international reputation for collaborative project support. The following factors were dependent on disciplinary differences: Reputation of repository, cost, and appropriateness for data type. Cost and appropriateness for data type were important factors in disciplines where disciplinary repositories were more commonly used (Table 6.7).

Table 6.7 Factors that influence choice of repositories in different disciplines

Subject category (n=previously data shared)	*Disciplinary norms*	*Cost*	*East of use*	*Reputation of a repository*	*Appropriateness for data type*	*Data curation services offered*
Social Sciences (*n*=312)	145 (46.5%)	104 (33.3%)	178 (57.1%)	156 (50%)	124 (39.7%)	48 (15%)
Arts and Humanities (*n*=157)	67 (43%)	54 (34%)	86 (55%)	73 (46%)	57 (36%)	18 (11%)
Business and Economics (*n*=193)	79 (41%)	56 (29%)	87 (45%)	79 (41%)	54 (28%)	17 (9%)
Physical Sciences (*n*=160)	66 (41%)	65 (41%)	98 (61%)	74 (46%)	73 (46%)	27 (17%)
Biomedical Sciences	60 (43%)	61 (43%)	79 (56%)	77 (55%)	78 (55%)	23 (16%)

(*n*=141)						
Medicine (*n*=65)	21 (32%)	23 (35%)	28 (43%)	35 (54%)	27 (42%)	9 (14%)
Environmental Sciences (*n*=176)	63 (36%)	66 (38%)	87 (49%)	73 (41%)	71 (40%)	27 (15%)
Earth and Planetary Sciences (*n*=132)	49 (37%)	57 (43%)	68 (52%)	50 (38%)	53 (40%)	20 (15%)
Engineering (*n*=75)	24 (32%)	26 (35%)	41 (55%)	41 (55%)	27 (36%)	5 (7%)
Chi-square test result	$X^2 = 12.87$, $p = 0.16$	$X^2 = 18.33$, $p = 0.03$	$X^2 = 16.38$, $p = 0.06$	$X^2 = 17.27$, $p = 0.04$	$X^2 = 31.2$, $p < 0.001$	$X^2 = 11.59$, $p = 0.24$

6.3.5 Data reuse across disciplines and research experience

Overall, 54.3% (*n*=1,769) of the participants self-reported that they had previously reused existing datasets and among them 82.5% (*n*=1,460) of researchers would like to know how others reused their datasets. Data reuse frequency was dependent on researchers' experience ($X^2 = 8.88$, $p = 0.03$). The proportion of previous data reuse experience increased with research experience: 47% in 0-3 years, 49% in 3-6 years, 53% in 6-9 years, and 56% in 10+ years.

Data reuse experience significantly varied between disciplines as well ($X^2 = 152.03$, $p < 0.001$). Over 80% of respondents in Physical Sciences and Earth and Planetary Sciences, and 56~60% respondents in Business and Economics, Environmental Sciences, and Engineering had reused existing data (Figure 6.5). This rate was lower among Arts and Humanities (42%), Medicine (44%), Social Sciences (47%), and Biomedical Sciences (49%). Between 1-3% of participants in all disciplines responded that their research does not use data; Arts and Humanities was an exception (16%).

Data reuse varies within specific disciplines in Business and Economics, as well as in Environmental Sciences. 72% (*n*=180) of researchers in Economics and Econometrics reuse secondary data and 24% (*n*=59) mentioned that they use only primary data. The opposite picture is seen in Business and International Management, where 54% (*n*=105) use their primary data and data reuse is comparatively low (38%, *n*=74). Within Environmental Sciences data reuse is higher in Ecology (63%, *n*=93) than Pollution (49%, *n*=53). This trend is similar to data sharing behaviour in these fields as data sharing was lower in the fields that rely on using primary data only.

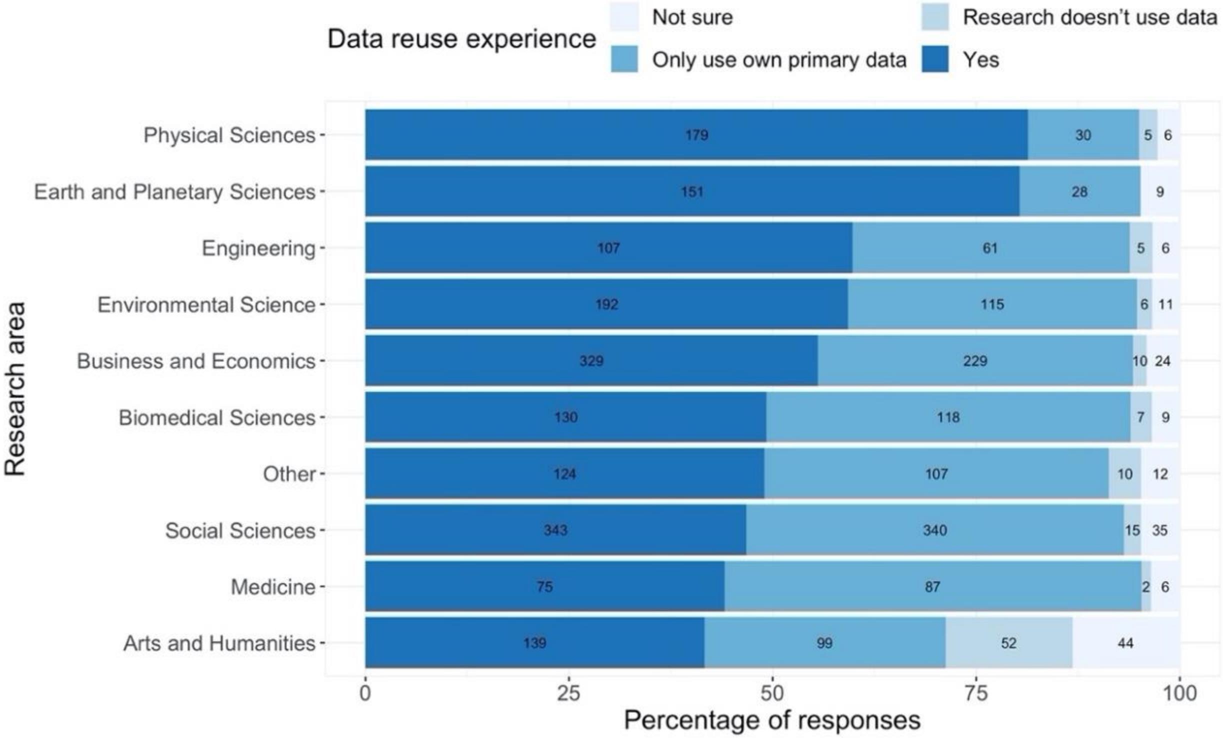

Figure 6.5 Data reuse across disciplines (labels on bars represent numbers of responses)

Similar to data sharing, a binomial multivariable logistic regression was performed with discipline and research experience as predictors (Table 6.8). 3,095 observations were included excluding "I don't know/ Not sure" values and aggregating these two groups: "I only use my primary data" and "My research doesn't use data" since both groups are not reusing data from any sources. Research experience did not have significant effect on data reuse and the complex model did not increase the accuracy of model (p=0.11 for Likelihood ratio test). Therefore, the simple model for disciplines is reported. This indicates that when compared to data reuse in Arts and Humanities, the chances of data reuse increase by 1.5 times in Business and Economics; 5.86 times in Earth and Planetary Sciences; 1.76 times in Engineering; 1.72 times in Environmental Sciences; and 5.56 times in Physical Sciences.

Table 6.8 Logistic regression of data reuse outcomes (*n*=3,095)

Predictor	*Estimate (β)*	*Std. Error*	*z-value*	*p-value*
Intercept	-0.08281	0.11754	-0.704	0.48114
Biomedical Sciences	0.12203	0.17178	0.710	0.47748
Business and Economics	0.40240	0.14505	2.774	0.00553 ***
Earth and Planetary Sciences	1.76788	0.23697	7.460	8.62e-14 ****
Engineering	0.56598	0.19368	2.891	0.00383 ***
Environmental Sciences	0.54451	0.16520	3.296	0.00098 ****
Medicine	-0.08834	0.19592	-0.451	0.65206
Other	0.14091	0.17444	0.808	0.41919
Physical Sciences	1.71484	0.21903	7.829	4.91e-15 ****
Social Sciences	0.04842	0.13982	0.346	0.72912
Overall model evaluation				
Likelihood ratio test	X^2 = 5.96	df = 10		0.1136
Significance codes: 0 '****'; 0.001 '***'; 0.01 '**'; 0.05 '*'				

6.3.6 Data sharing vs data reuse

The proportions of researchers who share or reuse data varied within disciplines. For example, data reuse was more frequent in Engineering and Business and Economics than data sharing on the web. However, sharing and reuse of data were dependent overall (X-squared = 181.11, p < 0.001). This was the same within all individual subject categories except Medicine and Engineering (Table 6.9).

Table 6.9 Comparison between data sharing and reuse across disciplines

Subject category	*Previously shared data*	*Previously reused data*	*Chi-square test results (Data sharing vs reuse)*
Social Sciences	312 (43%)	343 (47%)	X^2 = 34.594, p < 0.001
Arts and Humanities	157 (47%)	139 (42%)	X^2 = 7.839, p = 0.02
Business and Economics	193 (33%)	329 (56%)	X^2 = 19.175, p < 0.001
Physical Sciences	160 (73%)	179 (81%)	X^2 = 13.559, p < 0.001
Biomedical Sciences	141 (53%)	130 (49%)	X^2 = 35.29, p < 0.001

Medicine	65 (38%)	75 (44%)	$X^2 = 0.008$, p = 0.93
Environmental Sciences	176 (54%)	192 (59%)	$X^2 = 10.737$, p = 0.001
Earth and Planetary Sciences	132 (70%)	151 (80%)	$X^2 = 4.813$, p = 0.03
Engineering	75 (42%)	107 (60%)	$X^2 = 2.082$, p = 0.149

Those who responded “Yes” to previous data reuse were more likely to share data (56.8%, *n*=1,004). In contrast, those who responded “I only ever use my own primary data for my research” were less likely to share data (32.6%, *n*=396), except for Earth and Planetary Sciences, where 50% of the researchers who only use their primary data for research had previously shared data on the web (Figure 6.6). Data sharing among those who rely on own data was relatively common in Arts and Humanities (44%), Physical Sciences (43%), and Environmental Science (41%) as well. It is possible that those who reuse data shared by other researchers are more aware of data sharing practices in their field, but those who only use their own primary data for research are less so.

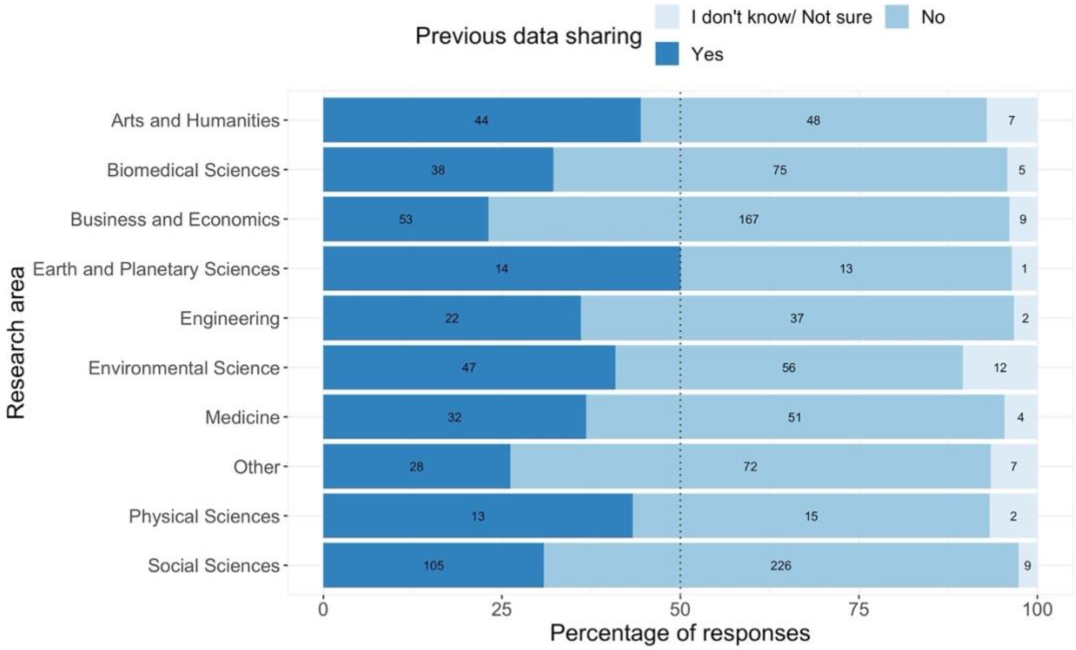

Figure 6.6 Data sharing among those who only use own primary data (labels on bars represent numbers of responses)

In disciplines where data reuse was more common overall, reuse was higher even among those who did not share data. At least half of the respondents in the top five disciplines reported that they have reused data from other sources even though they did not have previous data sharing experience (Figure 6.7). This includes Earth and Planetary Sciences, Physical Sciences, Engineering, Environmental Science, and Business and Economics. This could be because of the type of data required to conduct research in those disciplines, which leads to higher reuse of data even among those who do not have previous experience of sharing data on the web.

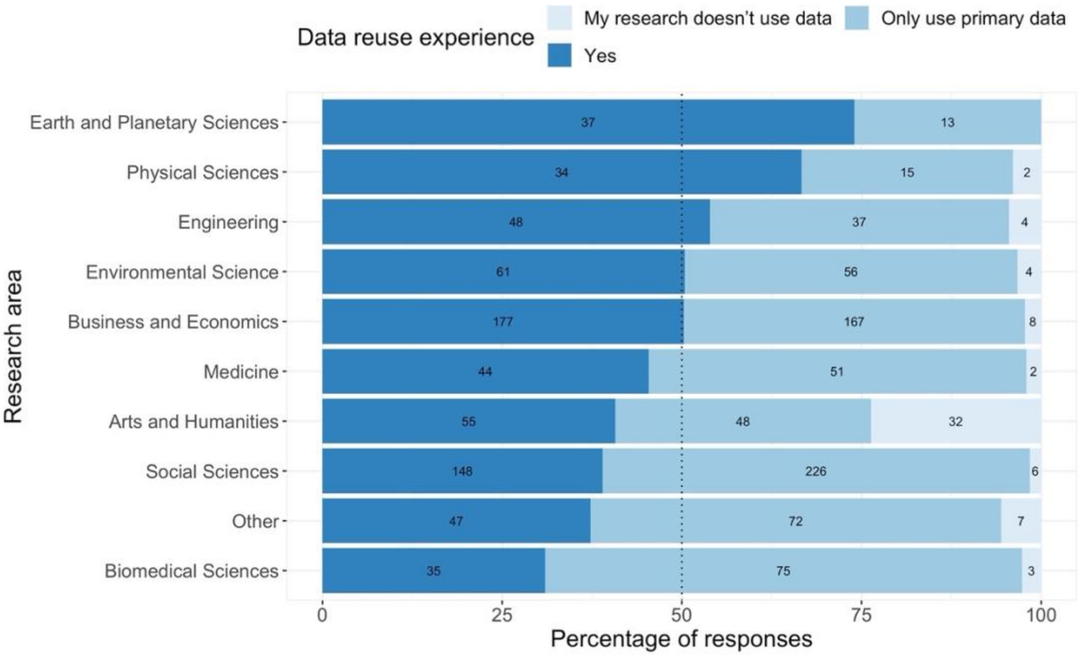

Figure 6.7 Data reuse among those who do not share data (labels on bars represent numbers of responses)

6.3.7 Data reuse purposes

Overall, 63.1% (*n*=1,116) of researchers reported that they combine multiple existing datasets to answer novel research questions; 50.7% (*n*=897) reuse data for comparing or ground truthing, i.e., calibrate, compare, confirm; and 46.6% (*n*=825) analyse a single dataset to answer novel research questions (multiple choice). However, types of data reuse were dependent on disciplines ($p < 0.001$ across all three types) (Figure 6.8). Comparing with the dotted lines in Figure 6.8 that show the average in each category of data reuse, analysis of a single dataset was more common in Medicine (59%); Business and Economics (55%); Social Sciences (53%); Physical Sciences (52%), and least common in Environmental Sciences (29%). Combining multiple datasets to answer new research questions was high overall: Earth and Planetary Sciences (80%); Physical Sciences (77%); and Environmental Sciences (71%) being the highest. Comparative data analysis was most common in

Engineering (71%); Earth and Planetary Sciences (66%); Physical Sciences (63%); Arts and Humanities (60%); and least common in Business and Economics (28%).

Other reuse types include testing and validating machine learning models, historical data analysis, teaching (e.g., master's student projects), analyzing evolution through time, quantifying long-term climate conditions, trying new statistical methods on existing datasets, replicating findings in diverse populations, reusing existing linguistic corpora as creating new corpora is time consuming, systematic review and meta-analysis, using GIS data to correlate with image files, inductive methods, and discourse analysis.

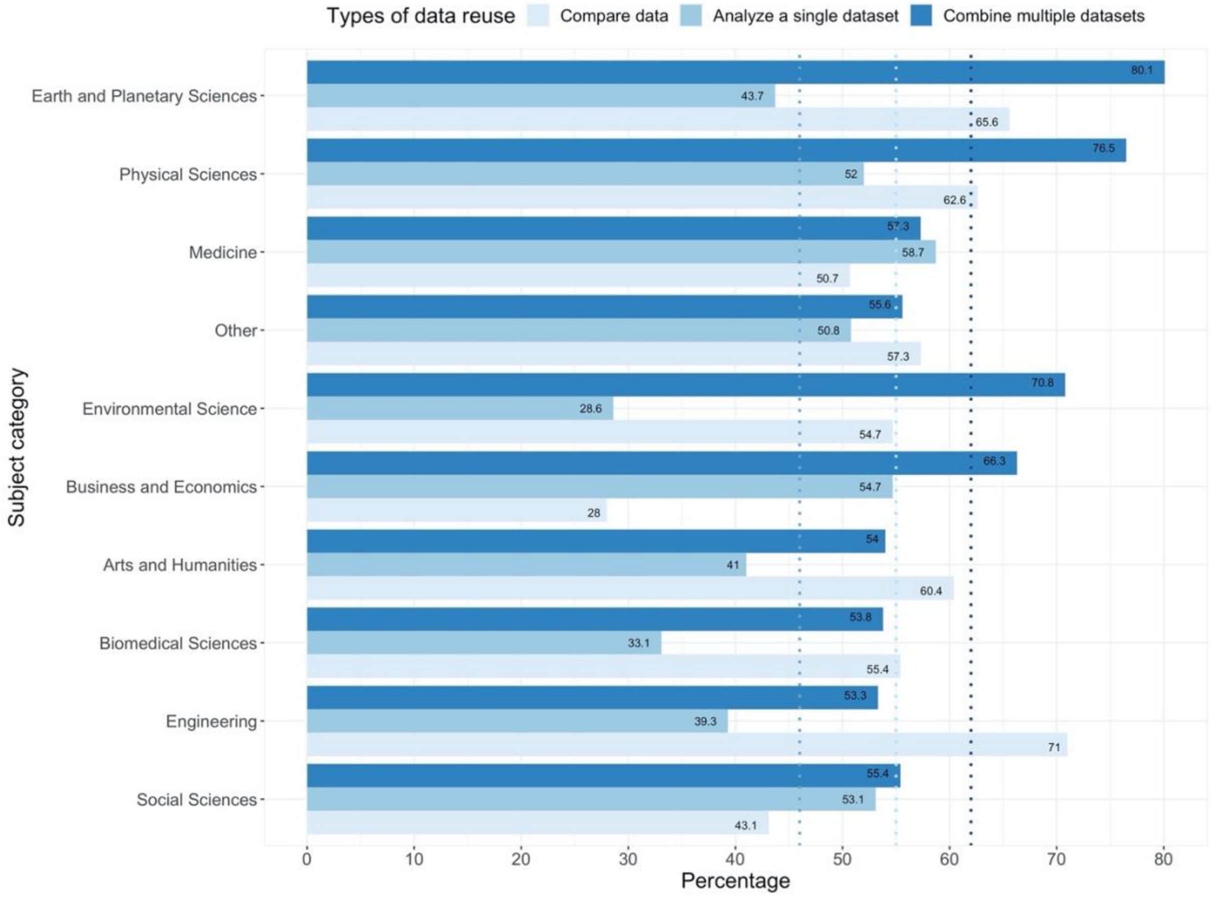

Figure 6.8 Data reuse types across disciplines

Specific examples of data reuse in each subject category from the open-text responses are included below, in the order of areas where data reuse is most common to least common (Figure 6.5):

Physical Sciences (*n*=113)

This subject category includes Astronomy and Astrophysics, and Organic Chemistry besides others. Researchers frequently use previous observations (e.g., telescope observations), all-sky maps, catalogues, images, measurements (e.g., photon flux, brightness, cosmological distance, galaxy clustering) in their research. Types of data reused include atomic reaction, spectroscopic, photometric, radioastronomical, optical, radio wavelength, numerical simulations, satellite, fMRI, EEG, and experimental data. Common astronomical databases used as a source of open data are: SIMBAD, Gaia archive, Sloan Digital Sky Survey, Galaxy Evolution Explorer, NASA Extragalactic Database, Hubble and Spitzer archives, Gemini Observatory Science archive, MAST archive of the space telescope. Chemical Abstract Service (CAS) and Cambridge Crystallographic Data Bank (CCDB) are used by researchers in Organic Chemistry. Researchers use these data in miscellaneous ways: to provide reference values for simulations, to test and tune simulations, for variability analysis (by comparison of targets against control sample), to compare own numerical simulations with data derived from astronomical observations, to test theoretical models, to confirm calibration, and to use astronomical observations in combination with newly acquired data to improve model accuracy. Some researchers mentioned reusing data associated with journal articles to combine with their own data and publish new findings. Others published articles entirely based on reuse of existing data. The following response explains the culture of data reuse in this area:

"*Most astronomical research requires data from multiple wavebands, and most observatories now archive all observations, so if one has, e.g., a new radio image of a galaxy, one can search for optical, X-ray, infrared, etc., images and spectra to combine with the new data.*"

Earth and Planetary Sciences (*n*=106)

This subject category represents Geology, Oceanography, and other fields. Researchers reuse different types of data: optical photometric, meteorological observations, metagenomic sequences, geochemical compositional and isotopic data, glacier inventory, climate data, remote-sensing, GIS data, digital photographs, trajectory data, maps, and statistics. Single or multiple types of data from different sources are frequently combined for research purposes. Data from multiple (40 or more) sources were used in one instance for input into models, validation, and assimilation. Other examples include collecting data from multiple radar stations from all over the world;

spectroscopic databases of natural targets (plant leaves and soils) combined with biogeophysical information; remote-sensing data from Landsat, MODIS, etc. combined with map land features; and combining physical oceanography data with fisheries survey data and coastal community socio-economic data from different sources. Researchers also compare their own data with other published studies, e.g., comparing their own geochemical data with those of others on similar rocks. Some apply "*new analysis techniques to historical time series data sets to see what additional information can be gleaned from older data sets*". The response below demonstrates the diverse use cases for existing research data in this field:

"I use a lot of data that are publicly available in national repositories that are collected by NOAA and NASA. E.g., weather buoy data from NOAA, and satellite ocean vector wind or satellite sea-surface temperature data from NASA. I also use public data collected by National Science Foundation […]. I also use large data sets collected previously by other oceanographers for other purposes […]. I (and my collaborators/advisors) have published many journal articles answering novel research questions using these data sets that were collected for other purposes (monitoring or to answer other questions)."

Engineering (*n*=70)

Researchers in various engineering fields reuse data to validate models and apply machine learning algorithm; develop and evaluate new analysis methodologies; use data as a benchmark for comparison; use existing video and image datasets to train baseline classification algorithms; and for comparison and ground truthing. Earthquake data, vibration data from NASA tests, electrophysiological measurements, EEG data, MR, CT and echographic DICOM data along with patient's clinical information, wind tunnel, flood damage data, and audio signals are some examples of datasets used by researchers in engineering. The use of existing data for modelling was most commonly mentioned. For example, "*I used results from laboratory experiments and results from numerical models to compare with the results of my own numerical model, to both calibrate/compare/confirm the validity of my model and to answer novel research questions with my model.*"

Environmental Sciences (*n*=137)

Disciplines included in Environmental Sciences are Ecology, Pollution, and others. Types of data/ datasets used include spatial datasets, DNA sequence data, meteorological, hydrogeological, water quality data, rainfall, temperature, satellite data, oceanographic, bathymetric data, GIS, remote sensing, and real time sensor data. Combinations of data from multiple disciplines and sources, as well as extraction of data for meta-analyses are common in these fields. Examples include combining biodiversity, land use, climate data; satellite imagery combined with ground measurements; combining multiple published phylogenetic trees into a single larger phylogenetic tree. One such example is, "*I use data from sources such as the IUCN Red List and combine these data with other databases on species traits such as EltonTraits 1.0 to answer questions such as the traits that drive species extinction risk*".

Another researcher used a combination of point data and data sets from several disciplines from over 300 different locations to answer novel questions in relation to population health and costs on the continental scale. Additionally, researchers frequently obtain data from other countries or regions to get wider geographical coverage in order to conduct cross-country analyses. Historical data in combination with field data are used to analyse long-term ecological trends, which is impossible without using prior data. An example given by a researcher is, *"My co-authors and I used water quality data collected since 1940 [...] from the US Geological Survey. We plotted the data to look at changes in sodium and chloride with time to look at increases due to road salt application".*

Business and Economics (*n*=175)

Within Business and Economics, data reuse was more common in Economics and Econometrics than Business and International Management. Administrative data, surveys, interview data, large-scale data collected by government sponsored programs, statistical offices, archival data about online transactions, micro and macroeconomic data, time series, panel data, firm-level or territorial-level data, and labour market data are all frequently used by researchers in these disciplines. The following sources were mentioned by the respondents: Bankscope, Wharton research databases, Amazon Review dataset, NHANES (National Health and Nutrition Examination Survey), FAO, world governance indicators, UK WERS data, Bloomberg data, EDGAR database, GEM, OECD data. The following example shows how a researcher combines and reuses various data for different purposes: "*Financial data on asset pricing, macro data on growth, data on climate research such as temperature, emissions, etc. – these were used both in my teaching for purpose of illustration and my research to calibrate parameters and evaluate model predictions*".

Biomedical Sciences (*n*=86)

Types of data reused by biomedical researchers include DNA sequences, screening, X-ray data, MRI, genes and genomes (nucleotide sequence), data on biological activities of compounds, RNA sequence, cognitive measures, and medical imaging data. Such data are used to answer new novel questions, test new techniques on existing datasets when developing new methods, combine with primary data to increase sample size or power of analysis. For example, one researcher mentions, "*I have used a large dataset of transcriptions from children with speech disorders, all of whom received intervention, to answer novel questions about intervention efficacy*".

The following example demonstrates how different types of data were reused by one of the respondents: "*Biological activity data (clustered, compared, analysed); chemical structure data (clustered, computed descriptors, examine variations, normalize, analyse); biomedical text (entity recognition, co-associations, frequency, analysed); terminology (compared, capture synonymy, analysed)*".

Social Sciences (*n*=198)

Social scientists frequently reuse the following types of data: census data, survey data, citation data, social media datasets, quantitative data from previously published peer-reviewed articles, socioeconomic, geographical data, public health data, eye-tracking data. Example sources include Afrobarometer survey to explore women's participation in politics, Talk bank for transcriptions from a wide variety of populations, the CHILDES database, the NIH Health and Retirement Study, household survey data from UNICEF, the GTAP database, the ICMA dataset, NCDS, SWAN, ALSPAC data, and the American Community Survey. Researchers often combine data from multiple sources with their own data (integrative data reuse), as mentioned by a researcher, "*I've collected my own survey and reaction-time data and compared important variables to the National*

Election Studies (NES) or General Social Survey (GSS) in the United States. Further, I often use census data and administrative data online".

Usage of corpus data is common among researchers in Linguistics and Language, including the British National Corpus, and the Corpus of Contemporary American English. Such use cases are diverse. For example, researchers have used corpus data from other dialects or other registers of the same dialect to analyse by itself, compare, confirm, establish diachronic relations; and have built collections of interactional phenomena drawing from different spoken corpora. One respondent mentioned, "*there are unlimited numbers of research questions that can be investigated using such data. Mine often involve comparing how particular linguistic forms are used in different varieties of English*".

Medicine (*n*=49)

Data reuse is comparatively less common in Medicine, perhaps because medical data often contain personally identifiable information that cannot be openly shared. Researchers mentioned reusing de-identified patient data, X-ray images, MRI/fMRI data, ultrasound data, epidemiological data, historical nucleotide sequence data, population health survey data, and genome-wide association studies summary statistics. Use cases include meta-analysis, cross-country comparisons, assessment of machine learning algorithms, estimate disease burden, risk factors, and treatment effects. One researcher reported, "*I usually use data from previous RNAseq analysis. I find candidate genes that could be induced/repressed, and I compare with results using different methods. I also get relevant information that provide new hypothesis about the pathway I am studying"*.

Arts and Humanities (*n*=74)

Arts and Humanities researchers mentioned reusing survey data, text corpora, interviews, text citations, digitized music scores and manuscripts, language acquisition data, archaeological reports, audio, and video recordings. Comparisons with existing data are commonly performed to strengthen the researcher's own findings. Usage of historical datasets is also frequent. For example, one researcher mentioned using crowd-sourced collections of musical metadata and digitised music scores to identify and investigate individual pieces of music in their cataloguing project.

6.3.8 Finding datasets to reuse

Among the researchers who previously reused datasets, 60.9% (*n*=1,078) found datasets by reading relevant papers. Also popular were web searches, such as Google Dataset Search (46.1%, *n*=816), and disciplinary repository searches (45.6%, *n*=806). However, all methods to find datasets varied in popularity between disciplines (Table 6.10). Searching disciplinary repositories was more common in Physical Sciences (69.4%) and Earth and Planetary Sciences (50%), whereas interdisciplinary repository search was higher in Arts and Humanities (35%) and Social Sciences (28%). Similarly, web search was a common choice in Engineering (63%), Arts and Humanities (60%), and Business and Economics (51.7%). These methods were not dependent on research experience.

Table 6.10 How researchers find datasets to reuse in different disciplines

Subject category (n=previously reused data)	*Search disciplinary repositories*	*Search inter-disciplinary repositories*	*Web search (e.g., Google Dataset Search)*	*Read relevant papers*	*By accident*
Social Sciences (*n*=343)	148 (43.1%)	96 (28%)	164 (47.8%)	177 (51.6%)	66 (19%)
Arts and Humanities (*n*=139)	66 (47%)	49 (35%)	84 (60%)	81 (58%)	38 (27%)
Business and Economics (*n*=329)	142 (43.2%)	76 (23%)	170 (51.7%)	181 (55%)	54 (16%)
Physical Sciences (*n*=170)	118 (69.4%)	22 (13%)	62 (36%)	142 (83.5%)	30 (18%)
Biomedical Sciences (*n*=130)	58 (45%)	35 (27%)	44 (34%)	79 (61%)	28 (22%)
Medicine (*n*=75)	28 (37%)	13 (17%)	23 (31%)	39 (52%)	9 (12%)

Environmental Sciences (*n*=192)	69 (36%)	40 (21%)	77 (40%)	114 (59.4%)	24 (13%)
Earth and Planetary Sciences (*n*=151)	75 (50%)	33 (22%)	71 (47%)	115 (76.2%)	17 (11%)
Engineering (*n*=107)	39 (36%)	25 (23%)	67 (63%)	73 (68%)	17 (16%)
Chi-square test result	$X^2 = 46.94$, $p < 0.001$	$X^2 = 31.38$, $p < 0.001$	$X^2 = 55.41$, $p < 0.001$	$X^2 = 63.12$, $p < 0.001$	$X^2 = 21.36$, $p = 0.01$

The following factors were considered most important by researchers when searching for existing datasets to reuse: proper documentation (67.39%, *n*=2,195), data being open (51.52%, *n*=1,678), and information on usability of data (42.22%, *n*=1,375). Availability of data in a universal standard format (35.7%, *n*=1163) and evidence that the dataset has an associated publication (34%, *n*=1,107) were of moderate importance. Evidence of prior reuse was considered important by a small percentage (8.3%, *n*=270). These factors were dependent on disciplines, and higher response rates within each factor are marked in bold in Table 6.11. All factors except evidence of prior reuse were of high importance to researchers in Physical Sciences. In contrast, researchers in Arts and Humanities have the lowest preference in all areas except open data. This could be because more researchers in this category responded that their research do not use data (16%), or they were unsure (13%). Information on the usability of data was considered important by researchers in all disciplines except Biomedical Sciences, and Arts and Humanities. However, such information is often not available via dataset records.

Table 6.11 Important factors when searching for existing datasets

Subject category	*Documentation*	*Open data*	*Info on usability*	*Associated publication*	*Universal format*	*Evidence of reuse*
Social Sciences (*n*=733)	493 (67.3%)	364 (49.7%)	326 (44.5%)	221 (30.2%)	257 (35.1%)	59 (8%)
Arts and Humanities (*n*=334)	183 (54.8%)	161 (48.2%)	110 (32.9%)	78 (23%)	64 (19%)	14 (4%)

Business and Economics (*n*=592)	381 (64.4%)	285 (48.1%)	239 (40.4%)	153 (25.8%)	201 (34%)	**63 (11%)**
Physical Sciences (*n*=220)	171 **(77.7%)**	151 **(68.6%)**	99 (45%)	112 **(50.9%)**	111 **(50.5%)**	10 (5%)
Biomedical Sciences (*n*=264)	189 **(71.6%)**	136 (51.5%)	96 (36.4%)	104 (39.4%)	103 (39%)	23 (9%)
Medicine (*n*=170)	111 (65.3%)	68 (40%)	88 **(52%)**	66 (39%)	**74 (44%)**	**19 (11%)**
Environmental Sciences (*n*=324)	233 **(71.9%)**	176 (54.3%)	142 (43.8%)	138 (42.6%)	111 (34.3%)	22 (7%)
Earth and Planetary Sciences (*n*=188)	146 **(77.7%)**	122 **(64.9%)**	91 **(48.4%)**	79 (42%)	**91 (48%)**	17 (9%)
Engineering (*n*=179)	120 (67%)	91 (51%)	82 **(46%)**	65 (36%)	59 (33%)	**20 (11%)**
Chi-square test results	$X^2 = 51.93$, $p < 0.001$	$X^2 = 55.18$, $p < 0.001$	$X^2 = 29.53$, $p=0.001$	$X^2 = 89.41$, $p < 0.001$	$X^2 = 81.52$, $p < 0.001$	$X^2 = 20.99$, $p=0.013$

Despite evidence of increasing data reuse in all disciplines, most researchers reported that it is difficult to find datasets to reuse (Figure 6.9). Physical Sciences was an exception, where over 50% researchers could easily find datasets to reuse. This percentage was slightly higher in Earth and Planetary Sciences (33%, *n*=57 out of 174) and Biomedical Sciences (29%, *n*=58 out of 199) as well.

Finding datasets becomes slightly easier with experience (Figure 6.10). 24% (*n*=511 out of 2,090) of researchers with over 10 years of research experience found it difficult to find datasets to reuse, compared to 26~28% of those with less experience.

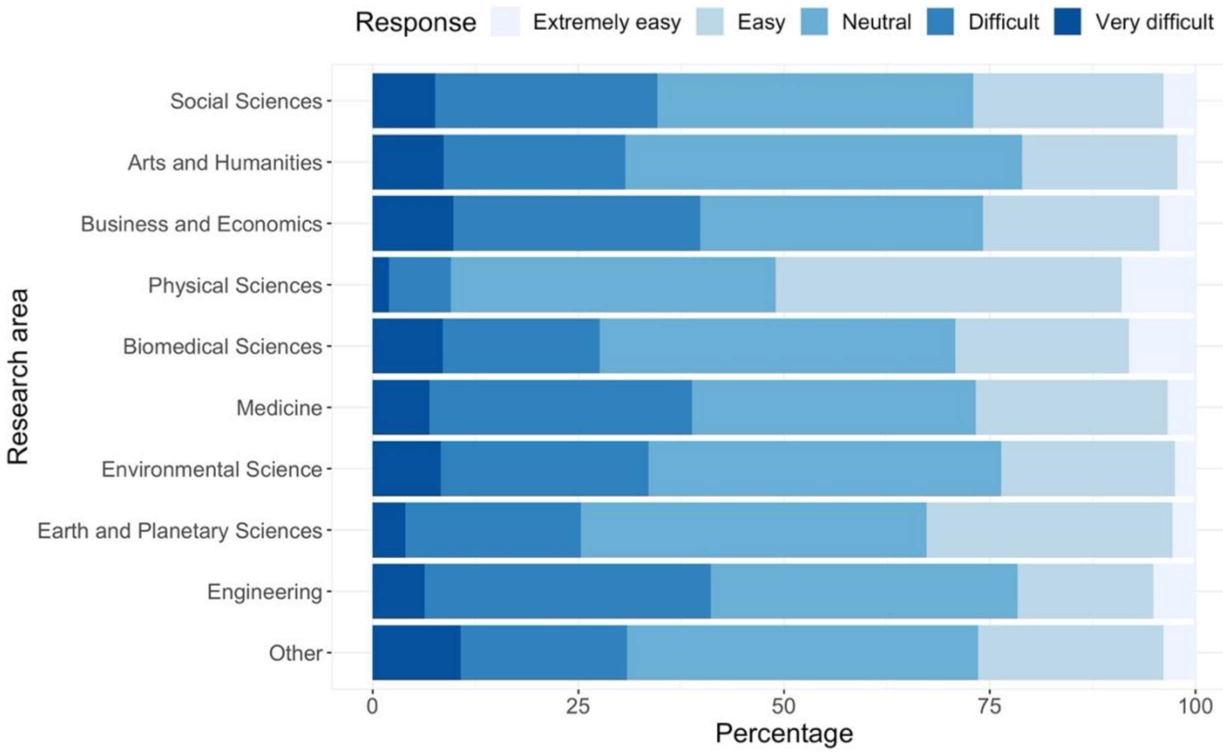

Figure 6.9 Ease of finding datasets to reuse by discipline

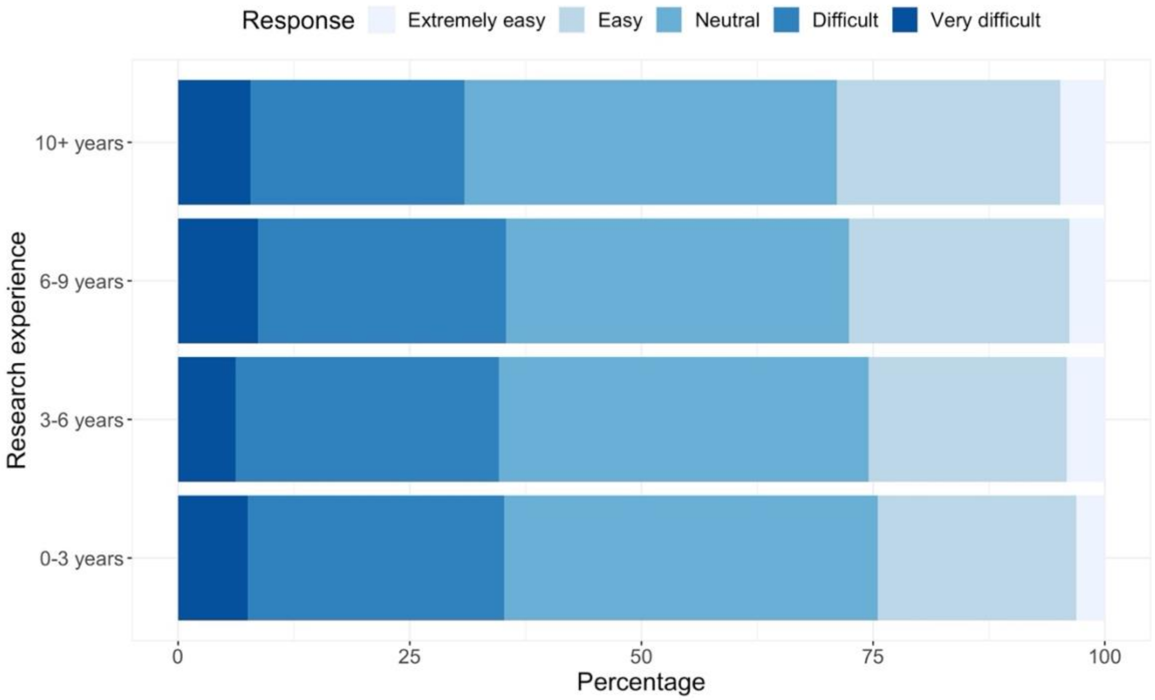

Figure 6.10 Ease of finding datasets to reuse by research experience

6.3.9 Promotion of data and future improvements

When asked whether they actively promote datasets, only 28.8% (*n*=510) responded yes (among 1,769 participants). Among these, 25.9% (*n*=132) promote their data when teaching, 33.3% (*n*=170) promote on social media platforms, 80.2% (*n*=409) promote within research groups and collaborator channels, and 16% (*n*=83) use other means of promotion. Twitter is the most used social media (61.8%, *n*=105), followed by Facebook (31%, *n*=53), blog posts (25%, *n*=42), and other platforms (31%, *n*=53). Among the participants who responded in the open-text field for "other means of promotion", most mentioned speaking about published datasets at conferences and workshops, as well as publishing data descriptor articles and reports about datasets. Others mentioned linking datasets on personal and project websites, in addition to institutional and research group websites. Academia.edu, Researchgate, Publons, and LinkedIn were mentioned by a few participants as some of the platforms they use to promote their datasets.

1,831 open-text responses were received for the question on future improvements in current systems to promote data reuse. A content analysis of these responses identified 23 recommendations in eight themes within three categories: 1. Issues around data, 2. Technological solutions, and 3. Cultural and policy changes (Table 6.12).

Table 6.12 Future improvements needed to promote data reuse

Category	Theme	Recommendations
Data related issues	Availability of data	Increased data sharing with available code (where applicable)
		Data is easily available and accessible with a DOI
	Handling of data	Better data management during research lifecycles
	Data citation	Formalize data citation to ensure datasets and associated articles are linked
	Usability of data	Data quality - reliable data with adequate documentation and in a standard format supported in individual's discipline

		Publish data paper/ data descriptor articles to enhance the usability of datasets
		Information on usability of data with some case examples (some datasets are produced for a single use)
Technological solutions	Search system	A single trusted portal or federated search system to search across multiple repositories and disciplines
		Enhanced search system with better tagging feature
		User-friendly data repository interfaces with fast data retrieval (for disciplines producing big data)
	New search system feature	A recommendation system for datasets
		Availability of data extraction and analysis-support tools in the same platform used to access data
		Alert system to notify when relevant datasets are made available publicly
Cultural and policy changes	Awareness and acceptance	Readiness, awareness, and acceptance within the scientific community to support secondary data analysis and publish in journals
		Promotion of data and repositories within scientific communities via conferences, webinars, training for early career researchers
	Incentives	Credit data creators/ reward data sharing in a similar way to publishing journal articles
		Create incentives such as data badges, data reuse indicators to promote data reuse
		Increased funding for secondary data analysis projects and to rescue historical data
	Collaboration	Form collaborations between data creators and users and their institutions (in some cases data are not reusable without contextual explanation)

	Guidelines and documentation	Streamlined IRB rules on how to handle qualitative/ medical data to share at the end of research
		Adequate guidelines on how to anonymize qualitative and health data to ensure data privacy
		Adequate legal and copyright information in place to access and reuse data
		Reduce bureaucratic application procedure for data access to avoid extended waiting periods

The most mentioned barrier to data reuse was a lack of knowledge about where and how to search for datasets. Therefore, a single trusted portal or federated search system across disciplines is needed that allows easy discovery of data:

"Perhaps more universal/federated searching mechanisms or portals--ArchiveGrid (https://researchworks.oclc.org/archivegrid/) was a game-changer for my research when it was released--now I no longer have to think "where might records about X person be?" and go to each individual institution and search."

Legal constraints about cultural data can be an impediment to data reuse in Arts and Humanities. A response from a humanities researcher outlines different policy issues and the need for incentives:

"Before we can improve data reuse, we need to improve communications between disciplines, accept the resource costs of making data reusable, reward people who do make their data reusable, and of course work with legal systems and institutions (archives, libraries, publishers etc) who 'own' cultural data to make reuse for research more fluid".

A few responses pointed out that not all datasets can have multiple use cases, because some are created for a single use only. Therefore, information on the usefulness of data can be helpful to external users. Streamlined IRB rules are critical to data sharing for reuse purposes in research where human participants are involved. Journals often request data to be made available, but IRB rules do not necessarily align. Adequate contextual information is key to successful reuse of data, along with researchers' commitment to share data and application of proper data curation methodology. Collaboration between data creators and reusers were recommended by multiple participants. One participant mentions:

"More opportunities for those offering qualitative data for reuse and those who might wish to reuse it to interact. There could be events in which the producers of the data explain what uses they had thought of, as well as potential reusers discussing their ideas to identify whether the data is really suitable for the purpose and what the ethical issues are. Most of the instances of successful reuse of qualitative data have involved interaction between the producers and reusers - the context in which the data is gathered is more important and the ethical issues are more nuanced (I think) than for quantitative data sets."

Changes in research culture and policies were mentioned by several participants where they indicated that secondary data analysis may not be considered as 'original enough' by journals to be accepted for publication. Environment Sciences researchers mentioned that data is often very difficult to collect in this area, and due to 'publish or perish' nature of academia, scientists are often reluctant to share their data. Incentives such as data badges, data reuse indicators, more funding for secondary data analysis projects, and rescuing of historical data were recommended to promote more data sharing and reward data creators. As suggested by one participant:

"…data work is nowadays high-quality scientific work as well, i.e., the reputation for data work needs to be increased (co-authorship for data work; establish "data"-chairs at universities and research institutes, etc.)"

6.4 Limitations

The precision of these results is affected by differing subgroup sample sizes. The sample sizes of researchers in different experience groups varied, with over 60% in the 10+ years' experience group. This could be due to the topic of this survey since the results indicate that those with more experience tend to share and reuse datasets more frequently. In addition, four disciplines had fewer than 30 responses. Disciplinary differences for these disciplines were not reported separately as the results may not accurately represent these groups. The participant recruitment method may also have impacted this (i.e., sample selection bias) as more experienced researchers tend to publish more and are listed as the first author more frequently. The results also have an unknown survey self-selection bias related to the 4.65% response rate. Unlike similar studies (Unal et al., 2019;

Tenopir et al., 2020), researchers' geographic location was not considered in this survey due to its focus on web-based data sharing and reuse. However, language barrier (i.e., for researchers in non-English spoken countries) and data sharing culture in different countries could have affected researchers' responses to these questions.

6.5 Discussion

Data sharing is known to be increasing in some disciplines to comply with funding body and institutional requirements. However, research data may not always be shared in a meaningful way that can lead to long-term accessibility and reuse. In this study, both data sharing and reuse were dependent on researchers' experience; those with more than 10 years of experience tended to share and reuse data more often. This supports the positive association between data sharing and a longer career reported by Gregory et al. (2020). Disciplinary differences exist in how researchers share data on the web, presumably driven by the culture of data sharing in a discipline – Physical Sciences, Earth and Planetary Sciences, and Environmental Sciences are more likely, whereas Business and Economics, Medicine, and Engineering are less likely to share data. Institutional repositories were frequently used in all disciplines, followed by journal-supported repositories. This could be because of rapid growth of institutional repositories and research data services in higher education institutions to comply with funder mandates (Cragin et al., 2010; Cox et al., 2017). Many journals are also mandating data accessibility statements and have associated data repositories, such as Mendeley Data by Elsevier. These results extend the previously known patterns in Tenopir et al. (2015) to a wider range of disciplines, (e.g., Business and Economics) and demonstrate increased usage of such repositories in recent years.

Disciplinary repositories have emerged to support domain specific data, such as in astronomy and astrophysics, zoology, and social science (Wallis et al., 2013; Faniel & Yakel, 2017). Data sharing and reuse activities are relatively common in these subject areas because researchers tend to be more aware of frequently used repositories in their field. This is in line with the good data practices reported by Tenopir et al. (2020) for Earth and Planetary Sciences, and Environmental Sciences. However, the current results suggest that disciplinary repository usage has increased in Physical Sciences in recent years, compared to Tenopir et al. (2015). Despite growing number of data repositories, personal websites were frequently used for data sharing in many disciplines except

Medicine, perhaps because of sensitive personal health data. This aligns with the findings of Tenopir et al. (2020). The examples of commonly used repositories reported by participants in this study demonstrate a lack of established data sharing methods in Engineering, Business and Economics, and Arts and Humanities, which could be one of the reasons for less frequent data sharing in these subject areas. The Registry of Research Data Repositories (re3data.org) currently lists 951 repositories under Humanities and Social Sciences, including 207 for Economics, and 517 repositories for Engineering Sciences among other disciplines. However, searching re3data.org was not a preferred method to find repositories; being used by under 5% of researchers across all disciplines.

This study supports existing evidence of growing data reuse in most disciplines, in line with the findings from biodiversity case study in Chapter 3 (Bishop & Kuula-Luumi, 2017; Borgman et al., 2019). Researchers in Physical Sciences and Earth and Planetary Sciences most frequently reuse data. Self-reported data reuse was more common than data sharing in Engineering, as well as in Business and Economics. In contrast to Curty et al. (2017), it reports that sharing and reuse of data are dependent within the study sample, except for Engineering and Medicine. This suggests that the nature of data sharing and reuse activities is evolving within disciplines and that the previous finding was limited to that study sample.

Despite increasing data reuse, researchers in most disciplines except Physical Sciences usually struggle to find datasets to reuse. Hrynaszkiewicz et al. (2021) report similar findings for their overall study population. These study results support the findings of Kratz and Strasser (2015) that most researchers read relevant papers to find reusable datasets, followed by web searches, and disciplinary repository searches and extends them to show disciplinary differences. Searching disciplinary repositories was common in Physical Sciences and Earth and Planetary Sciences, compared to other disciplines. This supports the findings from Kim and Yoon (2017), where the availability of data repositories was one of the main factors influencing data reuse at the disciplinary level. Even though reading papers is opted for by over 60% researchers, recent studies reported that only a small percentage of journal articles share data in a meaningful and accessible way (Federer et al., 2018; Thelwall et al., 2020). This may increase the difficulty and decrease efficiency of finding datasets from relevant articles. Therefore, it is important that data is shared in a standard data repository that assigns DOIs to ensure long-term access.

6.6 Summary

The survey results reported in this chapter have revealed the extent to which data production, sharing and reuse vary between disciplines. While self-reported data sharing is increasing, significant disciplinary differences remain in the adoption of standard data sharing methods. Data sharing is usually less common in disciplines that involve human participants (e.g., education, medicine), perhaps partly due to a lack of knowledge and preparedness among qualitative researchers (Mozersky et al., 2020). The need for adequate data anonymization guidelines in these disciplines was mentioned by researchers, along with streamlined IRB rules on handling qualitative or medical data to share at the end of research. In absence of appropriate guidance, new guidelines need to be developed, and existing guidelines should be reviewed by experts and funders regarding best practices for de-identifying data. For example, in 2012 the U.S. Department of Health and Human Services published a guidance on de-identification standards in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Such guidance is useful but may not be applicable in all circumstances and should be adapted depending on country and discipline specific rules.

The high usage of institutional repositories in this study sample indicates that institutional support can play an important role in data sharing. At the institutional level, research data management training programs and curated resources (e.g., lists of relevant data repositories) can help researchers in all disciplines adopt best practices for data production, management and sharing. Early career researchers will especially benefit from this, since the results indicate limited data sharing and reuse among less experienced researchers. Resources developed by the community of researchers, such as FAIRsharing.org (https://fairsharing.org/) can be used in training to help find a suitable data repository. Further guidance on responsible data sharing is provided by Alter and Gonzalez (2018); and Figueiredo (2017).

Standard data sharing and citation makes data findable and accessible in the long-term and can help reduce the burden of finding data to reuse, as echoed by researchers’ responses about future improvements needed to promote data reuse. Although the results show that sharing data as supplementary materials in a journal or personal website is still common in many disciplines, this does not ensure better discoverability and accessibility, and journal editors can ensure any related research data are deposited in a standard manner, adhering to FAIR principles. Since web search (e.g., Google Dataset Search) was the second most common method used to find reusable datasets, data repository managers can help researchers by adopting the Schema.org metadata standard to be indexed by Google Dataset Search and make their datasets more easily discoverable and reusable (Patel, 2019). This will particularly help siloed institutional repositories as these results show that researchers are more likely to search well-known disciplinary repositories to find datasets to reuse. Librarians should also be aware of the capabilities of Google Dataset Search so that they can help researchers find data that are not available in disciplinary repositories.

Future studies can examine researchers' attitudes and needs in specific subject areas identified in this study, such as Arts and Humanities, Business and Economics, and Engineering to further explore why data sharing is particularly low in these subject areas despite relatively frequent data reuse, which will help identify areas that need new policies, guidance, and infrastructure development. Furthermore, researchers in this study mentioned different incentives to support data sharing and promote data reuse, such as rewarding data creators in a formal manner similar to article publishing and implementing data reuse indicators or data badges to visualize impact of data sharing. However, incentives-based studies of data sharing are currently rare (Rowhani-Farid et al., 2017) and can be useful to understand and assess their impact on data sharing in these disciplines.

7. Conclusion

"I believe that promotion of data sharing and reuse could be encouraged more. This could be at an institutional level and by journals/databases. The focus in research design is often the generation of new data sets - almost as though using existing sets is 'less' professional or academic. By promoting reuse, communities of interest may not suffer the level of research fatigue they currently do. Also, it would be useful to promote the reuse of data not as an alternative to generating primary data, but as a complimentary asset. Research could use existing datasets and supplement them with contextual nuances through primary research."
(Survey respondent, 2020).

Technological advances in the past three decades have accelerated the growth of digital and digitized data. The numbers of data repositories, policies and guidelines have rapidly increased in the last 10 years to support systematic sharing of these data, a future that was envisioned by Ceci and Walker (1983). Despite the potential to promote new research and collaboration, as well as to reduce the duplication of efforts to collect the same data, differences between research cultures and disciplines can result in varied data sharing and reuse practices between individuals, within research groups, and across disciplines. Therefore, in this dissertation I addressed the guiding research question: what can be improved in the current systems and policies to support and promote data sharing and reuse? In this chapter, I review and summarize my findings from each empirical chapter and link these back to my guiding research question. I conclude by identifying the implications of my studies, limitations, and future directions of research in this area.

7.1 Key findings

In this section I identify the key findings from each individual chapter, which answer the specific research questions, and together these address the overarching research question.

RQ1: How do researchers reuse and cite secondary datasets in journal articles? Which metrics are informative about the impacts of secondary datasets?
The case study in Chapter 3 examined reuse cases of open biodiversity data indexed and published via GBIF by performing a content analysis of citing journal articles. The results show that openly

shared data create unique research opportunities and are frequently reused for background and foreground research in this area. Chapter 6 illustrated a more detailed picture with self-reported data reuse examples from researchers in nine subject categories: researchers most frequently combine multiple existing datasets to answer novel research questions, followed by reusing data for comparing or ground truthing, and analyzing a single dataset to answer novel research questions. However, content analysis results of citing journal articles in both Chapter 3 and Chapter 4 demonstrate that best practices of data citation are not common yet, however improving in recent years. The usage of multiple subsets further complicates data citation in biodiversity.

The majority of researchers (82.5%, n=1,460 out of 1,769) expressed an interest to learn about reuse cases of their data. Among different data reuse metrics, most repository managers considered citation counts to be of the greatest importance, followed by links to datasets from external websites, and download counts (Figure 5.1). A correlation test for download and citation counts for a random sample of cited GBIF datasets found a very strong positive correlation, suggesting that download counts may reflect similar kinds of impact for biodiversity datasets. From analysing altmetric mentions of GBIF datasets, it was found that blog mentions of *Checklist* datasets were more common, although most blogs referred to articles associated with datasets instead of the dataset itself. Blog posts on *Occurrence* datasets were more informative about the usability of these datasets, even though such blog posts were rare. In contrast, tweets and Facebook posts were mostly promotional. Even though altmetric scores may not be used as impact metrics, social media mentions of the usefulness of a dataset can potentially lead to further reuse.

RQ2: Is the major data citation tracker (Scholexplorer) able to automatically capture data citation and help identify data reuse?

While GBIF has developed its own semi-automated citation-tracking and article-linking system, academic institutions and institutional repositories often lack the capability to find links between their publications and dataset links, and vice versa. The results show that the Scholexplorer API can find previously unknown article-dataset links and reduce the manual labour of data repository staff by automatically identifying links between datasets and journal articles to a certain extent (Chapter 4). The results also show that many of the currently exposed article-dataset links did not result from consistent data citation practices and automated linking on the journal side. Also, the citation format was inconsistent in the majority of citing articles that reused data. Furthermore, the coverage of the API varied, with limitations in its metadata schema. Therefore, its current version may not completely replace the manual linking of articles to datasets by data librarians. However, following the recommended improvements in technical functionalities, it has the potential to identify more data reuse in the future.

RQ3: How do data repositories vary in data support services, and the technological and operational challenges they face when providing these services?

Chapter 5 explored the current practices, challenges, and future needs of data repositories. The survey results identified heterogeneity in the technical frameworks used by different repositories. Discipline-specific repositories develop more bespoke services, perhaps to support discipline-specific data types, whereas institutional repositories tend to use the technical frameworks available. A lack of engagement from users and a shortage of human resources were the top challenges for all repositories, especially for institutional repositories. Inadequate funding was more challenging for discipline-specific repositories. Despite their perceived usefulness, repository managers struggled to track dataset citations. However, repository managers are mostly reluctant to display download counts even though most tracked this metric. The main recommendations for future repository systems are: integration and interoperability between data and systems, improving research data management tools, implementing tools that allow computation without downloading datasets, and enabling automated systems (Table 5.3).

RQ4: How do data sharing and data reuse practices by researchers vary by research experience and between disciplines?

Chapter 6 identified that disciplinary differences exist in researchers' data sharing and reuse behaviours. Researchers with longer careers (10+ years) are more likely to share and reuse data than early-career researchers. Disciplines vary in the types and formats of data they produce for research. Publishing data via repositories is increasing but researchers still use personal websites to publish data. In cases when researchers are not already aware of a repository, they consult with colleagues and experts. Researchers self-reported various cases of reusing data, and combining multiple secondary datasets was the most common approach. However, most researchers usually find it difficult to locate datasets to reuse. Data sharing and reuse were dependent within the study sample. Data sharing was 1.7 times higher among those who had previous data reuse experience than those who only used their primary data for research. To promote further data reuse, 23 recommendations within eight themes and three categories were made: 1. Data related issues, 2. Technological solutions, and 3. Cultural and policy changes (Table 6.12). These should be considered to support data sharing and encourage data reuse in the future.

7.2 Contributions and significance of the study

This thesis contributes to the understanding of key areas of improvements needed in technology and policies to support data sharing and reuse, as well as to develop and implement reliable impact metrics for research data. Together, the four areas of investigation in this dissertation illustrate a broad picture of the current practices and gaps in data sharing, reuse, and citation.

By examining data citation methods in journal articles and the attribution of citations, the first case study of biodiversity (Chapter 3) identifies that appropriate data citation needs to be encouraged to automatically capture citation counts. Based on the findings, it recommends an alternative citation attribution method for GBIF, where usage of large subsets of data may result in erroneous citation counts. It also examines how datasets are mentioned in social media and suggests that blogs, Facebook, and Twitter could be used to promote datasets for further data reuse. The second case study (Chapter 4) examines the potentials of Scholexplorer to find links between datasets and publications and its applicability to identify data reuse cases. Further enhancement of the Scholix schema and enrichment of the Scholexplorer metadata using controlled vocabularies were recommended to assist in identifying data reuse in the future.

The survey studies of data repository managers and researchers in Chapter 5 and Chapter 6 present two of the largest surveys conducted in this area. The results depict heterogeneity in the current landscape. Studying different repository systems at this scale was challenging as there is no central database of repository managers and previous studies recruited participants in an ad-hoc basis via personal and professional channels. A significant contribution of the data repository managers' survey is that it identifies the gaps and needs in repository systems and outlines the type of tools and policies needed in the future to efficiently support data sharing by researchers and track data reuse. Similarly, the second survey of researchers explored data sharing and reuse practices across a large sample of 20 disciplines in nine subject categories including less represented disciplines and suggests 23 recommendations in three key areas for further improvement in systems, policies, guidance, and training programs.

I consider the primary contributions of this dissertation to be the following:

1. Recommendation of an alternative citation attribution system for biodiversity data subsets in GBIF (Section 3.4).
2. Recommendations for specific improvements of Scholexplorer to identify data reuse and an open-source code that can be reused by other institutions or repository services, as well as recommendations for improving metadata capture by the registry of data repositories (Section 4.5).
3. Identification of key areas of technological and policy-related interventions based on responses from data repository managers and researchers. These include nine recommendations for future repository systems to ensure efficient use of data repositories (Table 5.3) and 23 recommendations to support sharing and reuse of data by researchers (Table 6.12).

These areas play important roles in the complex ecosystem of research data creation, dissemination, and promotion, and can help establish an efficient data sharing and reuse model, which in turn can guide to an effective impact measure system.

7.3 Summary of findings

Chapter 3 explored data citation and reuse practices in the context of biodiversity datasets published in different data repositories and aggregated by the federated biodiversity infrastructure, GBIF. GBIF has been a leading platform in biodiversity that has a bespoke system of attributing citations to its indexed datasets and provides access of dataset metadata via an API, making it an ideal source for this study compared to other individual repositories. The investigation of nearly 44k datasets within the dataset creation period 2007 and April 2019 suggested that data citation depends on data quality, hence *Occurrence* datasets are the most frequently cited by articles that reused GBIF datasets. 642 unique articles published between 2013 and 2019 cited at least one GBIF indexed dataset, increasing from 0.6% in 2013 to 40.5% in 2018. This indicates high reuse value of open biodiversity datasets. A content analysis of a random sample of 100 citing articles demonstrated various background (18%) and foreground (81%) data reuse cases. Standard data citation practices have grown in recent years, with 48% articles linking to datasets in data access statements or references along with citations to the dataset within the article text. However, data citation practices are yet to be standardized in GBIF as the usage of large numbers of subsets complicates the citation process. When a subset is cited in a literature, GBIF automatically attributes this citation to the parent datasets from which these subsets were derived. However, given downloaded data subsets are usually cleaned and many observations are dropped in the final dataset, this can result in erroneous citation attributions. Careful consideration is needed to use citation counts as an impact measure in this case. An alternative solution was recommended in that chapter to correct this feature. Altmetric mentions of GBIF datasets in blogs, tweets, Facebook posts, and Wikipedia were explored as an alternative metric to understand whether these are indicative of the societal impact of biodiversity datasets. Moderate correlations were found between citation counts and blog and Twitter mentions for *Occurrence* datasets, and blog posts are the most indicative of the impact of such datasets even though these are rare. Facebook posts were mostly from data creators for promotional purposes, but these posts often contained information on the usefulness of a dataset compared to tweets, perhaps due to the character count limitation of Twitter.

Given the importance of attributing citations to datasets, Chapter 4 explored Scholexplorer as a technological solution to automatically find links between datasets and articles. This is of particular importance to institutional repositories since disciplinary federated repositories, such as GBIF, have the capacity to create in-house systems. This first empirical study of the Scholexplorer API searched for nearly 32k University of Bath associated research output DOIs and identified 1,501 new dataset links published in external data repositories - a 31-fold increase. However, it identified gaps in the API coverage. For example, white papers or reports were not covered at the time of this study, and most links were generated by DataCite instead of Crossref. Comparisons of the author names of 121 datasets linked to 41 journals in the initial sample identified 10 data reuse cases, among which only three articles included citation and reference to the datasets in a standard manner. This finding was reinforced by 319 article-dataset links identified for 269 datasets published via the University of Bath Research Data Archive. All of these were primary publications associated with the datasets, which are generally manually linked by the archive staff. It is likely that most of these links were captured because of manual dataset-article linking rather than inclusion of citations and references to datasets in a standard manner. These findings show that further enhancement of the Scholix schema, as well as the application of a standard vocabulary and naming convention can be useful to identify data reuse cases. Standard data citation practices are key to the automatic identification of journal-dataset links. More links from Crossref would be helpful as this can help to reduce the manual effort of data repository staff.

Chapter 5 identified the current challenges and needs for improving data repository functionalities and user experiences by surveying data repository managers. Among 189 responses received, 47% (*n*=89) were discipline specific and 34% (*n*=64) were institutional repositories. 71% of the repositories reported that their software used bespoke technical frameworks, with DSpace, EPrint, and Dataverse being commonly used by institutional repositories. Few data repository managers reported that they were able to track data reuse: 25% of institutional and 38% of disciplinary repository managers reported tracking some form of secondary data reuse, while 64% of institutional, 49% of disciplinary, and 41% of cross-disciplinary repository managers would like to implement this. Among data reuse metrics, citation counts were considered the most important by the majority (57%), followed by links to the data from other websites (44%) and download counts (41%). This was particularly important to institutional repository managers (61%). Despite their perceived usefulness, repository managers struggle to track dataset citations: 46% of discipline-specific, 33% of institutional, and 23% of cross-disciplinary repositories reported tracking citations. Those who are currently not tracking citation counts reported a lack of standard data citation practices by researchers, absence of reliable technological solutions, and shortcoming of DCI coverage as reasons for not being able to do so. This is an important gap, which can be addressed by the enhancement of new solutions like Scholexplorer as discussed in Chapter 4.

Most repository managers support dataset and metadata quality checks via librarians, subject specialists, or information professionals. A lack of engagement from users and shortage of human resources were the top two challenges. Insufficient funding was challenging for nearly half of the discipline-specific repositories, perhaps because institutional repositories often rely on academic institutions' funding. Key motivators for data sharing were outreach (69%), which was mentioned by repositories across all groups. This was followed by funder policies (59%) and training programs (56%). Funder policies were a strong motivator for institutional repositories (77%) as academic institutions need to demonstrate compliance with funder mandates. Ensuring FAIR data (49%), providing user support for research (36%), and developing best practices (29%) were the top three priorities for repository managers among 11 categories identified. Among nine recommendations for future repository systems, the highly mentioned ones are - integration and interoperability between data and systems (30%), better research data management tools (19%), tools that allow computation without downloading datasets (16%), and automated systems (16%).

This study identified gaps in current data sharing and data citation practices, as well as issues with the available technological solutions that assist with data discovery and track secondary data reuse. The recommendations from this study can inform technology providers, e.g., software designers of repository framework, Scholexplorer, re3data.org, and Google Dataset Search, about the needs of academic institutions and researchers in different disciplines. Most importantly, data should be cited in a correct manner when used in research publications. Researchers, journal editors, and data repositories have their own roles to ensure a sound data-article linking system that can assist in defining automatic and reliable methods to capture data citations.

Finally, Chapter 6 explored data sharing and reuse practices across a large sample of researchers in 20 disciplines under nine subject categories, including less represented disciplines, and suggested areas of improvement in terms of policy, guidance, and technological solutions. Types and formats of data produced varied between disciplines. Surveys are most frequently produced by researchers across all disciplines, followed by observations. Social Sciences and Arts and Humanities generate more qualitative data. Therefore, text format was the most common in these two subject categories, along with multimedia. Collection of samples are common in Environmental Sciences, Earth and Planetary Sciences, and Biomedical Sciences; simulations are frequently generated in Physical Sciences and Engineering; and images in Biomedical Sciences and Physical Sciences. Among the 3,257 researchers who responded, 46.8% (n=1,523) self-reported sharing data on the web and 54.3% (n=1,769) had experience of reusing existing data. Sharing and reuse of data increased by research experience and varied between broader subject categories, as well as within specific disciplines under these categories. Data sharing and reuse practices were most common in data-intensive disciplines with long-standing culture of data sharing, such as Physical Sciences, and Earth and Planetary Sciences. In contrast, data sharing was less common in Business and Economics, Engineering, and Medicine, even though data reuse was more frequent in the first two subject areas. Examples of differences in specific disciplines include Oceanography, where data sharing is more usual than Geology, similarly higher data sharing rate in Economics and Economics than Business and International Management. However, standard data sharing via a repository is still not universal and many still share data by posting on personal website. Use of disciplinary repositories was more common in STEM disciplines, but institutional repositories were most frequently used for data sharing across all disciplines, followed by journal-supported repositories and personal websites. Most researchers were already aware of a repository or consulted with colleagues. Ease of use (53.8%), repository reputation (46.9%), disciplinary norms (41.1%), and appropriateness for data type (40.5%) were the main factors that influenced their choice of a repository. Cost was a main factor for disciplinary repositories. This is understandable as data repository managers reported lack of funding as a major challenge (Chapter 5), and therefore charge for repository usage.

The findings on data reuse from this survey supports the findings in Chapter 3: open data unfolds new opportunities in research. Data sharing and reuse were dependent in this study sample, where data sharing increased with prior data reuse experience, and those who only used their primary data in research reported a lower rate of data sharing in general. Perhaps this is because not all disciplines have similar data sharing cultures, and when researchers reuse data from other sources it raises more awareness. Most researchers read relevant papers to find datasets to reuse. Disciplinary repository searches were more common in disciplines with established disciplinary repositories, such as Physical Sciences, and Earth and Planetary Sciences. Interdisciplinary repository searches were higher in Arts and Humanities, as well as in Social Sciences. Similarly, web searches were a common choice in Engineering, Arts and Humanities, and Business and Economics. However, researchers find it relatively difficult to find datasets to reuse except those in Physical Sciences.

Among different types of data reuse cases, the most common method was to use a combination of datasets to find answers to novel research questions, followed by comparative use for ground truthing, and analysis of a single dataset to answer new questions. This study also provides specific examples of data reuse methods, along with common sources of data researchers reuse in each subject category (Chapter 6). Access to open data allows using datasets for teaching and learning, e.g., examine new statistical models, benchmark against existing data. These types of educational use cases demonstrate societal impact and are considered important by data repository managers as well (Chapter 5). However, currently there is no systematic method to capture such information. A set of 23 recommendations around data, technology, and research policy were made in Chapter 6 to further support data sharing and promote data reuse.

7.4 Implications

This study unveils the complex ecosystem of research data by exploring the current landscape and challenges from different stakeholder perspectives. It identifies the gaps and necessities, as well as provides recommendations to improve current technologies, policies, and research culture. Results from this study confirm that data citation practices remain inconsistent (Federer et al., 2018; Thelwall et al., 2020), but have been improving in recent years. The use of data subsets poses an additional challenge in disciplines with large or aggregated datasets (Silvello, 2018). This is the first study to conduct a content analysis of articles citing biodiversity datasets indexed by GBIF and expose how articles that use large number of subsets find it challenging to cite data in an appropriate manner. A revised metadata storage and citation model was recommended to update the current citation attribution system. This recommendation can be generalized in similar systems and has the benefits of assigning citations to the correct datasets only, be more informative of dataset content, and encourage learning from various use cases of GBIF datasets, which can in turn lead to the generation of newer ideas.

Standard data citation practices in journal articles and the availability of systematic journal-article links via Crossref would be beneficial to automatically capture these relationships with Scholexplorer. Journal mandates to include data citations and agreeing to a universal data citation practice by journals in all disciplines would be essential for its successful implementation. This would be valuable for data repositories, especially for institutional repositories, to help identify whether and where their researchers publish datasets and to demonstrate compliance to funder mandates by reducing repository staffs’ manual effort to link datasets with publications. The case study of the Scholexplorer API is the first known empirical study that identifies the gaps in its coverage and recommends further enhancement of the Scholix schema. The open-source code developed as a part of this study has already been reused by external partners in libraries and by the Spanish National Research Council to use this method for finding any existing links between datasets and research articles that they are currently unaware of (Khan, 2020).

The following model shows the ecosystem of research data with the essential roles of different stakeholders (Figure 7.1).

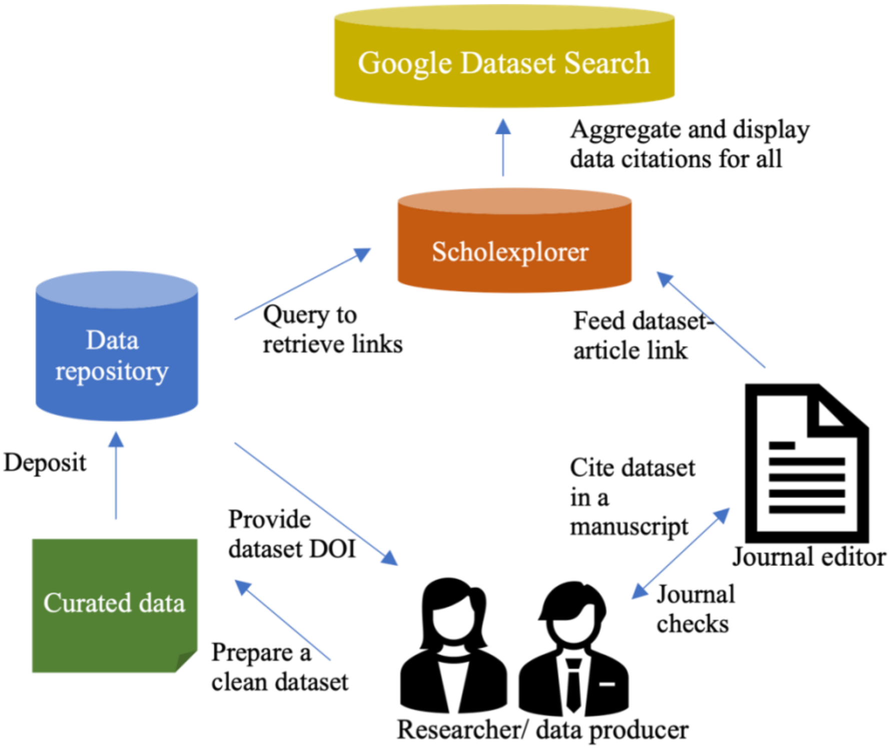

Figure 7.1 Research data ecosystem

The two surveys performed in this research suggest key areas of improvement by comparing data repository services and identifying disciplinary differences in data sharing and reuse practices. Integration and interoperability between data and systems, and a central data portal or discovery system was recommended by both repository managers and researchers in different disciplines. The registry of research data repositories, re3data.org indexes data repositories across the world and is considered one of the main sources to learn about available data repositories in a discipline.

However, the researcher survey revealed that this is rarely used by academics to find repositories. This research also identified gaps in re3data metadata. Recommendations for further enhancement and refinement of the re3data schema would be beneficial to both researchers who are interested to find data repositories and those who are interested in studying data repositories.

The researcher survey also highlighted that there are differences in data sharing and reuse rates in most disciplines even within the same subject category. These findings are important when implementing policy changes, such as rewarding or penalizing researchers for sharing or not sharing their data. Data sharing and data reuse are separate actions, and there may not be a direct cause-effect relationship. Researchers are expected to share their data with the broader scientific community for greater dissemination, usability, and reproducibility. However, data reuse can vary in many ways and does not always link to a new research publication. Therefore, it can be expected that researchers may have data reuse experience even if they did not have data sharing experience in the past. Increasing data sharing with reuse in this sample may indicate that when reusing data researchers get more familiar with data sharing. It should also be considered that if researchers heavily rely on secondary data to answer their research questions and those data are already available from other databases, they cannot re-deposit the existing data since there will be legal issues. Hence more data reuse in that case may result in less data sharing.

Disciplines in some subject categories such as Business and Economics, and Arts and Humanities have not been well-represented in previous studies. This study identified that secondary data analysis is more common in Business and Economics, but the data sharing rate was much lower among those who only use their primary data for research (23%, 53 out of 229). However, researchers mentioned a lack of open data and known or usable databases in this area. This study lays the groundwork for future research: is data sharing less common because of lack of repositories or because data already exists in a database from which data is reused, and therefore, cannot be shared?

While data repositories are instrumental for data sharing, too many repositories make it cumbersome to find datasets. This study sets out six areas of improvement: 1. A single trusted portal or federated search system to search across multiple repositories and disciplines; 2.

Enhanced search system with a better tagging feature; 3. User-friendly data repository interfaces with fast data retrieval (in cases of disciplines producing big data); 4. A recommendation system for datasets; 5. Availability of data extraction and analysis-support tools in the same platform used to access data; and 6. Alert system to notify when relevant datasets are made publicly available. Launched in September 2018, Google Dataset Search is trying to fulfil the need of a central data portal allowing discoverability across repositories. However, the service is still in its infancy and not widely known or used by researchers in this study sample. Moreover, a comparison of the citation counts of same datasets on Google Dataset Search and GBIF showed a major discrepancy (Chapter 3). Suggested features in this study will not only facilitate the search process but also support serendipitous finding.

A few studies have explored badges as an incentive for data sharing where journals in psychology and life sciences participated in pilot studies (Kidwell et al., 2016; Pearce, 2018; Rowhani-Farid et al., 2020). A recent randomised control trial to study the effectiveness of badges did not produce positive results, however, and better designs have been suggested (Rowhani-Farid et al., 2020). Researchers in this survey study suggested similar incentives and new types, such as a data reuse indicator (Chapter 6). Fear (2013) suggested an alternative indicator for social sciences data as well, but given disciplinary differences in how data are reused, this may need to be further tested. As suggested by this thesis and supported by previous data sharing studies, not all datasets that are shared will receive a citation or secondary reuse. Data sharing is important, however, and researchers need to be rewarded for their time and effort. Therefore, I suggest an alternative model that combines the findings of this research with previous studies and can be incorporated into a formal academic evaluation system. This was mentioned by respondents in the survey that it is of utmost importance that data sharing is being rewarded in a formal manner by linking it to performance evaluation. Badges were tested by journals only, whereas data quality checks are often performed by the repository staff. Therefore, I recommend a functionality to apply badges to a dataset that has fulfilled the usability indicators identified as important in this study. It consists of: 1. Documentation (in line with the FAIR principles) (GO FAIR, 2022), 2. Data is open, 3. Information on usability, 4. Universal format, and 5. Associated publication. A reuse indicator can include citation count as evidence of prior reuse, number of links from educational websites, and blog mentions (upon further investigation of altmetric mentions in multiple disciplines). Together, these can form a scoring system for researchers, which can be used for academic evaluation. A data authorship model would assist in the scoring process by supplementing citation counts. Furthermore, in most repositories it is not possible to discover why a user has downloaded a dataset, which often devalues download counts as an impact metric. Data repositories can include a simple checklist based on common data reuse cases (Gregory et al, 2020) when someone clicks to download the data, which can be included on the dataset page to inform other users when searching for datasets with a particular use case in mind.

The promotion of research data is still rare, as seen in altmetrics results and the researcher survey. However, more promotion about the usability of a dataset can help disseminate this information within the broader community, as seen in Twitter, Facebook, and blog posts of biodiversity datasets. Altmetric numbers should not be used as an impact metric yet, unless a post describes a specific use case or the importance of data, but if more users share about a dataset and their usability on the social web, this can indicate an 'interest score' to attract more users. A combined score, as suggested in this thesis, will reward data sharing and reflect data reuse. This can be refined and tested out in the future to make data sharing a standard practice. Findings from this study will inform funders and policy makers in this area of current trends and needs to advance this process.

7.5 Challenges and limitations

A key challenge to explore the reuse value of data published in a data repository is the lack of openly available metadata that can be accessed from repositories in an easy way. As it is found from the data repository managers' survey in Chapter 5, data reuse metrics are still difficult to capture in a meaningful way. When such data are available from data repositories, this often requires writing codes for specific APIs to retrieve these data, rather than being easily downloadable. This was the process for GBIF, which makes it complicated to study data repositories. A lack of standard data citation practices is a major challenge as identified in Chapter 3 and Chapter 4. Technical solutions to automatically capture this information are hindered by this. For example, DCI is currently only tracking data citation mentioned in references rather than performing a full-text analysis. The case study of Scholexplorer in Chapter 4 demonstrated that links between journals and datasets are captured even when a dataset was not cited in a standard manner. It is likely that these links were sourced from the DataCite, so repository staff had manually linked the publications to the dataset record.

Besides citation counts, altmetric mentions can be informative of societal impact or help users learn about the usability of a dataset. Chapter 3 results demonstrate that few datasets have altmetric mentions even though these are growing. However, at present data from the Altmetric API only provides the number of mentions, not including what was mentioned. Therefore, each mention needs to be manually accessed to collect its contextual information, which makes it time consuming. A limitation of this study is that it only explored data from Altmetric among other sources, such as PlumX Metrics from Plum Analytics.

Another major challenge is the sporadic development of data repositories as found in re3data.org and the lack of adequate metadata to enable data repository studies. Previous studies recruited participants in an ad-hoc manner through personal channels which has resulted in low response rates. This research used contact information from the re3data API for this purpose, but various types and formats of contact information, e.g., web form, general email address made it challenging and time consuming. Furthermore, how repositories manage data quality is an important issue (Downs, 2021). An optional field to indicate data quality management was filled by only 0.6% repositories in re3data. The survey study in Chapter 5 addressed this in a limited manner by asking how repositories manage data quality (automated checks, librarian or information professional support, no curation, other). A key limitation of the researcher survey is that it did not investigate why researchers did not share data, which could be informative of any cultural or technological barriers. Additionally, this study mainly focused on data reuse cases within research, which could be expanded further. Even though geographic location was not considered in both surveys, differences in languages and research cultures may have impacted the responses received in these surveys. As discussed in the literature review chapter, very few studies have explored incentives to increase data sharing (Rowhani-Farid, 2017). Different incentives were recommended by researchers in the survey as well. However, this was not in the scope of this study and outlined in the future directions.

7.6 Directions for future research

While this research suggests that open biodiversity data encourages advances in science by aiding data reuse, it did not examine individual sources of datasets or data creators since GBIF aggregates datasets from various repositories. However, previous studies expressed concerns about the loss of small datasets created by researchers in small projects as a large proportion of available datasets in GBIF was created from major incentives (Costello et al., 2013; Hampton et al., 2013). Even though results from the researcher survey in Chapter 6 demonstrated a higher data sharing rate in Ecology (63%), further exploration of data sources and data creators of biodiversity datasets published in GBIF, or another relevant data repository could inform whether and how this landscape has changed since 2013.

Citation counts are still considered the main data reuse metric by repository managers and researchers. Given that researchers reuse data for various purposes, it would be useful to explore for which data reuse cases researchers cite a dataset in article references. Additionally, future studies can expand on the findings from altmetric mentions of datasets by comparing datasets in different disciplines, published in different repositories. Previous studies are based on quantitative analysis rather than contextual analysis of social web mentions. Further studies will reveal new information regarding disciplinary differences in this area. In addition, despite the prevalence of data reuse for educational purposes, it is unclear how we can accurately capture educational use cases. System developers, repository managers, researchers, and policy makers can work together to find new solutions to fill this gap.

Given that researchers in two subject categories (Business and Economics, and Engineering) more frequently reused than shared data, further studies can explore the following aspects: common data sources for researchers in these disciplines, the availability of repositories to deposit their data, and awareness of researchers in terms of data sharing practices. Other relevant questions to investigate across disciplines include: Is data sharing by institutional repositories, journal-supported repositories, and personal websites more common because of a lack of disciplinary repositories? Does a weak data sharing culture in certain disciplines affect the discoverability of new data? A study of specific data reuse types and sources of these data would also add value to this understanding. Future studies should also examine the effect of incentives, such as badges, data reuse indicators, and data authorship models (Bierer et al., 2017) in multiple disciplines and journals.

The suggested new research in this area will strengthen the results from this thesis and inform different stakeholders about the needs and gaps in the current research data support systems. This will help shape the policies and guidelines, build systems with improved features, as well as shift research cultures in a positive direction to support data sharing across disciplines and promote reuse of shared data.

7.7 Data and code availability

Datasets for Chapter 3, 5 and 6 are available in the following dataset collection in figshare, DOI: https://doi.org/10.6084/m9.figshare.c.5946145. Dataset and code for Chapter 4 are available in the University of Bath Research Data Archive, DOI: https://doi.org/10.15125/BATH-00739.

References

Ali-Khan, S. E., Harris, L. W., & Gold, E. R. (2017). Point of view: motivating participation in open science by examining researcher incentives. *Elife*, 6, e29319. DOI: https://doi.org/10.7554/eLife.29319.001

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, 73(2), 146.

Altmetric (2022). What are Altmetrics? Available at: https://www.altmetric.com/about-altmetrics/what-are-altmetrics/ (accessed April 15, 2022)

Anagnostou, P., Capocasa, M., Milia, N., & Bisol, G. D. (2013). Research data sharing: Lessons from forensic genetics. *Forensic Science International: Genetics*, 7(6), e117-e119.

Antelman, K. (2004). Do open-access articles have a greater research impact?. *College & research libraries*, 65(5), 372-382.

Arregoitia, L. D. V., Cooper, N., & D'Elía, G. (2018). Good practices for sharing analysis-ready data in mammalogy and biodiversity research. *Hystrix, the Italian Journal of Mammalogy*, 29(2), 155-161.

Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are scientific data repositories coping with research data publishing?. *Data Science Journal*, 15:6, 1-24. DOI: http://doi.org/10.5334/dsj-2016-006

Ball, A., & Duke, M. (2015). How to Track the Impact of Research Data with Metrics. Retrieved July 17, 2017, from https://www.dcc.ac.uk/guidance/how-guides/track-data-impact-metrics.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, 323(5919), 1297-1298.

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2, 8-14.

Bertagnolli, M. M., Sartor, O., Chabner, B. A., Rothenberg, M. L., Khozin, S., Hugh-Jones, C., Resee, D. M., & Murphy, M. J. (2017). Advantages of a truly open-access data-sharing model. *The New England journal of medicine*, 376(12), 1178-1181.

Bierer, B. E., Crosas, M., & Pierce, H. H. (2017). Data authorship as an incentive to data sharing. *The New England journal of medicine*; 376(17), 1684-1687. DOI: https://doi.org/10.1056/NEJMsb1616595

Bishop, L. (2009). Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues*, 44(3), 255-272.

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting qualitative data reuse: A decade on. *Sage Open*, *7*(1), 2158244016685136. DOI: https://doi.org/10.1177/2158244016685136

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., Scharnhorst, A., & Golshan, M. S. (2019), “Digital data archives as knowledge infrastructures: Mediating data sharing and reuse”, *Journal of the Association for Information Science and Technology*, 70(8), 888-904.

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics*, *103*(3), 1123-1144.

Brickley, D., Burgess, M., & Noy, N. (2019, May). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference* (pp. 1365-1375).

Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M. and Schindler, U. (2017a). The data-literature interlinking service: Towards a common infrastructure for sharing data-article links. *Program: electronic library and information systems*, 51(1), 75-100. DOI: https://doi.org/10.1108/PROG-06-2016-0048

Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U., & Authr, C. (2017b). The Scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine*, *23*(1/2). DOI: https://doi.org/10.1045/january2017-burton

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. *Open Journal Systems*, 7(1). DOI: https://doi.org/10.2218/ijdc.v7i1.218

Callaghan, S. (2014). Preserving the integrity of the scientific record: data citation and linking. *Learned Publishing*, *27*(5), S15-S24.

Ceci, S. J., & Walker, E. (1983). Private archives and public needs. *American Psychologist*, 38(4), 414-423.

Ceci, S. J. (1988). Scientists' attitudes toward data sharing. *Science, Technology, & Human Values*, 13(1-2), 45-52.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, 12(15), 1-12.

Chamanara, J., Kraft, A., Auer, S., & Koepler, O. (2019). Towards Semantic Integration of Federated Research Data. *Datenbank Spektrum*, 19, 87–94. DOI: https://doi.org/10.1007/s13222-019-00315-w

Chen, X., & Wu, M. (2017). Survey on the needs for chemistry research data management and sharing. *The Journal of Academic Librarianship*, *43*(4), 346-353.

Coady, S. A., Mensah, G. A., Wagner, E. L., Goldfarb, M. E., Hitchcock, D. M., & Giffen, C. A. (2017). Use of the national heart, lung, and blood institute data repository. New England Journal of Medicine, 376(19), 1849-1858.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PloS one*, 15(4), e0230416.

Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013). The value of research data–Metrics for datasets from a cultural and technical point of view. *A Knowledge Exchange Report*, Leiden. Available from https://repository.jisc.ac.uk/6205/1/Value_of_Research_Data.pdf

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003-2019. https://arxiv.org/pdf/1401.4321.pdf

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, 28(8), 454-461.

Costello, M. J., & Wieczorek, J. (2014). Best practice for biodiversity data management and publication. *Biological Conservation*, 173, 68-73. DOI: https://doi.org/10.1016/j.biocon.2013.10.018

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, 18(1), 1-7. DOI: http://doi.org/10.5334/dsj-2019-009

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, 68(9), 2182-2200.

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.

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists’ data reuse. *PloS one*, 12(12), e0189288.

Drachen, T., Ellegaard, O., Larsen, A., & Dorch, S. (2016). Sharing data increases citations. *Liber Quarterly*, *26*(2).

Downs, R. R. (2021). Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories. *Data Science Journal*, *20*(1).

Edmunds, S. C., Pollard, T. J., Hole, B., & Basford, A. T. (2012). Adventures in data citation: sorghum genome data exemplifies the new gold standard. *BMC research notes*, 5(1), 1-5.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data—. Ecological Informatics, 11, 25-33.

Escribano, N., Galicia, D., & Ariño, A. H. (2019). Completeness of Digital Accessible Knowledge (DAK) about terrestrial mammals in the Iberian Peninsula. *PloS one*, *14*(3), e0213542.

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, 67(6), 1404-1416.

Faniel, I. M., & Yakel, E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. *Curating research data, volume one: Practical strategies for your digital repository*, 1, 103-126.

Fear, K. M. (2013). Measuring and Anticipating the Impact of Data Reuse. [Thesis]. http://hdl.handle.net/2027.42/102481

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing?. *PloS one*, 10(2), e0118053.

Federer, L. M., Lu, Y. L., Joubert, D. J., Welsh, J., & Brandys, B. (2015). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PloS one*, 10(6), e0129506.

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y. L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: an analysis of data availability statements. *PloS one*, 13(5), e0194768.

Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M. and Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, 6(1), 1-9.

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nature neuroscience*, 17(11), 1442-1447. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728080/

Figueiredo, A. S. (2017). Data sharing: convert challenges into opportunities. *Frontiers in public health*, 5, 327, 1-6. DOI: https://doi.org/10.3389/fpubh.2017.00327

Fink, A. (2003), *How to design survey studies*, Thousand Oaks, USA: Sage Publications.

Gazra, K. & Fenner, M. (2018). Glad You Asked: A Snapshot of the Current State of Data Citation [Blog post]. Retrieved from https://doi.org/10.5438/h16y-3d72

Ghavimi, B., Mayr, P., Vahdati, S., & Lange, C. (2016). Identifying and improving dataset references in social sciences full texts. *arXiv preprint arXiv:1603.01774*.

Gherghina, S., & Katsanidou, A. (2013). Data availability in political science journals. *European Political Science*, 12, 333-349.

Gibson, C. (2019). From Couch to Almost 5K: Raising Research Data Visibility at The University of Manchester [Blog post]. Retrieved from: https://blog.research-plus.library.manchester.ac.uk/2019/02/

GO FAIR (2022). FAIR Principles. Available at: https://www.go-fair.org/fair-principles/ (accessed April 15, 2022)

Goldstein, S. (2017), “The Evolving Landscape Of Federated Research Data Infrastructures”, available at: https://www.rd-alliance.org/sites/default/files/attachment/The_Evolving_Landscape_of_Federated_Research_Data_Infrastructures.pdf

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*, *2*(2).

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3), 156-162.

Hansson, K., & Dahlgren, A. (2021). Open research data repositories: Practices, norms, and metadata for sharing images. *Journal of the Association for Information Science and Technology*. 73(2), 303-316.

Henneken, E. A., & Accomazzi, A. (2011). Linking to data-effect on citation rates in astronomy. *arXiv preprint arXiv:1111.3618*.

Henneken, E. (2015). Unlocking and sharing data in astronomy. *Bulletin of the Association for Information Science and Technology*, 41(4), 40-43.

Hersh, G. (2017). Making Open Access/Open Data/Open Science A Reality. *Against the Grain*, 29(3), 43.

Holdren, J. (2013). Increasing Access to the Results of Federally Funded Scientific Research. Office of Science and Technology Policy. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Hripcsak, G., & Heitjan, D. F. (2002). Measuring agreement in medical informatics reliability studies. *Journal of biomedical informatics*, 35(2), 99-110.

Hrynaszkiewicz, I. (2019). Publishers' responsibilities in promoting data quality and reproducibility. In *Good Research Practice in Non-Clinical Pharmacology and Biomedicine* (pp. 319-348). Springer, Cham.

Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021). A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal*, 20(1). DOI: http://doi.org/10.5334/dsj-2021-031

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. *Conservation Letters*, 5(5), 399-406.

ICPSR (2021). Timeline. Available at: https://www.icpsr.umich.edu/web/pages/about/history/timeline.html (accessed 23 September 2021)

Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics*, *12*(15), 1-10.

Ivanović, D., Schmidt, B., Grim, R., Dunning, A. (2019). FAIRness of Repositories & Their Data: A Report from LIBER's Research Data Management Working Group. Available at https:// doi.org/10.5281/zenodo.3251593

Joo, S., Kim, S., & Kim, Y. (2017). An exploratory study of health scientists' data reuse behaviors: Examining attitudinal, social, and resource factors. *Aslib Journal of Information Management*, 69(4), 389-407.

Khan, N., & Thelwall, M. (2019a). Dataset supporting "Data Citation and Reuse Practice in Biodiversity". Figshare. Dataset. https://doi.org/10.6084/m9.figshare.8181098.v1

Khan, N., & Thelwall, M. (2019b). Dataset supporting "Measuring the Impact of Biodiversity Datasets: Data Reuse, Citations and Altmetrics". Figshare. Dataset. https://doi.org/10.6084/m9.figshare.11357693

Khan, N. (2020). Dataset for "Linking Datasets and Articles – Potentials and Challenges of Scholix Framework". Bath: University of Bath Research Data Archive. https://doi.org/10.15125/BATH-00739.

Khan, N., Thelwall, M., and Kousha, K. (2021a). Dataset supporting "Are data repositories fettered? A survey of current practices, challenges and future technologies". Figshare. Dataset. https://doi.org/10.6084/m9.figshare.14191739.v2

Khan, N., Thelwall, M., and Kousha, K. (2022). Survey data on disciplinary differences in data sharing and reuse practices. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.19596967.v1

Khan, N. (2022). NK PhD Thesis datasets. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.5946145.v1

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S. & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS biology*, *14*(5), e1002456.

Kiley, R., Peatfield, T., Hansen, J., & Reddington, F. (2017). Data sharing from clinical trials—a research funder's perspective. *The New England Journal of Medicine*, 377(20), 1990-1992.

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library & Information Science Research*, 37(3), 189-200.

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, 67(4), 776-799.

Kim, Y., & Yoon, A. (2017), "Scientists' data reuse behaviors: A multilevel analysis", *Journal of the Association for Information Science and Technology*, 68(12), 2709-2719.

Konkiel, S. (2013). Tracking citations and altmetrics for research data: Challenges and opportunities. *Bulletin of the American Society for Information Science and Technology*, *39*(6), 27-32.

Konkiel, S. (2020), "Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators", *Scholarly Assessment Reports*, 2(1).

Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, 3.

Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific data*, 2(1), 1-5.

Kratz, J. E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLoS One*, 10(2), e0117619.

Lafia, S., Ko, J. W., Moss, E., Kim, J., Thomer, A., & Hemphill, L. (2021). Detecting Informal Data References in Academic Literature. Preprint. DOI: https://dx.doi.org/10.7302/1671

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*, *6*(2), 4-37.

Li, R., von Isenburg, M., Levenstein, M., Neumann, S., Wood, J., & Sim, I. (2021). COVID-19 trials: declarations of data sharing intentions at trial registration and at publication. Trials, 22(1), 1-5.

Limani, F., Latif, A., & Tochtermann, K. (2018). Linked Publications and Research Data: Use Cases for Digital Libraries. In *International Conference on Theory and Practice of Digital Libraries* (pp. 363-367). Springer, Cham.

Luther, J. (2018). The Evolving Institutional Repository Landscape. ACRL/Choice. Available at: http://choice360.org/ librarianship/whitepaper

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in ecology & evolution*, 25(10), 574-582.

Mathiak, B., & Boland, K. (2015). Challenges in matching dataset citation strings to datasets in social science. *D-Lib Magazine*, *21*(1/2), 23-28.

Mayernik, M. S. (2013). Bridging data lifecycles: Tracking data use via data citations workshop report. NCAR Library.

Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. *International Journal of Digital Curation*, 11(1), 150-155.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276-282.

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017). Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics*, *26*(1), 38.

Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., Cockerill, M., & Chavan, V. (2011). Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics*, *12*(15), 1-10.

Mozersky, J., Walsh, H., Parsons, M., McIntosh, T., Baldwin, K., & DuBois, J. M. (2020). Are we ready to share qualitative research data? Knowledge and preparedness among qualitative researchers, IRB Members, and data repository curators. *IASSIST quarterly*, *43*(4).

Noesgaard, D. (2019). Improving Impact Metrics of Open and Free Biodiversity Data through Linked Metadata and Academic Outreach. Biodiversity Information Science and Standards. DOI: https://doi.org/10.3897/biss.3.35723

Office for Civil Rights (2012), “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule”, available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.J., Gundlach, J., Schirmbacher, P. and Dierolf, U. (2013). Making research data repositories visible: the re3data. org registry. *PloS One*, 8(11), e78080.

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, 111(1), 443-461.

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, 16(8).

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review*, 1(2).

Patel, D. (2019). How Google's Dataset Search Engine Work. Available at: https://towardsdatascience.com/how-googles-dataset-search-engine-work-928fa5237787 (accessed 31 March 2021).

Pearce, R. (2018). Springer Nature launches Open data badges pilot. Available at: https://blogs.biomedcentral.com/bmcblog/2018/10/08/springer-nature-launches-open-data-badges-pilot/ (accessed 20 September 2021).

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2015). Research data explored: Citations versus altmetrics. *arXiv preprint arXiv:1501.03342*.

Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS One*, 9(12), e114734.

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS one*, 6(7), e18657.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3), e308. 0000308

Piwowar, H. (2013). Value all research products. *Nature*, 493(7431), 159-159. Available at: http://eprints.icrisat.ac.in/12069/1/value-all-research-products.pdf.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175.

Prieto, A. G. (2009). From conceptual to perceptual reality: trust in digital repositories. *Library Review*, 58(8), 593-606. DOI: https://doi.org/10.1108/00242530910987082

REF (2019), "Guidance on submissions (2019/01) – REF 2021", available at: https://www.ref.ac.uk/publications/guidance-on-submissions-201901/ (accessed 13 July 2021).

re3data.org - Registry of Research Data Repositories. Available at: https://doi.org/10.17616/R3D (accessed 17 November 2020).

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, 67(12), 2964-2975.

ROR: Research Organization Registry (2020). Retrieved from https://ror.org/about/. [Accessed 23 Jan. 2020]

Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017). What incentives increase data sharing in health and medical research? A systematic review. *Research integrity and peer review*, *2*(1), 1-10.

Rowhani-Farid, A., Aldcroft, A., & Barnett, A. G. (2020). Did awarding badges increase data sharing in BMJ Open? A randomized controlled trial. *Royal Society open science*, 7(3), 191818.

Sands, A., Borgman, C. L., Wynholds, L., & Traweek, S. (2012). Follow the data: How astronomers use and reuse data. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1-3.

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

Scholix (2019). Scholix: A Framework for Scholarly Link Exchange. Retrieved from http://www.scholix.org/home. [Accessed 18 Nov. 2019].

Shearer, K., & Furtado, F. (2017), "COAR Survey of Research Data Management: Results", available at: https://www.coar-repositories.org/files/COAR-RDM-Survey-Jan-2017.pdf

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. *Journal of the Association for Information Science and Technology*, *65*(5), 1018–1027.

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1), 6-20.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, *19*(3), 387-420.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R., Duerr, R., Haak, L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J., Lin, J., Nielsen, L., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M. and Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1, e1. DOI: https://doi.org/10.7717/peerj-cs.1

Syrotiuk, N. (2019). sefnyn/scholix. Retrieved from https://github.com/sefnyn/scholix. [Accessed 18 Nov. 2019].

Tay, A. (2018). How does Scopus find and link to related research data? Or an attempt to understand how to link datasets to articles via Scholix [Blog post]. Retrieved from http://musingsaboutlibrarianship.blogspot.com/2018/10/how-does-scopus-find-and-link-to.html

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one*, 10(8), e0134826.

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R. & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one*, 15(3), e0229003.

The FREYA Project: The PID Graph. (2019). Retrieved from https://www.project-freya.eu/en/pid-graph/the-pid-graph [Accessed 18 Jan. 2020]

The OpenAIRE Scholexplorer: The Data Literature Interlinking Service (2020). Retrieved from https://scholexplorer.openaire.eu/#/. [Accessed 17 Jan. 2020].

The R Project for Statistical Computing (2019). Retrieved from https://www.r-project.org/. [Accessed 18 Nov. 2019].

Thelwall, M., Munafò, M., Mas-Bleda, A., Stuart, E., Makita, M., Weigert, V., Keene, C., Khan, N., Drax, K. and Kousha, K. (2020), "Is useful research data usually shared? An investigation of genome-wide association study summary statistics", *Plos One*, 15(2), e0229578. https://doi.org/10.1371/journal.pone.0229578

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS one*, 8(5), e64841. DOI: https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., & Kousha, K. (2017). Do journal data sharing mandates work? Life sciences evidence from Dryad. *Aslib Journal of Information Managemen*t, 69(1), 36-45. http://wlv.openrepository.com/wlv/bitstream/2436/620330/1/dryad_preprint.pdf

Troudet, J., Vignes-Lebbe, R., Grandcolas, P., & Legendre, F. (2018). The increasing disconnection of primary biodiversity data from specimens: how does it happen and how to handle it?. Systematic Biology, 67(6), 1110-1119.

Uhlir, P. F. (2012). *For Attribution--: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: National Academies Press.

Unal, Y., Chowdhury, G., Kurbanoglu, S., Boustany, J., & Walton, G. (2019). Research data management and data sharing behaviour of university researchers. Available at: http://hdl.handle.net/11655/23736

University of Bath Research Data Archive. (2019) https://researchdata.bath.ac.uk/. [Accessed 18 Nov. 2019].

Van Dalen, H. P., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. Journal of the American Society for Information Science and Technology, 63(7), 1282-1293.

Velden, T. (2013). Explaining field differences in openness and sharing in scientific communities. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 445-458).

Vertesi, J., & Dourish, P. (2011, March). The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 533-542).

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current biology*, 24(1), 94-97. http://www.sciencedirect.com/science/article/pii/S0960982213014000.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. PloS one, 7(1), e29715.

Wiley, C. (2018). Data sharing and engineering faculty: An analysis of selected publications. *Science & technology libraries,* 37(4), 409-419.

Wallis, J. C., Rolando, E., & Borgman, C. L. (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.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9.

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012, June). Data, data use, and scientific inquiry: Two case studies of data practices. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 19-22).

Xafis, V., & Labude, M. K. (2019). Openness in big data and data repositories. *Asian Bioethics Review*, *11*(3), 255-273.

Xafis, V., Schaefer, G. O., Labude, M. K., Brassington, I., Ballantyne, A., Lim, H. Y., ... & Tai, E. S. (2019). An ethics framework for big data in health and research. *Asian Bioethics Review*, *11*(3), 227-254.

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology*, 53(1), 1-6.

Yoon, J., Chung, E., Lee, J. Y., & Kim, J. (2019). How research data is cited in scholarly literature: A case study of HINTS. *Learned Publishing*, 32(3), 199-206.

Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., & Balaban, E. (2018). Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal of documentation*, 74(5), 1053-1073. DOI: https://doi.org/10.1108/JD-09-2017-0126

Zhao, M., Yan, E., & Li, K. (2018). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, *69*(1), 32-46.

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, technology, & human values, 33*(5), 631-652.

Appendices

Appendix A

Questionnaire for repository managers

Section 1: Data sharing and reuse

Q1. What type of data do you collect and support?

(a) Discipline specific data (please specify), (b) Any data produced by the institutional researchers, (c) Specific format of data (please specify), (d) Data from any discipline

Q2. (i) Which technical framework do you use for your repository?

(a) Dspace, (b) Eprint, (c) Fedora, (d) Others (please specify)

Q2. (ii) Why did you choose this particular framework?

Q3. What kind of persistent identifiers do you support in your repository?

(a) DOI, (b) Handle, (c) others, (d) None supported

Q4. (i) Do you use any of these services for data support in your repository?

(a) DataCite, (b) Data Citation Index

Q4. (ii) If yes, what are the advantages of using the(se) service(s)?

Q5. (i) Which of these metrics do you track and expose?

(a) Downloads, (b) Citations to a dataset, (c) Views, (d) Citations to your repository as a whole, (e) None.

Q5. (ii) If you are not currently tracking and/or exposing any of these metrics, please explain the main reason why not.

Q6. Would the following give you useful information about the impact of your data?

(a) Citations/ papers published with it, (b) Downloads, (c) Links to the data from other websites, i.e. educational use, Wikipedia, (d) Landing page views

Section 2: About Research Data Support

Q7. What do(es) Research Data Support role(s) look(s) like in your institution?
(a) Single-person support, (b) Small department consisting of 2-3 people, (c) Larger department with more than 3 people, (d) No structured support provided, (e) Other

Q8. How do you maintain dataset and metadata quality in your repository? (a) Librarian/ information professional support, (b) Depend on the researchers for quality check, (c) Other

Q9. How do you motivate your researchers to deposit data?

(a) Outreach, (b) Funders' policy, (c) Teaching/ training workshops, (d) Others

Q10. Currently, what are your main priorities in providing research data management support?

Q11. What are the most challenging factors in supporting research data in your repository?

Q12. Which type of tools or research data support system do you envision for future?

Appendix B: Survey invitation letter template for the repository survey

Subject: Invitation to complete a survey on the broader impact of research data sharing

Dear [Recipient's name or Repository Manager],

I am reaching out to you regarding a survey I am conducting as a part of my doctoral research. The aim of this study is to explore the impact of research data sharing and the barriers to track and expose metrics via repositories.

The survey consists of 13 questions and should not take more 10 minutes to complete. We would greatly appreciate if you could take the time to complete the survey.

[Personalized survey link if sent through the survey platform or public survey URL to contact via web form]

If you have any concerns or wish to delete your data from our system at any point during the study, please contact us with the unique receipt number you receive at the end of the survey. To store the receipt number for future reference, please use either the 'Print' or 'Email' option.

Thank you in advance for your help with this study.

Nushrat

Nushrat Khan
Doctoral Researcher,
Statistical Cybermetrics Research Group, University of Wolverhampton.
Research Data Librarian, Library, University of Bath.
Email-address: n.j.khan@wlv.ac.uk/ n.j.khan@bath.ac.uk

Appendix C

Questionnaire for researchers

Data production

Q1: Please indicate your research experience in terms of years (A research career would normally start when a PhD starts).
0 – 0-3 years
1 – 3-6 years
2 – 6-9 years
3 – 10+ years

Q2: Please select your main current research area from the options below. If your research area is not included in the list, then please include it under ‘Others’.

1. Physical Sciences – a. Astronomy and Astrophysics, b. Organic Chemistry
2. Biomedical Sciences – a. Neurology, b. Pharmacology
3. Social Sciences – a. Linguistics and Language, b. Education, c. Library and Information Science
4. Arts and Humanities – a. Visual and Performing Arts, b. Literature and Literary Theory
5. Earth and Planetary Sciences – a. Geology, b. Oceanography
6. Engineering – a. Aerospace Engineering, b. Biomedical Engineering, c. Environmental Engineering
7. Environmental Science – a. Ecology, b. Pollution
8. Medicine – a. Radiology, Nuclear Medicine and Imaging, b. Infectious Diseases
9. Business and Economics – a. Business and International Management, b. Economics and Econometrics
10. Other (Please specify)

Q3: What type of data do you produce in your research? [Please give specific examples, e.g., survey data, type of samples/ observations]

Q4: What are the most important formats of data that you produce in your research? [Select all that apply]

1. Text, 2. Images, 3. Multimedia (Audio and/or Video), 4. Software/ code, 5. Numerical data (Any type of quantitative measurements), 6. Others (specify) 7. None/ Not sure

<u>Data sharing</u>

Q5. Have you ever shared your data in a repository? [If no then skip to data reuse question 7]

1. Yes, 2. No

Q5 (If yes): How did you first find a repository to share your data?

1. I was already aware of the popular/ relevant repositories in my field
2. Searched re3data.org (Registry of Research Data Repositories)
3. Web search
4. Consulted with colleagues or senior researchers
5. Consulted with the experts in my institution, e.g. Research data support services
6. Others (Please specify)

Q6: How do you usually share your data? [‘Please specify’ will be included in all except website]

- a. Institutional repository (e.g., university repository)
- b. Discipline-specific repository (e.g., Inter-university Consortium for Political and Social Research (ICPSR), PANGAEA)
- c. Interdisciplinary repository (e.g., Zenodo, UCLA Center for Embedded Networked Sensing (CENS))
- d. A journal supported repository (e.g., PLOS ONE)
- e. Personal website.
- f. Other
- g. I prefer not to share data.
- h. I don’t know/ Not sure

Q7: Which of these factors influence your choice of repositories to share your data from? [select all that apply]

1. Discipline norms, 2. Cost, 3. Ease of use, 4. Reputation of the repository, 5. Appropriateness for data type, 6. Data curation services offered, 7. Other factors [specify], 8. None of the above

<u>Data Reuse</u>

Q8: Have you ever reused existing datasets created by other people in your research?

1. Yes [Please select the best option that applies]
    a. I use my own primary data but sometimes combine it with data from existing data sources
    b. I never use my own primary data but only ever use data from existing sources (e.g. datasets published in repositories) to answer new research questions
2. I only ever use my own primary data for my research
3. My research doesn't use data
4. I don't know/ Not sure

(Those who answer 'Yes' proceed to next questions. Skip to question 12 for option 2, 3,4)

Q9: How do you find datasets to reuse? [check all that apply]

1. Search disciplinary repositories
2. Search interdisciplinary repositories
3. Web search (e.g., Google Dataset Search)
4. Read relevant papers and then check if the authors shared data
5. By accident – I noticed the dataset (e.g., in the original paper) and decided to use it.
6. Other [Specify]
7. I don't know/can't remember

Q10: For which purposes do you reuse existing data? [Specify use case and type of data] [check all that apply]

1. Ground truthing: calibrate, compare, confirm (Comparative reuse) [Specify]
2. Analyse a single existing dataset to answer novel research questions (Integrative reuse) [Specify]
3. Combine multiple existing datasets to answer novel research questions (Integrative reuse) [Specify]
4. Other [Specify]
5. I don't know/can't remember

<u>Measuring data reuse</u>

Q11: Would you like to know whether someone else has reused your published data?

1. Yes, 2. No, 3. Not sure

Q12: Do you ever actively promote your published datasets?

1. Yes, 2. No 3. Not sure

Q12 (If Yes): How do you promote your datasets?

1. In classrooms
2. Using social media platforms – (i) Twitter, (ii) Facebook, (iii) Blog posts
3. Promote within research groups and collaborators' channels
4. Other (specify)

Incentive

We are investigating the type of incentives that can improve the search experience and usage of research data. The following questions identify the factors that may assist in such decision-making process.

Q13: When searching for existing datasets in a repository, which of the following factors you consider important for the decision to use one? [check all that apply]

1. Proper documentation for the dataset
    a. Type of data
    b. Subject of data
    c. Data collection method
2. The data is open (no application procedure)
3. Information on the usability of the data
4. Evidence that the data is from an associated publication
5. The data is in a universal standard format
6. Evidence that the data has been reused before
7. Others (Please specify)

Q14: How easy is it for you usually to find relevant datasets for reuse? [Likert scale]

1. Extremely, 2. Very, 3. Neutral, 4. Difficult, 5. Very difficult or often impossible, 6. I don't know/does not apply

Q15: What can be improved in current systems to encourage and promote data reuse? (open-ended)

Appendix D: Personalised survey invitation letter template for the researcher survey

Subject: Invitation to participate in a survey on disciplinary differences in data sharing and reuse

Dear [Researcher's name],

This is Nushrat Khan. I am currently conducting a survey as a part of my doctoral research. We are exploring the disciplinary differences in data sharing and reuse practices and effect of incentives to promote data reuse.

The survey consists of 15 questions and should not take more than 10 minutes to complete. We would greatly appreciate if you could take the time to complete the survey.

[Personalized URL sent through the survey platform]

If you have any concerns or wish to delete your data from our system at any point during the study, please contact us with the unique receipt number you receive at the end of the survey. To store the receipt number for future reference, please use either the 'Print' or 'Email' option.

Thank you in advance for your help with this study.

Nushrat

--
Nushrat Khan
Doctoral Researcher, Statistical Cybermetrics Research Group, University of Wolverhampton.
Research Data Librarian, University of Bath, UK.
Email-address: [e-mail addresses redacted].

Appendix E: Format of data produced by subject category

Subject category	Text	Images	Multimedia	Software/ code	Numerical data
Social Sciences (*n*=733)	540 (73.7%)	142 (19.4%)	142 (19.4%)	93 (13%)	524 (71.5%)
Arts and Humanities (*n*=334)	**295 (88.3%)**	116 (34.7%)	**70 (21%)**	22 (7%)	83 (25%)
Business and Economics (*n*=592)	246 (41.6%)	69 (12%)	20 (3%)	87 (15%)	508 (85.8%)
Physical Sciences (*n*=220)	84 (38%)	112 (50.9%)	14 (6%)	**99 (45%)**	186 (84.6%)
Biomedical Sciences (*n*=264)	113 (42.8%)	**146 (55.3%)**	18 (7%)	58 (22%)	206 (78%)
Medicine (*n*=170)	81 (48%)	61 (36%)	13 (8%)	29 (17%)	144 (84.7%)
Environmental Sciences (*n*=324)	122 (37.7%)	135 (41.7%)	15 (5%)	58 (18%)	**281 (86.7%)**
Earth and Planetary Sciences (*n*=188)	68 (36%)	85 (45%)	3 (2%)	67 (36%)	161 (85.6%)
Engineering	82 (48%)	61 (36%)	13 (8%)	29 (17%)	144 (84.7%)

(*n*=179)					