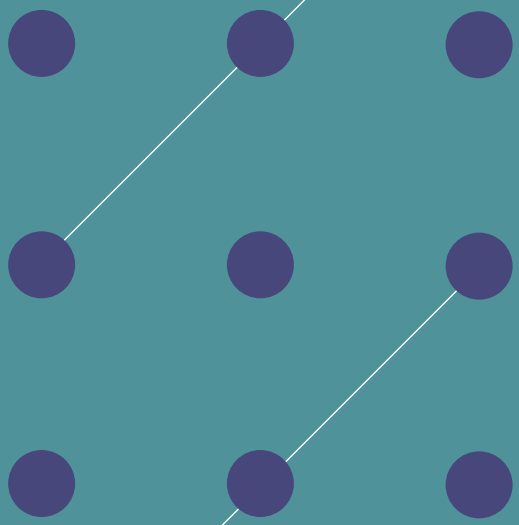
OPEN DATA

THE RESEARCHER PERSPECTIVE



CWTS Meaningful metrics

ELSEVIER

Universiteit Leiden

Preface

A year ago, in April 2016, Leiden University's Centre for Science and Technology Studies (CWTS) and Elsevier embarked on a project to investigate open data practices at the workbench in academic research. Knowledge knows no borders, so to understand open data practices comprehensively the project has been framed from the outset as a global study. That said, both the European Union and the Dutch government have formulated the transformation of the scientific system into an open innovation system as a formal policy goal. At the time we started the project, the Amsterdam Call for Action on Open Science had just been published under the Dutch presidency of the Council of the European Union. However, how are policy initiatives for open science related to the day-to-day practices of researchers and scholars?

With this report, we aim to contribute to bridging the gap between policy on the one hand, and daily research practices from a global perspective on the other hand. As we show, open data practices are less developed than anticipated, with the exception of fields where data practices are integrated in the research design from the very beginning. While policy has high expectations about open science and open data, the motive force comes not from the policy aims, but in changing practice at the grass roots level. This requires we confront the harsh reality that the rewards for researchers and scholars to make data available are few, and the complexity in doing so is high.

This report is produced in close collaboration between CWTS and Elsevier. Elsevier and CWTS have been long-time partners, and both partners are able to draw on deep knowledge of - and networks in - the world of research. This project was developed as a research project, and it benefits from a well-designed public-private partnership. The project team has enjoyed in-depth discussions on matters at the very heart of open data and data sharing, bringing together a team that was built on closely working together in data collection, analysis and writing the report.

Now we are ready to share our insights for policy leaders, researchers, funders and publishers alike, bringing the message that at the interface of policy and practice more efforts are needed to make open data a responsible research and innovation action.

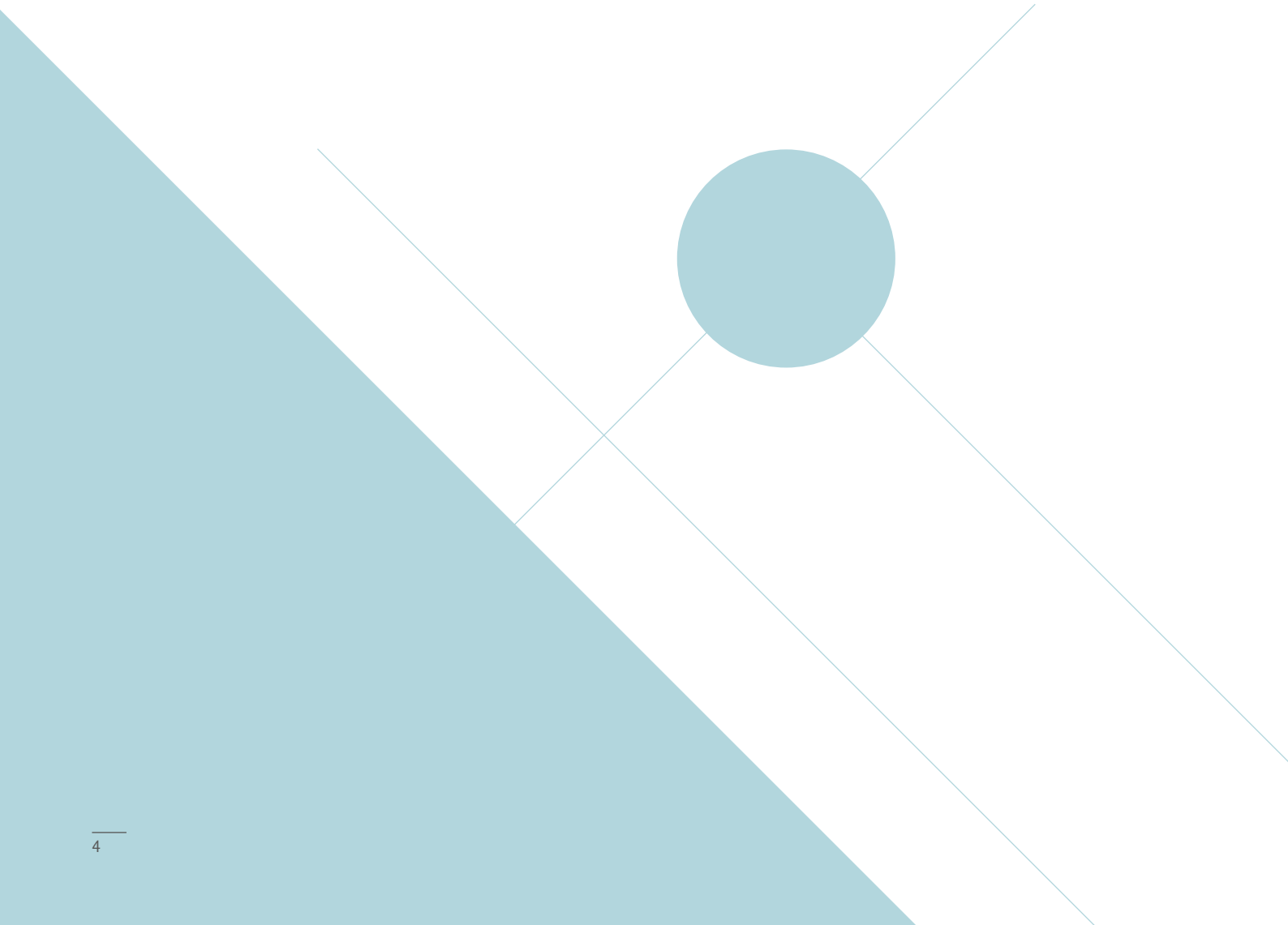
Paul Wouters
Professor of Scientometrics,
Director of CWTS,
Leiden University

Wouter Haak
Vice President,
Research Data Management,
Elsevier

Executive Summary

Open data practices facilitate collaboration, drive data analysis, and promote transparency and reproducibility. Yet the research community has not uniformly embraced open data or data sharing practices. This report describes the findings of a complementary methods approach to examine the practices, motivations, and obstacles to data sharing as well as perceived advantages among researchers across disciplines worldwide. Combining information from a bibliometric analysis, a survey and case studies, this report examines how researchers share data, the attitudes of researchers toward sharing data, and why researchers might be reticent to share data.

> *"Our study suggests that the concept of open data speaks directly to basic questions of ownership, responsibility, and control."*



Data-sharing practices depend on the field: there is no general approach

For fields in which data sharing is integral to the research being done, the incentive to follow open data practices is embedded into the research design and execution. Researchers in these fields are often members of collaborative groups that have mechanisms in place for sharing data with their colleagues throughout the research process, such as data repositories or cloud-based archives. This is illustrated in three case studies of open data practices in Soil Science, Human Genetics, and the emerging field of Digital Humanities. In other fields, where transfer of data amongst collaborators is less essential for data analysis or interpretation, open data practices are less uniform and, in some cases, may be absent. Data remains proximal to the researcher, in personal, departmental, or institutional archives. In these fields, data sharing is something that takes place independent of the research itself, for example, through publication after the research has been completed.

Researchers acknowledge the benefits of open data, but data sharing practices are still limited

Attitudes towards data sharing are generally positive, but open data is not yet a reality for most researchers. A global online survey of 1,200 researchers found that many perceive data as personally owned. Public data sharing primarily occurs through the current publishing system; less than 15% of researchers share data in a data repository. The survey also revealed that when researchers share their data directly, most (>80%) share with direct collaborators. This type of collaborative sharing is mainly direct (i.e., person-to-person), suggesting that trust is an important aspect of sharing data. Collaborative research is a common driver of data sharing in all fields. Our study suggests that the concept of open data speaks directly to basic questions of ownership, responsibility, and control.

Executive Summary continued

Barriers to sharing slow the uptake of open data practices

The survey also found that while most researchers recognize the benefits of sharing unpublished research data, fewer are willing to share data or have shared data. This might be because there is a lack of training in data sharing and because sharing data is not associated with credit or reward. Research data management and privacy issues, proprietary aspects, and ethics are barriers common to all fields. In intensive data-sharing fields, the reticence to sharing data depends on ethical and cultural limitations and boundaries. Financial and legal issues could also hamper sharing. Research data management plans mandated by funders (or publishers) are not considered to be a strong incentive.

Analysis of publication in data journals reveals scattered practices

A lack of consistency in referring to datasets makes it difficult to analyze data sharing through a quantitative analysis. Therefore, we analyzed publication in and citation of data journals—journals dedicated to publishing research data. A quantitative analysis of data journals found that while the number of data journals is still limited, they play an increasingly important role in terms of the number of articles they publish and the citations they receive. However, our survey shows that data sharing still occurs more often in traditional ways, such as through publication or presentation of data aggregated into tables and annexes, or data is not published at all (34%).

How can open data be seen as a responsible and rewarding practice in research?

Although data sharing seems to have a global benefit, cultural and national factors pose a significant challenge to a one-size-fits-all approach. Regardless of the benefits, deciding what data can be shared, how it should be shared, and making it usable by others requires additional effort, training, and resources. Furthermore, freeing up data for reuse and sharing depends on accommodation or coordination of disciplinary, cultural, and local differences with respect to data privacy and licensing.

Open data mandates from funders or publishers are only a starting point when it comes to sharing research data. Policies that incentivize the use of open data practices are needed, as are formal training programs on data sharing, management, and reuse. Departments and institutions can highlight the benefits of open data to the research enterprise, encourage publication of research data, and provide tools and guidance to support data sharing. To this end, solutions and tools should not be seen as storage tools, but as working tools that provide an environment that fits into the researcher workflow and makes it possible to directly and rapidly reuse data.

Bridging the gap will require both researchers and policymakers

In the future of open data, there are many stakeholders involved including but not limited to the research communities, funding bodies, publishers and research institutions. Researchers feel they are at heart of the practice of sharing and re-use of data. Therefore, open data development would benefit from taking a bottom-up approach. A change in the scientific culture is needed, where researchers are stimulated and rewarded for sharing data and where institutions implement and support research data sharing policies, including mandates. Given that open data guidelines and standards have been developed, all actors should now try to bridge the gap between policy and practice and ensure researchers are in a position to implement them. While open data mandates provide an initial set of instructions, guidance should be given on implementation and sharing should be incentivized

Table of Contents

01: Introduction

1. Introduction

Across all fields, researchers and knowledge users are increasingly aware of the need for more efficient data access and sharing (Borgman, 2012). Important policy efforts are invested in promoting data sharing across a wide front, as stated in numerous declarations (e.g., The Hague Declaration on Knowledge Discovery in the Digital Age, the Brussels declaration, and the Joint Declaration of Data Citation Principles).

Open data can be defined as "... data that can be freely used, re-used and redistributed by anyone" (Open Data Handbook, Open Knowledge Foundation) and can be accessed on equal terms by the international research community at the lowest possible cost (OECD Principles and Guidelines for Access to Research Data). Furthermore, the openness of data applies to all components of the research process, not just to research outcomes. Open data needs to be embedded in the research process from start to finish. Such changes will likely impact the entire research cycle and its organization, from the inception of research to its publication. In the research system as a whole, this shift toward open data may also result in the rise of new disciplines, alternative ways of evaluating the quality and impact of research, new pathways in publishing, and different scientific reputation systems.

Open data suitable for data sharing applies to data collection, data curation (e.g., metadata, identifiers), and data dissemination (e.g., searchable archives). Not much is known about how data is used and reused. It is also important to consider what constitutes data (raw, processed, summarized, or aggregated) and to distinguish between machine-readable data (the focus of linked open data initiatives) and human-readable data. Furthermore, we need to make the distinction between "big science" and "little science," as the data that results from the former (e.g., climate data, genomic data such as from the Human Genome Project, large-scale clinical trial data, data from publicly funded scientific infrastructure projects such as the Hubble Space Telescope and the Large Hadron Collider, etc.) may be considered a public good and may be subject to different legal and moral sharing obligations than the data resulting from relatively small-scale studies conducted in laboratories and field sites around the world.

Nowadays, scholars can conceptualize, collect data, analyze these data, and write and publish their results openly. Hence, the question arises: at what point in the research workflow, and to what extent, is open data already a part of current academic practice? It is important to understand how the research community views the opportunities and challenges of open data. A general challenge related to open data is the relatively low commitment to data sharing among scholars, which is related to both their own perceptions and the cultural environment in which they work (Tenopir et al., 2011, Costas et al., 2013). Open data practices are also not evenly spread across the various academic disciplines.

What elements factor into the willingness or capacity to incorporate open data practices among researchers? In this study, we investigate whether and to what extent open data sharing is practiced in academic life. We examine the barriers and drivers that affect further development of open data practices from the researchers' point of view, across a wide range of disciplines, and from a global perspective. We address the following research questions.

> *"Open data needs to be embedded in the research process from start to finish."*

1.1 Research questions

1) How are researchers sharing data?
This question relates to formalized or ad hoc data-sharing workflows, the infrastructures being used or created, whether researchers are adding metadata to published datasets, and if these datasets are stored in repositories. When datasets are in a repository, is it a local or commercial one, and is the metadata accessible and by whom? Also relevant to this question is how researchers acknowledge or cite such activities done by others.

2) Do researchers themselves actually want to share data and/or reuse shared data?
This question relates to practices in fields in which data sharing is more prominent, whether we can identify fields that are opening up or closing down with regard to data sharing, and if there are discipline-specific perceptions about open data.

3) Why might researchers be reticent to share their own data openly?
This question relates to why some researchers are more oriented toward publishing data (vs. publishing articles) and whether these researchers share a common research profile or disciplinary background. Also, why and when do researchers feel that they need access to others' data?

4) What are the effects of new data-sharing practices and infrastructures on knowledge production processes and outcomes?
This question relates to whether we can identify changes in research organization, governance, and/or funding associated with data sharing. Likewise, are new roles, forms of expertise, and/or authority being created?

These questions are explored using a complementary methods approach to capture as broad a view of data sharing practices as possible.

1.2 Complementary methods approach

Our complementary methods approach, which allows for triangulation of findings, consists of three parallel, interdependent studies.

Quantitative analysis of publication data

We analyzed the acknowledgments section in scientific articles to gain insight into data sharing practices. We also analyzed data articles and the citations they receive.

Large-scale global survey

We conducted a large-scale survey of researchers worldwide regarding their data sharing practices and perceptions of open data. The aim of the survey was to examine the drivers and influences on the communication of scholarly research data, data sharing, and data management practices.

Case studies

Three case studies illustrate specific aspects of data sharing within three fields: Soil Science, Human Genetics, and Digital Humanities. We focused on the ways in which data is shared during the course of conducting research.

02: Results

2. Results

2.1 Quantitative analysis of bibliometric data

In this study we explore several quantitative bibliometric approaches. The most informative results come from our analysis of data journals, discussed below. Results from an analysis of the acknowledgments section in scientific articles are also presented below.

2.1.1 Analysis of data journals

Data journals, such as Scientific Data published by Springer Nature and Data in Brief published by Elsevier, are a recent phenomenon. These journals seem to reflect an evolving perspective on scientific data. Traditionally, scientific articles discuss data as part of a broader research project, but articles focusing exclusively on data are becoming more common. Sometimes these articles appear in dedicated data journals, while in other cases they appear in traditional journals.

To study the phenomenon of data journals, we focus on articles published in these journals and citations given to these articles. A survey of data journals is provided by Candela et al. (2015), and a list of data journals can be found at http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList. Our analysis does not include mixed journals that publish both data articles and regular research articles. Only data journals, which just publish data articles, were taken into account. However, even data journals do not represent a perfectly homogeneous category. For instance, Scientific Data publishes data with occasional contextualization, whereas Data in Brief strictly publishes data descriptors and metadata.

The table below shows the number of articles published in different data journals as well as the number of citations given to these articles. Many data journals are not indexed in bibliographic databases such as Web of Science™ and Scopus®. We therefore obtained an approximation of each journal's number of articles by browsing through the journal's archive on the website of the publisher. All articles published until the end of 2016 were counted. Although most journals are not indexed in the Web of Science™ database, we were able to count the number of citations given to these journals by other journals that are indexed in the Web of Science™ database. In this way, the citation counts reported in the table were obtained. Citations given in the last months of 2016 may not be included in the citation counts in the table.

Table 1. Articles and their citations in data journals

JOURNAL	APPROX. NO. OF ARTICLES	NO. OF CITATIONS
Data in Brief (Elsevier)	1200	433
Biodiversity Data Journal (Pensoft)	400	187
Scientific Data (Springer Nature)	250	786
Journal of Open Psychology Data (Ubiquity Press)	60	16
Geoscience Data Journal (John Wiley and Sons)	30	98
Dataset Papers in Science (Hindawi)	20	21
Journal of Open Archaeology Data (Ubiquity Press)	20	5
Open Health Data (Ubiquity Press)	20	5
Open Journal of Bioresources (Ubiquity Press)	15	8

The three data journals that published the largest number of articles are Data in Brief, Biodiversity Data Journal, and Scientific Data. Together, these journals published approximately 1850 articles, of which almost two-thirds were published by Data in Brief. The same three data journals also received the largest number of citations. Together, they were cited approximately 1400 times. More than half of the citations were received by Scientific Data, even though this journal is the smallest of the three in terms of its number of articles. The six remaining journals listed in the table are quite small. In total, these journals published fewer than 200 articles and were cited fewer than 200 times.

“Data journals are still a relatively small-scale phenomenon.”

While data journals are a recent addition to the literature, their popularity is increasing quite rapidly. The yearly number of articles published in these journals has increased steadily. We do not have precise statistics available for all years and all data journals, but 60% of the total number of articles of the three largest journals (i.e., Data in Brief, Biodiversity Data Journal, and Scientific Data) appear in the most recent year (i.e., 2016). We do have statistics on the growth in the yearly number of citations given to data journals. These statistics, presented in the table on the right, confirm the increasing popularity of data journals. As already mentioned, citation counts for 2016 are incomplete. We expect that the total number of citations given to data journals in 2016 will be approximately 1250.

Table 2. Citation growth to data journals

YEAR	NO. OF CITATIONS
2012	3
2013	1
2014	50
2015	425
2016	1028

"The popularity of data journals and citations to these journals is growing rapidly."

Table 3. Citations to data journals in different fields of science

YEAR	NO. OF CITATIONS
Life sciences	563
Medical sciences	294
Earth and environmental sciences	246
Multidisciplinary journals	164
Chemistry, physics, and astronomy	146
Engineering sciences	34
Mathematics, statistics, and computer science	20
Social sciences	20
Health sciences	9
Culture	4
Economics, management, and planning	2
Information and communication sciences	2
Law	2

The table left shows the number of citations given to data journals by journals in different fields of science. Based on these statistics, data journals seem most popular in the life, medical, earth, and environmental sciences and in the fields of chemistry, physics, and astronomy.

The main conclusion that can be drawn from our analysis is that data journals are still a relatively small-scale phenomenon; however, our analysis also indicates that the popularity of data journals is growing quite rapidly. In a few years, these journals may be a significantly more important part of the publication landscape and the practice of scientific research (Park et al., 2017). A limitation of our analysis is that we are not able to study data articles published in mixed journals containing both regular research articles and data articles. Hence, our analysis offers only a partial insight into the phenomenon of data journals.

2.1.2 Analysis of acknowledgment sections

The acknowledgment section in scientific articles potentially offers evidence of data sharing. Researchers who publish work in which they have used data made available by other researchers may mention the contribution of these other researchers in the acknowledgment section. In this way, acknowledgment sections may provide insight into data sharing practices.

In our analysis of acknowledgment sections for evidence of data sharing practices, we included articles that appeared in 2014 and that were indexed in the Web of Science™ database. We focused on articles classified as a research article or a review article in the Web of Science™ database. Other types of articles, such as letters, editorials, book reviews, and corrections, were not taken into account. We analyzed the acknowledgment sections in research and review articles in which funding information is included in the acknowledgment section.

A total of 1.51 million research and review articles were published in 2014. Of these, 0.93 million articles have funding information included in the acknowledgment section. For these articles, the full text of the acknowledgment section is available. We performed a search of the acknowledgment section of these 0.93 million articles for the word "data" and either the verb "provide" or the verb "share". A total of 29,637 articles (3.2%) have an acknowledgment section that includes our search words. In the text box, some examples are provided of acknowledgments that were identified using our search strategy.

- "We thank '...' and '...' for fruitful discussions and name for sharing data with us"
- "We are grateful to '...' for sharing unpublished data on Gata3 neurons"
- "NLDN data are provided by Vaisala, Inc. (name, email)"
- "We thank '...' and '...' for providing their observational data"
- "We gratefully acknowledge '...' and '...' for data collection, and '...' for sharing experience of the validation of C-EdFED-Q"
- "Financial support for this work provided by Grand Challenge Canada Stars in Global Health (LS) ... The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript"

For the first four examples in the box, the acknowledgment provides clear evidence of data sharing. However, this is not true for the last two examples, which we consider to be false positives. They meet our search criteria, but they do not reflect the sharing of data. From our analysis, we draw the following conclusions:

- Acknowledgment sections can provide evidence of data sharing, but they do so only for a small share of all publications. Data sharing is mentioned in an acknowledgment in only a limited number of articles.
- Identifying acknowledgment sections that mention data sharing is not straightforward. Our search strategy suffers from the identification of false positives. Accurate identification of acknowledgments mentioning data sharing requires a more sophisticated text mining approach.
- Acknowledgment sections in which data sharing is mentioned seem to refer mostly to informal data sharing. They typically describe situations in which one researcher or research team informally shares data with another researcher or research team. Detailed information on the way in which data is shared usually is not available in an acknowledgment.

2.1.3 Highlights

Key finding 1
The introduction of data journals is a recent development. Data journals are still a small-scale phenomenon, but their popularity is growing quite rapidly and it is detectable in strong growth of citations over time.

Key finding 2
Open data is largely driven by disciplinary culture given the significant differences between scientific fields in the adoption of data journals.

Key finding 3
The lack of consistency in reporting data sharing in the acknowledgment section of scientific articles highlights a lack of reporting standards.

2.2 Large-scale global survey

Open data is generally operationalized as open sharing and reuse of (research) data; however, this description does not include a clear definition of what constitutes research data. Therefore, before conducting our survey, we explored among researchers the best way to interpret the phrase "research data." The following definition was most commonly recognized by researchers and was used in our survey: "Recorded factual material generated (and commonly retained) and accepted in the research community as necessary to derive and validate research findings." This is a definition paraphrased from: *www.epsrc.ac.uk/about/standards/researchdata/scope/*

The survey addresses topics around data production, data management, data sharing and using other researchers' data aiming primarily on the researcher's perspective. Questions focus on actual practices, policies and perceptions around data sharing. The on-line survey was delivered in June-July 2016 to researchers worldwide in all scientific fields. 1,162 researchers responded, representing a 2.3% response rate which is to be expected for a survey like this one. Responses are weighted to be representative of the researcher population (UNESCO counts of researchers, 2013). Margin of error for 1,162 responses is ± 2.87% at 95% confidence levels.

A selection of the survey results is presented below (full and raw data results can be found at doi:10.17632/bwrnfb4bvh.1), covering the following general topics:

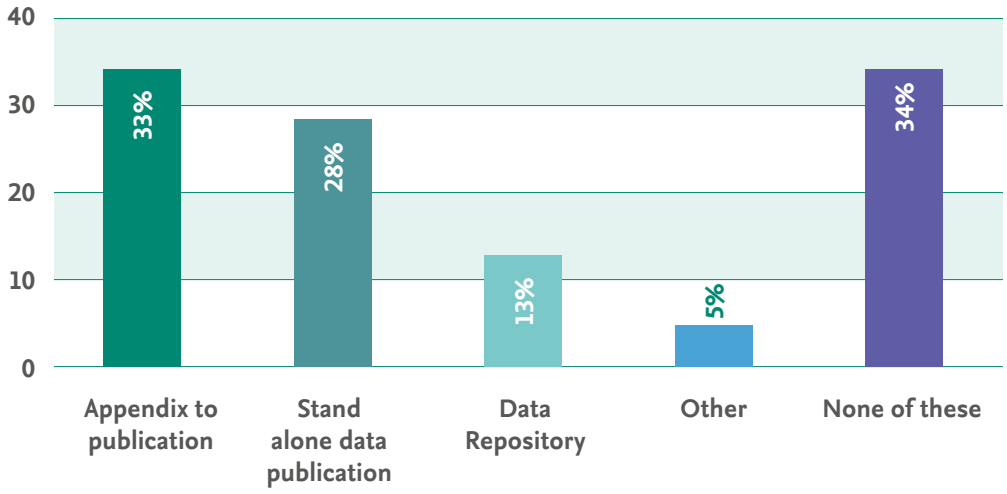
- How and why are researchers sharing data?
- Why are researchers reticent to share their data?
- What is the role of research data management in research data sharing?
- How do researchers perceive reusability?

2.2.1 How and why are researchers sharing data?

Research data that are Findable, Accessible, Interoperable, and Re-usable (FAIR, Wilkinson et al., 2016) are ideally kept in repositories and archives, often related to specific fields or disciplines (e.g., GenBank or NASA Distributed Active Archive Centers). These repositories can be global, national, or institutional, and the same dataset can be stored at different levels. Most researchers prefer to store their data close to their "home," for example, in departmental or institutional archives (37% and 34%, respectively). Twenty percent of researchers stored data in various forms including personal archive and cloud facilities, and one in eight (12%) indicate that they do not archive their data at all.

Currently, public data sharing mainly occurs through publishing avenues (see figure 1) though notably, one third of the researchers do not publish their data at all. Dissemination of research data most often takes place as an appendix or supplement to a research article or as stand-alone article in a data journal. By contrast, less than 15% of researchers publish their data in a data repository. These findings are in line with the literature, which reports that 13% of research articles with original data make these data available to others (Womack 2015). Researchers noted that they prefer to publish their data alongside a research article in a data journal rather than in a repository because, as authors, they receive several benefits: more collaboration possibilities, greater reproducibility of research, higher likelihood of being cited, and encouragement of others to reciprocate and make their data available. Interestingly, compliance with journal or publisher requirements or funder mandates to share data is not often perceived as being important.

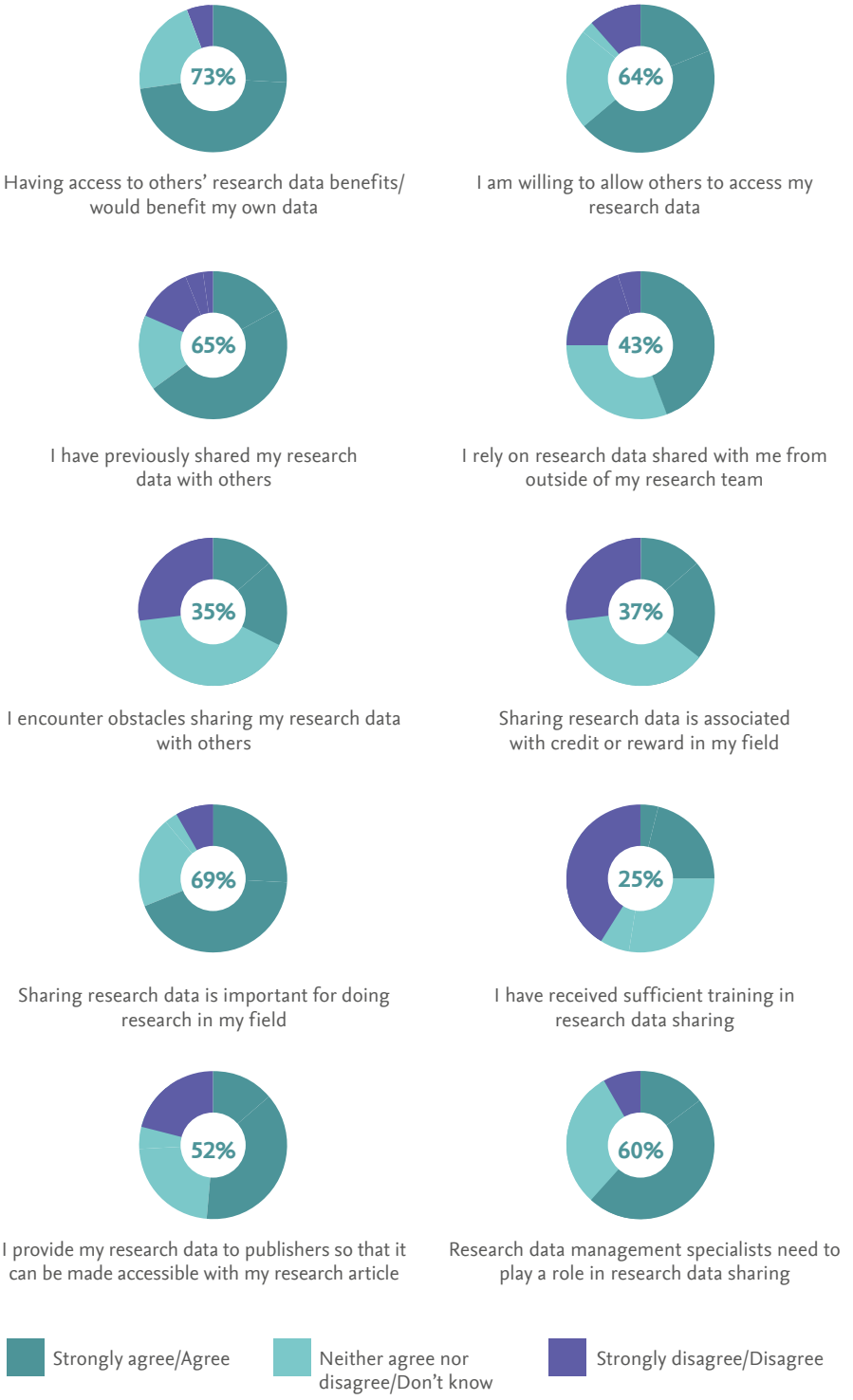
Figure 1. Dissemination of research data (%, n=1162)



Of those researchers who share their data directly (i.e., person-to-person), most (>80%) share with direct collaborators and 39% share with external parties, but only 14% share data directly with researchers they do not know when they are working on a project. Sharing therefore seems to be connected to collaboration (see also 2.3), and suggests that trust is an important aspect of sharing data. A third of researchers have not shared data from their last project, indicating that there is significant room for growth in data-sharing behaviors between researchers.

When asked about why they share data, or more specifically, their attitude towards sharing of unpublished data, somewhat more researchers agree that having access to other researchers' data would benefit them (73%) than agree that they are willing to share their data (64%), or have shared data (65%). Most researchers acknowledge that the field benefits from sharing data (69%). These opinions are especially strong in the fields of computer science, physics, and astronomy, which have the most positive view of data sharing. Thus, there is a gap between the perceived benefit of data sharing and the actual practice. Lack of reward or training may explain this gap (figure 2). The benefits of sharing identified by researchers relate mainly to the impact of their work (i.e., combining data increases the validity and reproducibility of the research), research efficiency (i.e., saving time and costs), generation of new ideas and contributions to the field, and transparency and collaboration. These answers confirm previous findings that data sharing helps develop a democratic society (Baack, 2015), enhances the transparency of scientific research (Parmesan & Yohe, 2003), allows for reproducing and validating research (Nosek et al., 2015), and unleashes the potential of data to solve complex societal issues (Figshare 2016; AWTI 2016).

Figure 2. Attitudes towards sharing of research data (%, n=1162)

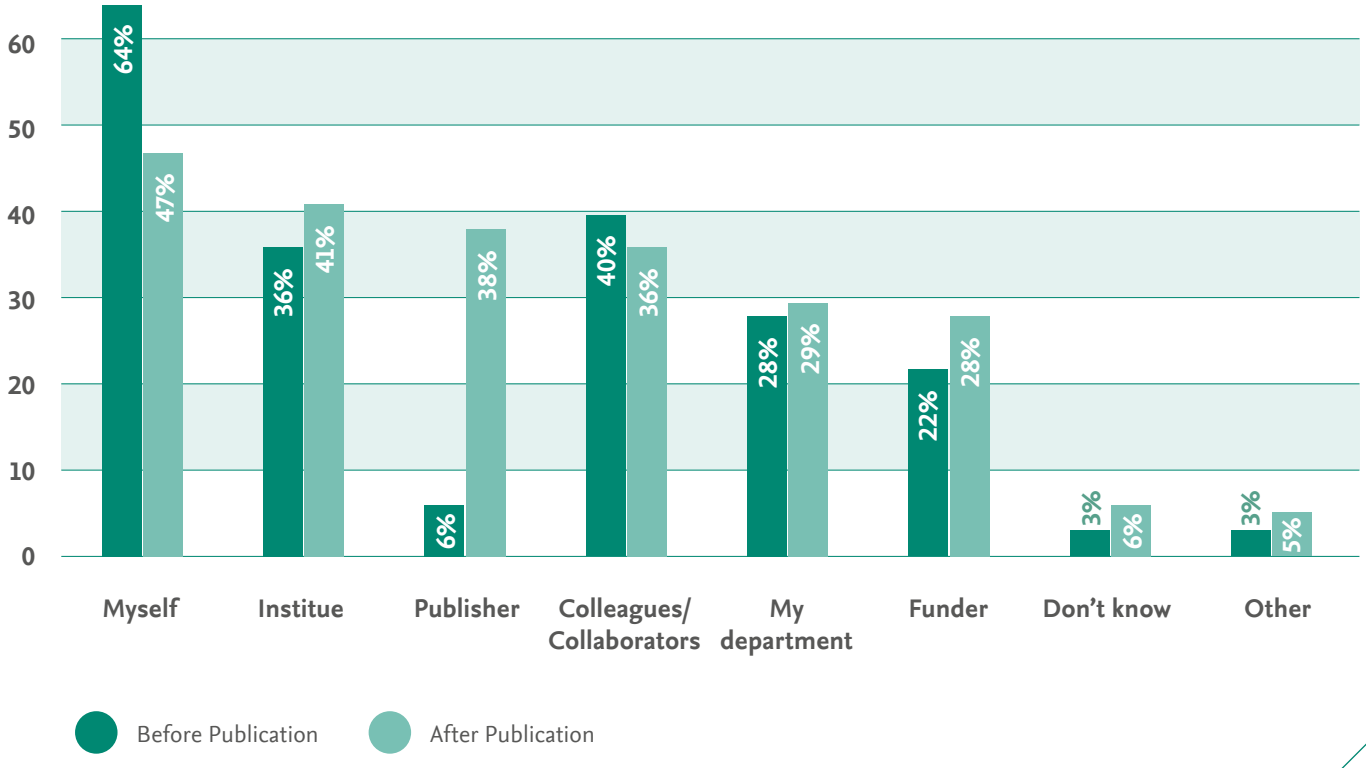


2.2.2 Why are researchers reticent to share their own data openly?

While the benefits of sharing data may be recognized, the barriers are clear as well (Piwowar, 2011; Longo and Drazen, 2015). The survey shows that a third of researchers did not share data from their last project. This reflects the finding that 34% of researchers do not publish their data at all (see figure 1, and Kratz & Strasser, 2015). Legal barriers to open data include privacy concerns, ethical issues, and intellectual property rights. These all relate to the basic question of ownership, responsibility, and control of data, about which there is a lack of agreement according to Borgman (2015). Researchers, however, have clear beliefs about who owns data (see figure 3). Two-thirds mention "myself" as the data owner prior to publication, followed by their colleagues and collaborators, indicating that a person has more ownership over data than an institute, department, or funder. The perception of ownership was also seen in e-infrastructure where almost half of the researchers reported they took the data with them when leaving their institute (e-infrastructure Austria, 2015). After publication of data, many researchers feel (incorrectly) that ownership is transferred to the publisher. The option "society" was not presented as a choice in the survey, but in the category "other," a substantial number of researchers responded that the wider society or taxpayer is an owner of the data.

Legal and ethical concerns are cited as reasons for not publishing research data alongside an article: a substantial proportion of the survey answers on this topic mention that data is proprietary or that researchers do not have consent to share data. Also, respondents answered that they do not like the idea that others might abuse or misinterpret their data (let alone take credit for it).

Figure 3. Research data ownership before and after publication (%, n=1162)



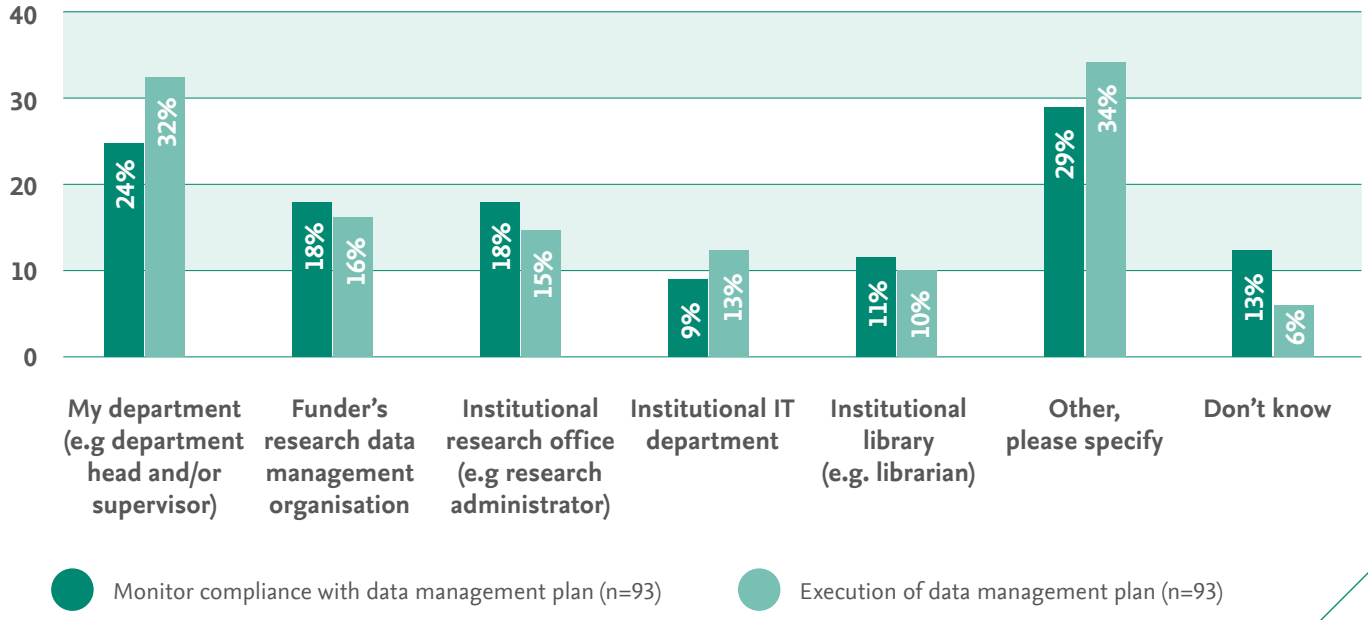
2.2.3 What is the role of research data management in research data sharing?

For research data to be open, it must be managed, stored, and curated, along with any contextual information needed for access and retrieval. Together, these processes comprise data management. The question is, who takes responsibility for this aspect of data sharing? From the survey researchers describe that research data management typically requires some (59%) to a lot (25%) of effort. The main reasons for this level of effort include the need to navigate legal issues (e.g., confidentiality, legislative issues), format the data (i.e., presenting it clearly), develop logistics (e.g., where to upload), and perform data cleaning (i.e., making the data usable).

Data management practices vary considerably among researchers, as more than half report that they do not consistently manage their data for future use and a quarter do not structure their management approach. At present, data archiving is driven by researchers' individual opinions or the culture of their specific field. Researchers believe funders only mandate archiving in a minority of circumstances. When asked who usually does the archiving of research data, 76% of researchers said they do it themselves. Only approximately a quarter believe institutions provided funding for archiving in 2016, and in those that did, more funds are expected to be available in 2017, which may reflect a response to a policy need. The annual spend towards data management also varies.

The execution and monitoring of research data management plans are organized in various ways (see figure 4). Data management planning tends to be implemented at the departmental level or by the individual researcher rather than at the institutional level, as was highlighted from specified answers in the 'other' category which most researchers selected.

Figure 4. Execution and monitoring of research data management (%)



“Open research data is a reality for policy makers, but has not yet become a reality for researchers”

2.2.4 How do researchers perceive reusability?

Responses to open questions reveal that researchers access others’ research data via email or through direct personal contacts. Alternatively, they access data through articles, appendix/supplementary material from the original article, institution archives, conferences, personal websites, and a wide range of specific websites. Nearly half of researchers (48%) report that they made use of other (or “third party”) data on their last research project. Good documentation is the most important factor for trusting another researcher’s data, but institutional reputation or data having been cited elsewhere add credibility. The use of others’ data appears less dependent on personal acquaintance or researcher reputation.

Although many researchers have reused data, a large proportion of researchers (45%) do not feel there are clear standards for citing others’ data. This means that reuse might not always be associated with appropriate attribution, which may lead to concerns about plagiarism and a lack of credit for data sharing. However, a similar proportion of researchers (41%) agree that there are clear data citation standards in their field and that most people follow them. In addition, researchers are not actively thinking about reuse licenses they can assign to their data. When asked which creative commons license they would make their data available under, 62% answered that they didn’t know. Where researchers did provide an answer, they tended to favor more restrictive licenses. Overall, this suggests that researchers might have a lack of knowledge regarding the sharing and reuse of data, which may affect their willingness to do so.

Taken together, these findings clearly demonstrate that real-world research data sharing practices do not live up to the expectations of many policy makers at national or international levels. Open research data is not an established practice in many fields. What emerges is a picture of very scattered practices across and within fields that, if they take place at all, are happening primarily at the individual level, with the funders and publishers significantly removed from the process. Even with regard to research data management plans, few activities are in place and the focus is on data production. There is little activity in the area of data searching and standardization at the researcher level, and infrastructure, training, organization, and funding are lagging behind. Open research data is a reality for policy makers, but has not yet become a reality for researchers.

2.2.5 Highlights

How is data shared?
Key finding 1

Dissemination of data is primarily contained within the current publishing system, even though one third of the researchers do not publish their data at all. The preferred method of research data dissemination is publication—be it in data articles in traditional journals, as an appendix to a research article, or in a dedicated data journal (although still infrequent)—though a large amount of data might remain unshared. Given the limited number of data journals, most data sharing is likely occurring through appendices or supplemental materials to research articles. Depositing data in archives, which is generally viewed as the gold standard, is not standard practice as of yet, but may be more likely in fields where open data has already become an integral part of the field (see 2.3).

How is data managed?
Key finding 2

Data management requires significant effort, and training and resources are required. Research data management typically requires a lot of effort related to navigating legal aspects, presenting data clearly, deciding where to upload, and making the data usable by adding metadata or identifiers. Data management practices are currently quite variable; many researchers do not consistently manage their data for future use and do not follow a structured management approach. While training related to open data is generally understood as beneficial and/or desired, this training is largely missing. Open data mandates from funders or publishers are not perceived as a driving force to improving data management training or planning.

How do researchers perceive data sharing?
Key finding 3

Research data is perceived as personally owned and decisions on sharing are driven by researchers, not by institutes or funders. It is important to be aware that the concept of open data speaks directly to basic questions of ownership, responsibility, and control. Researchers often consider themselves and their colleagues or collaborators to be the owners of data resulting from research projects. Institutions, funders, or publishers are viewed as entities that are neither responsible for data nor able to address ethical concerns, provide consent, or prevent abuse or misinterpretation of shared data.

How do researchers perceive reusability?
Key finding 4

Researchers have little awareness of reuse licenses and proper attribution, thereby making it less rewarding to make data reusable.

“Researchers feel they are at heart of the practice of sharing and reuse of data.”

2.3 Case studies

Studies of open data often focus on the status and potential of making data publicly available for reuse by academic actors not involved in data generation or by public actors not directly associated with academic research (Borgman 2012). This framework for describing open data addresses the widest practical range of potential users and reusers, and implies that significant effort is needed to prepare data for use by actors unknown to those who created the data. Yet, there has been little assessment of the data sharing practices that take place in fields that have a tradition of data sharing—many of these practices might not be considered "open data" using the above framework.

2.3.1 Conceptualization of case studies

We shifted the focus from concerns of public access to include a stratified account of data sharing practices. We expanded our study of openness to include real-world concerns, such as the practicalities of making data reusable, issues of transparency and validity, globalization of research, and commodification (commercialization) of data (Leonelli, 2013).

To achieve this, we investigated data sharing practices within three scientific disciplines: Soil Science, Human Genetics, and Digital Humanities, by interviewing key individuals involved in data collection, analysis, and deposition. Conceptually, we adapted Leonelli's framework for open data to understand six dimensions of data sharing practices:

(a) Data situated: As the case studies were selected to provide a diversity of research contexts, we use the "data situated" dimension to explore ways in which data is conceptualized in daily practice. The aim here is to explore contextual factors associated with data sharing.

(b) Pragmatics of data sharing and/or reuse: The activities involved with making data shareable require coordination of the tools, procedures, and standards, as well as communication among collaborators, which together enable transfer of datasets through different stages of the research process.

(c) Incentives for sharing and/or reuse: As sharing data is often an activity apart from the academic reward system (and/or researcher evaluation), this dimension focuses on both internal (e.g., collaboration) and external (e.g., policy) factors associated with data sharing.

(d) Governance and accountability: With increased political attention to open data, we focus on the role of mandates, data management procedures, and training associated with preparing data for sharing.

(e) Commodification: Data sharing and reuse often involve third party entities to provide services in support of data process, analysis, and storage. The aim of this dimension is to examine the role of licensing, commercial data services, and commercial funders in data sharing.

(f) Globalization: Sharing and reuse of data within a local research team necessitates coordination among collaborators. International collaborations have the potential to further complicate the pragmatics of data sharing.

Soil mapping

The site described in this case study is an international center dedicated to gathering information on world soil. Over a period of decades, outside scientists' willingness to share their data with the center has meant they have accumulated a variety of data pertaining to soil properties of particular regions. The center receives donations, external grant funds, and block funds for projects mapping soil classes and properties of particular nations, regions, or across the entire globe. These maps are then used for modeling various ecosystem characteristics, such as climate change, soil erosion, soil nutrients, land use capacity, and soil biodiversity. Large-scale mapping initiatives tend to be the domain of dedicated centers, including national governmental agencies.

Human genetics

The research center selected for this case study is organized into several co-located biomedical genetics labs. A centralized bioinformatics group provides data processing and analysis expertise to multiple labs in the research center, coordinating their activities with several projects simultaneously. We focus on the application of data sharing to clinical genetics research on rare diseases. Modes of data sharing and reuse are organized by local collaborations required for diagnosing rare diseases and with other clinics also investigating rare diseases. Collaboration teams are comprised of a clinical geneticist, a wet-lab technician, a bioinformatician who prepares and analyzes the data, and often additional researchers to assist with interpreting the outcomes from the data analysis.

Digital humanities

Many digital humanities research projects in the Netherlands are linked through a national level network. For this case study, we focus on researchers whose work straddles the traditional humanities and computational science. An important part of sharing in this context is the ability to transfer data among project participants, which enables later reuse (for instance, to extend or compare with existing data). What is shared is the data itself along with the analysis and processing tools. In addition, transparency and reproducibility are important, but these can be very expensive to implement and researchers are not incentivized to do so. While this is a significant issue for the field of computational science, it is less so in the humanities.

2.3.2 Analysis of data sharing dimensions across three cases

The full text case studies are presented at doi:10.17632/bwrnfb4bvh.1. Here we focus on the six analytical dimensions for each case.

a) Data situated

Soil mapping: The first "inputs" into the soil mapping process are soil surveys collected or shared by other scientists. These surveys provide both field information and information based on subsequent laboratory analysis (e.g., chemical and physical properties of soil sampled from the field). To produce the maps of soil classes and properties, traditionally soil mappers used mental models to draw spatial distributions of the soils, based on their interpretations of landscape maps. Today, statistical and geostatistical methods are applied to spatially predict soil classes and properties. To use soil survey data in a model, the soil descriptions must be typed into a table (e.g., Excel), which is imported into software to run models and create maps for various characteristics (e.g., soil erosion, soil fertility).

Human genetics: In the genetics lab, a collaboration is built between people with particular expertise needed to fulfill tasks. As such, the mode of sharing data is defined by the need to locally share or transfer the data between these individuals, working at different stages of data processing, analysis, and interpretation. Data in this context is digital or digitized versions of genetic source material. Sets of data are also stored and processed in databases throughout the course of the research project. In this way, transport of data creates an apparent epistemic distance between the source material and the object of analysis. Metadata is added to a dataset at each stage of data processing. As a consequence, metadata takes on increased importance both as the object of analysis and for enabling distribution of datasets among collaborators.

Digital humanities: The composition of project participants and the distribution of research labor varies across digital humanities projects. There is often an ongoing tension associated with converging fields and their respective communities of practice. Sometimes more senior humanities scholars do not have an interest in what they refer to as "technological" issues and prefer "intellectual" work. It is mainly PhD students who realize the need to develop computational science skills and set out to develop expertise across domains. Data gathering and processing involves computer scientists who collect "raw" data from the web, write scripts, and store the data locally on a server that 10-12 people can access. At some point, the pre-processed dataset, usually stored in a database, is "frozen" if it is to be used for research.

b) Pragmatics of sharing and reuse

Soil mapping: A small research team digitizes soil reports using a standard table format. As soil data is collected in different countries and at different periods of time ("legacy data"), different soil classification systems and laboratory measurement methods may be used. Regional or global model harmonization is needed to make the data usable. This process involves harmonizing data according to standards laid out in the classification of soil taxonomy (USA) and the FAO (Food and Agricultural Organization of the Unite Nations) classification. Some surveys also lack metadata information (e.g., coordinates of soil location, information on the collection method), which requires ad hoc solutions within the teams.

When maps are produced, they are made freely available. The center also attempts to make the data on which the maps are based shareable via their web portal. The center has developed its own centralized and user–focused database, which is a central figure in the work practices of the center. Strikingly, the very definition of data draws on its anticipated future sharing and reuse; these capabilities can only be achieved via systematization through databases.

Human genetics: Sharing and reuse of data is integral to the research objectives of each genetics lab at the center and is embedded in the research design. The data tasks typically begin with a digital version of sequence data, which then undergoes many layers of analysis according to the intended research inquiry. In most cases, this involves analysis in a semi-automated bundle of routines referred to as the "pipeline."

Bioinformaticians are tasked with developing and maintaining a suite of analytical routines while also pursuing increased efficiency of the pipeline. The skill set needed for this function is a combination of consultancy, the ability to work with the genetics researcher to understand the specific data needs related to the "biological question," and data modeling, the ability to design and implement a data model as a component of the research method. The pragmatics of data sharing and reuse are included in the translation of the biological question into an appropriate data model.

Digital humanities: Sharing and reuse of data in digital humanities is generally bounded by the configuration of a particular collaboration. Humanities scholars, information scientists, and computer scientists, for example, work together to analyze traditional humanities research objects with the benefits of digitized content and computational methods. The pragmatics of sharing and reuse across cases depend on sequential, analytical processes, where metadata often becomes the central object of analysis. Metadata is crucial for sharing among project participants and for understanding, interpreting, and reusing the data. Datasets are usually stored in a database or as a large file that can be zipped, and they are typically stored locally. Others can request access to datasets, which are then provided via a web server.

> *"Datasets become shareable due to the internal logic of a research design."*

c) Incentives for sharing

Soil mapping: Research data sharing and reuse is commonplace, and is governed by informal yet widely adhered to expectations about citing one's data sources. Practices of acknowledging data reuse are not necessarily taught through formal training or "awareness" courses or communicated via policy instruments, but simply are introduced to young researchers entering the field. Students learn to cite other publications through reading and submitting written work. Nonetheless, there is no standardized means of referencing data sources, and the priority is for the reference simply to be made, not for a particular referencing style to occur.

Human genetics: Incentives for sharing are largely embedded in how the research is organized. With the centralized organization of data processing and analysis, transfer of data among collaborators is a necessity for carrying out research. There are benefits to sharing data, for the medical fields in general and for patients in particular, beyond the boundaries of a particular project. Sharing frequency data for rare diseases is particularly illustrative of a mode of sharing that has been successful. Though datasets may be conditioned for sharing within a collaborative group, additional tasks are needed for sharing data externally. Barriers to sharing include: lack of professional credit for making the datasets sharable, additional time and expertise needed, and resistance from others to share their data.

Digital humanities: Obstacles to data sharing in the digital humanities case include: tensions in the distribution of labor and publications not specifying which processing tools or version of the datasets were used. While sharing data is valued, the career benefits to doing so are uncertain. There are indications that this is also changing, with funding agencies introducing preferences (if not mandates) related to data management and transparency. This also relates to emerging hierarchies in what is perceived as "actual" academic work versus intellectual contributions.

d) Governance and accountability

Soil mapping: When sharing data via the center's web portal, the licensing agreement provides instructions on how the data should be cited. While users tick a box agreeing to this as part of the terms and conditions on the web portal, there is no way of policing this and few users adhere to the correct citation format in practice.

From the perspective of their own accountability as a center, a lack of consistent data citing practices means that accrediting committees are unable to evaluate the number of times the center's data has been reused in publications. While users are not made accountable through data citations, the web portal does enable the center to collect information on those who download the data, which can be used both for strategic monitoring and external accountability.

Human genetics: Execution of pipeline analytics and interpreting the outcomes is often an iterative process that can involve multiple people with specializations in areas such as computer science, software coding, statistics, and data analysis, as well as an understanding of the biological system(s) at the center of the investigation. Once a particular genetic variant is identified and the rare disease has been diagnosed, the data is stored for local reuse and made available for sharing through a linked network of other rare disease datasets. However, sharing of genetic data must comply with strict privacy measures. While it is not common, it is possible to identify individuals from genetic data. Given these privacy and security concerns, the sharing of rare disease data is configured to provide "frequency" queries without providing access to the genetic data itself. Moreover, the database maintains rigorous security measures to comply with data security.

Digital humanities: Training related to open data is generally understood as beneficial and/or desired, but appears to be largely missing. Instead, collaboration of humanities scholars with computer scientists and information scientists serves as an important dimension for establishing the needed skillsets for data-intensive research. Humanities scholars cite GitHub, which is commonly used by computer scientists to share software, as an important means for sharing data. This illustrates the transfer of practices between disciplines and the utilization of resources with collaborators rather than the use of typical repository-oriented resources associated with the broader open data movement.

e) Globalization and licensing

Soil mapping: Governmental organizations are willing to share data with the center for a number of reasons: official policies of the governmental organizations to share data, the lack of a commercial threat in allowing others to use their data, the possible benefits to their own countries or regions of having their soil landscapes mapped, and not being concerned with gaining priority for discoveries (which can restrict the sharing of data among researchers). However, some frictions stem, in part, from diverse national and regional differences surrounding data privacy, licensing, and bureaucratic structures. Noted examples include: strict privacy laws that prevent inclusion of geographical coordinate points (France), restrictions on the scale of data that can be shown (China), bureaucratic practices that prolong and may prevent access to data (India), and diverging expectations over whether monetary exchange should occur (Netherlands and United Kingdom).

These examples show that even in a relatively mature field like soil mapping, where global data sharing is commonplace and established, national differences continue to introduce challenges.

Human genetics: The dimension of globalization did not seem particularly relevant for this case. Although present in the human genetics projects, data practices associated with globalization appeared to be embedded in the normal course of research collaborations. For example, international research collaborations are common, as is the distribution of tasks among participants on the basis of differentiated expertise across international collaborators.

Digital humanities: Intellectual property rights are an issue of particular concern in this field. Sharing data, for example, often involves addressing licensing of literary texts. Data is relatively easy to access with the use of web scraping tools and techniques, but it is hard to penetrate commercial and legal attributes related to platform owners like Amazon or Library Thing, as well as content owners such as university libraries or content publishers.

> *"Freeing up data depends on disciplinary, cultural, and local differences with respect to data privacy and licensing."*

f) Commodification

Soil mapping: For the soil mapping case, the commodification dimension is closely related to diverse national and regional differences outlined above in the globalization section. As noted above, some friction is encountered in the exchange of data across different international contexts. See web link for additional details doi:10.17632/bwrnfb4bvh.1

Human genetics: The commodification dimension is present in the Human Genetics case, but generally follows conventional practices outlined in past studies (e.g., Costas et al. 2013). For example, there are numerous biological data repository services available. For more details see web link doi:10.17632/bwrnfb4bvh.1

Digital humanities: As a new cross-disciplinary field, there is variability in the ways commodification is addressed by the digital humanities. There is increasing awareness of commercial opportunities, and commercial parties are present at conferences. We find instances of both enthusiasm and skepticism associated with commercial partners. Some welcome the opportunity to partner with commercial third parties, with an eye toward long-term development and sustainability of the humanities discipline. For others, commodification is still a dirty word, though this position is changing as a result of priority-setting by funding agencies (e.g., requirement for bringing in the creative industry). Commodification is happening, but at a very slow pace due to specific obstacles; for example, start-ups cannot provide matching funds that are often required to participate in conferences.

> *"Data is not always considered as a public good, but as something to pay for."*

2.3.3 Highlights

Key finding 1
In all three fields, a distance is created between the object of study (soil, genetic material, or texts) and the object of analysis (metadata associated with datasets). In other words, the object of study is turned into the object of analysis through sequential data analysis procedures. Of particular relevance to open data are the ways in which source materials are transformed into shareable and reusable entities on the basis of research needs. This is especially apparent in complex, data-intensive analysis that requires application of specialized expertise, such as computer science, in addressing, for example, a biological question. Datasets become shareable due to the internal logic of a research design, rather than from external influences. In this way, properties of sharing and reuse appear to be embedded in the concept of the data.

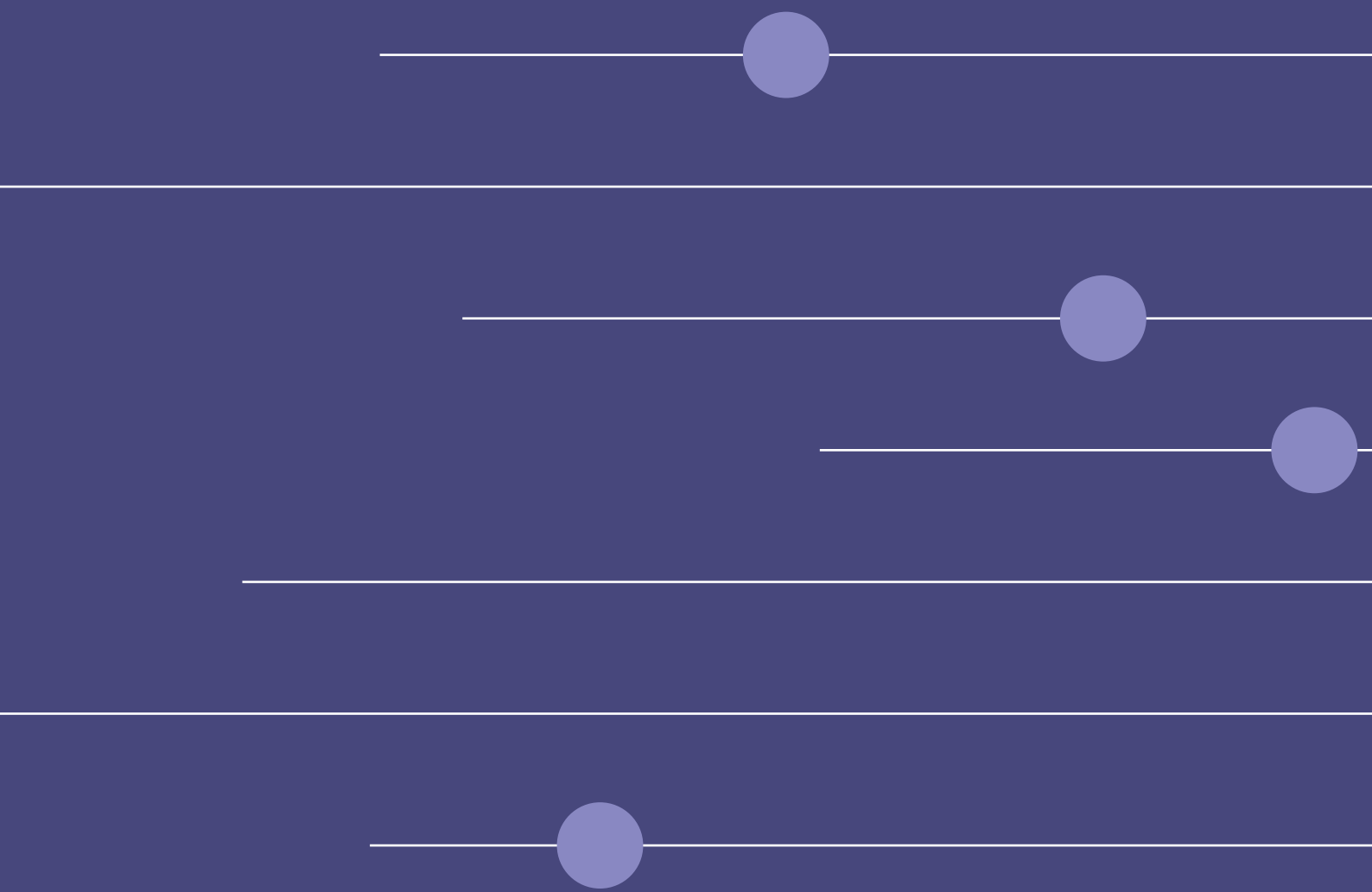
Key finding 2
Freeing-up data for reuse and sharing is hindered by national and regional differences with respect to data privacy and licensing. The case study material illustrates potential globalization challenges regarding "late stage" data sharing and reuse practices. Friction from national differences is evident, including diverse national and regional laws surrounding data privacy and licensing. Also striking are diverging "cultural" assumptions about the terms under which data should be exchanged (e.g., should monetary exchange occur, or should data sharing be understood as part of a "gift economy"). Finally, differing bureaucratic structures across countries and regions pose particular challenges with respect to releasing data for reuse and sharing.

"Open data might benefit from solutions in fields where sharing data is already an integral part of the research design."

Key finding 3
Data is only integrally configured for sharing and reuse in collaborative research projects, if incentives for sharing are embedded in the research design itself. In our case studies, sharing and reuse of data is shaped significantly by collaborative research configurations. Incentives for sharing and reuse are embedded in a research design that involves distribution of expertise across collaborators. As such, this mode of sharing is shaped by the need to share or transfer the data among collaborators across the different stages of data processing, analysis, and interpretation. A notable exception to this is an account of compliance with open data mandates from a publisher or journal, in which researchers are required to share/deposit data as a condition of publication of an article.

Key finding 4
Training related to open data is generally understood as beneficial and/or desired, but is still largely missing. For the digital humanities, collaboration with computer science and information science is an important dimension for establishing the needed skillset for data-intensive research. There is some indication of transfer of skills/practices between disciplines in development of data-sharing practices (e.g., use of GitHub by humanities scholars).

“Training and support facilities for open data-sharing practices need to be provided.”



03: Key Findings

3. Key Findings

In this complementary methods study, the survey portion addresses a broad, international multidisciplinary community about data sharing attitudes, while the case studies (carried out in the Netherlands) focus on specific instances of data sharing practices. The outcomes of the quantitative analysis, survey, and case study suggest a relationship between collaborative research and the potential for open data. Common barriers include efficient research data management practices and the need for awareness and training in open data practices. Data sharing practices vary across different research contexts. In a first step toward answering our research questions, the following are two data sharing scenarios we identified in this study.

Intensive data-sharing scenario

In fields where transferring datasets among collaborators is important for data processing and analysis, the basic conditions for facilitating open data are already embedded in the research design.

- Datasets accumulate layers of metadata associated with each sequential step in processing and/or analysis and describe the data and the various processing steps.
- Databases typically provide both the central transport medium between steps and serve as platform for analytical procedures.
- Portability of datasets is important to the research project.
- The features associated with this form of data portability are features that hold strong potential for open data.
- Datasets are already configured for storage in a repository.

Restricted data-sharing scenario

In fields where data processing does not necessitate the transfer of datasets among collaborators, basic conditions for facilitating open data are generally missing or only partially formed.

- Tasks associated with open data are unlikely to have significance and therefore have little overlap with the research design.
- Increased effort is needed to prepare datasets for sharing (occurs after or apart from the research).
- To offset an increased effort of data preparation for sharing, incentives for open data need to be external to the research project.

Overall, a bimodal picture emerges with respect to the data sharing scenarios outlined above. In contexts where data sharing is embedded in the research design, we observe specific data practices associated with transporting data through a particular data analysis sequence.

Attitudes are generally positive, but open data is not yet a reality for most researchers. We speculate that the recent Figshare report, The State of Open Data, included a cross-section of researchers from more intensive data-sharing fields, which may account for some of the differences between their findings and ours.

In contexts where data sharing is not necessary for conducting research, we find generally scattered practices. Amongst researchers in the latter contexts, there is confusion about terminology and expectations. There is clearly overlap in attitudes, challenges, and opportunities between intensive and restricted data-sharing fields.

3.1 Answering the research questions

1) How are researchers sharing data?
The application of data sharing principles is dependent on the field and practices in that field: intensive data-sharing fields are advanced in the areas of data curation, storage, and sharing, whereas restricted data-sharing fields predominantly keep data to themselves and share it through publication or collaboration, making it less accessible or open. In this case, trust in the receiving researcher is important.

2) Do researchers themselves want to share data and/or reuse shared data?
In the intensive data-sharing fields, data practices are an integral part of the research and consequently, the largely collaborative activities in these fields hold strong potential for open data and are supported in the field. In restricted data-sharing fields, there is less incentive to share data, although the benefit of it is recognized. Benefits include an increase in the impact, validity, reproducibility, efficiency, and transparency of scientific research. Ownership of data is considered to be personal in restricted data-sharing fields, and data is stored in proximity to the researcher, with access by collaborators upon request. Sharing occurs in more traditional ways, such as publication and presentation of data aggregated into tables and annexes, although there is increasing interest in stand-alone data journals. The latter may increase the number and value of data citations in the future, but these data journals are currently still very limited. Collaborative research is a common driver of data sharing in all fields.

3) Why are some researchers reticent to share their own data openly?
In intensive data-sharing fields, the reticence to sharing data depends on ethical and cultural limitations and boundaries. Financial and legal issues could also hamper sharing. In restricted data-sharing fields, reticence is based on a combination of factors relating to legal and ethical issues (proprietary nature of the data, informed consent); ownership, control, and responsibility; and preventing abuse or misinterpretation. An increased effort is needed in restricted data-sharing fields to prepare datasets for open data sharing, as this activity occurs after or apart from the research. This complicates data preparation and requires additional time, funds, capacity, and training. Research data management plans mandated by funders (or publishers) are not considered to be a strong incentive. Research data management and privacy issues, proprietary aspects, and ethics are barriers common to all fields.

4) What are the effects of new data-sharing practices and infrastructures on knowledge production processes and outcomes?
In the intensive data-sharing fields, new practices arise around access and use of data from databases, which is usually global in nature (albeit with local barriers). The restricted data-sharing fields are more traditional in terms of knowledge production and dissemination. They are aware of digital platforms such as GitHub, but use these in a personal, random way. For all fields, the scientific reward system does not include valuation of data-sharing practices.

Summary of the current situation

1. Data-sharing practices depend on the field: there is no general approach. General policy initiatives towards open data might benefit from encouraging bottom-up solutions in fields where open data is already an integral part of the research design.

2. Although data sharing seems to have a global benefit, cultural and national factors pose a significant challenge to a one-size-fits-all approach.

3. Freeing up data for reuse and sharing depends on accommodation or coordination of disciplinary, cultural, and local differences with respect to data privacy and licensing.

4. The role of funders and publishers in mandating data practices is limited compared to the role of researchers themselves. Open data mandates would benefit from better alignment with researcher incentive and evaluation structures (i.e., linked to the academic reputation).

5. In both intensive and restricted data-sharing fields, training and support facilities for open data-sharing practices need to be provided.

6. Data journals are still a relatively small-scale phenomenon, but their popularity is growing rapidly.

7. Data is not always considered as a public good, but as something to pay for. This perception could be a threat to open data.

8. Where open data management is occurring, it is often perceived as a burden, and not as a responsibility.

3.2 Challenges and opportunities

Challenges

...in data sharing

- Most sharing currently occurs among collaborators
- Data practices occur in a variety of different contexts and are therefore not easily standardized
- Privacy and ethical issues hinder the transfer of data practices from closed to more open situations

...in data management

- Researchers are not aware of data sharing mandates
- Data management plans are not used consistently
- Staff is needed for taking care of repeated and iterative data handling, e.g., bioinformaticians

...in perceptions on sharing and reuse

- Standards for citing another researcher's data are not universally understood
- Global versus local: global and national differences need to be addressed
- Researchers feel they are the drivers of data sharing (and are alone in recognizing its importance)
- Licensing issues and formats are not well understood

Opportunities

...in data sharing

- Researchers recognize the importance of data sharing
- Researchers are already sharing data in ways that can be optimized, e.g., defining "data pipelines" in research fields
- Collaborative practices can be used to further streamline data sharing
- Cross-disciplinarity provides an opportunity for open data in emerging intensive data-sharing research fields

...in data management

- Data sharing practices can be better facilitated by offering training and awareness programs, e.g., in research data management (RDM), Creative Commons (CC) licencing practices, and sharing mandates
- There is a need for increasing funding of data management activities

...in perceptions on sharing and reuse

- The scientific credibility system could reward participation in open data practices, e.g., through publishing in data journals, which is a recent and growing development
- There is a need for improving standardization and harmonization of processes, e.g., citation practices of data

3.3 The next step

The findings in this study provide a benchmark of the state of open data in research. As open data practices develop over time, we need to keep track and investigate possible changes. We would support an initiative which could monitor data sharing practices and implementation of more explicit research data policies. In the future of open data, there are many stakeholders involved including but not limited to the research communities, funding bodies, publishers and research institutions.

Researchers feel they are at the heart of the practice of sharing and reuse of data. Therefore, open data development would benefit from taking a bottom-up approach. A change in the scientific culture is needed, where researchers are stimulated and rewarded for sharing data and where institutions implement and support research data policies, including mandates in some cases. With this shift in culture, the perception of open data practices will transform. Rather than being seen as an extra effort removed from the research itself, research data management may be recognized as an integral part of the daily work of researchers. Currently, researchers have many responsibilities and data sharing is not perceived as a responsibility that will help their careers. Given that open data guidelines and standards have been developed, policy makers should now try to bridge the gap between policy and practice and ensure researchers are in a position to implement them.

However, while open data mandates provide an initial set of instructions, guidance should be given on implementation and sharing should be incentivized. Furthermore, solutions such as the European Open Science Cloud and the NIH Data Commons, which are currently being developed, should not be seen as storage tools, but as working tools that provide an environment that fits into the researcher workflow and makes it possible to directly and rapidly reuse data.

Finally, it is critical to recognize that, ultimately, all stakeholders are working towards a common goal. Researchers recognize that open data empowers research analysis and results, reduces unnecessary experiments, and promotes transparency and collaboration. When more people become involved and recognize the benefits of open data, policy and practice will continue to converge.

Bibliography

Bibliography

AWTI. (2016) Dare to Share: Open Access and Data Sharing in Science. The Hague, The Netherlands: The Netherlands Advisory Council for Science, Technology and Innovation. *www.english.awti.nl/documents/publications/2016/01/20/dare-to-share*

Baack, S. (2015) Datafication and empowerment: how the open data movement re-articulates notions of democracy, participation, and journalism. Big Data Soc July–December: 1–11. doi:10.1177/2053951715594634

Borgman, C.L. (2012) The Conundrum of Sharing Research Data. JASIST 63 (6): 1059–78. doi:10.1002/asi.22634.

Borgman, C.L. (2015) Big data, little data, no data: scholarship in the networked world. Cambridge, MA: The MIT Press. *www.Mitpress.mit.edu/big-data.*

Candela, L., Castelli, D., Manghi, P., Tani, A. (2015) Data journals: a survey. Adv Inform Sci. 66 (9): 1747-62. doi:10.1002/asi.23358

Costas, R., Meijer, I., Zohreh, Z., and Wouters, P. (2013) The Value of Research Data: Metrics for Datasets from a Cultural and Technical Point of View. A Knowledge Exchange Report. *www.knowledge-exchange.info/event/value-reseach-data-metrics.*

e-infrastrcutures Austria: Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung (2015) *phaidra.univie.ac.at/detail_object/o:407513*

European Open Science Cloud: *ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud*

Figshare/Digital Science Report. (2016) The State of Open Data: A Selection of Analyses and Articles About Open Data, curated by Figshare. *www.dx.doi.org/10.6084/m9.figshare.4036398*. ISBN: 978-0-9956245-1-1

Katz, J.E., and Strasser, C. (2015) Making data count. Sci Data 2: 150039. doi:10.1038/sdata.2015.39

Leonelli, S. (2013) Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. Bull Sci Technol Soc 33 (1-2): 6–11. doi:10.1177/0270467613496768

Longo, D.L., and Drazen, J.M. (2016) Data sharing. N Engl J Med 374 (3): 276-7. doi:10.1056/NEJMe1516564

National Iinstitutes of Health (NIH) Data Commons: *datascience.nih.gov/commons*

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., et al. (2015) Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. Science 348 (6242): 1422-5. doi: 10.1126/science.aab2374

Bibliography

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007): *www.oecd.org/science/sci-tech/38500813.pdf*

Open Knowledge Foundation (2012): *www.opendatahandbook.org/guide/en/*

Park, H. & Wolfram, D. An examination of research data sharing and re-use: implications for data citation practice. Scientometrics (2017). doi:10.1007/s11192-017-2240-2

Parmesan, C., and Yohe, G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. Nature 421 (6918): 37-42. doi:10.1038/nature01286

Pasquetto, I.V., Sands, A.E., and Borgman, C.L. (2015) Exploring openness in data and science: what is "open," to whom, when and why? ASIST 2015, 6-10 November 2015, St. Louis, MO, USA.

Piwowar, H. (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLoS ONE 6 (7): e18657. doi:10.1371/journal.pone.0018657

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., et al. (2011) Data sharing by scientists: practices and perceptions. PLoS ONE 6 (6): e21101. doi:10.1371/journal.pone.0021101

The Brussels declaration: *www.stm-assoc.org/public-affairs/resources/brussels-declaration/*

The Hague Declaration on Knowledge Discovery in the Digital Age: *www.thehaguedeclaration.com*

The Joint Declaration of Data Citation Principles: *www.force11.org/group/joint-declaration-data-citation-principles-final*

UNESCO Science Report: towards 2030 (2015): *unesdoc.unesco.org/images/0023/002354/235406e.pdf*

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018. doi: 10.1038/sdata.2016.18

Womack, R.P. (2015) Research data in core journals in biology, chemistry, mathematics and physics. PLoS ONE 10 (12): e0143460. doi:10.1371/journal.pone.0143460

Project Team

This project would not have been possible without the hard work and expertise of many at CWTS and Elsevier. We are truly grateful for all the insights and valuable contributions we received, as well as the relentless enthusiasm, dedication, and professionalism of each and everyone involved. We list here the key contributors to the report in alphabetical order, while acknowledging the support of many others.

Stephane Berghmans
Vice President, Academic and Research Relations, Elsevier
orcid.org/0000-0001-5414-8674

Helena Cousijn
Senior Product Manager Research Data Management Solutions, Elsevier
orcid.org/0000-0001-6660-6214

Gemma Deakin
Research Manager, Elsevier
orcid.org/0000-0003-1429-0398

Ingeborg Meijer
Senior Researcher, CWTS
orcid.org/0000-0003-1481-1739

Adrian Mulligan
Research Director, Elsevier
orcid.org/0000-0003-3585-5183

Andrew Plume
Director – Market Intelligence, Elsevier
orcid.org/0000-0002-4942-1426

Alex Rushforth
Researcher, CWTS
orcid.org/0000-0003-3352-943X

Sarah de Rijcke
Deputy Director, CWTS
orcid.org/0000-0003-4652-0362

Clifford Tatum
Researcher, CWTS
orcid.org/0000-0002-2212-3197

Stacey Tobin
Writer and Editor, The Tobin Touch, Inc.
orcid.org/0000-0003-1674-9844

Thed van Leeuwen
Senior Researcher, CWTS
orcid.org/0000-0001-7238-6289

Ludo Waltman
Deputy Director, CWTS
orcid.org/0000-0001-8249-1752

04/2017