2016-07-17

What Do We Know About The Stewardship Gap?

York, Jeremy

What Do We Know About The Stewardship Gap?

Jeremy York[1]*, Myron Gutmann[1], Francine Berman[2]

[1] University of Colorado Boulder
[2] Rensselaer Polytechnic Institute
* jjyork@umich.edu

Acknowledgements

We would like to thank the members of the Stewardship Gap project advisory board for their many useful comments and suggestions on drafts of this paper. Members include George Alter, Christine Borgman, Philip Bourne, Vint Cerf, Sayeed Choudhury, Elizabeth Cohen, Patricia Cruse, Peter Fox, John Gantz, Margaret Hedstrom, Brian Lavoie, Cliff Lynch, Andy Maltz, Guha Ramanathan. Any errors are the responsibility of the authors alone. We would also like to thank the Alfred P. Sloan Foundation for their funding of the project.

Abstract

In the 21st century, digital data drive innovation and decision-making in nearly every field. However, little is known about the total size, characteristics, and sustainability of these data. In the scholarly sphere, it is widely suspected that there is a gap between the amount of valuable digital data that is produced and the amount that is effectively stewarded and made accessible. The Stewardship Gap Project (http://bit.ly/stewardshipgap) seeks to investigate characteristics of and measure the stewardship gap for sponsored scholarly activity in the United States. This paper presents a preliminary definition of the stewardship gap based on a review of relevant literature and investigates areas of the stewardship gap for which metrics have been developed and measurements made, and where work to measure the stewardship gap is yet to be done. The main findings presented are 1) there is not one stewardship gap but rather multiple "gaps" that contribute to whether data is responsibly stewarded; 2) there are relationships between the gaps that can be used to guide strategies for addressing the stewardship gap; and 3) there are imbalances in the types and depths of studies that have been conducted to measure the stewardship gap.

Background

In the 21st century, digital data drive innovation and decision-making in nearly every field.[1] Key questions today center not on whether data can add value in these areas, but rather on how to obtain access to more data, how best to leverage data given a variety of concerns,

[1] Podesta et al. 2014, Big Data Value Association (BDVA) 2015, Manyika et al. 2011, Manyika et al. 2013, Holdren 2013, Science and Technology Council 2007, Vickery 2011, Houghton and Gruen 2014, Obama 2013, Kalil and Miller 2015.

including privacy and security, and how much value can be gained by using data.[2] This is true in both the public and private sectors, where data are increasingly seen as an asset to be used to promote innovation, economic growth and trust in or accountability of government,[3] and to further the arts and advance and verify scientific discovery.[4]

Desires for greater transparency and access to data have been particularly high for data 1) that are produced at public expense, whether as part of publicly-funded research or other public initiatives, and 2) that are used in or produced as part of sponsored scholarly inquiry ("sponsored research data") whether publicly or privately funded. Demand for access to the former is driven by public interest and opportunity for public benefit.[5] Demand for access to the latter is driven by public interest and principles of scholarship, especially those that advocate for open availability of knowledge to support further inquiry.[6] Demand is also driven by a desire for increased accountability. Several high profile cases of suspected or actual fraud, or a lack of supporting data have brought greater scrutiny to the availability of data to replicate and verify research results.[7]

Data used in or produced as part of sponsored scholarly inquiry are of primary interest for our research project, which we call the Stewardship Gap study (http://bit.ly/stewardshipgap). Included in this category are data sponsored by for-profit corporations that follow conventions of public or not-for-profit sponsored research; that is, data that support publications that are disseminated through accepted scholarly channels, and that rely on public or not-for-profit support for continued access.[8] Data resulting from sponsored creative work (such as the visual and performing arts) are in scope for a study of the stewardship gap, but expectations surrounding data sharing and the funding and policy environments for creative projects are different from those for sponsored research data, leading us to give them separate consideration. The literature reviewed in this paper relates primarily to public data and sponsored research data.

Sponsored research data, then, are data for which enduring access is in the public interest or interest of scholarship and for which public or scholarly bodies have the agency to preserve and make accessible, without being inhibited by proprietary or private interests. For our

---

[2] Manyika et al. 2013, Cummings et al. 2008, NSF 2007, Vickery 2011, 2012, Office of Management and Budget (OMB) 2012, Obama 2009 and 2011, Thompson Reuters 2013.
[3] BDVA 2014, Podesta et al. 2014, Berman et al. 2010, Ubaldi 2013, Organization for Economic Co-operation and Development (OECD) 2015, Sveinsdottir et al. 2013.
[4] National Research Council (NRC) 2003, National Academy of Sciences (NAS) 2009, Borgman 2012, OECD 2015, Wallis et al. 2013, NSF 2007, The Royal Society 2012, Association of Research Libraries 2006, Sveinsdottir et al. 2013, Tenopir et al. 2011.
[5] Vickery 2011, 2012 Podesta et al. 2014, Manyika et al. 2013, Borgman 2012, Holdren 2013, Obama 2013, OMB 2012, Ubaldi 2013.
[6] NRC 2003, NAS 2009, Borgman 2012, Royal Society 2012, OECD 2015, Holdren 2013, Research Councils UK 2015, Tenopir et al. 2011.
[7] See for example the Climatic Research Unit email controversy https://en.wikipedia.org/wiki/Climatic_Research_Unit_email_controversy, Vogel 2011, and Wicherts et al. 2011, which found a significant rate of errors in the statistical results of studies reported in four high-ranked journals published by the American Psychological Association.
[8] See NRC 2003 for rationale about sharing research data, however generated; see Berman et al. 2010 for more on public support for research data.

particular study, we additionally limit the scope to data produced through activities sponsored by the United States government or non-profit or corporate entities incorporated within the United States (referred to as "US sponsored research data"), though this paper will review practices and relevant studies and trends from other countries.

Data Stewardship

The demand for greater access to sponsored research data has focused attention on the chain of activities that lead to data access, including what data are saved by those who create them, where and how those data are stored and preserved, how they are described, what support for their reuse is available, and how they can be discovered and accessed. Taken together, these activities to maintain the integrity of and preserve access to data are commonly known as data stewardship. The National Academies, in an important 2009 study (NAS 2009, p. 27), defines stewardship as:

> "…the long-term preservation of data so as to ensure their continued value, sometimes for unanticipated uses. Stewardship goes beyond simply making data accessible. It implies preserving data and metadata so that they can be used by researchers in the same field and in fields other than that of the data's creators. It implies the active curation and preservation of data over extended periods, which generally requires moving data from one storage platform to another. The term "stewardship" embodies a conception of research in which data are both an end product of research and a vital component of the research infrastructure."

The importance of data stewardship to leveraging sponsored research data for a variety of purposes, both now and in the future, is reflected in the initiatives and policies that have been created in recent years in countries around the world, particularly in the United States and Europe, but in other places as well, to increase access to sponsored research data.[9]

While there is general agreement about the actions that must be taken and roles that must be played to steward sponsored research data, there is a lack of clarity about who should be responsible for carrying out these activities and funding them, and a lack of incentives for responsible parties to act. These deficiencies are reflected in the diversity of and lack of alignment among policies, deposit requirements, and data sharing practices across the landscape of those who fund, manage, create and reuse sponsored research data. A lack of agreement about responsibility for fundamental aspects of stewardship, and different understandings of what constitutes research data and which data that should be saved and made available contribute to a somewhat fractured and diffuse environment for stewardship of sponsored research data.[10] In

[9] OMB 2002, OMB 2013, Holdren 2013, Obama 2013a, Obama 2013b, Willetts et al. 2013, NASA n.d., OECD 2007, The Royal Society 2012, Engineering and Physical Sciences Research Council (EPSRC) 2011.
[10] Wynholds et al. 2012, Borgman 2015. See also the discussion of primary versus derived data in Pepe et al. 2014 and the discussion of which data researchers handed over to data managers in Thaesis and van der Hoeven 2010.

fact, despite the large number of data repositories, stewardship initiatives, and policies at multiple levels of the research data landscape, little is known about the total amount, characteristics, or sustainability of sponsored research data. Based on a lack of comprehensive planning for data stewardship and stories from different industries and disciplines, we and others suspect that there is or eventually could be a "stewardship gap" between the amount of valuable sponsored research data that is produced, and the amount that is effectively stewarded.[11]

The value that sponsored research data could have for advancing scholarly inquiry and providing public benefit, combined with the lack of answers to fundamental questions such as how much data exist and how it is cared for, bring urgency to the task of defining and measuring the stewardship gap. This paper explores the extent of the stewardship gap through a review of relevant literature, and summarizes efforts that have been made to develop metrics for, and measure, aspects of the stewardship gap.

Defining the Stewardship Gap

While numerous studies and reports have defined data stewardship, identified stewardship needs, put forth strategies to improve stewardship, and undertaken measurement and analysis of key factors that contribute to data stewardship (these are discussed below), there do not exist metrics for or measurements of the stewardship gap as a whole.

Measuring the stewardship gap is complex not only because it is difficult to measure the amount of sponsored research data that exists, but because a simple quantified measure of data would not provide critical information about the stewardship environment, prospects for stewardship, or other indicators that could yield insight into the likelihood that data will be stewarded either in the short or long term. Measuring the stewardship gap involves taking stock of a wide variety of component issues or "gaps" and the ways these interrelate and affect one another.

We show the scale of the issue in Appendix Table 1, in which we identify fourteen gap areas, drawn from eighty-seven articles, reports, and other works related to data stewardship. The areas include Culture; Knowledge; Responsibility; Commitment; Legal and Policy Issues; Funding; Collaboration; Human Resources; Infrastructure and Tools; Curation, Management and Preservation; Sharing and Access; Discovery; and Reuse.

Many of the articles and reports that we examined also indicate a relationship between gap areas. For instance, a statement may identify the effect that existing policies have on data sharing by researchers, or the impact that a lack of agreement about responsibility has on the ability to curate, manage, or preserve data. Some examples of statements and the gap relationships identified from them are shown in Table 1. The arrow indicates the direction of impact. As the fourth and fifth statements indicate, the influences are not always unidirectional

---

Berman et al. 2010, Thaesis and van der Hoeven 2010, Downs and Chen 2013, Esanu et al. 2004, ARL 2006, Borgman 2015, Borgman 2012.

[11] Berman 2014, Pienta 2006, STC 2007, Berman 2008 and Hilbert and López 2011, Gantz et al. 2008, Turner et al. 2014.

(e.g., knowledge can affect sustainability planning and vice versa).

Statement 1: "data sharing is even less systematic in domains where few common-pool resources exist" (Borgman 2015) **Relationship: infrastructure → sharing**

Statement 2 : "who will pay the costs associated with this curation and quality control" (Reuters 2013) **Relationship: funding → curation and quality**

Statement 3: "[there is] no group whose mainstream mission it is to plan and coordinate the data infrastructure needed for the research community (Berman 2014) **Relationship: responsibility → commitment → planning → infrastructure**

Statement 4: "Data curation faces the challenge that the data must be housed, managed, and made accessible prior to use, but actual uses may not be known until after sizeable investments are made (Wynholds et al. 2012) **Relationship: knowledge → planning**

Statement 5: "not possible to know costs until a preservation strategy is chosen" (Lavoie 2006, Eakin et al. n.d.) **Relationship: planning → knowledge**

Figure 1. Statements about gap areas and the relationships between them.

Figure 2 shows the relationships between the fourteen gap areas as identified from more than three hundred relationship statements like the ones above within seventy-three of the eighty-seven works.[12] The figure is arranged to show that gap areas in each column impact the gap areas in the rows below them. For instance, Culture (in the first column) impacts Knowledge, Commitment, Legal and Policy Issues, etc. (the gap areas in the rows of that column) The relationships shown are direct relationships from the statements. That is, one might infer that legal and policy issues would have an impact on how much we know in certain areas, or who is responsible for which aspects of stewardship. Since these relationships are not explicitly indicated in the literature, however, they are not represented here. The figure, then, does not attempt to represent comprehensive or definitive relationships between the gap areas. It does, however, represent what has been written about in a fairly large sample of widely cited literature about research data stewardship and as such is revealing.

[12] A complete data file containing the gap areas, relationships, and statements from literature these were derived from is linked to from the project website: http://bit.ly/stewardshipgap.

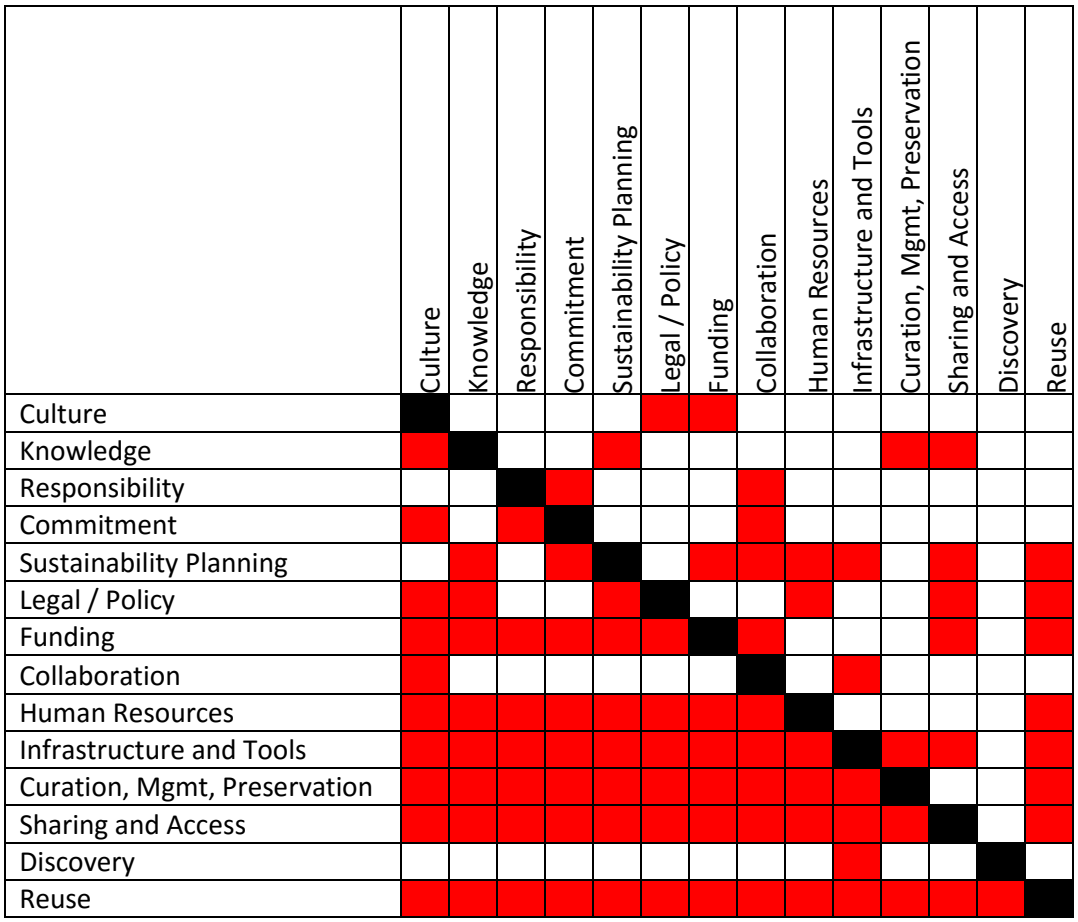

Figure 2. Gap areas and relationships between them

Horizontal rows with significant amounts of red indicate areas where many factors are at play. For instance, there are many factors that affect funding for data stewardship, the seventh row from the top (e.g., Culture, Knowledge, Responsibility, Commitment, etc.). Rows with significant white space indicate areas that may be difficult to address because there are not a lot of identified factors that influence them. For instance, Responsibility is impacted by Collaboration and Commitment, but there are few factors that affect Collaboration and Commitment themselves (and two of the factors that affect Commitment are bi-directional relationships with Responsibility and Collaboration).

One of the findings from this analysis is that many gap areas that have the largest impact on other areas ("strong gaps") are affected by relatively few factors. Some examples are Collaboration, Culture, Knowledge, Responsibility and Commitment. The analysis suggests that changes in these areas could significantly benefit data stewardship, but also that they may be more difficult to influence. The flip side of this analysis is that there are at least some factors that influence strong gaps (e.g., Collaboration is impacted by Infrastructure and Tools and Culture by Funding and Legal and Policy Issues) and could potentially be leveraged in efforts to change the size and nature of the gap in those areas.

A second finding is the scarcity of references to factors that impact Discovery of data, or

that Discovery impacts. Discovery is only mentioned in a couple of contexts in the reviewed literature, mainly in connection with infrastructure (e.g., that infrastructure is needed for discovery). Many sources talk about curation, management and preservation influencing reuse of data, but notably skip the step of how it is made known that data are available for reuse.

Stewardship Gap Measurements and Metrics

Our initial review of the literature led us to identify fourteen areas relevant to the stewardship gap. An expanded view of these gap areas is helpful for purposes of analyzing the relationships between them. However, we have collapsed the areas into the six categories below for ease of discussion and representation:

1. Culture (including Legal and Policy Issues)
2. Knowledge
3. Responsibility
4. Commitment
5. Resources (including Infrastructure and Tools, Human Resources and Funding)
6. Stewardship Actions (including Curation, Management and Preservation, Sustainability Planning, Collaboration, Sharing and Access, Discovery, and Data Reuse)

With these gaps areas identified, we set out to explore in which areas measurements had been taken or metrics developed, considering that a complete measurement of the gap, or metrics for measuring it, would at the least comprise the sum of existing measurements and metrics. For the purposes of our investigation, we considered studies to be measurements if they gathered information relevant to a stewardship gap area (whether through case studies, interviews, surveys, ethnography, or another method), and to develop or articulate metrics if they stated criteria that could be used as a basis for measurement.

Examples of Measurement and Metrics Studies

Fecher et al.'s 2015 article "What Drives Academic Data Sharing" is an example of a study including both measurements and metrics. In it, Fecher and colleagues describe a framework for understanding data sharing in academia (metrics). The framework comprises six categories of factors that impact data sharing. These are, as described in the paper:

- Data donor, comprising factors regarding the individual researcher who is sharing data (e.g., invested resources, returns received for sharing)
- Research organization, comprising factors concerning the crucial organizational entities for the donating researcher, being the own organization and funding agencies (e.g., funding policies)
- Research community, comprising factors regarding the disciplinary data-sharing practices (e.g., formatting standards, sharing culture)
- Norms, comprising factors concerning the legal and ethical codes for data sharing

(e.g., copyright, confidentiality)

- Data recipients, comprising factors regarding the third party reuse of shared research data (e.g., adverse use)
- Data infrastructure, comprising factors concerning the technical infrastructure for data sharing (e.g., data management system, technical support)

In order to develop the framework, Fecher et al. conducted a systematic review of literature and survey of secondary data users (measurement). For our analysis, we consider literature reviews by themselves as measurement if authors aggregated information across literature to show trends (e.g., in attention to data sharing as Fecher et al. do), or take measurements (e.g., of value and impact as in Vickery 2011).

In their survey, Fecher et al. explored questions such as why researchers do not share data, what returns or awards are received from data sharing, whether data sharing is encouraged by employers or funding agencies, what would motivate researchers to share data, and what value is gained from data sharing. They related the results from their survey to findings of other studies on data sharing in order to build the data sharing framework, which they believed had both theoretical and practical use. From a theoretical standpoint, their findings indicated that, in contrast to theoretical representations of open science or crowd science, "[r]esearch data is in large parts not a knowledge commons." On the contrary, their results pointed to "a perceived ownership of data (reflected in the right to publish first) and a need for control (reflected in the fear of data misuse). Both impede a commons-based exchange of research data." This finding, they argued, had practical implications for policy:

> "Considering that research data is far from being a commons, we believe that research policies should work towards an efficient exchange system in which as much data is shared as possible. Strategic policy measures could therefore go into two directions: First, they could provide incentives for sharing data and second impede researchers not to share."

Overall, Fecher et al. argued that the framework they presented helped "to gain a better understanding of the prevailing issues and [provide] insights into underlying dynamics of academic data sharing."[13]

Fecher et al.'s study is out of the norm among studies on data sharing in addressing both measurement and metrics. Out of 45 reviewed studies, only one other study on data sharing articulated metrics. This was "A game theoretic analysis of research data sharing."[14] In their paper, Pronk et al. develop a game theoretic model in which there is a cost associated with sharing datasets and a benefit associated with reusing datasets. The model includes such parameters as the time-cost to prepare a dataset for sharing and for reuse, citation benefit, and the

[13] All quotes are from Fecher et al. 2015, p.19.
[14] Pronk et al. 2015.

probability of finding an appropriate dataset to reuse, and the percentage of scientists sharing their research data. Pronk et al. ran simulations (for the purposes of our study considered measurements) with varying parameter values and found that not sharing data is the best option for researchers individually. That is, "it would be logical if all individual researchers would choose not to share and eventually end up getting the average impact by citations depicted at zero percent sharing."[15] They described this as a classic example of the prisoners dilemma phenomenon (individuals are better off not sharing, but all would be better off if all shared). Along with suggesting strategies to improve each of the parameters (time spent to access and reuse a dataset, chances of finding a dataset to reuse, time associated with sharing a dataset, etc.), Pronk et al. concluded that while policies are useful tools, reducing the costs of sharing and introducing a 'citation benefit' with shared datasets would result in the greatest sharing of data and overall benefit for researchers both individually and as a community.

The majority of studies regarding data sharing took measurements only, exploring questions such as attitudes towards data sharing, whether data is shared and how, limits on data sharing (e.g., privacy, intellectual property, or security concerns), incentives for data sharing, and problems encountered when trying to share data.

Measurement and Metrics Across the Literature

Figure 3 shows a visualization of the distribution of one hundred forty-five studies, reports, and projects (referred to as "studies") that either measure or provide metrics for measuring aspects of the stewardship gap. There are fifty-six distinct gap areas within the six major categories, and we identified studies conducted in forty-six of the fifty-six areas.[16] Many studies were relevant to more than one gap area. The total distribution of studies is as follows: Culture: 73; Knowledge: 47; Responsibility: 18, Commitment: 2, Resources: 37, Actions: 87.

The studies reviewed do not comprehensively represent all written works related to stewardship gap areas. However, as in Figure 2, they represent a large subset of works that are frequently referenced and that we viewed as highly relevant to our investigation of the stewardship gap.[17] The bibliography is posted online (http://bit.ly/stewardshipgap_bibliography), and we expect to add to it over time. The size of the rectangles in Figure 3 indicates the number of studies that have been undertaken to measure aspects of the stewardship gap, and the color

[15] Pronk et al. 2015, p.7.

[16] No studies were found in the following areas: Tradeoffs between data management for short or long term, Lack of critical mass, Support structures [for collaboration], Archive mandates and objectives, Imbalance in funding, Unequal access to resources and expertise, Different timescales of infrastructure development and maturity, and Provenance and authenticity. Although no studies were identified, these areas are represented in the treemap (as the smallest boxes) to provide a picture of the overall landscape.

[17] As an example, for the most part we scoped our references to those having to do with research data in particular (as opposed, for instance, to research publications). However, there are a number of studies, particularly related to amounts of data, that go beyond the scope of research data). These were included because they are significant works in the field (e.g., Gantz 2007) and because we viewed aspects of the methodology and overall area of study as highly relevant to characterizing and understanding the stewardship gap.

indicates the number of studies that have articulated metrics (grey indicates no studies were found, blue indicates one to four studies, red indicates five to nine studies, and yellow indicates 10-14 studies). Figure 3 does not show the impact or influence that certain studies may have had. A dynamic version of the visualization is available from the Stewardship Gap Project website (http://bit.ly/stewardshipgap).

The studies we examined were limited for the most part to those dealing explicitly with research data (as opposed, e.g., to preservation of digitized cultural materials), plus a few others. These include studies that investigated the total amount of digital information,[18] studies targeted toward digital curation skills broadly (but that include consideration for research data)[19] and some studies that investigated public sector or government information.[20]

No measurement studies were found in the following areas (and thus neither measurement or metrics studies are represented in Figure 3): Tradeoffs between data management for the short or long term; Lack of critical mass for collaboration; Support structures for collaboration; Duration of commitment (one metrics study); Archive mandates and objectives; Provenance and authenticity; Reuse possibilities; Imbalance in funding; Unequal access to resources and expertise; Different timescales of infrastructure development and maturity.

[18] E.g., Lyman and Varian 2000 and 2003, Gantz 2007, Manyika 2011 and others.
[19] E.g., the DigCurV project (Cirrinnà et al. 2013) and the DigCCurr initiative (Hank et al. 2010).
[20] Vickery 2011, Ubaldi 2013.

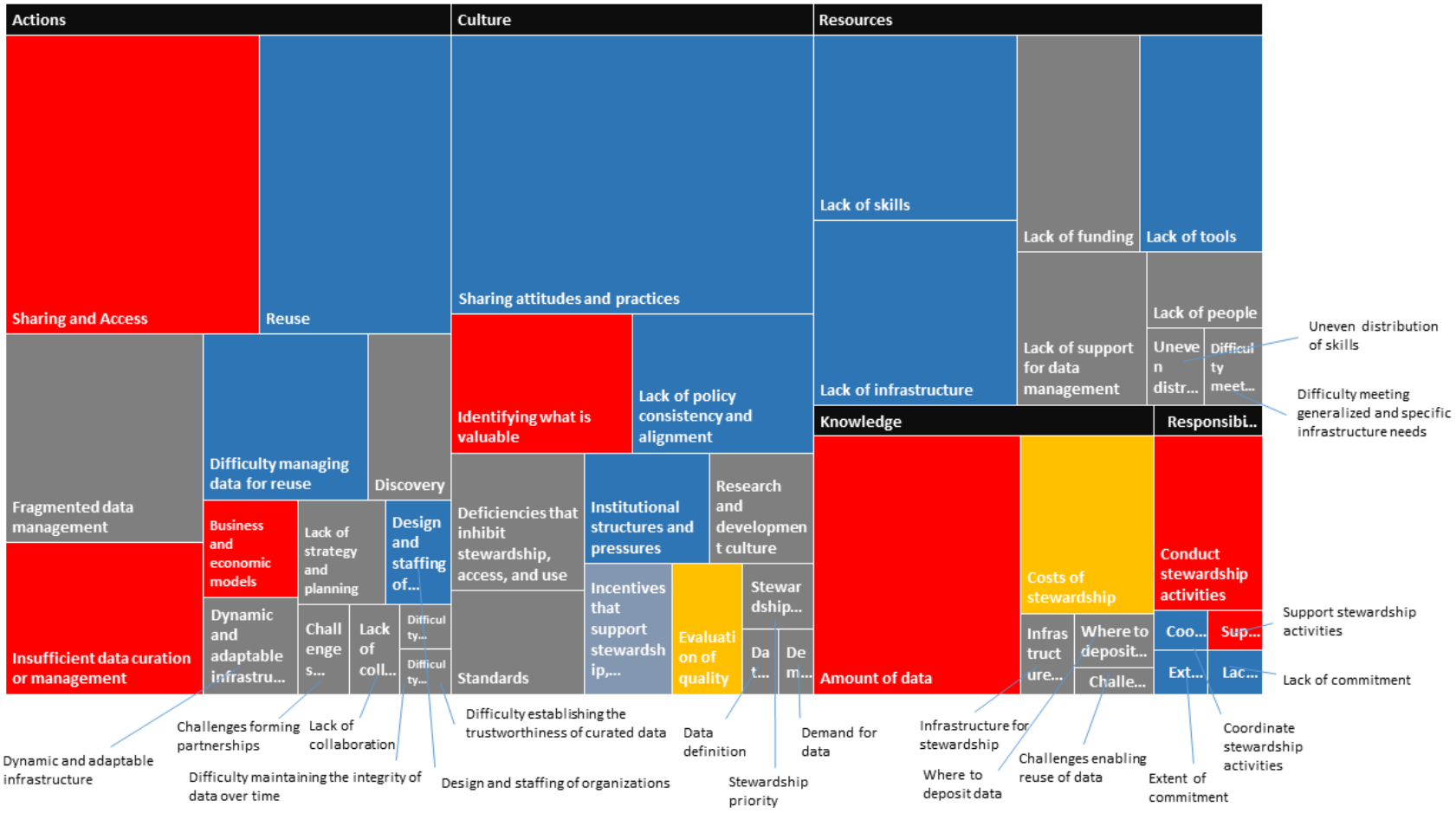

Figure 3. Distribution of studies relevant to the stewardship gap

Results

Many stories could be told from the results presented in Figure 3. The results most pertinent from the perspective of measuring the stewardship gap are imbalances and differences in the numbers and types of studies in the different gap areas. These are, more specifically:

- Imbalances in the attention given to different gap areas
- Imbalances between the number of measurements and metrics studies
- Differences in the depth of investigation undertaken

Imbalances in attention to different areas

The size of the boxes in Figure 3 shows the differing amounts of attention that have been given to measuring aspects of the stewardship gap. The small amount of attention given to Commitment and Collaboration are particularly striking since, as discussed above, these are two areas where deficiencies or strengths have the greatest impact on other gap areas. The large number of studies on sharing and access issues (both those under Culture and Access) in comparison those on Planning, Legal and Policy issues, Funding, and issues having to do with Curation, Management, and Preservation, is also notable given the influence these areas have on data sharing (as shown in Figure 1).

Figure 3 also illustrates, by color, the differing attention given to metrics across the gap areas. Some of the most surprising results relate to areas where no metrics were found (thirty out of fifty-six areas). These include metrics for understanding when data are managed in a fragmented way, when there is a lack of strategy and planning, how to identify dynamic and adaptable infrastructure, how to determine whether data are discoverable, understanding a variety of issues related to collaboration (identifying critical mass for collaboration, challenges forming partnerships, evaluating support structures for collaboration, and identifying when there is a lack of collaboration), what constitutes adequate funding or staff support, how to identify different cultures of research and development, and several others. A lack of metrics in these areas may also indicate a lack of common targets for individuals or organizations to aim to achieve, and a deficiency in means of evaluating progress.

Imbalances in measurements and metrics studies

There are several areas where the contrast between measurement and metrics studies is particularly striking. These include metrics for sharing attitudes and practices (45 measurement studies to 2 articulating metrics), reuse of data (26 to 2), fragmented data management (21 to 0), lack of skills (19 to 4), lack of infrastructure (19 to 2), lack of tools (16 to 1), lack of funding (12 to 0), difficulty management data for reuse (14 to 2), lack of support for data management (10 to 0) and incentives and deficiencies in, and alignment among legal and policy issues (5 to 1, 10 to 0, and 11 to 3, for incentives, deficiencies, and alignment, respectively).

One of the common challenges mentioned in studies of stewardship gap areas is the

difficulty in applying findings from one study to another.[21] It is possible that a greater focus on metrics in the areas above, and generally in areas where metrics have been articulated but not widely agreed upon, would facilitate greater exchange of results and findings across disciplinary and other boundaries.

Differences in the depth of studies

Throughout our review it became clear that there were two general types of studies that were undertaken to measure stewardship areas. In one type, which we termed "targeted," the entire investigation was focused in one or two closely related areas.[22] The other, comprising "wider" studies, investigated several different gap areas at once, often in the context of a campus or in some cases national or international study designed to gather information to better understand issues in or improve services for research data management and stewardship.[23] There are 117 studies (including both measurement and metrics studies) of the 145 reviewed that are targeted, and 28 that are wider.

Wider studies are frequently in-depth for their intended purpose, but may ask only one or a few questions related to a given gap area (e.g., about how much data researchers store, attitudes towards sharing, or research data management skills) in the context of a broader survey. A raw count of studies can thus overestimate the depth of investigation that has occurred in a particular area. Table 1 shows stewardship gap areas with a large proportion of wider as opposed to targeted studies (see in particular the Wider column under Measurement).

	Measurement		
Gap SubArea	**Targeted**	**Wider**	**Total**
Fragmented data management	7	14	21
Lack of infrastructure	4	16	19
Lack of skills	4	15	19
Difficulty managing data for reuse	3	11	14
Insufficient data curation or management	4	14	18
Lack of funding	2	10	12
Lack of tools	4	12	16
Identifying what is valuable	4	7	11
Lack of support for data	0	10	10

[21] See Ashley 2012, Beagrie and Houghton 2014, Borgman et al. 2014.

[22] E.g., Akmon 2014, Atkins 2003, Ayris et al. 2010, Beagrie and Houghton 2013a and 2013b, Borgman et al. 2014, Cirrinnà et al. 2013 and many others.

[23] Some examples are Alexogiannopoulos et al. 2010, Gibbs 2009, Hoeflich Mohr et al. 2015, Jerrome and Breeze 2009, Martinez-Uribe 2009, Mitcham et al 2015, Open Exeter Project Team 2012, Parsons et al. 2013, Perry 2008, Peters and Dryden 2011, Thornhill and Palmer 2014, UNC-CH 2012, Kuipers and van der Hoeven 2009, Waller and Sharpe 2006.

	Measurement		
Gap SubArea	**Targeted**	**Wider**	**Total**
management			
Conduct stewardship activities	0	9	9
Deficiencies that inhibit stewardship, access, and use	0	10	10
Standards	1	7	8
Incentives that support stewardship, access, and use	0	5	5
Evaluation of quality	1	4	5
Lack of people	1	4	5
Lack of strategy and planning	1	3	4

Table 1. Gap areas with a large proportion of "wider" studies

The proportion of targeted versus wider studies is not indicated in Figure 3, but is an important factor in understanding the universe of research relevant to the stewardship gap. In sixteen of the forty-seven gap subareas where some type of study was found, two thirds or more of the measurement studies were not in-depth studies in that area.[24]

Conclusion

This paper has reported the results of our efforts to understand the nature and characteristics of the stewardship gap through a review of relevant literature. In the process of our review we came to understand that there is not a single stewardship gap, but rather numerous and diverse components that contribute to and influence whether research data are responsibly stewarded. We identified fourteen gap components or areas from the literature and the relationships between them. We further categorized these components into six major areas, Culture, Knowledge, Responsibility, Commitment, Resources, and Actions, and identified studies that had been conducted to measure or develop metrics in these areas and corresponding subareas. Our desire to measure the stewardship gap led us to focus on three primary results: imbalances in the attention given to different gap areas in the reviewed literature, imbalances in

[24] We do not assert a broad correlation between the depth of investigation and the breadth of the investigated areas. We did find a number of wider studies, however, that did not treat individual gap areas at depth and it is this finding we wish to represent. The degree of depth was particularly noticeable in the studies listed in Table 2. We did find some wider studies that covered topics in greater depth than many wider surveys, though they were rare. Some examples are 1) Tenpoir et al. 2012 whose investigation covered the following areas: Coordinate stewardship activities, Lack of strategy and planning, Lack of collaboration, Lack of people, Lack of skills, Lack of support for data management, Lack of infrastructure, and Lack of tools; and 2) Wynholds et al. 2011, whose investigation covered the following areas: Identifying what is valuable, Demand for data, Dynamic and adaptable infrastructure, Incentives that support stewardship, access, and use, Fragmented data management, Insufficient data curation or management.

the number of measurement versus metrics studies, and differences in the depth at which studies investigated gap areas.

Our review has shown the stewardship gap literature to be rich with descriptions of challenges to effective stewardship, but that measurement of those challenges is haphazard and imbalanced. At the same time, the literature is also rich with descriptions of the relationships between challenge or gap areas, and these relationships can provide guidance to institutions and organizations, acting individually or cooperatively, to prioritize and affect gap areas that are most relevant to their situations and needs. Some key questions going forward are:

- What strategies are most effective for addressing particular gaps or combinations of gaps, and over what timescales?
- How might these strategies differ depending on discipline, cultures of practice, or levels of knowledge, responsibility or commitment?
- How can we improve ongoing measurement and evaluation of gap areas to adjust strategies appropriately over time?
- How can we stay abreast of changes to the gap areas themselves to ensure meaningful and accurate measurement?

Appendix

The following table provides descriptions of identified stewardship gap areas and citations to the sources they were identified from.

Culture		
	Culture	**Description:** Gap arising from differences in attitudes, goals, practices, and priorities among disciplines and communities that impact data stewardship and reuse **Indicators:** Differences in attitudes and practices for sharing data,[25] priority and value of stewardship,[26] cultures of research and development,[27] definitions of data,[28] use of standards,[29] mandates and objectives of archives,[30] appraisal criteria (identifying what is valuable);[31] differing means of determining quality;[32] differing amounts of interest in, and demand for, reusing data[33]
	Legal/Policy	**Description:** Gap between current regulations and policies that govern data stewardship and reuse and those that would maximally facilitate stewardship and reuse **Indicators:** Lack of incentives to undertake activities that directly or indirectly support stewardship (such as data sharing, curation and preservation, planning and deposit,[34] experimentation,[35] partnership,[36] and investment in stewardship,[37]); deficiencies in policy and legal environments

[25] Fecher et al. 2015, NRC 2003, OECD 2015, Ubaldi 2013, NSF 2007, Wallis et al. 2013.
[26] Esanu 2004, NAS 2009, NRC 2003, Borgman 2015, NSF 2007, STC 2007, Wynolds et al. 2011, Pepe et al. 2014, Berman et al. 2010, Berman 2014.
[27] Borgman et al. 2014, Finholt and Birnholtz 2006, NSF 2007.
[28] Atkins et al. 2003, Ubaldi 2013, Esanu et al. 2004, Borgman 2012, Read et al. 2015, Sturges et al. 2015.
[29] Esanu 2004, Ubaldi 2013, OECD 2015, Borgman 2015; Lack of common (cross-domain) standards for annotation, description, quality, curation, formats, citation, infrastructure: NRC 2003, McDonough 2012, Reeves et al. 2015, STC 2007, Pepe et al. 2014, Science Staff 2011.
[30] Esanu et al. 2004.
[31] Esanu et al. 2004.
[32] Esanu et al. 2004.
[33] Borgman 2012, Wallis et al. 2013.
[34] Borgman 2012, Wallis et al. 2013, OECD 2015, ARL 2006, Thompson Reuters 2013, Manyika et al. 2011, The Economist 2010, Pienta et al. 2010, Hedstrom and Niu 2008, Lavoie 2003, Bradley 2005, Eakin n.d., Pepe et al. 2014.
[35] NSF 2007 (in the social science and humanities).
[36] Berman et al. 2010.
[37] Ubaldi 2013.

		(especially regarding privacy, confidentiality, security, and intellectual property) that inhibit data storage and sharing,[38] data reuse,[39] competition,[40] and innovation;[41] lack of consistency and alignment among regulations and policies from journals, funders, and the federal government regarding data sharing;[42] institutional structures and pressures that pit the creation of new data against preservation of existing data or otherwise curtail abilities to support stewardship[43]
Knowledge		**Description:** Gap between what is known and what needs to be known to effectively plan for and ensure effective data stewardship **Indicators:** Lack of knowledge about costs and benefits of curation, preservation and provision of access;[44] what data exist;[45] skills and qualifications for relevant jobs and education programs;[46] what will have value in the future and how long data should be retained;[47] preservation practices and infrastructure;[48] what data are reused, by whom, and for what purposes;[49] the challenges of enabling reuse;[50] the future of scholarly communication;[51] lack of awareness about where to deposit and possibilities of data reuse;[52] lack of information about provenance and authenticity that inhibits data curation;[53] lack of information about preservation commitments by repositories[54]
Responsibility		**Description:** Gap between who currently has responsibility

[38] NRC 2003, NSF 2007, Borgman 2015, Ubaldi 2013, OECD 2015.
[39] OECD 2015, Ubaldi 2013, Marchionini et al. 2012.
[40] OECD 2015, Manyika et al. 2011.
[41] Manyika et al. 2011, Ubaldi 2013.
[42] Sturges et al. 2013, Reeves et al. 2015, Borgman 2012, Ubaldi 2013, Thaesis and van der Hoeven 2010.
[43] Berman 2014, Ember and Hanisch 2013, NAS 2009, NSF 2007, Downs and Chen 2013, Edwards et al. 2013, Ubaldi 2013.
[44] Brown et al. 2015, Weigert 2015, Eakin et al., n.d., Lavoie 2006, Hendley 1998, Russell and Weinberger 2000, STC 2007, Ubaldi 2013, Houghton and Gruen 2014, Kupiainen 2015, Lyon 2007.
[45] Weigert 2015, Ubaldi 2013; Ubaldi speaks in particular about the data in possession by governments.
[46] Borgman 2015, Lynch 2008.
[47] Brown et al. 2015, Esanu et al. 2004, NAS 2009, Borgman 2015, Read et al. 2015, STC 2007. Note: in Figure 2, studies in this category have been recorded in the Culture category along with studies that deal with identifying what is valuable.
[48] STC 2007.
[49] Wynholds et al. 2012. Note: In Figure 2, these studies are listed under Reuse.
[50] Brown et al. 2015.
[51] NSF 2007.
[52] Sturges et al. 2015, Wallis et al. 2013.
[53] Tenopir et al 2015, Kuipers and van der Hoeven 2009, Thaesis and van der Hoeven 2010.
[54] Lavoie and Malpas 2015.

		for stewardship and who is best placed to steward data over time **Indicators:** Lack of clarity about who should maintain data over time,[55] and who should be responsible for funding and supporting stewardship activities (such as data storage, curation, quality control, preservation, sharing, and access);[56] lack of clarity about who should coordinate stewardship activities[57]
Commitment		**Description:** Gap between the stewardship commitments that exist on valuable data and the commitments necessary to ensure long-term preservation and access **Indicators:** Mismatches in the duration of funding (short) and preservation commitment (long);[58] difficulty of building institutional commitments;[59] differences between the scale of need for on-campus data curation and the commitment to address it[60]
Resources		
	Human Resources	**Description:** Gap between the human effort and skills needed to steward and make data accessible, and the effort and skilled workers that are available **Indicators:** Lack of skills for digital preservation, data management and curation, and reuse of data;[61] difficulty in retaining skilled workers; mismatch between job classifications like archivist and librarian, and the actual demands of the data world; [62] skills are unevenly distributed in organizations and across domains;[63] data curation is time consuming and competes with other priorities;[64] unequal access to the resources and expertise necessary to create and operate a digital data collection;[65] insufficient effort and support for data management;[66] challenges in designing and

[55] Berman et al. 2010, Sturges et al. 2015, Borgman 2015, Downs and Chen 2013, Thompson Reuters 2013.
[56] Berman 2010, Downs and Chen 2013, Thompson Reuters 2013.
[57] Berman et al. 2010, Berman 2014.
[58] Ember and Hanisch 2013, Kupiainen 2015, Dillo et al. 2015.
[59] Borgman 2015.
[60] Giarlo 2012 (commitment by academic libraries specifically).
[61] Brown et al. 2015, Downs and Chen 2013, Wallis et al. 2013, Eakin et al. n.d., OECD 2015, Ember and Hanisch 2013, Borgman 2015.
[62] Ember and Hanisch 2013.
[63] OECD 2015, Atkins et al. 2003.
[64] Pampel and Dallmeier-Tiessen 2014, Borgman 2015, Wallis et al. 2013, OECD 2015, McDonough 2012.
[65] NSB 2005.
[66] Wynholds et al. 2011, NAS 2009.

		staffing of organizations that work with faculty to provide access services;[67] lack of available staff to pursue repository certification[68]
	Infrastructure and Tools	**Description:** Gap between the infrastructure available to steward and reuse data and infrastructure needed to maximize stewardship and reuse capabilities **Indicators:** Lack of infrastructure for large-scale, long-term digital preservation;[69] lack of shared community resources to support scholarship;[70] difficulty in meeting both general and specialized or local needs for infrastructure;[71] insufficient infrastructure and support to adapt to new digital strategies and business models;[72] lack of uptake of developed tools;[73] lack of tools to facilitate open access to government data;[74] few robust systems for making decisions about what to preserve; [75] different pieces of infrastructure evolve on different time scales (e.g., technology fast, universities and publishers more slowly)[76]
	Funding	**Description:** Gap between the funding needed for effective stewardship and the funding available **Indicators:** Lack of funding for data stewardship;[77] mismatch between budget levels and funding to retain skilled people;[78] dearth of public funding; disproportionate private funding in the social sciences and humanities;[79] lack of funding for innovation;[80] threat of funding cuts[81]
Actions		
	Curation, Management,	**Description:** Gap between the ways data is managed and prepared for preservation and reuse and ways that would

[67] Lynch 2008.
[68] Edwards et al. 2013.
[69] Downs and Chen 2013, Kuipers and van der Hoeven 2009.
[70] Borgman 2015.
[71] Cummings et al. 2008, Atkins et al. 2003, Lynch 2008, Borgman 2015.
[72] Johnson et al. 2015.
[73] 4C Project Executive Briefing Note 2013.
[74] Ubaldi 2013.
[75] Berman et al. 2010.
[76] Borgman 2015.
[77] For libraries and museums (Downs and Chen 2013, Berman et al. 2010, Erway 2013), domain repositories (Ember and Hanisch 2013), and for general storage and preservation activities (Pepe et al. 2014, Ember and Hanisch 2013, Borgman 2015).
[78] Ember and Hanisch 2013, Borgman 2015.
[79] NSF 2007.
[80] Kupiainen 2015, Dillo et al. 2015.
[81] Pepe et al. 2014.

	and Preservation	maximize its potential for preservation and reuse **Indicators:** Data is managed in a fragmented way or not made available to be curated and preserved;[82] difficulty in managing data for reuse;[83] insufficient curation resulting in data that is difficult to use or unusable; [84] difficulty in establishing the trustworthiness of curated data;[85] tradeoffs between data management for immediate use or long-term preservation;[86] difficulty in maintaining the integrity of data over time[87] This gap includes attributes of the end results of curation, management, and preservation such as interoperability and quality that affect preservation and reuse of data.[88]
	Sustainability Planning	**Description:** Gap between planning that is done to ensure adequate resources for stewardship and the planning that is needed **Indicators:** Lack of business and economic models for stewardship and access;[89] shortcomings in the design of archives in light of continual change in stewardship landscape;[90] lack of strategy and planning[91]
	Collaboration	**Description:** Gap between the collaboration needed for effective stewardship and the collaboration that takes place **Indicators:** Lack of critical mass to develop and sustain shared data resources;[92] lack of institutional structures to ensure accountability, transparency, quality of data and

[82] Borgman 2015, Addis 2015, Pepe et al. 2014, Thaesis and van der Hoeven 2010, Gantz 2007.
[83] Berman 2008, Science Staff 2011, Esanu 2004, STC 2007, Borgman 2015, Rauber 2012.
[84] Sturges et al. 2015, Noorman et al. 2014, Rauber 2012, Eakin et al. n.d., Wynholds et al. 2011, Kuipers and van der Hoeven 2009.
[85] Rauber 2012.
[86] McDonough 2012.
[87] Podesta et al. 2014.
[88] Some specific gaps in interoperability include a lack of or low level of interoperability among data sources (Wynholds et al. 2011, Ubaldi 2013), and the phenomenon that interoperability among certain data and stakeholders of data can result in lack of interoperability with other data and stakeholders (Borgman 2015). Some gaps in quality include disparate procedures and metrics among archives for data quality (Esanu et al. 2004);[88] quality issues can be difficult to monitor; it can be difficult to trust data quality (Tenopir et al. 2015).
[89] Downs and Chen 2013, Thompson Reuters 2013, ARL 2006, Ubaldi 2013, OECD 2015, Brown et al. 2015, Berman 2014, Edwards et al. 2013, Borgman 2015.
[90] Esanu et al. 2004, Rosenthal et al. 2005 (cited in STC 2007), Borgman et al. 2014, Atkins et al. 2003, Burda and Teuteberg 2013, Palmer 2015, Downs and Chen 2013.
[91] Edwards et al. 2013, NSB 2005, Lynch 2008, Kuipers and van der Hoeven 2009, Ubaldi 2013 (Ubaldi is speaking about open government data in particular).
[92] Borgman 2015.

		responsibility in a collaborative context or to span needs for local and national resources and build innovate infrastructures;[93] lack of formal digital communities and platforms for online collaboration in the humanities;[94] managing needs for infrastructure and changing responsibilities of researchers, data managers, users, and funders can only be done collaboratively given the scale of the issues, the costs, and the interdependency involved;[95] lack of information sharing regarding the construction and operation of virtual organizations[96]
	Sharing and Access	**Description:** Gaps between the amount of data that are shared or made accessible and the amount of data that is not **Indicators:** Many indicators of a sharing gap carry over from other gap areas (for instance, lack of incentives and demand, challenges of dealing with issues of privacy, confidentiality, security, and intellectual property, and difficulties arising from legal and policy considerations). Indicators of an access gap similarly carry over from other areas (e.g., deficiencies in commitment and responsibility, insufficient infrastructure, human resources, or curation). The implication is that more sharing and provision of access might occur if there were not gaps in these other areas.
	Discovery	**Description:** Gap between the amount of accessible data that is discoverable and the amount that is not **Indicators:** Similar to sharing and access, indicators of a discovery gap arise from other areas, including infrastructure and tools to facilitate discovery, lack of data shared and able to be discovered, lack of sufficient metadata or curatorial activities that would facilitate discovery.
	Reuse	**Description:** Gap between the data that is available for reuse and the data that is used **Indicators:** Difficulty of reusing data (due to poor metadata quality,[97] lack of required expertise,[98] inability to assess

[93] Ubaldi 2013, Lynch 2008.
[94] NSF 2007.
[95] NSB 2005, OECD 2015, Borgman et al. 2014, Borgman 2015, Ubaldi 2013.
[96] Cummings et al. 2008.
[97] Eakin et al. n.d.
[98] Borgman 2012, Borgman 2015.

		integrity of and understand data,[99] challenges of certain approaches to reusing data, such as computational analysis,[100] reticence to trust data[101])

Table 1. Stewardship gap areas, descriptions, and indicators

[99] Borgman 2012, Wynholds et al. 2011.
[100] Borgman 2015.
[101] Wynholds et al. 2011, Tenopir et al. 2015.

References

4C Project. (2013). *4C Project: Executive Briefing Note*. Retrieved from http://4cproject.eu/component/docman/doc_download/2-march-2013-briefing-note?Itemid=

Addis, M. (2015). *Estimating Research Data Volumes in UK HEI*. Retrieved from http://figshare.com/articles/Estimating_Research_Data_Volumes_in_UK_HEI/1575831

Akmon, D. (2014). *The Role of Conceptions of Value in Data Practices: A Multi-Case Study of Three Small Teams of Ecological Scientists*. University of Michigan, Ann Arbor, MI.

Alexogiannopoulos, E., McKenney, S., & Pickton, M. (2010). *Research Data Management Project: a DAF investigation of research data management practices at The University of Northampton*. University of Northampton. Retrieved from http://nectar.northampton.ac.uk/2736/

Ashley, K. (2012). Generic Data Quality Metrics – what and why. Presented at the Prepared for the NSF III #1247471 workshop, Curating for Quality: Ensuring Data Quality to Enable New Science, Arlington, VA. Retrieved from http://dl.acm.org/citation.cfm?id=2582001

Association of Research Libraries Workshop on New Collaborative Relationships. (2006). *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering* (A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe). Arlington, VA. Retrieved from http://www.arl.org/storage/documents/publications/digital-data-report-2006.pdf

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-molina, H., Klein, M. L., Messerschmitt, D. G., … Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on*. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf

Ayris, P., Wheatley, P., Aitken, B., Hole, B., McCann, P., Peach, C., & Lin, L. (2010). *The LIFE3 Project: Bringing digital preservation to LIFE*. Retrieved from http://www.life.ac.uk/3/docs/life3_report.pdf

Beagrie, N., & Houghton, J. (2013a). *The Value and Impact of the Archaeology Data Service: A Study and Methods for Enhancing Sustainability*. Bristol and London: Joint Information Systems Committee. Retrieved from http://repository.jisc.ac.uk/5509/1/ADSReport_final.pdf

Beagrie, N., & Houghton, J. (2013b). *The Value and Impact of the British Atmospheric Data Centre*. Bristol and London: Joint Information Systems Committee. Retrieved from http://repository.jisc.ac.uk/5382/1/BADCReport_Final.pdf

Beagrie, N., & Houghton, J. (2014). *The Value and Impact of Data Sharing and Curation*. Retrieved from http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf

Beagrie, N., Houghton, J., Palaiologk, A., & Williams, P. (2012). *Economic Evaluation of Research Data Infrastructure*. London: Economic and Social and Research Council. Retrieved from http://www.esrc.ac.uk/files/research/evaluation-and-impact/economic-impact-evaluation-of-the-economic-and-social-data-service/

Beagrie, N., Lavoie, B., & Woollard, M. (2010). *Keeping Research Data Safe 2*. Charles Beagrie Limited. Retrieved from

http://www.webarchive.org.uk/wayback/archive/20140615221405/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Becker, C., Antunes, G., Barateiro, J., & Vieira, R. (2011). A Capability Model for Digital Preservation: Analysing Concerns, Drivers, Constraints, Capabilities and Maturities. Presented at the 8th International Conference on Preservation of Digital Objects (IPRES 2011), Singapore. Retrieved from https://www.researchgate.net/publication/235922140_A_Capability_Model_for_Digital_Preservation_Analysing_Concerns_Drivers_Constraints_Capabilities_and_Maturities

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

Berman, F. (2010). We need a research data census, *53*(12), 39. http://doi.org/10.1145/1859204.1859220

Berman, F. (2014). Despite Growing Data, Infrastructure Stands Still – Why the gap puts research data at risk. *IEEE*. Retrieved from http://theinstitute.ieee.org/ieee-roundup/opinions/ieee-roundup/despite-growing-data-infrastructure-stands-still

Berman, F., Lavoie, B., Ayris, P., Choudhury, G. S., Cohen, E., Courant, P., … Van Camp, A. (2010). Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information, 110.

Big Data Value Association. (2015). *European Big Data Value Strategic Research & Innovation Agenda Big Data Value Strategic Research & Innovation Agenda*. Brussels, Belgium: Big Data Value Europe. Retrieved from http://www.bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final.pdf

Borgman, C. L. (2012). The conundrum of sharing research data, *63*(6), 1059–1078. http://doi.org/10.1002/asi.22634

Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world*.

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. *Proceedings of the Joint Conference on Digital Libraries, 2014 (DL2014)*. Retrieved from http://works.bepress.com/borgman/321

Bradley, K. (2005). *APSR Sustainability Issues Discussion Paper*. National Library of Australia. Retrieved from http://apsr.anu.edu.au/documents/APSR_Sustainability_Issues_Paper.pdf

Brown, S., Bruce, R., & Kernohan, D. (2015, March 30). Directions for Research Data Management in UK Universities [Publication]. Retrieved September 3, 2015, from http://repository.jisc.ac.uk/5951/

Cirrinnà, C., Fernie, K., & Lunghi, M. (2013). *Digital Curator Vocational Education Europe (DigCurV): Final report and Conference Proceedings*. Retrieved from http://www.digcur-education.org/eng/Events/Final-Report-and-Conference-Proceddings

Climatic Research Unit email controversy. (2015, September 30). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Climatic_Research_Unit_email_controversy&oldid=683477293

Cummings, J., Finholt, T., Foster, I., Kesselman, C., & Lawrence, K. A. (2008). Beyond Being

There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations. *Technology*, *3*(2). Retrieved from http://web.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

Dillo, I., Hodson, S., & de Waard, A. (2015). *Income Streams for Data Repositories*. White paper by RDA Interest Group on Publishing Data Cost Recovery for Data Centres.

Downs, R., Duerr, R., Hills, D. J., & Ramapriyan, H. K. (2015). Data Stewardship in the Earth Sciences. *D-Lib Magazine*, *21*(7/8). http://doi.org/10.1045/july2015-downs

Downs, R. R., & Chen, R. S. (2013). Towards Sustainable Stewardship of Digital Collections of Scientific Data. Presented at the GSDI World Conference (GSDI 13) Proceedings, Quebec City, Canada. Retrieved from http://www.gsdi.org/gsdiconf/gsdi13/papers/130.pdf

Eakin, L., Friedlander, A., Schonfeld, R., & Choudhury, G. S. (n.d.). *A Selective Literature Review on Digital Preservation Sustainability*. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjkpsvSlKDKAhUCcD4KHa6tCA0QFggdMAA&url=http%3A%2F%2Fbrtf.sdsc.edu%2Fbiblio%2FCost_Literature_Review.pdf&usg=AFQjCNHHvphn-rWRt96eruRL8XSuixkEoQ

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges* (Working Paper). Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97552

Ember, C., & Hanisch, R. (2013). *Sustaining Domain Repositories for Digital Data*. Retrieved from http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

Engineering and Physical Sciences Research Council. (2011). EPSRC policy framework on research data. Retrieved September 3, 2015, from https://www.epsrc.ac.uk/about/standards/researchdata/

Erway, R. (2013). *Starting the Conversation: University-wide Research Data Management Policy*. Dublin, Ohio: OCLC Research. Retrieved from http://www.oclc.org/research/publications/library/2013/2013-08r.html

Esanu, E., Davidson, J., Ross, S., & Anderson, W. (2004). Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ERPANET/CODATA Workshop. *Data Science Journal*, *3*(December 2003), 226–232. http://doi.org/10.2481/dsj.3.227

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *PLoS ONE*, *10*(2), e0118053. http://doi.org/10.1371/journal.pone.0118053

Finholt, T. A., & Birnholtz, J. P. (2006). If We Build It, Will They Come? The Cultural Challenges of Cyberinfrastructure Development, 89–101. http://doi.org/10.1007/1-4020-4107-1_7

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J. W., & Rasmussen, B. (2008). *Identifying benefits arising from the curation and open sharing of research data produced within UK Higher Education and research institutes: exploring costs and benefits*. Loughborough University, Victoria University. Retrieved from http://www.webarchive.org.uk/wayback/archive/20140615075522/http://www.jisc.ac.uk/med

ia/documents/programmes/digitalrepositories/jiscdataproposal-public.pdf

Gantz, J. F., McArthur, J., Minton, S., Reinsel, D., Chute, C., Schlichting, W., … Manfrediz, A. (2007). The Expanding Digital Universe [White Paper]. http://doi.org/10.1002/humu.21252

Gantz, J. F., Minton, S., Reinsel, D., Chute, C., Schlichting, W., Toncheva, A., & Manfrediz, A. (2008). The Diverse and Exploding Digital Universe [White Paper]. http://doi.org/10.1002/humu.21252

Giarlo, M. J. (2012). Academic Libraries as Data Quality Hubs. In *Curating for Quality: Ensuring Data Quality to Enable New Science* (pp. 101–106). Arlington, VA. Retrieved from http://dl.acm.org/citation.cfm?id=2582001

Gibbs, H. (2009). *Southampton Data Survey: Our Experience and Lessons Learned*. University of Southampton. Retrieved from http://www.disc-uk.org/docs/SouthamptonDAF.pdf

Hank, C., Tibbo, H. R., & Lee, C. A. (2010). *DigCCurr I Final Report, 2006-09*. Retrieved from http://www.ils.unc.edu/digccurr/digccurr_I_final_report_031810.pdf

Harvey, R. (2008). *Appraisal and Selection* (Briefing Papers: Introduction to Curation). Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/appraisal-and-selection

Hedstrom, M., & Niu, J. (2008). Incentives for Data Producers to Create "Archive-Ready" Data: Implications for Archives and Records Management. Presented at the Society of American Archivists Research Forum. Retrieved from http://files.archivists.org/conference/2008/researchforum/M-HedstromJ-Niu-SAA-ResearchPaper-2008.pdf

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, *332*(6025), 60–65. http://doi.org/10.1126/science.1200970

Hofelich Mohr, A., Bishoff, J., Johnston, L., Braun, S., Storino, C., & Bishoff, C. (2015). Data Management Needs Assessment - Surveys in CLA, AHC, CSE, and CFANS. Retrieved from http://conservancy.umn.edu/handle/11299/174051

Holdren, J. P. (2013). *Increasing Access to the Results of Federally Funded Scientific Research* (Executive Office of the President Office of Science and Technology Policy Memo). Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Jerrome, N., & Breeze, J. (2009). *Imperial College Data Audit Framework Implementation: Final Report*. Imperial College London. Retrieved from http://ie- repository.jisc.ac.uk/307/

Johnson, L., Adams Becker, S., Estrada, V., & Freeman, A. (2015). *NMC Horizon Report: 2015 Library Edition*. Austin, Texas: The New Media Consortium. Retrieved from http://www.nmc.org/publication/nmc-horizon-report-2015-library-edition/

Kalil, T., & Miller, J. (2015, July 29). Advancing U.S. Leadership in High-Performance Computing. Retrieved from https://www.whitehouse.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing

Kuipers, T., & van der Hoeven, J. (2009). *PARSE.Insight: Insight into Digital Preservation of Research Output in Europe: Survey Report*. Retrieved from http://www.parse-

insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Kupiainen, I. (2015, March). *Cost Recovery Models - RDA/WDS IG*. Presented at the Research Data Alliance Plenary 5, San Diego, CA.

Lavoie, B. (2003). *The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making*. Dublin, Ohio: OCLC Online Computer Library Center, Inc. Retrieved from http://www.oclc.org/content/dam/research/activities/digipres/incentives-dp.pdf

Lavoie, B., & Malpas, C. (2015). *Stewardship of the Evolving Scholarly Record: From the Invisible Hand to Conscious Coordination*. Dublin, Ohio: OCLC Research. Retrieved from http://www.oclc.org/research/publications/2015/oclcresearch-esr-stewardship-2015.html

Lesk, M. (1997). How Much Information Is There In the World? Retrieved October 26, 2015, from http://www.lesk.com/mlesk/ksg97/ksg.html

Lyman, P., & Varian, H. R. (2000). How Much Information? *Journal of Electronic Publishing*, *6*(2). Retrieved from http://groups.ischool.berkeley.edu/archive/how-much-info/summary.html

Lyman, P., & Varian, H. R. (2003). *How Much Information? 2003*. Retrieved from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Lynch, C. (2008, October 23). The Institutional Challenges of Cyberinfrastructure and E-Research. *Educause Review*, *43*(6). Retrieved from http://www.educause.edu/ero/article/institutional-challenges-cyberinfrastructure-and-e-research

Lyon, L. (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships: Consultancy Report*. University of Bath. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Manyika, J., Byers, A. H., Chui, M., Brown, B., Bughin, J., Dobbs, R., … McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity, 156.

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., Doshi, E. A., & McKinsey Global Institute. (2013). Open data: Unlocking innovation and performance with liquid information, 103.

Marchionini, G., Lee, C. A., Bowden, H., & Lesk, M. (2012). Curating for Quality: Ensuring Data Quality to Enable New Science, 119.

Martinez-Uribe, L. (2009). *Using the Data Audit Framework: An Oxford Case Study*. Retrieved from http://www.disc-uk.org/docs/DAF-Oxford.pdf

McDonough, J. (2012). Start Making Sense: Quality, Context & Meaning. Presented at the Prepared for the NSF III #1247471 workshop, Curating for Quality: Ensuring Data Quality to Enable New Science, Arlington, VA. Retrieved from http://dl.acm.org/citation.cfm?id=2582001

Mitcham, J., Awre, C., Allinson, J., Green, R., & Wilson, S. (2015). *Filling the Digital Preservation Gap. A JISC Research Data Spring Project. Phase One Report*. Retrieved from http://figshare.com/articles/Filling_the_Digital_Preservation_Gap_A_Jisc_Research_Data_Spring_project_Phase_One_report_July_2015/1481170

National Academy of Sciences. (2009). Ensuring the integrity, accessibility, and stewardship of research data in the digital age, *325*(5939), 368. http://doi.org/10.1126/science.1178927

National Aeronautics and Space Administration. (n.d.). Data & Information Policy. Retrieved December 10, 2015, from http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/

National Research Council. (2003). *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/10613

National Research Council. (2007). *Environmental Data Management at NOAA: Archiving, Stewardship, and Access*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/12017

National Science Board. (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

National Science Foundation, Cyber Infrastructure Council. (2007). Cyberinfrastructure Vision for 21st Century Discovery. *Director*, (March). Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

Noorman, M., Kalaitzi, V., Angelaki, M., Tsoukala, V., Linde, P., Sveinsdottir, T., … Wessels, B. (2014). *Institutional barriers and good practice solutions*. Policy RECommendations for Open access to research Data in Europe (RECODE). Retrieved from http://recodeproject.eu/wp-content/uploads/2014/09/RECODE-D4.1-Institutional-barriers-FINAL.pdf

Obama, B. (2009). *A Strategy for American Innovation: Driving Towards Sustainable Growth and Quality Jobs*. Retrieved from https://www.whitehouse.gov/administration/eop/nec/StrategyforAmericanInnovation/

Obama, B. (2011). *A Strategy for American Innovation: Securing Our Economic Growth and Prosperity*. Retrieved from https://www.whitehouse.gov/administration/eop/nec/StrategyforAmericanInnovation/

Obama, B. (2013a). *Executive Order -- Making Open and Machine Readable the New Default for Government Information*. The White House. Retrieved from https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Obama, B. (2013b, May 9). Open Data Policy-Managing Information as an Asset. Retrieved from https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

Office of Management and Budget. (2012). *Digital Government: Building a 21st Century Platform to Better Serve the American People*. Retrieved from https://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html

Open Exeter Project Team. (2012). *Summary Findings of the Open Exeter Data Asset Framework Survey*. Exeter, UK: University of Exeter. Retrieved from https://ore.exeter.ac.uk/repository/bitstream/handle/10036/3689/daf_report_public.pdf?sequence=1

Organization for Economic Co-operation and Development. (2015). *Making Open Science A*

Reality. Retrieved from https://www.innovationpolicyplatform.org/content/open-science

Pampel, H., & Dallmeier-Tiessen, S. (2014). Open Research Data: From Vision to Practice. In *Opening Science* (pp. 213–224). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-00026-8_14

Parsons, T., Grimshaw, S., & Williamson, L. (2013). *Research Data Management Survey*. University of Nottingham. Retrieved from http://admire.jiscinvolve.org/wp/files/2013/02/ADMIRe-Survey-Results-and-Analysis-2013.pdf

Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S. (2015). A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets. *Data Science Journal*, *13*(0). http://doi.org/10.2481/dsj.14-049

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

Perry, C. (2008). Archiving of publicly funded research data: A survey of Canadian researchers. *Government Information Quarterly*, *25*(1), 133–148.

Peters, C., & Dryden, A. (2011). Assessing the Academic Library's Role in Campus-Wide Research Data Management: A First Step at the University of Houston. *Science & Technology Libraries*, *30*(4), 387–403. http://doi.org/10.1080/0194262X.2011.626340

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/78307

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J., & Zients, J. (2014). *Big Data: Seizing Opportunities, preserving values*. Executive Office of the President. Retrieved from https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Rauber, A. (2012). Digital Preservation in Data-Driven Science: On the Importance of Process Capture, Preservation and Validation. In *Proceedings of the 2nd International Workshop on Semantic Digital Archives (SDA 2012)*. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjj19aUm6DKAhUGNSYKHdG8AC4QFggdMAA&url=http%3A%2F%2Fceur-ws.org%2FVol-912%2Fpaper0.pdf&usg=AFQjCNEhDcKIh9quVYWf07HsgUg3jQwAXA

Read, K. B., Sheehan, J. R., Huerta, M. F., Knecht, L. S., Mork, J. G., Humphreys, B. L., & NIH Big Data Annotator Group. (2015). Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. *PLoS ONE*, *10*(7), e0132735. http://doi.org/10.1371/journal.pone.0132735

Reeves, J., Brodeur, J. J., Daniels, M. G., Nicholls, N., & Turnator, E. (2015). Libraries and the Research Data Management Landscape. In *The Process of Discovery: The CLIR Postdoctoral Fellowship Program and the Future of the Academy* (pp. 82–102). Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub167/

Research Councils UK. (2015, July). RCUK Common Principles on Data Policy. Retrieved

September 3, 2015, from http://www.rcuk.ac.uk/research/datapolicy/

Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V., & Morabito, S. (2005). Requirements for Digital Preservation Systems: A Bottom-Up Approach. *D-Lib Magazine*, *11*(11). http://doi.org/10.1045/november2005-rosenthal

Science and Technology Council. (2007). The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials. *Academy of Motion Picture Arts and Sciences*. Retrieved from http://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma

Science Staff. (2011). Challenges and Opportunities. *Science*, *331*(6018), 692–693.

Sturges, P., Bamkin, M., Anders, J. H. S., Hubbard, B., Hussain, A., & Heeley, Melanie. (2015). Research data sharing: developing a stakeholder-driven model for journal policies. *Journal of the Association for Information Science and Technology*. http://doi.org/10.1002/asi.23336

Sveinsdottir, T., Wessels, B., Smallwood, R., Linde, P., Kala, V., Tsoukala, V., & Sondervan, J. (2013). *Stakeholder values and relationships within open access and data dissemination and preservation ecosystems*. Policy RECommendations for Open access to research Data in Europe (RECODE). Retrieved from http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf

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

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE*, *10*(8), e0134826. http://doi.org/10.1371/journal.pone.0134826

Thaesis, & van der Hoeven, J. (2010). *PARSE.Insight: Insight into issues of Permanent Access to the Records of Science in Europe. Final Report*. Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

The Economist. (2010). *Data, data, everywhere*. Retrieved from https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf

The Royal Society. (2012). *Science as an Open Enterprise*. London. Retrieved from http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE-Summary.pdf

Thompson Reuters. (2013). *Unlocking the Value of Research Data: A Report from the Thompson Reuters Industry Forum*. Retrieved from http://researchanalytics.thomsonreuters.com/m/pdfs/1003903-1.pdf

Thornhill, K., & Palmer, L. (2014). An Assessment of Doctoral Biomedical Student Research Data Management Needs. Retrieved from http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1075&context=escience_symposium

Turner, V., Reinsel, D., Gantz, J. F., & Minton, S. (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. Retrieved from http://idcdocserv.com/1678

Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government

Data Initiatives. *OECD Publishing*, (No. 22). http://doi.org/http://dx.doi.org/10.1787/5k46bj4f03s7-en

University of North Carolina Chapel Hill. (2012). *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership*. Chapel Hill, North Carolina: University of North Carolina Chapel Hill. Retrieved from http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf

Vickery, G. (2011). *Review of Recent Studies on PSI Re-use and Related Market Developments*.

Vickery, G. (2012). *Review of Recent Studies on PSI Re-use and Related Market Developments (revised and abridged)*. Retrieved from http://www.scb.se/statistik/_publikationer/NR9999_2012A01_BR_X76BR1201.pdf

Vogel, G. (2011, October 31). Report: Dutch "Lord of the Data" Forged Dozens of Studies (UPDATE). *Science Insider*. Retrieved from http://news.sciencemag.org/europe/2011/10/report-dutch-lord-data-forged-dozens-studies-update

Waller, M., & Sharpe, R. (2006). *Mind the gap: Assessing digital preservation needs in the UK*. York, United Kingdom: Digital Preservation Coalition. Retrieved from http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk

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. http://doi.org/10.1371/journal.pone.0067332

Wang, D., Strodl, S., Kejser, U. B., Ferreira, M., Borbinha, J., Proença, D., … Stokes, P. (2015). *4C Project: From Costs to Business Models*. Retrieved from http://www.4cproject.eu/d4-5-from-costs-to-business-models

Weigert, V. (2015, April 23). Supporting universities to implement effective research data management. Retrieved from http://www.researchinformation.info/news/news_story.php?news_id=1893

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE*, *6*(11), e26828. http://doi.org/10.1371/journal.pone.0026828

Willetts, D., Livanov, D., Schütte, G., Harayama, Y., Carrozza, M. C., Goodyear, G., … Geoghegan-Quinn, M. (2013). *G8 Science Ministers Statement*. London, UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). 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). New York, NY, USA: ACM. http://doi.org/10.1145/2232817.2232822

Wynholds, L., Fearon, D. S., Jr., Borgman, C. L., & Traweek, S. (2011). When Use Cases Are Not Useful: Data Practices, Astronomy, and Digital Libraries. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 383–386). New York, NY, USA: ACM. http://doi.org/10.1145/1998076.1