

RESEARCH PAPER

What Do We Know about the Stewardship Gap

Jeremy York¹, Myron Gutmann² and Francine Berman³¹ University of Michigan, US² University of Colorado Boulder, US³ Rensselaer Polytechnic Institute, USCorresponding author: Jeremy York (jjyork@umich.edu)

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>) investigates characteristics of, and measures, 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 various stewardship gaps; and 3) there are imbalances in the types and depths of studies that have been conducted to measure the stewardship gap.

Keywords: data stewardship; digital curation; research data; digital preservation; data sustainability

Introduction

In the 21st century, digital data drive innovation and decision-making in nearly every field (Big Data Value Association (BDVA) 2015, Holdren 2013, Houghton and Gruen 2014, Kalil and Miller 2015, Manyika et al. 2011, Manyika et al. 2013, Obama 2013, Podesta et al. 2014, Science and Technology Council 2007, Vickery 2011). 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 concerns about privacy and security, and how much value can be gained by reusing data (Cummings et al. 2008, Manyika et al. 2013, NSF 2007, Obama 2009 and 2011, Office of Management and Budget (OMB) 2012, Thompson Reuters 2013, Vickery 2011, Vickery 2012). These questions are being raised in both the public and private sectors, where stakeholders increasingly see data as an asset that can be leveraged to spur creativity, innovation and economic growth, as well as to increase trust (e.g., in government or the results of scientific research) (Association of Research Libraries 2006, BDVA 2014, Berman et al. 2010, Borgman 2012, National Academy of Sciences (NAS) 2009, National Research Council (NRC) 2003, NSF 2007, Organization for Economic Co-operation and Development (OECD) 2015, Podesta et al. 2014, Sveinsdottir et al. 2013, Tenopir et al. 2011, The Royal Society 2012, Ubaldi 2013, Wallis et al. 2013).

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” or “research data”), whether publicly or privately funded. Demand for access to the former is driven by public interest and opportunity for public benefit (Podesta et al. 2014, Borgman 2012, Holdren 2013, Manyika et al. 2013, Obama 2013, OMB 2012, Ubaldi 2013, Vickery 2011, Vickery 2012). 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 (Borgman 2012, Holdren 2013, NAS 2009, NRC 2003, OECD 2015, Research Councils UK (RCUK) 2015, Royal Society 2012, Tenopir et al. 2011). Demand is also driven by a desire for increased reproducibility and accountability. Several high profile cases of a lack of supporting data or suspected or actual fraud have brought greater scrutiny to the availability of data to replicate and verify research results (see for example Climatic Research Unit email controversy 2015, Vogel 2011, Wicherts et al. 2011).

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 a 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 (Holdren 2013, National Aeronautics and Space Administration n.d., Obama 2013a, Obama 2013b, OECD 2007, OMB 2002, OMB 2013, RCUK 2015, The Royal Society 2012, Willetts et al. 2013).

While there is general agreement about the actions that must be taken and roles that must be played to steward research data, there is a lack of clarity about who should have responsibility for fundamental aspects of stewardship, and different understandings of what constitutes effective stewardship. This contributes to a fractured and diffuse environment for stewardship (Wynholds et al. 2012, Borgman 2015; see also Pepe et al. 2014, ARL 2006, Borgman 2012, Borgman 2015, Downs and Chen 2013, Esanu et al. 2004, Thaesis and van der Hoeven 2010, Thaesis and van der Hoeven 2010, Berman et al. 2010). In fact, despite the large number of data repositories, stewardship initiatives, and policies across the research data landscape, we know relatively little about the total amount, characteristics, or sustainability of stewarded research data (Berman 2008, Berman 2014, Gantz et al. 2008, Hilbert and López 2011, Pienta 2006, STC 2007, Turner et al. 2014).

What we do know gives us pause. For instance, in 2015, Read et al. conducted a study of the number of datasets mentioned in journal articles resulting from National Institutes of Health (NIH)-funded research that were deposited in “well-known, publicly-accessible data repositories” (Read et al., 2015, p. 3). They found mentions of such deposit in only 12% of published articles. Based on the number of datasets Read et al. identified in articles where no deposit of datasets was mentioned, they estimated that between 200,000 and 235,000 datasets resulting from NIH-funded research in 2011 were “invisible” (not found in one of the well-known repositories). The PARSE.Insight project found similar results in its wide-ranging study to “gain insight into the practices, needs and requirements of research communities” (Kuipers and van der Hoeven, 2009, p. 9). It found in a survey of more than 1,300 researchers across multiple disciplines that only 20% of respondents deposited data in a digital archive (Thaesis and van der Hoeven, 2010).

This research raises important questions. Are stewardship arrangements sufficient? Do researchers, research sponsors, and research institutions adequately understand what they need to do? Are public policies appropriate? These are questions worth answering.

The starting point for answering these questions is a very substantial published literature about research data stewardship. In this article we explore what we know about research data stewardship through the lens of that literature, allowing us to characterize the important questions that previous researchers have asked. It also allows us to show areas that will require additional research in the future. We return to those

unanswered questions at the end of this article, so that we can propose valuable lines of future research that need to be explored.

Data Collection and Analysis

Our literature review explores three different samples of literature, which we used to conduct the different analysis presented in the paper. For the purposes of developing our samples, we defined data stewardship according to the National Academy of Sciences definition provided above. Also as above, we defined “sponsored research data” or “research data” as data used in or produced as part of sponsored scholarly inquiry, whether publicly or privately funded.

The first sample (Sample A) is a body of 87 works, including literature reviews, reports, and empirical research that we analyzed to discover what scholars and practitioners identify as challenges to data stewardship. A list of these works can be found in York et al. (2018b).

We conducted descriptive coding of this sample, from which we identified three levels of stewardship gap “areas” and “sub-areas.” At the highest level we defined six gap areas. We arrived at these by distilling 14 broader gap areas, which we in turn aggregated from 56 more granular gap sub-areas. The areas and sub-areas are described below. The different levels of gap areas and subareas, as well as the papers we identified them from are available to browse at York et al. (2018d and 2018e).

The second sample (Sample B) comprises 74 works selected out of the 87 works from Sample A. These are listed at York et al. (2018c). In addition to identifying challenges to data stewardship, the authors of these 74 works also identified relationships between the challenges (e.g., challenges that cause or exacerbate others). The data in **Figures 1** and **2** refer to this sample.

The third sample (Sample C) is a set of 142 works, some of which are included in Sample A and Sample B, that explicitly seek to measure stewardship gap areas and sub-areas, or articulate metrics for measuring them. These are listed at York et al. (2018d). Sample C excludes reports and other works that, for instance, articulate strategies or ideas for addressing stewardship challenges but do not conduct empirical research to measure at least one of our identified gap areas, or theoretical research to identify what might be measured.

We limited works in Sample C for the most part to those dealing explicitly with research data (as opposed, for example, to preservation of digitized cultural materials), though there were a few others. These include studies that investigated the total amount of digital information (e.g., Lyman and Varian 2000 and 2003, Gantz et al. 2007, Manyika 2011), studies targeted toward digital curation skills broadly (but that include consideration for research data) (e.g., Cirinnà et al. 2013, Hank et al. 2010) and some studies that investigated public sector or government information (e.g., Ubaldi 2013, Vickery 2011).

We conducted initial stages of coding using a combination of spreadsheets and the Web-based tool Workflowy. We subsequently kept track of article codes using spreadsheets and a Web-based database platform (Drupal) where data from the project are available (see <http://www.stewardshipgap.net>; for data in tabular form see York et al. 2018a).

We identified the works in all three samples through a variety of methods, including searching for topics related to stewardship and curation in and across databases (e.g., using services such as Google Scholar and cross-database aggregation services such as Summon), and analyzing cited references in relevant articles, reports, and projects. The works have a geographic bias towards North America and Europe and are biased as well to those in English. We describe our analyses using these samples below.

Defining the Stewardship Gap

Identifying Gap Areas

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 (described below), no community-wide metrics for or measurements of the stewardship gap as a whole exist (one method for identifying the existence of a stewardship gap is described in York et al. 2016).

Measuring the stewardship gap is complex not only because it is difficult to measure the amount of sponsored research data that exist, 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.

Table 1: Stewardship gap areas, descriptions.

Gap Area	Description
Culture	Gap arising from differences in attitudes, goals, practices, and priorities among disciplines and communities that have an impact on data stewardship and reuse
Legal/Policy	Gap between current regulations and policies that govern data stewardship and reuse and those that would maximally facilitate stewardship and reuse
Knowledge	Gap between what is known and what needs to be known to effectively plan for and ensure effective data stewardship
Responsibility	Gap between who currently has responsibility for stewardship and who is best placed to steward data over time
Commitment	Gap between the stewardship commitments that exist on valuable data and the commitments necessary to ensure long-term preservation and access
Human Resources	Gap between the human effort and skills needed to steward and make data accessible, and the effort and skilled workers that are available
Infrastructure and Tools	Gap between the infrastructure available to steward and reuse data and infrastructure needed to maximize stewardship and reuse capabilities
Funding	Gap between the funding needed for effective stewardship and the funding available
Curation, Management, and Preservation	Gap between the ways data is managed and prepared for preservation and reuse and ways that would maximize its potential for preservation and reuse
Sustainability Planning	Gap between planning that is done to ensure adequate resources for stewardship and the planning that is needed
Collaboration	Gap between the collaboration needed for effective stewardship and the collaboration that takes place
Sharing and Access	Gaps between the amount of data that are shared or made accessible and the amount of data that is not
Discovery	Gap between the amount of accessible data that is discoverable and the amount that is not
Reuse	Gap between the data that is available for reuse and the data that is used

We show the scale of the issue in **Table 1**, in which we identify 14 gap areas, drawn from 87 articles, reports, and other works related to data stewardship (Sample A).

While in this paper we discuss all 14, in some of our analysis we combined these into six categories as listed below:

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)

Further information about each gap area is provided in the Appendix.

Identifying Relationships Between Gaps

Many of the articles and reports that we examined also indicate a relationship between gap areas—for instance, that deficiencies or gaps in policies for archiving data affect the quantity of data that are shared. Examples of statements indicating such relationships are shown in **Figure 1**. The arrow indicates the direction of the relationship. As the fourth and fifth statements indicate, the influences are not always unidirectional (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 and Tools → Sharing and Access**

Statement 2: “who will pay the costs associated with this curation and quality control” (Reuters 2013) **Relationship: Funding → Curation, Management, and Preservation**

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 → Sustainability Planning → Infrastructure and Tools**

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 → Sustainability Planning**

Statement 5: “not possible to know costs until a preservation strategy is chosen” (Lavoie 2006, Eakin et al. n.d.) **Relationship: Sustainability Planning → Knowledge**

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

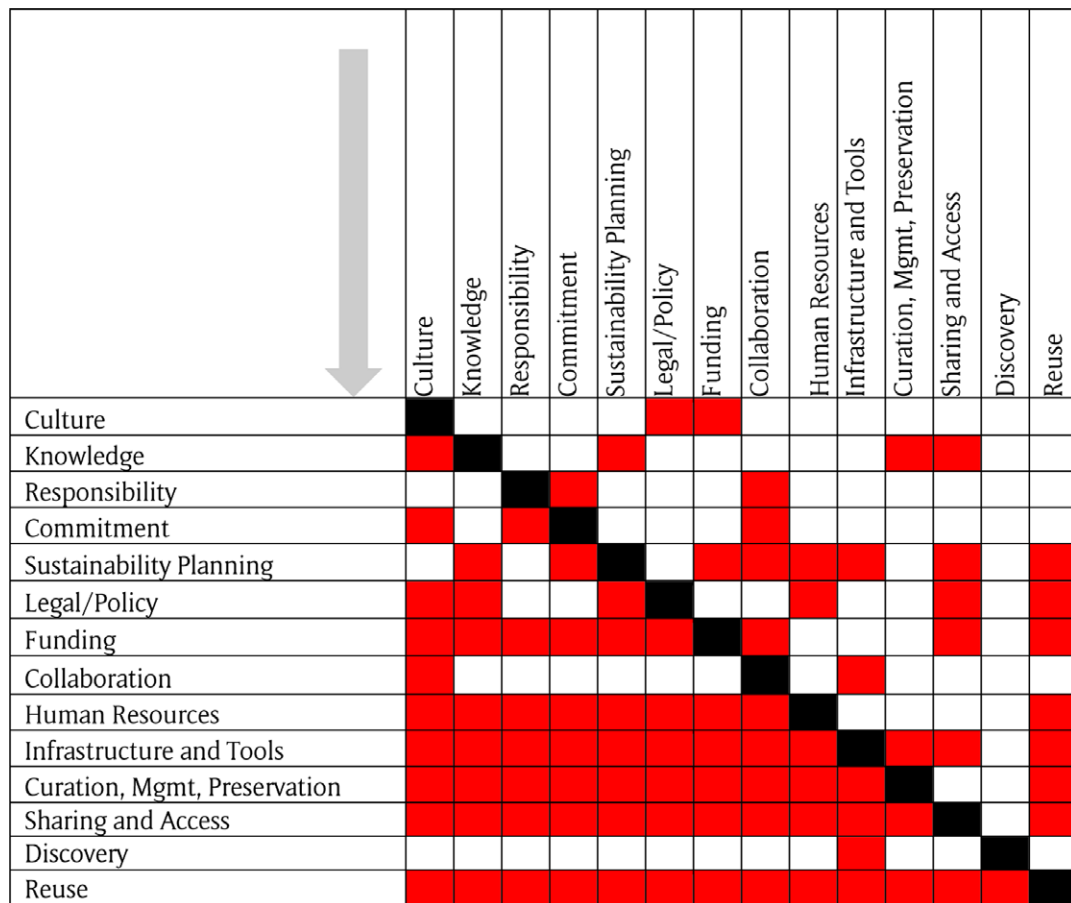


Figure 2: Gap areas and relationships between them. The figure is arranged to show that gap areas in each column impact the gap areas in the rows below them. For instance, Culture (in the first column) impacts Knowledge, Commitment, Legal and Policy Issues, etc. (the gap areas in the rows of that column).

Figure 2 shows the relationships between the 14 gap areas as identified from nearly 300 relationship statements like the ones above within 74 of the 87 works we reviewed (Sample B). 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 drawn from the statements, and not something that we have inferred. One might infer, for example, 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.

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 shaped 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 finding from this analysis is that many gap areas that have the largest impact on other areas are affected by relatively few factors. This implies that changes in such areas, including Collaboration, Culture, Knowledge, Responsibility and Commitment, could benefit data stewardship, but may also be difficult to effect. On the positive side, our analysis shows that there are at least some factors that *do* influence these gaps (e.g., Collaboration is impacted by Infrastructure and Tools and Culture by Funding and Legal and Policy Issues) and these factors could potentially be leveraged in efforts to change the size and nature of some gaps.

A second significant finding is the scarcity of references to factors that have an impact on Discovery of data, or vice versa. Discovery is only mentioned in a couple of contexts in the 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 skip the step of how it is made known that data are available for reuse.

Our effort to define the stewardship gap leads us to believe that there are multiple gaps, that the gaps are not isolated from one other but rather relate to and impact each other in different ways, and that while a number of such relationships have been identified in the literature, the relationships between some may be better understood than others (e.g., more is known about factors that affect Infrastructure and Tools than Discovery). It follows from these beliefs that the development of effective strategies to address apparent stewardship gaps will depend on an analysis of the gap areas that are most relevant in a particular context, an understanding of which other gap areas could be targeted to help reduce or eliminate the observed gaps, and reliable means of measuring the extent of the gaps (in order to calibrate levels of investment). We turn our attention now to the last of these—means of measurement.

Stewardship Gap Measurements and Metrics

How do we measure the stewardship gap? The stewardship literature includes many studies that define ways to measure the gap (“metrics”), or that actually measure the gap itself (“measurements”). 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. The value of measurements is that they help us understand specific attributes of stewardship that can be measured, for example the amount of resources or the size of archives; the value of metrics is that they help us understand how to measure stewardship or define measurements.

Our initial review of the literature led us to identify the 14 areas identified in the previous section relevant to the stewardship gap. We discuss measurement of these areas across the literature below, following some examples of what we considered to be measurement and metrics studies.

Examples of Measurement and Metrics Studies

Fecher et al.’s (2015) article “What Drives Academic Data Sharing” is an example of a study that includes both measurements and metrics. Fecher and colleagues describe a framework for understanding data sharing in academic settings, which we consider metrics. The framework comprises six categories of factors that contribute to 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 their 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 and colleagues conducted a systematic review of the literature and a survey of secondary data users, which we consider measurement. In their research, they 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. 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.” 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.” (Fecher et al. 2015, p. 19)

Overall, they argued that their framework helped “to gain a better understanding of the prevailing issues and [provide] insights into underlying dynamics of academic data sharing” (Fecher et al. 2015, p. 19).

Fecher’s study is out of the norm in addressing both measurement and metrics, and we found only one other study, “A game theoretic analysis of research data sharing,” by Pronk et al. (2015) that articulated metrics for data sharing. This study describes 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, the benefits of gaining citations, the probability of finding an appropriate dataset to reuse, and the percentage of scientists sharing their research data. The authors ran simulations with varying parameter values and found that not sharing data is always the best option for researchers individually; however, both researchers who share data and those who do not are better off when more researchers share, and more researchers can thus gain the benefits associated with reusing data. Pronk et al. note that this is a classic example of the prisoner’s dilemma. They conclude from their experiments that introducing a “citation benefit” for papers that are accompanied by a shared dataset is a more effective means of incentivizing and increasing rates of sharing than, for instance, reducing the costs of data sharing or making sharing obligatory through the use of policies.

The majority of studies regarding data sharing concentrated on measurement alone, focusing on attitudes towards data sharing, whether and how data are shared, 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

Table 2 shows the number of studies, reports, and projects (hereafter referred to as “studies”) out of 142 investigated in Sample C that either measure or provide metrics for measuring aspects of the stewardship gap. There are 56 distinct gap sub-areas within the 14 gap areas described above and our final six areas. These gap areas and sub-areas are represented as Level 3, Level 2, and Level 1, respectively, in **Table 2**. We identified some type of study (related to either measurement or metrics) in 48 of the 56 areas.

Many studies were relevant to more than one gap area. The overall distribution of studies is as follows: Culture: 77; Knowledge: 47; Responsibility: 18, Commitment: 2, Resources: 37, Actions: 87.

Table 2: Three levels of coding of gap areas and sub-areas and how many studies for each we identified in the literature. We identified some works as both measurement and metrics studies and some fell into multiple gap areas and sub-areas. The rows with totals include the distinct number of studies (out of 142) in each Level 1 gap area.

Level of coding aggregation			Measurement Studies	Metrics Studies
Level 1	Level 2	Level 3		
Culture	Culture	Sharing attitudes and practices	45	2
		Standards	8	0
		Research and development culture	6	0
		Evaluation of quality	5	10
		Stewardship priority	2	0
		Demand for data	1	0
		Data definition	1	0
		Intellectual property	1	0
		Archive mandates and objectives	0	0
	Identifying what is valuable	11	5	
	Legal and Policy	Lack of consistency and alignment	11	3
		Deficiencies that inhibit stewardship, access, and use	10	0
		Institutional structures and pressures	6	1
		Incentives that support stewardship, access, and use	5	1
		Culture Measurement and Metrics Study Total		
Knowledge		Knowledge	Amount of data	27
	Costs of stewardship		14	10
	Infrastructure for stewardship		2	0
	Where to deposit data		2	0
	Challenges of enabling data reuse		1	0
	How to preserve		1	0
	Provenance and authenticity		0	3
	Reuse possibilities		0	0
	Knowledge Measurement and Metrics Study Total			47
Responsibility	Responsibility	Conduct stewardship activities	9	8
		Coordinate stewardship activities	1	1
		Support stewardship activities	1	7
	Responsibility Measurement and Metrics Study Total			18
Commitment	Commitment	Lack of commitment	1	1
		Extent of commitment	1	1
		Duration of commitment	0	1
	Commitment Measurement and Metrics Study Total			2

(contd.)

Level of coding aggregation			Measurement Studies	Metrics Studies	
Level 1	Level 2	Level 3			
Resources	Human Resources	Lack of skills	19	4	
		Lack of support for data management	10		
		Lack of people	5	0	
		Uneven distribution of skills	2	0	
		Unequal access to resources and expertise	0	0	
	Infrastructure and Tools	Lack of infrastructure	19	2	
		Lack of tools	16	1	
		Difficulty meeting generalized and special needs	2	0	
		Different timescales of infrastructure development and maturity	0	0	
		Funding	Lack of funding	12	0
		Imbalance in funding	0	0	
	Resources Measurement and Metrics Study Total				37
	Actions	Curation, Management, and Preservation	Fragmented data management	21	0
			Insufficient data curation or management	18	7
			Difficulty managing data for reuse	14	2
Difficulty establishing the trustworthiness of curated data			1	2	
Difficulty maintaining the integrity of data over time			1	9	
Tradeoffs between data management for short or long term			0	0	
Sustainability Planning		Business and economic models	5	8	
		Dynamic and adaptable infrastructure	4	0	
		Lack of strategy and planning	4	0	
		Design and staffing of organizations	3	2	
Collaboration		Lack of collaboration	3	0	
		Challenges forming partnerships	2	0	
		Support structures	0	0	
		Lack of critical mass	0	0	
Sharing and Access		Sharing and Access	36	7	
Discovery	Discovery	7	0		
Reuse	Reuse	26	2		
Actions Measurement and Metrics Study Total				87	

We did not find any measurement studies in the following areas in our sample: 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. As can be seen in **Table 2**, there were many more areas for which we did not find metrics studies as well. We discuss these later.

The studies reviewed do not comprehensively represent all written works related to the stewardship gap, but they constitute a large subset of such works. The bibliography on which this analysis is based is posted online (see York et al. 2018a and York et al. 2018d), and we expect to add to it over time. A dynamic visualization of the data in **Table 2** is available from York et al. (2018e).

Results

Many stories could be told from the results presented in **Table 2**. 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 differing amounts of attention given to measuring different aspects of the stewardship gap that we discovered in our sample is clear from the counts of studies in **Table 2**. The small amount of attention given to Commitment and Collaboration is particularly striking because these are two areas where deficiencies or strengths have the greatest potential impact on other gap areas (as identified in **Figure 2**). The large number of studies that focus on sharing and access (under the umbrellas both of Culture and Sharing and Access) in comparison to the smaller numbers on Sustainability Planning, Legal and Policy, Funding, and Curation, Management, and Preservation, is also notable given the influence the latter areas have on data sharing (also as shown in **Figure 2**).

Table 2 also illustrates the differing attention given to metrics across the gap areas. Some of the most striking results point to areas where no metrics were found (30 out of 56 areas). These include fragmented data management, lack of strategy and planning, dynamic and adaptable infrastructure, discoverability, kinds of collaboration, adequate funding or staff support, and different cultures of research and development. A lack of metrics in these areas may indicate a lack of common targets for individuals or organizations to achieve, or a deficiency in means of evaluating progress.

Future research needs to address the importance of areas that have been little studied until now, and direct attention to those that will have the greatest impact on future stewardship.

Imbalances in measurements and metrics studies

There are several areas where the contrast between measurement and metrics studies is particularly pronounced. These include metrics for sharing attitudes and practices (45 measurement studies to two that articulate metrics), reuse of data (26 to two), fragmented data management (21 to zero), lack of skills (19 to four), lack of infrastructure (19 to two), lack of tools (16 to one), lack of funding (12 to zero), difficulty in management of data for reuse (14 to two), lack of support for data management (10 to zero) and incentives and deficiencies in, and alignment among legal and policy issues (5 to one, 10 to zero, and 11 to three, for incentives, deficiencies, and alignment, respectively).

One of the common challenges encountered by studies of stewardship gap areas is the difficulty of obtaining comparable results across different academic domains, especially at large scale. For example, Borgman et al. (2014) note that while the case study method they use to investigate research data infrastructures could be used in other domains, large-scale surveys would likely be less effective due to the importance of local context. Similarly, in their study of the value and impact of research data, Beagrie and Houghton (2014) describe the challenges of conducting their study in different contexts: “The data collection and economic analysis are time consuming and need to be tailored to the specific nature of operation and use of each data centre.” It is possible that a greater focus on metrics in the areas above (e.g., what indicates a lack of infrastructure; what it means for data management to be fragmented; how the difficulty of managing data can be quantified) as

well as areas where metrics have been articulated but not widely agreed upon, would result in the collection of more consistent information in different contexts and domains. This could in turn result in more consistent measurement and comparison of research findings across disciplinary boundaries and at scale.

The imbalance between measurement and metrics studies suggests that future research should emphasize metrics, effectively setting broadly applicable standards for measuring discrete aspects of effective stewardship in order to understand how to improve stewardship, both in specific research and data domains, and more generally across the board.

Differences in the depth of studies

The literature contains multiple types of studies. In one common type, which we termed “targeted,” the entire investigation is focused in one or two closely related areas, such as resources or specific actions like curation (e.g., Akmon 2014, Atkins 2003, Ayris et al. 2010, Beagrie and Houghton 2013a and 2013b, Borgman et al. 2014, Cirrinnà et al. 2013). Another common type comprises “wider” studies, which investigate several different gap areas at once, often in the context of a single institution, a nation’s scientific enterprises, or a comparative international framework (e.g., Alexogiannopoulos et al. 2010, Gibbs 2009, Hoeflich Mohr et al. 2015, Jerrome and Breeze 2009, Kuipers and van der Hoeven 2009, Martinez-Urbe 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, Waller and Sharpe 2006). Of the 142 we reviewed, 115 studies were targeted, and 28 were wider.

Wider studies, though they may cover many topics (e.g., in the context of a survey), often have only one or a few questions about any specific given gap area (Wynholds et al. 2011 and Tenopir et al. 2012 are two exceptions that examine multiple gap areas in depth). A raw count of studies including both targeted and wider studies may thus overestimate the depth of investigation that has occurred in a particular area. **Table 3** shows 16 of the 50 overall gap sub-areas where we found either a measurement or metrics study. In all of these 16 there is a significantly higher proportion of wider studies (which did not necessarily investigate the indicated area in depth) than targeted. We include only measurement studies in the table as all but one metrics study, related to responsibility for conducting stewardship activities, were targeted.

Table 3: Gap sub-area measurement studies with a larger proportion of “wider” studies than “targeted”.

Gap Sub-area	Measurement		
	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 management	0	10	10
Conduct stewardship activities	0	9	9
Deficiencies that inhibit stewardship, access, and use [in legal and policy areas]	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

The proportion of targeted versus wider studies is an important factor in understanding the universe of research relevant to the stewardship gap. In many cases, such as those indicated in **Table 3**, not only more research, but more in-depth research is critical to advance our knowledge of the stewardship gap and to give guidance to policy makers, researchers, and research institutions about ways that they can ensure that the research data critical for future success is well stewarded.

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 14 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 effort 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 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 not necessarily balanced. 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?

It is important to note regarding the final two that the gap areas presented in this paper do not represent all gap areas, only those identified in the literature reviewed. In addition, our review does not cover all works that have been written that are relevant to the stewardship gap. Although it covers a significant subset, and has significantly guided the direction of our research, the stewardship gap bibliography is a work in progress that we expect will become more comprehensive over time through continuing investigation.

Data Accessibility Statement

The works reviewed in the samples of literature used in the study (samples A, B, and C) as well as information pertaining to the evidence of gaps, gap relationships, and study designations associated with each is titled "Stewardship Gap Project Bibliography" and is available at <https://doi.org/10.7302/Z2ZW1J47>.

Additional File

The additional file for this article can be found as follows:

- **Appendix.** Description of Gap Areas. DOI: <https://doi.org/10.5334/dsj-2018-019.s1>

Acknowledgements

We would like to thank the Alfred P. Sloan Foundation, the University of Colorado Boulder, and Rensselaer Polytechnic Institute for their support of the project. We would also like to thank the members of the Stewardship Gap project advisory board for critical driving discussions and comments 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.

Competing Interests

The authors have no competing interests to declare.

References

- 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. Available at: <https://deepblue.lib.umich.edu/handle/2027.42/107162> [Last accessed 17 May 2017].
- Alexogiannopoulos, E, McKenney, S and Pickton, M.** 2010. Research Data Management Project: a DAF investigation of research data management practices at The University of Northampton. Available at: <http://nectar.northampton.ac.uk/2736/> [Last accessed 17 May 2017].
- Association of Research Libraries Workshop on New Collaborative Relationships (ARL).** 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. <http://www.arl.org/storage/documents/publications/digital-data-report-2006.pdf> [Last accessed 17 May 2017].
- Atkins, DE, Droegemeier, KK, Feldman, SI, Garcia-Molina, H, Klein, ML, Messerschmitt, DG, Messina, P, Ostriker, JP and Wright, MH.** 2003. Revolutionizing Science and Engineering Through Cyberinfrastructure. Available at: <https://www.nsf.gov/cise/sci/reports/atkins.pdf> [Last accessed 17 May 2017].
- Ayris, P, Wheatley, P, Aitken, B, Hole, B, McCann, P, Peach, C and Lin, L.** 2010. The LIFE3 Project: Bringing digital preservation to LIFE. Available at: http://www.life.ac.uk/3/docs/life3_report.pdf [Last accessed 17 May 2017].
- Beagrie, N and Houghton, J.** 2013a. The Value and Impact of the Archaeology Data Service: A Study and Methods for Enhancing Sustainability. Joint Information Systems Committee, Bristol and London. Available at: http://repository.jisc.ac.uk/5509/1/ADSReport_final.pdf [Last accessed 17 May 2017].
- Beagrie, N and Houghton, J.** 2013b. The Value and Impact of the British Atmospheric Data Centre. Joint Information Systems Committee, Bristol and London. Available at: http://repository.jisc.ac.uk/5382/1/BADCRReport_Final.pdf [Last accessed 17 May 2017].
- Beagrie, N and Houghton, J.** 2014. The Value and Impact of Data Sharing and Curation. Available at: http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf [Last accessed 17 May 2017].
- Berman, F.** 2008. Got data?: a guide to data preservation in the information age. *Communications of the ACM*, 51: 50. DOI: <https://doi.org/10.1145/1409360.1409376>
- Berman, F.** 2014. Despite Growing Data, Infrastructure Stands Still – Why the gap puts research data at risk. IEEE. Available at: <http://theinstitute.ieee.org/ieee-roundup/members/achievements/despite-growing-data-infrastructure-stands-still> [Last accessed 17 May 2017].
- Berman, F, Lavoie, B, Ayris, P, Choudhury, GS, Cohen, E, Courant, P, Dirks, L, Friedlander, A, Gurbaxani, V, Jones, A, Kerr, A, Lynch, C, Rubinfeld, D, Rusbridge, C, Schonfeld, R, Smith Rumsey, A and Van Camp, A.** 2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information, 110. Available at: <http://www.oclc.org/research/publications/all/long-term-access-digital-information.html> [Last accessed 17 May 2017].
- Big Data Value Association (BDVA).** 2015. European Big Data Value Strategic Research & Innovation Agenda. Big Data Value Europe, Brussels, Belgium. Available at: http://www.bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final.pdf [Last accessed 17 May 2017].
- Borgman, CL.** 2012. The conundrum of sharing research data, 63: 1059–1078. DOI: <https://doi.org/10.1002/asi.22634>
- Borgman, CL.** 2015. Big data, little data, no data: scholarship in the networked world.
- Borgman, CL, Darch, PT, Sands, AE, Wallis, JC and 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)*. Available at: <http://works.bepress.com/borgman/321> [Last accessed 17 May 2017].
- Cirrinà, C, Fernie, K and Lunghi, M.** 2013. *Digital Curator Vocational Education Europe (DigCurV): Final report and Conference Proceedings*. Available at: <http://www.digcur-education.org/eng/International-Conference/DigCurV-2013-proceedings> [Last accessed 17 May 2017].
- Climatic Research Unit email controversy. 2015. Wikipedia, the free encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Climatic_Research_Unit_email_controversy&oldid=683477293 [Last accessed 17 May 2017].

- Cummings, J, Finholt, T, Foster, I, Kesselman, C and Lawrence, KA.** 2008. Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations. *Technology*, 3. Available at: http://web.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf [Last accessed 17 May 2017].
- Downs, RR and Chen, RS.** 2013. Towards Sustainable Stewardship of Digital Collections of Scientific Data. *Presented at the GSDI World Conference (GSDI 13) Proceedings*. Quebec City, Canada. Available at: <https://academiccommons.columbia.edu/catalog/ac:199366> [Last accessed 17 May 2017].
- Esanu, E, Davidson, J, Ross, S and Anderson, W.** 2004. Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ERPNANET/CODATA Workshop. *Data Science Journal*, 3: 227–232. Available at: <http://datascience.codata.org/articles/abstract/10.2481/dsj.3.227dsj.3.227/> [Last accessed 17 May 2017]. DOI: <https://doi.org/10.2481/dsj.3.227>
- Fecher, B, Friesike, S and Hebing, M.** 2015. What Drives Academic Data Sharing? *PLoS ONE*, 10: e0118053. DOI: <https://doi.org/10.1371/journal.pone.0118053>
- Gantz, JF, McArthur, J, Minton, S, Reinsel, D, Chute, C, Schlichting, W, Xheneti, I, Toncheva, A and Manfrediz, A.** 2007. The Expanding Digital Universe [White Paper]. Available at: https://www.tobbb.org.tr/BilgiHizmetleri/Documents/Raporlar/Expanding_Digital_Universe_IDC_WhitePaper_022507.pdf [Last accessed 17 May 2017].
- Gantz, JF, Minton, S, Reinsel, D, Chute, C, Schlichting, W, Toncheva, A and Manfrediz, A.** 2008. The Diverse and Exploding Digital Universe [White Paper]. Available at: <http://www.ifap.ru/library/book268.pdf> [Last accessed 17 May 2017].
- Gibbs, H.** 2009. Southampton Data Survey: Our Experience and Lessons Learned. University of Southampton. Available at: <http://www.disc-uk.org/docs/SouthamptonDAF.pdf> [Last accessed 17 May 2017].
- Hank, C, Tibbo, HR and Lee, CA.** 2010. DigCCurr I Final Report, 2006–09. Available at: http://www.ils.unc.edu/digccurr/digccurr_i_final_report_031810.pdf [Last accessed 17 May 2017].
- Hilbert, M and López, P.** 2011. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332: 60–65. DOI: <https://doi.org/10.1126/science.1200970>
- Hofelich Mohr, A, Bishoff, J, Johnston, L, Braun, S, Storino, C and Bishoff, C.** 2015. Data Management Needs Assessment – Surveys in CLA, AHC, CSE, and CFANS. Available at: <http://conservancy.umn.edu/handle/11299/174051> [Last accessed 17 May 2017].
- Holdren, JP.** 2013. Increasing Access to the Results of Federally Funded Scientific Research (Executive Office of the President Office of Science and Technology Policy Memo). Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [Last accessed 17 May 2017].
- Jerrone, N and Breeze, J.** 2009. Imperial College Data Audit Framework Implementation: Final Report. Imperial College London. Available at: <http://repository.jisc.ac.uk/307/> [Last accessed 17 May 2017].
- Kalil, T and Miller, J.** 2015. Advancing U.S. Leadership in High-Performance Computing. *Office of Science and Technology Policy*. Available at: <https://obamawhitehouse.archives.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing> [Last accessed 17 May 2017].
- Kuipers, T and van der Hoeven, J.** 2009. PARSE.Insight: Insight into Digital Preservation of Research Output in Europe: Survey Report. Available at: <http://docplayer.net/127428-Parse-insight-deliverable-d3-4-survey-report-of-research-output-europe-title-of-deliverable-survey-report.html> [Last accessed 17 May 2017].
- Lyman, P and Varian, HR.** 2000. How Much Information? *Journal of Electronic Publishing*, 6. Available at: <http://groups.ischool.berkeley.edu/archive/how-much-info/summary.html> [Last accessed 17 May 2017]. DOI: <https://doi.org/10.3998/3336451.0006.204>
- Lyman, P and Varian, HR.** 2003. How Much Information 2003? Available at: <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/> [Last accessed 17 May 2017].
- Manyika, J, Byers, AH, Chui, M, Brown, B, Bughin, J, Dobbs, R and Roxburgh, C, McKinsey Global Institute.** 2011. Big data: The next frontier for innovation, competition, and productivity 156. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation [Last accessed 17 May 2017].
- Manyika, J, Chui, M, Groves, P, Farrell, D, Van Kuiken, S and Doshi, EA, McKinsey Global Institute.** 2013. Open data: Unlocking innovation and performance with liquid information 103. Available at: <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information> [Last accessed 17 May 2017].
- Martinez-Uribe, L.** 2009. Using the Data Audit Framework: An Oxford Case Study. Available at: <http://www.disc-uk.org/docs/DAF-Oxford.pdf> [Last accessed 17 May 2017].

- Mitcham, J, Awre, C, Allinson, J, Green, R and Wilson, S.** 2015. Filling the Digital Preservation Gap. A JISC Research Data Spring Project. Phase One Report. Available at: http://figshare.com/articles/Filling_the_Digital_Preservation_Gap_A_Jisc_Research_Data_Spring_project_Phase_One_report_July_2015/1481170 [Last accessed 17 May 2017].
- National Academy of Sciences (NAS).** 2009. Ensuring the integrity, accessibility, and stewardship of research data in the digital age, 325: 368. DOI: <https://doi.org/10.1126/science.1178927>
- National Aeronautics and Space Administration (NASA).** n.d. Data & Information Policy [WWW Document]. NASA. URL: <http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/> (accessed 12.10.15).
- National Research Council (NRC).** 2003. Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. National Academies Press, Washington, D.C. Available at: <http://www.nap.edu/catalog/10613> [Last accessed 17 May 2017].
- National Science Foundation (NSF), Cyber Infrastructure Council.** 2007. Cyberinfrastructure Vision for 21st Century Discovery. Director. Available at: <http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf> [Last accessed 17 May 2017].
- Obama, B.** 2009. A Strategy for American Innovation: Driving Towards Sustainable Growth and Quality Jobs. Available at: http://www.politico.com/pdf/PPM41_9.20.09_innovation.pdf [Last accessed 17 May 2017].
- Obama, B.** 2011. A Strategy for American Innovation: Securing Our Economic Growth and Prosperity. Available at: <https://obamawhitehouse.archives.gov/sites/default/files/uploads/InnovationStrategy.pdf> [Last accessed 17 May 2017].
- Obama, B.** 2013a. Executive Order – Making Open and Machine Readable the New Default for Government Information. The White House. Available at: <https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government> [Last accessed 17 May 2017].
- Obama, B.** 2013b. Open Data Policy-Managing Information as an Asset. Available at: <https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf> [Last accessed 17 May 2017].
- Office of Management and Budget (OMB).** 2012. Digital Government: Building a 21st Century Platform to Better Serve the American People. Available at: <https://obamawhitehouse.archives.gov/sites/default/files/omb/egov/digital-government/digital-government.html> [Last accessed 17 May 2017].
- Open Exeter Project Team.** 2012. Summary Findings of the Open Exeter Data Asset Framework Survey. University of Exeter, Exeter, UK. Available at: https://ore.exeter.ac.uk/repository/bitstream/handle/10036/3689/daf_report_public.pdf?sequence=1 [Last accessed 17 May 2017].
- Organization for Economic Co-operation and Development (OECD).** 2015. Making Open Science a Reality. Available at: <https://www.innovationpolicyplatform.org/content/open-science> [Last accessed 17 May 2017].
- Parsons, T, Grimshaw, S and Williamson, L.** 2013. Research Data Management Survey. University of Nottingham. Available at: <http://admire.jiscinvolve.org/wp/files/2013/02/ADMIRE-Survey-Results-and-Analysis-2013.pdf> [Last accessed 17 May 2017].
- Pepe, A, Goodman, A, Muench, A, Crosas, M and 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: e104798. DOI: <https://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: 133–148. Available at: <http://www.sciencedirect.com/science/article/pii/S0740624X07000561> [Last accessed 17 May 2017]. DOI: <https://doi.org/10.1016/j.giq.2007.04.008>
- Peters, C and 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: 387–403. DOI: <https://doi.org/10.1080/0194262X.2011.626340>
- Podesta, J, Pritzker, P, Moniz, EJ, Holdren, J and Zients, J.** 2014. Big Data: Seizing Opportunities, preserving values. Executive Office of the President. Available at: https://obamawhitehouse.archives.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf [Last accessed 17 May 2017].
- Pronk, TE, Wiersma, PH, van Weerden, A and Schieving, F.** 2015. A game theoretic analysis of research data sharing. *PeerJ*, 3. DOI: <https://doi.org/10.7717/peerj.1242>


- Read, KB, Sheehan, JR, Huerta, MF, Knecht, LS, Mork, JG and Humphreys, BL, NIH Big Data Annotator Group.** 2015. Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. *PLoS ONE*, 10: e0132735. DOI: <https://doi.org/10.1371/journal.pone.0132735>
- Research Councils UK (RCUK).** 2015. RCUK Common Principles on Data Policy. Research Councils UK. Available at: <http://www.rcuk.ac.uk/research/datapolicy/> [Last accessed 17 May 2017].
- Science and Technology Council (STC).** 2007. The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials. Academy of Motion Picture Arts and Sciences. Available at: <http://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma> [Last accessed 17 May 2017].
- Sveinsdottir, T, Wessels, B, Smallwood, R, Linde, P, Kala, V, Tsoukala, V and 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). Available at: http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf [Last accessed 17 May 2017].
- Tenopir, C, Allard, S, Douglass, K, Aydinoglu, A, Wu, L, Read, E, Manoff, M and Frame, M.** 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6: e21101. DOI: <https://doi.org/10.1371/journal.pone.0021101>
- Thaesis, van der Hoeven, J.** 2010. PARSE.Insight: Insight into issues of Permanent Access to the Records of Science in Europe. Final Report.
- The Royal Society.** 2012. Science as an Open Enterprise. London. Available at: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE-Summary.pdf [Last accessed 17 May 2017].
- Thompson Reuters.** 2013. Unlocking the Value of Research Data: A Report from the Thompson Reuters Industry Forum. Available at: <http://researchanalytics.thomsonreuters.com/m/pdfs/1003903-1.pdf> [Last accessed 17 May 2017].
- Thornhill, K and Palmer, L.** 2014. An Assessment of Doctoral Biomedical Student Research Data Management Needs. Available at: http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1075&context=escience_symposium [Last accessed 17 May 2017].
- Turner, V, Reinsel, D, Gantz, JF and Minton, S.** 2014. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Available at: <https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf> [Last accessed 17 May 2017].
- Ubaldi, B.** 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Publishing, OECD Working Papers on Public Governance*. DOI: <https://doi.org/10.1787/5k46bj4f03s7-en>
- UNC-CH, Provost's Task Force on the Stewardship of Digital Research Data.** 2012. Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership. University of North Carolina Chapel Hill, Chapel Hill, North Carolina. Available at: http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf [Last accessed 17 May 2017].
- Vickery, G.** 2011. Review of Recent Studies on PSI Re-use and Related Market Developments. Available at: http://ec.europa.eu/newsroom/document.cfm?doc_id=1093 [Last accessed 17 May 2017].
- Vickery, G.** 2012. Review of Recent Studies on PSI Re-use and Related Market Developments (revised and abridged).
- Vogel, G.** 2011. Report: Dutch "Lord of the Data" Forged Dozens of Studies (UPDATE). Science Insider. Available at: <http://news.sciencemag.org/europe/2011/10/report-dutch-lord-data-forged-dozens-studies-update> [Last accessed 17 May 2017].
- Waller, M and Sharpe, R.** 2006. Mind the Gap: Assessing Digital Preservation Needs in the UK. Digital Preservation Coalition, York, United Kingdom. Available at: http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk [Last accessed 17 May 2017].
- Wallis, JC, Rolando, E and Borgman, CL.** 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, 8: e67332. DOI: <https://doi.org/10.1371/journal.pone.0067332>
- Wichert, JM, Bakker, M and 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: e26828. DOI: <https://doi.org/10.1371/journal.pone.0026828>

- Willets, D, Livanov, D, Schütte, G, Harayama, Y, Carrozza, MC, Goodyear, G, Fioraso, G, Falcone, P and Geoghegan-Quinn, M.** 2013. G8 Science Ministers Statement. London, UK. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf [Last accessed 17 May 2017].
- Wynholds, L, Fearon, DS, Jr., Borgman, CL and 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, JCDL '11*. ACM, New York, NY, USA, 383–386. DOI: <https://doi.org/10.1145/1998076.1998146>
- Wynholds, LA, Wallis, JC, Borgman, CL, Sands, A and 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, JCDL '12*. ACM, New York, NY, USA, 19–22. DOI: <https://doi.org/10.1145/2232817.2232822>
- York, J, Gutmann, M and Berman, F.** 2016. Will Today's Data Be Here Tomorrow? Measuring the Stewardship Gap. In: *Proceedings of the 13th International Conference on Digital Curation. Presented at the iPres 2016*. Swiss National Library, Bern, Switzerland. Available at: https://phaidra.univie.ac.at/detail_object/o:503172 [Last accessed 17 May 2017].
- York, J, Gutmann, M and Berman, F.** 2018a. Stewardship Gap Project Bibliography Data. DOI: <https://doi.org/10.7302/Z2ZW1J47>
- York, J, Gutmann, M and Berman, F.** 2018b. Stewardship Gap Bibliography Data, Sample A: Gap Evidence. Filter Keyword: gap_evidence. Available at: [https://stewardshipgap.net/biblio?f\[keyword\]=275](https://stewardshipgap.net/biblio?f[keyword]=275) [Last accessed 12 January 2018].
- York, J, Gutmann, M and Berman, F.** 2018c. Stewardship Gap Bibliography Data, Sample B: Gap Relationships. Filter Keyword: gap_relationships. Available at: [https://stewardshipgap.net/biblio?f\[keyword\]=276](https://stewardshipgap.net/biblio?f[keyword]=276) [Last accessed 12 January 2018].
- York, J, Gutmann, M and Berman, F.** 2018d. Stewardship Gap Bibliography Data, Sample C: Measurement and Metrics Studies. Available at: <http://stewardshipgap.net/all-studies-browse> [Last accessed 12 January 2018].
- York, J, Gutmann, M and Berman, F.** 2018e. Stewardship Gap Bibliography Data, Tree Map. Available at: <http://www.stewardshipgap.net/treemap> [Last accessed 12 January 2018].

How to cite this article: York, J, Gutmann, M and Berman, F. 2018. What Do We Know about the Stewardship Gap. *Data Science Journal*, 17: 19, pp. 1–17, DOI: <https://doi.org/10.5334/dsj-2018-019>

Submitted: 02 August 2017 **Accepted:** 30 July 2018 **Published:** 17 August 2018

Copyright: © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 