Will Today's Data Be Here Tomorrow? Measuring The Stewardship Gap

Jeremy York
University of Colorado Boulder
510 Miller Ave
Ann Arbor, MI 48103
jeremy.york@colorado.edu

Myron Gutmann
University of Colorado Boulder
483 UCB
Boulder, CO 80309-0483
myron.gutmann@colorado.edu

Francine Berman
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180
bermaf@rpi.edu

ABSTRACT

Stakeholders in scholarly research are paying increased attention to stewardship of digital research data[1] for the purposes of advancing scientific discovery, driving innovation, and promoting trust in science and scholarship. However, little is known about the total amounts, characteristics, and sustainability of data that could be used for these purposes. The Stewardship Gap is an 18-month project funded by the Alfred P. Sloan Foundation to understand issues in defining metrics for and measuring the stewardship gap: the potential gap between the amount of valuable data produced through sponsored projects in the United States and the amount that is effectively stewarded and made accessible. This paper reports on the first phase of the project, which sought to develop an instrument to gather information about research data sustainability from a broad variety of researchers and research disciplines and make progress toward the ultimate goals of 1) shedding light on the size, characteristics, and sustainability of valuable sponsored research data and creative work in the United States, and 2) recommending actions stakeholders can take to address the stewardship gap if one is found to exist.

Keywords

Digital curation, digital preservation, research data, data stewardship, data sustainability

1. INTRODUCTION

The explosion of digital information and the promise of using (and reusing) data to spur research innovation have focused attention in the past couple of decades on issues of appropriately curating, managing, and disseminating digital data for reuse. This is true in both the public and private sectors, where digital data are increasingly seen as an asset to be used to promote innovation, economic growth and trust in or accountability of government [12, 13, 40, 48, 58], and to further the arts and advance and verify scientific discovery [5, 14, 18, 36, 37, 38, 40, 55, 62].

Despite high interest in reuse of research data in the scholarly community, numerous challenges have inhibited the ability to understand the size and breadth of the research data universe, or to develop means to ensure that all data "of value" will be discoverable and usable at the appropriate time in the future. Challenges range from difficulty in defining the data of interest [49] and difficulty making measurements (e.g., due to the time required, complexity of social and technical factors, or lack of methods) [1, 8, 16, 17] to challenges in comparing results of different studies [4, 8, 15], poor understanding of the interplay between the many factors involved in data stewardship, and lack of knowledge about how to interpret what has been measured [1, 11].

Concerns that valuable data may not be adequately preserved come in part from studies such as "Sizing the Problem of Improving Discovery and Access to NIH-Funded Data," [49] which found that 88% of articles published in PubMedCentral in 2011 had "invisible datasets" (where deposit of data in a recognized repository was not explicitly mentioned). Other surveys and studies in recent years have similarly discovered small percentages of data deposited in public repositories. These studies have also uncovered information about data management and storage practices that raise concerns about data persistence, such as lack of future planning for preservation of project data, and significant amounts of research data archived on personal devices as opposed to institutional or community infrastructure [See for example 3, 6, 22, 29, 30, 39, 41, 44, 50, 56, 61].

The Stewardship Gap project was undertaken to investigate means of gathering information about these concerns. In particular, it aims to better understand the potential gap between the total amount of valuable data resulting from sponsored projects in the US that is being produced and the amount that is or will be responsibly stewarded and made accessible to others.

2. LITERATURE REVIEW

2.1 Stewardship Gap Areas

We conducted a survey of relevant literature to ascertain what is known about the stewardship gap. This survey revealed the presence of not one, but many gap areas that impact the ability to measure, analyze, plan for, and act to steward valuable research data. These areas include:

1. Culture (e.g., differences in attitudes, norms and values that affect data stewardship and reuse);
2. Knowledge (e.g., about how to preserve data, what skills are needed, how much data exist, of what kind, how much of it has value and for how long);
3. Commitment (e.g., commitments adequate to needs for stewardship and reuse);
4. Responsibility (e.g., who is responsible for funding stewardship and carrying out stewardship activities);
5. Resources (e.g., funding, infrastructure, tools, human resources);
6. Stewardship actions such as curating, managing, and preserving data, and activities that enable curation, management, and preservation such as making data

[1] Unless otherwise indicated, "data" and "research data" are used to refer to digital data throughout the paper, as opposed to analog data.

available (e.g., through data sharing or deposit in a data repository), long-term planning, and collaboration [64].

While all of these areas appeared crucial to understanding the stewardship gap as a whole, we designed a pilot project that would provide evidence of the presence of a gap and important elements of any gap we discovered. Based on background reading and focused interactions with the project advisory board, we hypothesized these elements to be the extent and duration of value that data have and the extent and duration of commitments made to steward valued data. We considered that if our study found e.g., that a given dataset had value for twenty years, but there was only a commitment to preserve the data for five, this could be an indication of a stewardship gap. We believed information about value and commitment would have greater value if combined with information about who could act to address a gap if one existed, and thus added stewardship responsibility as a third primary parameter in the study.

The first phase of the study was devoted to formulating questions about research data value, commitment, and stewardship responsibility that could be answered by researchers in a wide variety of disciplines about data of diverse sizes and types. Research in the first phase focused on data resulting from public- or non-profit-sponsored research conducted at institutions of higher education in the United States.

2.2 Review of Research Data Studies

There are two main types of studies that have sought to measure aspects of the stewardship gap for research data. The first comprises studies with a specific focus ("targeted" studies), for instance on research data sharing, deposit of data in repositories, or funding for stewardship [some examples include 23, 43, 46, 47, 54, 63]. Studies of the second type ("wider" studies) cover a range of topics at once, often at less depth for any given topic than a targeted study [e.g., 3, 21, 24, 26, 34, 35, 39, 41, 44, 45, 56, 59]. Many of the second type were conducted on university campuses to gather information to help establish or improve research data management services, though some studies extended across campuses as well [e.g., 18, 30].

Figure 1 shows the distribution of one hundred-seven studies reviewed for the stewardship gap project in the six gap areas described above. Studies related to data value are included under "Culture." However, they are also represented in the figure as a separate category since value is a main focus of the project. The figure also shows the number of targeted versus wider studies.

Figure 1 is not a comprehensive representation of all studies related to the stewardship gap. It does show the topical distribution among a significant subset, however, and shows in particular how our three areas of interest (data value, stewardship commitment, and stewardship responsibility) are represented in the broader landscape of studies.[2]

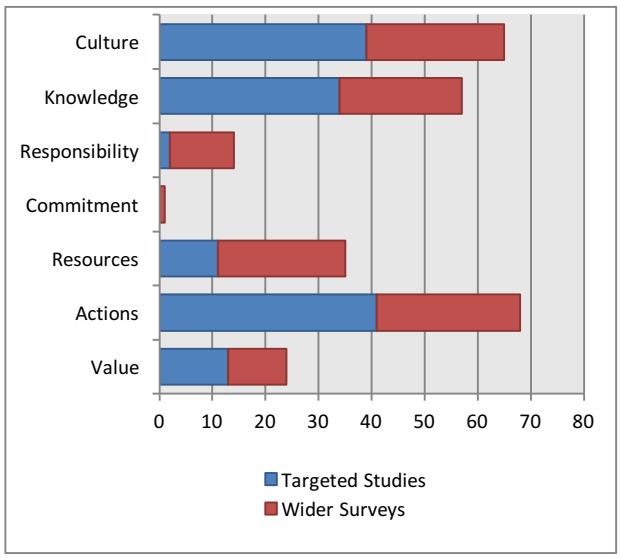

Figure 1. Prior measures of stewardship gap areas

2.3 Value, Commitment, and Responsibility

2.3.1 Value

We identified two general types of studies related to data value. The first focuses on means of understanding the value and impact of data and data stewardship, often in financial or business terms [some examples are 7, 9, 10, 20, 25, 28, 29, 31, 32, 51, 52, 57, 60]. The second type, which is most relevant to the stewardship gap project, investigates different kinds, degress, or durations of data value. One example is Akmon's 2014 investigation of how scientists understand the value of their data throughout the course of a project and the effect of conceptions of value on data practices [2]. Two other, wider ranging studies of this type[3] include a campus study that asked researchers whether their data would be valuable for projects other than the one they were gathered for (though the study did not ask about users of data or or reasons for data value) [50] and the PARSE.Insight project [30], which asked respondents to rate the importance of the following "well-known" reasons for data preservation:

1. If research is publicly funded, the results should become public property and therefore properly preserved
2. It will stimulate the advancement of science (new research can build on existing knowledge)
3. It may serve validation purposes in the future
4. It allows for re-analysis of existing data
5. It may stimulate interdisciplinary collaborations
6. It potentially has economic value
7. It is unique

The Stewardship Gap's investigation of value is most similar to the PARSE.Insight project in that it poses a range of different reasons for data value for researchers to respond to. It differs, however, in asking researchers to rate reasons for value, as value. However, early pilot studies encountered difficulty classifying value according to criticality of data to the institution as the framework specified [26, 34]. Explicit questions about data value do not appear to have been included in subsequent implementations, although there are questions about whether data should be preserved, whether they can be reused [3, 21], and how long data will be archived [3, 39, 45, 61]. Because they gather data that provide indicators of data value, these studies have been included in the tally of studies shown in Figure 1.

[2] See https://www.zotero.org/groups/data_stewardship_studies/items for a list of all studies represented.

[3] Several additional studies have been undertaken that used or were based on the Digital Asset Framework, a framework developed by the Humanities Advanced Technology and Information Institute at the University of Glasgow in association with the Digital Curation Centre to conduct an assessment of data assets in an institutional context. In its early instantiation, the DAF framework was designed to gather information about data

opposed to reasons for preservation—though our study does also ask researchers to indicate reasons for value that have had the greatest impact on decisions about preservation.

2.3.2 Commitment

Commitment is significantly understudied in comparison with other areas. Only one study was found that gathered information about commitment.[4] This was a 2005 survey conducted by Tessella for the Digital Preservation Coalition as part of an initiative to assess digital preservation needs in the United Kingdom [61]. The survey asked whether there was a high level of commitment to digital preservation in the respondent's organization. In contrast to this study, the Stewardship Gap asks about levels of commitment associated with research data from specific projects.

2.3.3 Responsibility

A number of the reviewed studies investigated questions of responsibility, including responsibility for:

- Storage, backup, and management of research data
- Funding of data management, storage, and preservation
- Decisions about stewardship, including which data are important to preserve and for how long, what constitutes compliance with regulations, licenses, and mandates, what descriptive metadata are appropriate, and provisions for short- and long- term storage and preservation [21, 26, 30, 33, 39, 45, 50, 61]

The Stewardship Gap did not introduce new questions in this regard. However, our primary purpose was to be able to compare information about responsibility with information about value and commitment in order to understand who could act to address a stewardship gap if one existed.

2.4 Common Themes In Reviewed Literature

Some common themes across reviewed studies that were relevant to our efforts to develop a strategy for measuring the stewardship gap included the following:

1) The significant amount of time that can be involved in conducting a study. Many studies employed a preliminary pilot phase to refine questions, followed by a broader survey and then follow-up interviews. This was done to balance the needs to survey a sufficiently large population but also gain important contextual information that interviews can provide.

2) The challenges of creating a common understanding of what "data" are for the purposes of the study. Three main challenges that surfaced were 1) addressing what type of materials are included in "data" (e.g., course materials, notes, structured data, images, non-digital files, etc.), 2) what constitutes a "dataset" (e.g., an individual file, a set of files in a particular format, or a set of related data regardless of format), and 3) what universe of data is being measured (e.g., all data held or created, all data held or created in a specific time frame, or data from a specific project).

3) The significance of the correlation between research discipline, research role, type of research (e.g., scientific or clinical), and level of experience and the amount of research data generated, attitudes and practices about data storage, data sharing and reuse, and beliefs about the primary reasons for and threats to preservation.

4) The broad diversity in sizes and formats of digital data generated and types of storage used, and the relatively small amounts of data that are deposited with disciplinary or other repositories outside the researcher's institution [30, 39, 41, 59].[5]

Regarding this last theme, the University of Nottingham investigated their results further and found that the majority of researchers stored their data in multiple locations [41]. This would seem to add a degree of confidence to concerns about adequate preservation of data. However, the Open Exeter Project found that much of the research data being held is not actively managed, raising additional concern [39]. This concern is supported by results from the University of Northampton that while most researchers intend to keep data beyond the completion of a project, and even indefinitely, this intention is not realized for a variety of reasons, including

- lack of data management strategies
- the need to store files that exceed computer hard-drive space on external media that are more prone to degradation and loss
- files and software stored on personal devices becoming out of sync with university resources that are needed to access and use them [3].

The considerations these common themes raised for our project and the ways we addressed them are described in section 3 below.

3. GOALS AND METHODOLOGY OF THE STEWARDSHIP GAP PROJECT

The first goal of our initial study was to test, across as broad a range of disciplines as possible, the performance and effectiveness of questions about data value, stewardship commitment, and stewardship responsibility. Because we wanted to be able to gather information about a measureable "gap", we also wanted to collect information about the size and characteristics of valued data. The second goal was to analyze responses in order to inform a more in-depth study of the stewardship gap in a second phase.

To accomplish these goals, we designed a questionnaire (see the question areas in Table 1) and conducted interviews with seventeen researchers in sixteen fields from thirteen US institutions over the course of November and December 2015. Interviewees were selected on the basis of their association with at least one of a range of academic disciplines, with the goal of achieving a wide range of disciplinary coverage. Most of the interviewees were known to or suggested by members of the project team or advisory board. Overall, thirty-one researchers were contacted, yielding a response rate of 55%.

3.1 Methodological Considerations

Some important considerations and decisions have made our study both similar to and different from preceding studies. These include:

1) Our study was preliminary, and in the context of other studies would fall into the preliminary pilot stage. The questions we developed were drawn out of our literature review and initial discussions with the project advisory board. We centered the questions around issues of value, commitment, and responsibility, and then added questions relevant to other gap

[4] A second study included metrics related to stewardship commitment [42], but did not undertake measurement.

[5] The PARSE.Insight project found that 20% of respondents submitted data to a data archive [30]; a University of North Carolina study found this number to be 17% [59]; at the University of Nottingham 2% of respondents said they stored data in an institutional repository (the only repository option) [41]; at the University of Exeter about 4% indicted they deposited data in a public repository when they have finished with it [39].

areas (e.g., infrastructure, sustainability planning) as they supplemented and supported these focal areas. Gathering relevant information in the least amount of time was a primary goal.

2) We decided to target project principle investigators (PIs) as subjects. We realized that PIs might describe their data and the way it is managed differently than others involved in the project,[6] but were concerned with learning about data value, stewardship commitments, and responsibility for stewardship and believed PIs to be primary sources for this information.

Table 1. Stewardship Gap Question Areas

Research Context	What is the purpose of the project? What domains of science or creativity are the resulting data in? Who are the project collaborators and funders? What are the characteristics and what is the overall size of the project data?
Commitment	For how much of the data is there: a formalized commitment to preserve; an intention to preserve; no intention to preserve (though no intention to delete); the data are temporary (and will be deleted)?
Stewardship	Who is currently stewarding the data? What is being done to take care of the data? Are there any concerns about the ability to fulfill the intention or commitment? What prospects exist when the current commitment or intention is over?
Value	Why are the data valuable and for how long? How does the valuation affect stewardship decisions? Would it be worthwhile to reassess the value of the data in the future?

3) We asked PIs about data from a single project, rather than data from all projects that they might be responsible for. This decision was made in order to have a coherent view of what it is we were discussing with researchers: a single research project, however broadly that might be defined. We asked researchers in particular to describe data from a project of their choosing where the project was:

- Funded by a public- or non-profit source
- One for which they were responsible for generating the digital data or creative content
- One for which they were able to speak confidently about questions of size, content characteristics, and preservation commitments related to the data.

4) We did not present interviewees with any definitions or parameters for understanding "data". As a pilot study, our concern in this and other areas was to hear researchers answer from their own perspective about the questions we raised (although we did define Steward and Preserve, two terms that were important to our framework for measuring commitments on data).[7]

5) For the purposes of analyzing results, we treated "datasets" as the researcher defined them. For instance, if a researcher designated three different datasets, one each of interviews, field samples, and GIS information, we understood these to be three datasets, regardless of the formats or types of data included in each.

6) We asked about specific types of value and value duration, and researchers' agreement with whether specific types applied to their data. We presented categories of value, but also gave researchers the opportunity to add their own (see Table 3 below, and following).

7) We asked researchers to place the data generated in their projects into one of four categories of commitment, associating a term of commitment with each where applicable. We choose these categories specifically to distinguish between formal and informal commitments on research data, and to look for patterns in the association of specific types of value with types and durations of commitment. The four categories are given in Table 1.

4. RESULTS

4.1 Research Context

Seventeen PIs were interviewed for the study. In the seventeen projects they described, PIs provided information about value and stewardship commitments on a total of 40 datasets. Table 2 shows the distribution of researcher fields, the number of datasets described in each area, total size of the datasets, and whether datasets included sensitive information (information that is private, proprietary, or confidential). Excepting environmental studies where two researchers were interviewed, only one researcher was interviewed in each discipline.

Table 2. Research Discipline and Dataset Details

Researcher Discipline	Number of Datasets	Size of all datasets	Sensitive data
Geography	5	< 5 GB	None
History	6	< 5 GB	None
Archaeology	2	< 5 GB	--[8]
Economics	1	< 5 GB	All
Political science	2	< 500 GB	A portion
Psychology	1	< 20 TB	A portion
Public administration	3	< 100 GB	All
Information	3	< 500 GB	A portion
Education	2	< .1 GB	A portion
Environmental studies	6	< 500 GB	A portion
Human physical performance and recreation	1	< 100 GB	A portion
Neuroscience	2	< .1 GB	None
Astronomy	1	< 50 TB	For a time[9]
Computer sciences	1	< .1 GB	None
Physics	3	< 50 TB	A portion
Statistics	1	< 500 GB	None

[6] Two studies [45, 46] found differences in data descriptions by principle investigators and researchers, and a third [31] found data created by researchers that were not passed on to data managers.

[7] We defined Steward as "to responsibly manage data that is in your care (including the wide variety of activities that might be involved in managing them)" and Preserve as "to execute a set of activities with the explicit goal of maintaining the integrity of data over time."

[8] Did not ask

[9] Data were restricted during a time of analysis, and then released to the public.

Five projects reported no sensitive information in resulting data, eight included some sensitive information or were restricted for a time, and all data were sensitive in two projects. In only one project where a portion of data were sensitive did a researcher make an explicit distinction between the value associated with the sensitive data and the non-sensitive data.

The start and end dates of the investigated projects spanned from 1948 to 2018 with most projects taking place between 2000 and 2015. Some of the projects had been continuously funded for decades, some were completed, and some were still ongoing. Many of the projects were conducted in multiple phases with funding from different sources, and some were continuations of or components of other projects. Despite these complexities, researchers did not have trouble identifying the specific data associated with the projects they selected for the interviews. Regardless of time, the number of funders, or changing collaborators, researchers had a strong sense of a cohesive activity that they viewed as a project, and its associated data assets.

We wished to cause as little disruption to researchers as possible and therefore notified them in advance that no research into the details of their data were required prior to the interview. We asked about details nonetheless in order to gauge what might be required to obtain this information if desired in a more in-depth study. We found that difficulty describing sizes and attributes of data varied across respondents. Many knew approximate sizes and formats offhand or had the information readily available during the interview. Others had a strong sense of what data were collected or produced (e.g., interview transcripts, images, etc.) but could not recall specific details.

A related issue we encountered, experienced in previous studies as well, resulted from researchers' understandings of what were considered "data". In most interviews, the researcher's description of his or her data evolved over time, as they remembered additional sets of data or provided more information in order to answer subsequent questions (for instance about the stewardship environment). In a few cases, however, we found that certain sets of data were not described initially because they were not considered as "project data" by the researcher. Some of these types of data included images taken on-site during field studies, audio of interviews, data that are produced as primary data are analyzed and refined, descriptive and contextual information about the data, and video recordings of study participants.

Whatever their challenges in remembering the details about data, interviewees had little difficulty answering questions about commitments on data, data value, or responsibility for stewardship.

4.2 Type of Value

The interview asked researchers to indicate their degree of agreement with eighteen different types of value (see Table 3). The degree choices were strongly agree, agree, neutral, disagree, and strongly disagree. Researchers could also indicate they were unsure or that the type of value did not apply, or specify "other" types of value. The eighteen value types can be grouped into four main categories, as shown in Table 3: value due to 1) reuse potential, 2) cost of reproduction, 3) impact, and 4) scholarly practice.[10] We also asked questions about the value of data over time. There were four datasets for which no information about value was obtained and one dataset for which only partial information about value was obtained, primarily due to time constraints on the interviews.

Table 3. Types of Value

Reuse potential: Audience	Value for the researcher's own research
	Value within the researcher's immediate community of research
	Value outside the researcher's immediate community of research
	Broadly applicable value (e.g., as cultural heritage or inclusion in a reference collection)
Reuse potential: Reasons for or factors that affect reuse	Value increases in combination with other data
	Data only has value when combined with other data
	Value due to the organization or usability of the data
	Value due to the timeliness or timely relevance of the data
	Value for use in support services (such as calibrations or search services)
	Value for audit purposes or because the data have been mandated to be kept
Cost	Value because the data would be costly to reproduce
Impact	Value due to demonstrated or potential impact (in terms of people, money, time, policy, transformative potential, or some other factor)
Scholarly practice	The data are retained in conformance with good scholarly practice
Change in value over time	The data have gained value over time
	The data will gain value over time
	The data have lost value over time
	The data will lose value over time[11]
	The data are timeless (will never lose their value)

Some of the "other" types of value respondents mentioned were:

- Historic value
- Value to facilitate research (training data)
- Value to facilitate policy-making
- Use for quotes in outreach
- Use as examples in teaching and executive development
- Repeatability, reference, transparency
- Longitudinal value
- Model for other studies
- Type of study: Resolution and context (moving between individual and societal analysis)

4.3 Type of Value and Term of Value

Figure 3 shows the main categories of value that researchers strongly agreed applied to their data, and the durations over which they believed the data would have value. Value for researchers' own use is represented separately from value for others' use. Reasons for or factors that affect reuse and information about change in value over time are excluded from the chart to focus the results on the high-level value categories investigated.

As Figure 3 indicates, researchers believed much of their data would have value for a long time. They most frequently

[10] Keeping data for reasons of good scholarly practice is not strictly a type of value. However, there was such strong agreement with this as a reason for keeping data that it is included alongside other results.

[11] Questions about lost value were only asked if respondents were neutral or negative about increase in value.

expressed strong agreement with the value data held for their own research, followed by value due to the cost involved in producing data, value as evidenced by reuse by others (including both in and outside their immediate community), and the demonstrated or potential impact the data could have. Researchers also strongly agreed that they retained their data in conformance with good scholarly practice.

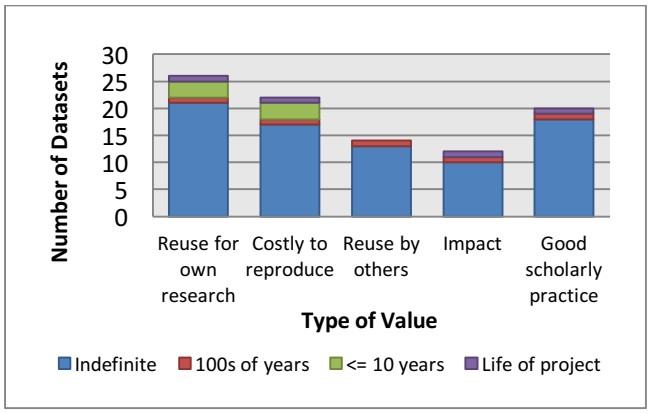

Figure 3. Type of value and term of value

By contrast, when researchers were asked which reasons for value had the greatest influence on decisions about preserving data, demand for data was most frequently cited, with difficulty of reproduction and use for their own research mentioned least frequently. These results show a difference between the types of value researchers most strongly agree that their data have and the reasons for value that have the greatest impact on decisions about data preservation.

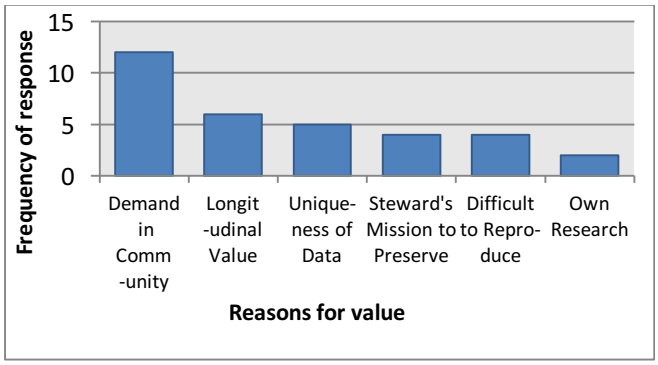

Figure 4. Reasons for value with the greatest impact on decisions about research data preservation

4.4 Commitment and Value

Figure 5 shows the types and terms of commitments that researchers associated with their project data. Our results indicate that researchers have strong intentions to preserve much of their data. While nearly ¾ of datasets carried either an intention or commitment of preservation, however, only two of the twenty datasets desired to be kept more than 10 years had a matching duration of commitment (see Figure 5 – there is a commitment term of more than five years for only two of the five datasets where commitments were expressed).

Juxtaposing type of commitment with term of value (see Figure 6) reveals a similar story, with only 5 out of 37 datasets believed to have value for more than 10 years carrying a commitment of any duration (three of the five commitments are for less than 5 years).

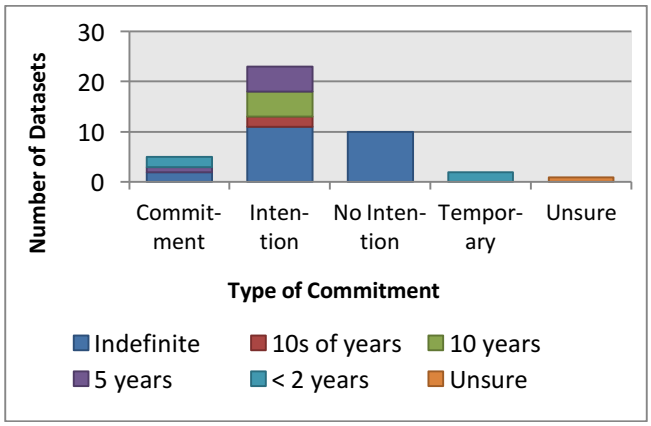

Figure 5. Type of commitment and term of commitment

These results raise an important question about whether intentions to preserve data translate ultimately into preserved data. It is notable that only one quarter of the datasets were identified as having indefinite value, but carried no preservation intention or commitment.

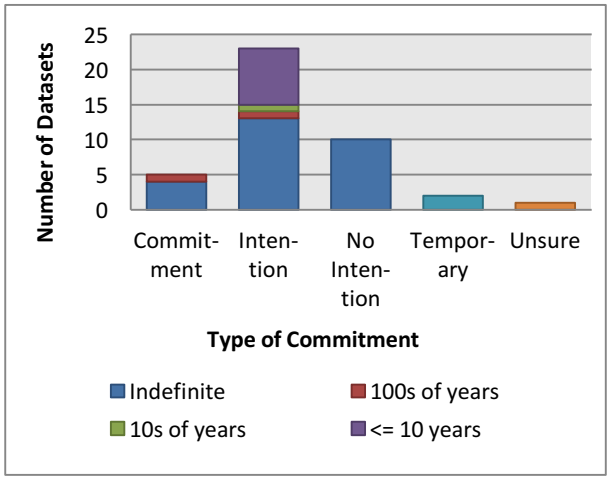

Figure 6. Type of commitment and term of value

4.5 Responsibility

The initial study gathered information about those responsible for funding the creation of research data, and those who have or might have ongoing responsibility for stewardship of the data, whether the role is as a funder or executor of stewardship activities. A tabulation of the most common funding sources for the projects investigated is given in Figure 7. Many of the projects had more than one funder.

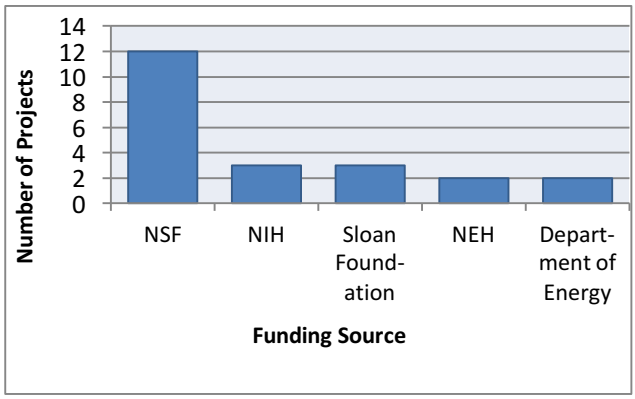

Figure 7. Sources of project funding

In only two of the seventeen projects had project data been transferred to someone other than the one who was originally

responsible for the data during the project. Figure 8 shows who researchers indicated was responsible for stewardship, separated into categories of personal stewardship (the data are on a personal computer, removable media, etc.), stewardship within an institution (within a lab or institutional repository) and multi-institutional or public stewardship (e.g., a repository that is operated on behalf of or for use by multiple institutions or the public). The figure also shows responses of researchers when asked how confident they felt in the ability of the person or entity stewarding the data to fulfill the commitment on intention that existed on the data.

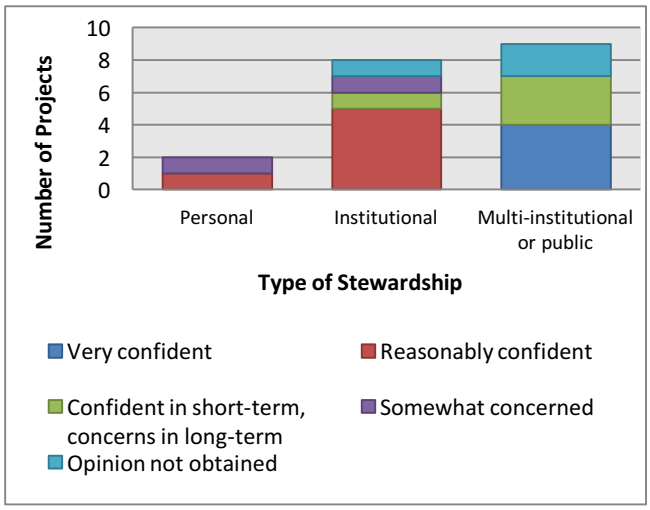

Figure 8. Responsibility and confidence in stewardship

No responses were obtained for three of the projects, but responses from the remaining projects were somewhat mixed. There is high confidence in multi-institutional or public repositories, but also concern about funding beyond the near term. There is reasonable confidence in institutional solutions, but also concern, including about long-term funding. There is both confidence and concern related to personal stewardship.

Some important questions in interpreting these results are the degree of knowledge researchers have about the environments where their data are stewarded, and how well founded their confidence is. No trend emerged in our interviews regarding the former. Some researchers displayed exceptional knowledge about the stewardship environments for their data while others were less knowledgeable, both about environments and which departments or staff were managing the data.

The question of confidence is also complicated, and relates to the issue of intentions translating into preserved data. As noted earlier, there are a number of considerations in determining the adequacy of management and preservation solutions. The ways we have determined to address issues of confidence in the second phase of research are given in the final section of the paper.

In addition to concerns about stewardship during the current period of commitment or intention, our results indicate that attention should be paid to stewardship after the period of commitment or intention is over. While it is a not a new notion that stewardship needs exist after the period of active data use, our results uncover not to a workflow issue (what happens with data when the project is over) but a commitment issue (what happens when the commitment or intention on the data is over). Responses to the question of what plans exist when the current commitment or intention is over are shown in Figure 9. We did not ask the question consistently across all interviewees. However, with only one researcher indicating definite plans for stewardship, the question bears broader investigation. Even if it could be demonstrated that data were secure while a researcher is active in the field, what plans for valuable data exist after the researcher has retired or no longer has an intention to keep the data?

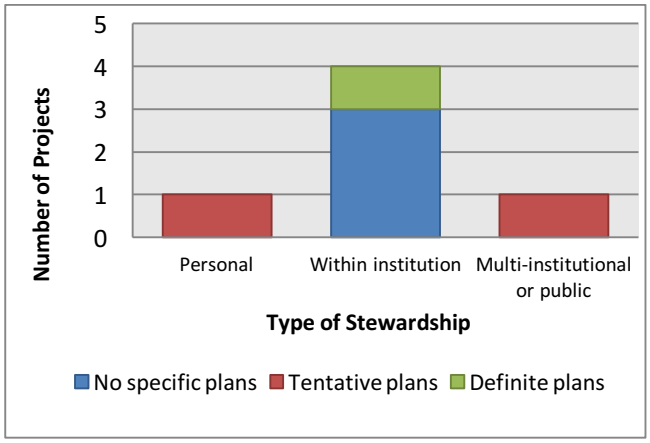

Figure 9. Stewardship plans after existing commitment is over

4.6 Size of Dataset and Data Value

We did not find any correlation between the size of data and the type or term of value, or the size of data and confidence in stewardship, a result that is relevant to future study of the stewardship gap. Researchers across the spectrum of projects believed strongly in the value of their data whether the data were larger or smaller, and types of value were distributed rather evenly across sizes. Similarly, our results did not show that larger sets of data were taken care of better than smaller sets or vice versa. This result has implications for the development of a strategy for measuring the stewardship gap going forward. If it holds true for a larger sample of projects, investigation of the stewardship gap should recognize the value and impact of large data and small alike.

5. DISCUSSION AND FUTURE WORK

5.1 Discussion

The preliminary interviews concluded with questions about whether the interview allowed respondents to describe their data in a way that was meaningful and accurate from their point of view, and what difficulty, if any, they experienced in answering the questions. The response to the first question was unanimously positive in fifteen of seventeen cases where it was asked,[12] and only minor difficulty was reported by two respondents in answering questions about commitment.

These results suggest that we successfully fulfilled the first goal of the initial study, which was to develop an instrument capable of gathering information about data value, stewardship commitments, and responsibility for stewardship across a range of disciplines.

The second goal of the study was to obtain information to inform a more in-depth study of the stewardship gap in a second phase, and our results provided this as well. In particular, they uncovered three main sets of questions, given below.

The first set of questions relates to whether the areas investigated are adequate to provide indicators of a stewardship gap. For example, preliminary results indicate that there could be a large amount of data regarded as being high in value that lack a

[12] We did not receive a response for two interviews due to time constraints.

sufficient commitment for stewardship. If this result were borne out in a larger sample of projects, however, would it point to a meaningful gap (that is, is there cause for concern or do intentions to preserve data in fact result in valuable data being preserved for future use)? What information might be needed to clarify or confirm this?

Our results also indicate a lack of correspondence between particular stewardship environments and confidence in stewardship. What additional information might be needed to validate researcher confidence, or assess the strength of existing stewardship environments?

A second set of questions relates to the selection of an appropriate sampling frame for a more structured study. A lack of correlation between data size on one hand and data value and confidence in stewardship on the other indicate a need to include a diversity of data sizes. How should a sample be structured to do this? Previous studies have found that discipline, researcher role, and level of experience have an impact on many factors such as amounts of data generated, attitudes about data sharing and management and preservation practices. Do all of these variables need to be represented in a sampling frame to provide meaningful results, and if so how can they be best represented?

A third set of questions relates to the granularity of information gathered. For instance, we did not investigate responsibility at the depth of some of the previous surveys (some of which included specific responsibility for data storage, management, etc.). However, the personal, disciplinary, institutional, and multi-institutional dimensions of responsibility appear to be appropriate high-level indicators of who can act to address a stewardship gap if it exists. Is this correct and do these levels provide adequate guidance as to who should act to address a gap if one is found to exist?

Similarly, the level of detail we obtained about data sizes and attributes appears to have been sufficient to associate distinct data with specific commitments, types of value, and responsible entities, our core indicators for determining the presence of a stewardship gap. However, many researchers were not entirely confident in their representation of specific formats used and sizes of data. How accurate do the descriptions of data need to be to provide a meaningful characterization of the stewardship gap?

As other studies have found, there is a direct relationship between the granularity of information that is gathered and the difficulty and amount of time needed to gather it. What is the optimum balance for obtaining meaningful results with minimum imposition on respondents?

5.2 Implications for Future Work

In light of the preceding questions and what we have learned from phase 1 of the study, the following are the major decisions and modifications we intend to make to the instrument in the second phase:

To further address the question of whether intentions translate into commitments for data stewardship we will clarify the purpose of projects, recognizing that there can be a difference between projects with a focus on data creation (e.g., with the explicit purpose of sharing with other researchers) and those where data are not explicitly intended for sharing. We will also gather information about what researchers expect will happen to their data when the current intention or commitment to preserve the data is over, and seek to better understand researchers' goals when transferring responsibility for data to others.

In the development of a sampling frame, we intend to focus first and foremost on stratification by researcher discipline and funding source. This is due to time considerations and the need to prioritize certain variables to keep interviews to a reasonable length. Additional factors may need to be explored more in a further study.

On the question of granularity, we intend to keep to the levels of information we have been gathering about data size and characteristics, stewardship environments (e.g., personal, institutional, multi-institutional), and responsibility for stewardship. While further study may indicate that greater granularity is needed, the current levels appear adequate to our purposes of investigating the presence of a stewardship gap and making general recommendations about how to address it if we find one exists. As above, more detailed analysis may be necessary to make targeted recommendations, depending on the results of the second phase of research.

6. ACKNOWLEDGMENTS

Our deep thanks are due to the Stewardship Gap project advisory board, including 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, and to the Alfred P. Sloan Foundation for their funding of the project.

7. REFERENCES

[1] Addis, M. 2015. *Estimating Research Data Volumes in UK HEI*. http://figshare.com/articles/Estimating_Research_Data_Volumes_in_UK_HEI/1575831

[2] 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.

[3] Alexogiannopoulos, E. et al. 2010. *Research Data Management Project: a DAF investigation of research data management practices at The University of Northampton*. http://nectar.northampton.ac.uk/2736/

[4] Ashley, K. 2012. Generic Data Quality Metrics – what and why. (Arlington, VA, 2012).

[5] 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*. http://www.arl.org/storage/documents/publications/digital-data-report-2006.pdf

[6] Averkamp, S. et al. 2014. *Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs*. http://ir.uiowa.edu/lib_pubs/153/

[7] Beagrie, N. et al. 2012. *Economic Impact Evaluation of Research Data Infrastructure*. Economic and Social and Research Council. http://www.esrc.ac.uk/files/research/evaluation-and-impact/economic-impact-evaluation-of-the-economic-and-social-data-service/

[8] Beagrie, N. and Houghton, J. 2014. *The Value and Impact of Data Sharing and Curation*. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf

[9] Beagrie, N. and Houghton, J. 2013. *The Value and Impact of the Archaeology Data Service: A Study and Methods for Enhancing Sustainability*. Joint Information Systems Committee. http://repository.jisc.ac.uk/5509/1/ADSReport_final.pdf

[10] Beagrie, N. and Houghton, J. 2013. *The Value and Impact of the British Atmospheric Data Centre*. Joint Information Systems Committee. http://repository.jisc.ac.uk/5382/1/BADCReport_Final.pdf

[11] Becker, C. et al. 2011. A Capability Model for Digital

Preservation: Analysing Concerns, Drivers, Constraints, Capabilities and Maturities. (Singapore, Nov. 2011). http://www.academia.edu/1249924/A_Capability_Model_for_Digital_Preservation_Analysing_Concerns_Drivers_Constraints_Capabilities_and_Maturities

[12] Berman, F. et al. 2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. (2010), 110. http://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf

[13] Big Data Value Association 2015. *European Big Data Value Strategic Research & Innovation Agenda*. Big Data Value Europe. http://www.bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final.pdf

[14] Borgman, C.L. 2012. The conundrum of sharing research data. 63, 6 (2012), 1059–1078. http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/epdf

[15] Borgman, C.L. et al. 2014. The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. *Proceedings of the Joint Conference on Digital Libraries, 2014 (DL2014)*. (2014). http://works.bepress.com/borgman/321

[16] Brown, S. et al. 2015. *Directions for Research Data Management in UK Universities*. http://repository.jisc.ac.uk/5951/

[17] Cummings, J. et al. 2008. Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations. *Technology*. 3, 2 (2008). http://web.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

[18] Fearon, D. et al. 2013. *ARL Spec Kit 334: Research data management services*. Technical Report #9781594079023. Association of Research Libraries. http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/

[19] Federer, L. et al. 2015. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *PLoS ONE*. 10, 6 (2015). http://dx.doi.org/10.1371/journal.pone.0129506

[20] Finn, R. et al. 2014. *Legal and ethical barriers and good practice solutions*. Policy RECommendations for Open access to research Data in Europe (RECODE). http://recodeproject.eu/wp-content/uploads/2014/05/D3.1-legal-and-ethical-issues-FINAL.pdf

[21] Gibbs, H. 2009. *Southampton Data Survey: Our Experience and Lessons Learned*. University of Southampton. http://www.disc-uk.org/docs/SouthamptonDAF.pdf

[22] Guindon, A. 2014. Research Data Management at Concordia University: A Survey of Current Practices. *Feliciter*. 60, 2 (2014), 15–17. http://connection.ebscohost.com/c/articles/95923248/research-data-management-concordia-university-survey-current-practices

[23] Hedstrom, M. et al. 2006. Producing Archive-Ready Datasets: Compliance, Incentives, and Motivation.

[24] Hofelich Mohr, A. et al. 2015. Data Management Needs Assessment - Surveys in CLA, AHC, CSE, and CFANS. (2015). http://conservancy.umn.edu/handle/11299/174051

[25] Houghton, J. and Gruen, N. 2014. *Open Research Data Report*. http://ands.org.au/resource/open-research-data.html

[26] Jerrome, N. and Breeze, J. 2009. *Imperial College Data Audit Framework Implementation: Final Report*. Imperial College London. http://ie- repository.jisc.ac.uk/307/

[27] Jones, S. et al. 2008. The Data Audit Framework: a toolkit to identify research assets and improve data management in research led institutions. (London, UK, Sep. 2008). http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf

[28] Kejser, U.B. et al. 2014. *4C Project: Evaluation of Cost Models and Needs & Gaps Analysis*. http://www.4cproject.eu/d3-1

[29] Kroll, S. and Forsman, R. 2010. *A Slice of Research Life: Information Support for Research in the United States*. Technical Report #1-55653-382-9 978-1-55653-382-2. http://www.oclc.org/content/dam/research/publications/library/2010/2010-15.pdf

[30] Kuipers, T. and Hoeven, J. van der 2009. *PARSE.Insight: Insight into Digital Preservation of Research Output in Europe: Survey Report*. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[31] Manyika, J. et al. 2011. Big data: The next frontier for innovation, competition, and productivity. (2011), 156. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[32] Manyika, J. et al. 2013. Open data: Unlocking innovation and performance with liquid information. (Oct. 2013), 103. http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

[33] Marchionini, G. et al. 2012. Curating for Quality: Ensuring Data Quality to Enable New Science. (2012), 119. http://dl.acm.org/citation.cfm?id=2582001

[34] Martinez-Uribe, L. 2009. *Using the Data Audit Framework: An Oxford Case Study*. http://www.disc-uk.org/docs/DAF-Oxford.pdf

[35] Mitcham, J. et al. 2015. *Filling the Digital Preservation Gap. A JISC Research Data Spring Project. Phase One Report*. http://figshare.com/articles/Filling_the_Digital_Preservation_Gap_A_Jisc_Research_Data_Spring_project_Phase_One_report_July_2015/1481170

[36] National Academy of Sciences 2009. Ensuring the integrity, accessibility, and stewardship of research data in the digital age. 325, 5939 (2009), 368. http://www.nap.edu/catalog.php?record_id=12615

[37] National Research Council 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. National Academies Press. http://www.nap.edu/catalog/10613

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

[39] Open Exeter Project Team 2012. *Summary Findings of the Open Exeter Data Asset Framework Survey*. University of Exeter. https://ore.exeter.ac.uk/repository/bitstream/handle/10036/3689/daf_report_public.pdf?sequence=1

[40] Organization for Economic Co-operation and Development 2015. *Making Open Science A Reality*. https://www.innovationpolicyplatform.org/content/open-science

[41] Parsons, T. et al. 2013. *Research Data Management Survey*. University of Nottingham. http://admire.jiscinvolve.org/wp/files/2013/02/ADMIRe-Survey-Results-and-Analysis-2013.pdf

[42] Peng, G. et al. 2015. A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets. *Data Science Journal*. 13, 0 (Apr. 2015). http://datascience.codata.org/articles/abstract/10.2481/dsj.14-049/

[43] Pepe, A. et al. 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 (2014), e104798. http://dx.doi.org/10.1371/journal.pone.0104798

[44] Perry, C. 2008. Archiving of publicly funded research data: A survey of Canadian researchers. *Government Information Quarterly*. 25, 1 (2008), 133–148. http://www.sciencedirect.com/science/article/pii/S0740624X07000561

[45] 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, 4 (2011), 387–403. http://dx.doi.org/10.1080/0194262X.2011.626340

[46] Pienta, A.M. et al. 2010. The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. (Nov. 2010). http://deepblue.lib.umich.edu/handle/2027.42/78307

[47] Piwowar, H.A. and Chapman, W.W. 2008. Identifying Data Sharing in Biomedical Literature. *AMIA Annual Symposium Proceedings*. 2008, (2008), 596–600. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655927/

[48] Podesta, J. et al. 2014. *Big Data: Seizing Opportunities, preserving values*. Executive Office of the President. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[49] Read, K.B. et al. 2015. Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. *PLoS ONE*. 10, 7 (Jul. 2015), e0132735. http://dx.doi.org/10.1371%2Fjournal.pone.0132735

[50] Scaramozzino, J. et al. 2012. A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University. *College & Research Libraries*. 73, 4 (Jul. 2012), 349–365. http://crl.acrl.org/cgi/content/abstract/73/4/349

[51] Sunlight Foundation and Keserű, J. 2015. We're still looking for open data social impact stories! *Sunlight Foundation*. http://sunlightfoundation.com/blog/2015/02/25/were-still-looking-for-open-data-social-impact-stories/

[52] Sveinsdottir, T. et al. 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). http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf

[53] Tenopir, C. et al. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*. 6, 6 (2011). http://dx.doi.org/10.1371/journal.pone.0021101

[54] Tenopir, C. et al. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE*. 10, 8 (2015). http://dx.doi.org/10.1371/journal.pone.0134826

[55] The Royal Society 2012. *Science as an Open Enterprise*. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE-Summary.pdf

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

[57] Turner, V. et al. 2014. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. http://idcdocserv.com/1678

[58] Ubaldi, B. 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Publishing*. No. 22 (May 2013). http://dx.doi.org/10.1787/5k46bj4f03s7-en

[59] UNC-CH 2012. *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership*. University of North Carolina Chapel Hill. http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf

[60] Vickery, G. 2011. *Review of Recent Studies on PSI Re-use and Related Market Developments*. http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk

[61] Waller, M. and Sharpe, R. 2006. *Mind the Gap: Assessing Digital Preservation Needs in the UK*. Digital Preservation Coalition. http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk

[62] Wallis, J.C. et al. 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 (Jul. 2013), e67332. http://dx.doi.org/10.1371%2Fjournal.pone.0067332

[63] Wynholds, L. et al. 2011. When Use Cases Are Not Useful: Data Practices, Astronomy, and Digital Libraries. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (New York, NY, USA, 2011), 383–386. http://escholarship.org/uc/item/4tk5d7hx

[64] York, J. et al. 2016. What Do We Know About The Stewardship Gap? https://deepblue.lib.umich.edu/handle/2027.42/122726