The Stewardship Gap: A Challenge in Long-Term Access to Data

Myron P. Gutmann[1*], Jeremy York[2], Francine Berman[3]

[1] Institute of Behavioral Science and Department of History, University of Colorado Boulder, Boulder, Colorado, United States of America

[2] School of Information, University of Michigan, Ann Arbor, Michigan, United States of America

[3] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, United States of America

* Corresponding Author

E-mail: Myron.gutmann@colorado.edu

Abstract

Despite broad consensus among many in the scientific research, data, and policy communities about the importance of preserving and sharing research data, there are significant concerns about the adequacy of measures being taken today to enable these activities. The difference between current activities and best or ideal policies and practices constitutes a gap that this article describes: the stewardship gap—a gap that will require innovative strategies by researchers, research organizations, and research sponsors to address. The authors interviewed 46 active researchers, drawn from a variety of scientific domains, to understand their perspectives on the value of their research data, the length of time their data would remain valuable, and the kind and extent of commitments in place to ensure ongoing preservation of valuable data. In all, the researchers provided descriptions, valuations, and prospective plans for 120 datasets produced in 46 projects. Four concepts are valuable for understanding our findings: the kinds of *commitment* researchers receive from data stewards; who takes *responsibility* for stewardship; the *value* of the data as perceived by the researcher and others; and the *length of time* over which data are valuable and commitments exist. Based on this study as a representation of the larger cohort of data created with federal and foundation R&D support, research data are "at risk." This is especially so when data are valuable and the length of time for which there is a preservation commitment is less than the length of time that the data will have value. Closing gaps in commitment and responsibility is essential if valuable data are to be effectively preserved. This calls for clear policy directives from government agencies and other research sponsors in partnership with research-performing institutions, designation and acceptance of responsibility, and supporting human and financial investments for the research data the community deems as valuable.

Introduction

There is broad consensus among many in the scientific research, data, and policy communities that effectively preserving and sharing research data is critical for advancing scientific progress. At the same time, there are significant concerns that adequate provisions are not in place to ensure the long-term availability of data that are valuable for future research. The difference between the current situation and best or ideal policies and practices constitutes a **gap** in stewardship practice for many research environments and is particularly critical when data are deemed valuable by the community. In the following, we refer to this as the *stewardship gap*.

Understanding the stewardship gap requires that we learn more about which data are most critical now and in the long term, and what stewardship practice is necessary and desirable in different contexts. This involves an assessment of value, commitment, and resources in the stewardship ecosystem: what gives the data value, how individuals and institutions make commitments about those data, which resources are employed, and how resources are distributed. Moreover, identifying gaps is not enough; *addressing* the gaps will require using what we learn about the gaps to develop targeted strategies and policies that help researchers, research-oriented institutions, and policy makers to effectively allocate resources that will ensure adequate stewardship.

In this paper we report on research designed to enhance our knowledge of the stewardship gap, identify vitally important aspects of the gap, and recommend some opportunities for improving policy and practice that will reduce the most significant aspects of the gap. Ultimately, the questions being asked are about investments: which data are most worth

preserving, what kinds of resources should be devoted to preserving them, and how those investments should be made.

The research we report here builds on a literature about data sharing and data preservation that has emerged over the last two decades, which has shed light on a growing body of knowledge about the context in which data sharing takes place and the attitudes of researchers about sharing their data. The literature on research data sharing that is most relevant for our work involved three distinct research modalities: examination of policies that mandate sharing established by research organizations and publications (1-7); examination of publications and research grants to see whether they produced or cited data and whether those data could reasonably by reused (2, 4, 8-14); and surveys of researchers, either in a single domain or across a wider span of the scientific world (6, 8, 15-20). In addition, there are a substantial number of studies that focused on a single research institution, and examined policies, practices, and attitudes of researchers working in that institution towards data sharing and preservation, in both the U.S. (20-28) and elsewhere.

These studies are diverse, and they draw a number of valuable conclusions. They confirm that there is much less data sharing than is desired or required by journal, institution, or sponsor policies, reflecting a lack of commitment by researchers and their organizations for sharing their data (2, 11-14, 16, 19). In addition, many research datasets are inadequately preserved, too often backed up on personal computers, or copied onto thumb drives, rather than deposited at expert repositories. Over time, these poor stewardship conditions lead to data that cannot be shared, because they are lost or corrupted (9). At the same time the authors of these studies learned that good policies, especially those of journals that require that data be properly stewarded and shared, lead to greater preservation and sharing, as compared to journals that have weak or non-

existent requirements (1, 3, 4, 10). Finally, what limited research explores change over time shows that commitments to data sharing and reuse have increased over time, while recognizing that there are differences between fields, with researchers who study human subjects the least likely to share their data (18, 19).

Our research was designed to draw on the conclusions of others we have noted, but to go further and in new directions. We are especially interested in getting more detailed knowledge about the gaps in data stewardship, and in the relationships between those gaps and the value researchers perceive in their data, the kinds of stewardship commitments they have received, and in how long researchers perceive that their data will have value and the stewardship commitments they receive will last. Our research has allowed us to do just that.

Moreover, our research builds on earlier work on our part, in particular on two publications as well as regular and in-depth deliberations with an expert Stewardship Gap Project Advisory Group, named in the Acknowledgements. In the first of our earlier publications (29) we summarized the considerable literature about data stewardship in an effort to identify important areas of research consensus as well as areas where future research would be valuable. That publication has an extensive bibliography of literature related to data stewardship, and points to an even larger on-line bibliography (https://doi.org/10.7302/Z2ZW1J47). We have not attempted to duplicate that discussion or the bibliography here. The most important conclusions of that article fall into four areas.

First, and most important, there are multiple gaps, not just one, and they represent a diverse array of stewardship elements. We identified fourteen in all, with six that are most important, listed in Table 1.

Table 1: Main Stewardship Gaps Identified

Stewardship Gap	Explanation
Culture	Gaps arising from differences in community attitudes, norms, and goals that affect data stewardship
Responsibility	Gap between who has responsibility for stewardship and who is accountable for stewardship
Resources	Gap between the people, money, infrastructure, and tools needed to steward data, and what is currently available
Knowledge	Gap between the knowledge needed to effectively steward data, and what is currently known
Commitment	Gap between existing commitments to steward valuable data and those necessary to ensure long-term stewardship
Actions	Gap between the actions taken to facilitate stewardship of data and the actions needed

Second, the literature pays different levels of attention to different gap areas, with noticeable "holes", calling for more research in some areas, including those related to dynamic and adaptable infrastructure, discoverability, collaboration, and levels of funding and staff support.

Third (and related to our second point), research has approached stewardship gaps with unequal depth, leaving us a shallower knowledge of the stewardship process than we might like. In particular, more in-depth research is needed for studies about fragmentation of data management, infrastructure and its shortages, needs for skills, data management for reuse, insufficient data curation, and the identification of what is valuable. This matters because it

provides evidence that supports prioritization of some stewardship investments over others, driving useful strategies for stakeholders in the stewardship ecosystem.

Fourth, while the stewardship literature encompasses studies about metrics (strategies for measurement) and measures (specific efforts to measure stewardship), the overwhelming focus is on specific measures of stewardship gaps. The relative shortage of metrics studies, which provide measurement strategies, calls for significant work on this large and generally unstudied area.

In a second publication (30), we presented results from an earlier version of the research reported here, based on a pilot of seventeen interviews with researchers in sixteen areas of research, representing thirteen U.S. research institutions. The in-depth interviews were conducted during November and December, 2015.

The main conclusions of that phase of our research demonstrated that the kind of in-depth interviews we conducted produce valuable information about the stewardship gap. Importantly, our findings showed that there were a diversity of stewardship arrangements, not easily correlated with size or type of data, or with scientific domain. A relatively small proportion of the data we heard about was well stewarded at the time of the interview, and long-term commitments were often lacking, something that confirms work done by others, and generates concern among researchers (12, 13, 31). We learned that more knowledge was needed to understand what factors gave researchers confidence in sustainable stewardship, and more analysis was needed in order to make confident recommendations about policy activities that would enhance stewardship in the future.

In this paper we turn our attention to a larger body of interviews (now 46), drawn from a broader range of scientific domains and encompassing the complete study. We drill down into

specific questions that emerged as especially important from our research. First, we are interested in understanding how researchers assess the value of their data, initially for themselves, but more critically in the long term, for others. We are also interested in knowing how they perceive the relationship between the value of their data and the stewardship responsibility that falls to them and others. From our perspective, value isn't just something that gets a "score" from one to ten, but rather a multidimensional indicator of the importance of the research activity to the community at large, with the capacity to give us an understanding of why specific data are meaningful (for use by others or because of the difficulty of replacement, for example), or how long the data will be important or useful. In Table 2 we list the reasons for data value provided by our respondents. Data could have more or less value for any one of these reasons (a high value for research inside the community, for example), while simultaneously having a lot of -- or little or no -- value for another of these reasons (for example as a longitudinal data series). In addition to different types of value, data collections can also retain value for different lengths of time.

Table 2: Characteristics of value researchers were asked to choose (controlled vocabulary)

Value for researcher's own research
Value because data would be difficult to recreate
Value due to characteristics of data organization
Value because of current or potential research impact
Value because of inclusion in reference collection
Value for reuse in immediate community
Value due to existence of longitudinal data series
Value because data are timeless (will never lose value)
Value for reuse outside immediate community

We are also interested in the kinds of commitments that researchers make and receive for the stewardship of their data. We assess the strength of promises to preserve and share data and

connect those promises to what we learned about researchers' assessments of value. We ask, for example, whether data that are valuable because they are useful for others are more likely to be preserved than data that are valuable just to the researcher. We make a similar assessment of the length of stewardship commitments and the length of time researchers predict that their data will have value: do data that are expected to have value for a long time have stewardship commitments that will last equally long? This often inadequate intersection of value and commitment provokes a significant finding of our research, one that is noteworthy but not entirely surprising, given the results of other research (12, 13). Even in the case of data for which the value is high, commitments are often not strong and may not last as long as the data have value.

Materials and Methods

Sampling

We gathered the data analyzed in this study from 46 interviews with principle investigators (PIs) of federally- or non-profit-funded research projects. We conducted 29 interviews during the spring and summer of 2016, in addition to the 17 that we reported on in York, Gutmann (30). The additional 29 interviews were conducted with PIs selected from a representative sample of research projects funded by U.S. federal agencies in 2010. We selected projects funded in 2010 to ensure that PIs would have generated at least some data that could be discussed in the interview. This approach differed from the opportunistic sample of researchers we used in our first phase of research (30), where we explored potential questions and evaluated the possibilities for this line of research. We recognize the challenges created by mixing interviews from two different sampling strategies with slightly different questions, specifically that they might have generated incompatible responses from our interview subjects. We

nonetheless believe that the commonality of approach harmonizes the differences, enabling the combined results to have significant value.

We constructed our sample using information on federal obligations for basic and applied research included in the National Science Board's *Science and Engineering Indicators 2016* (32). Appendix Table 4-24 of this report contains information on federal obligations for each U.S. federal agency by discipline. We used these numbers to calculate the percentage of federal funding obligated to each discipline and, correspondingly, the number of projects funded by each agency in each discipline that we would need to select to compose a representative sample of agency-funded projects by discipline (Table 3).

Table 3. Distribution of interview subjects based on funding obligations by federal agency and discipline.

	Percent of Fed. Obligations by S&E Area	Desired # of Cases	Dept. of Agriculture	Dept. of Commerce	Dept. of Defense	Dept. of Education	HHS	Dept. of Interior	Dept. of Trans.	Veteran's Admin.	NASA	NSF
% Federal Obligations by agency			3%	2%	10%	12%	50%	1%	1%	1%	9%	8%
Funder distribution (N=50)			2	1	5	6	25	1	1	1	5	5
Environmental Sciences	6%	3		1			1				1	1
Life Sciences	50%	25	1		1		21	1		1		
Computer Sciences and Mathematics	6%	3			1	1					1	1
Physical Sciences	11%	5			1	2					1	1
Psychology	3%	2					2					
Social Sciences	3%	2	1									1
Other Sciences nec	3%	2					1					1
Engineering	18%	9			2	2	1		1		2	1

We calculated the distribution based on an initial desire to conduct 50 interviews. For instance, because 84% of funding in Life Sciences is obligated from Health and Human services, we determined that 21 of 25 interviews in Life Sciences should be conducted with PIs from HHS-funded projects, given a target of 50 interviewees overall.

Once we established our sampling frame, we obtained data about projects funded by each federal agency in 2010 from the US Government's grant web site, https://grants.gov. Using the grants.gov web site's advanced search feature, we randomly selected grants from what the site calls "new assistance project grants" (so that we were sure to find grants in their first year) received by a "private" or "state controlled" institution of higher education. We then contacted the Principal Investigator (PI) of the selected grants by email. If the PI did not respond to our email, or declined our request for an interview, we randomly selected another project from the same sponsor and discipline, and contacted the PI of that project.

Table 4 shows a breakdown of the disciplines of the PIs we contacted and interviewed, the disciplinary response rate, our targeted percentage of interviews per discipline based on funding levels (% of Gov't Funding), and actual overall percentage of respondents. The important column here is the one labeled "Difference between % of Government Funding and % of Respondents." Cells with a negative value indicate disciplines where we interviewed fewer respondents than desired; cells with a positive value indicate disciplines with a greater number of respondents. Despite contacting a larger number of PIs in Life Sciences and Engineering than the other disciplines, these were the only areas where we interviewed fewer PIs than targeted. We determined the distribution of interview subjects (shown in Table 3) initially based on a desire to conduct 50 interviews. In the end, we conducted 26 interviews after contacting 207 researchers, a sign of the challenges that we and other researchers face in attempting to learn about the experiences researchers have in ensuring the sustainability of their data.

Table 4. Distribution of subjects interviewed, by Scientific Domain

Discipline	Contacted	Interviewed	Response Rate	% of Gov't Funding	% of Respon-dents	Difference between % of Gov't Funding and % of Respondents	# of Datasets	% of Datasets
Environmental Sciences	18	5	28%	7%	11%	4%	10	8%
Life Sciences	83	13	16%	50%	28%	-12%	32	26%
Computer Sciences and Mathematics	15	5	33%	6%	11%	5%	7	6%
Physical Sciences	22	6	24%	11%	11%	0%	23	19%
Psychology	10	3	30%	3%	7%	4%	6	5%
Social Sciences	13	8	62%	3%	17%	14%	25	21%
Engineering	38	1	2%	19%	4%	-15%	1	1%
Other	9	5	56%	3%	11%	8%	16	13%
Total	207	46	22%	~100%	100%		120	100%

Project Questions

The questions used for the initial sample of 17 and subsequent sample of 29 researchers were nearly identical, and those used in the later interviews are included in the supporting materials (S1 Appendix). The questions covered the following areas:

- Project Context: Purpose, domains of science, collaborators, funders, size and characteristics of data
- Commitment: The amount of data for which there is a) a commitment to preserve; b) an intention to preserve; c) no intention to preserve (though no intention to delete); or d) an intention or obligation to delete the data
- Stewardship: Who is stewarding data; what is being done to take care of them; concerns about stewardship; prospects for stewardship when the current commitment has ended

- Value: Why are the data valuable and for how long; how does the valuation affect stewardship decisions; the utility of reassessing the value of the data in the future

The main changes we made to the interview protocol between the first and second sets of interviews were the addition of questions to 1) distinguish the purpose of the project (i.e., primarily to test a hypothesis or to collect data in order to share them), 2) identify whether data collected by others were used in the research, 3) obtain more specific information about the degree of confidence PIs had in the current stewardship of project data and future stewardship plans, and 4) ask explicitly about reasons data have value as articulated in the course of the first set of interviews. We also added a working definition of "data." We removed a question about whether the data were being preserved "for yourself or someone else", because we found this question was confusing and did not add to our results given the other questions asked.

In both sets of interviews, we allowed respondents to select a project of their choice to discuss as long as they were the principle investigator and the project had generated digital data for which they were responsible at the time of the interview or previously. We also provided definitions for two terms: We used "steward" to refer to the responsible management of data (including the wide variety of activities that might be involved in managing them); and we used "preserve" to refer to the execution of a set of activities with an explicit goal of maintaining the integrity of data over time. In the second set of interviews, we defined "data" as digital outputs of research that did not exist beforehand and had not already been deleted. Further details about the rationale behind the interview questions and protocol are discussed in our earlier work (30).

Results

While our research sample was relatively small, our detailed interviews reveal a significant body of information about the characteristics of researchers and their data, their opinions about the value of their data, and the processes by which those data will be preserved and shared (or not). Including our earlier pilot sample, we interviewed a total of 46 respondents in 38 disciplines from 36 institutions. They had support from a variety of organizations (see Table 5), with the largest number from the National Science Foundation and the National Institutes of Health, and smaller numbers from NASA, the Alfred P. Sloan Foundation, the National Endowment for the Humanities, and the Department of Energy. The interviews also allowed us to see the extent to which U.S. researchers are engaging in work that crosses traditional disciplinary boundaries. In particular, the 46 respondents from 36 disciplines told their interviewer that their work covered a total of 79 domains of research. We list the research fields in which the respondents work and the domains of research they described in the supplemental materials (S2 Table).

Table 5: Funding Sources reported by Interview Respondents

Funding Source	Projects		Datasets	
	Number	Percent	Number	Percent
National Science Foundation	19	35.2	54	39.7
National Institutes of Health	16	29.6	35	25.7
NASA	4	7.4	10	7.4
National Endowment for the Humanities	2	3.7	8	5.9
Sloan Foundation	3	5.6	5	3.7
U.S. Department of Energy	2	3.7	4	2.9
Other Federal (1 project each)	5	9.5	15	11.1
Other Non-Federal (1 project each)	3	5.7	5	3.6
Total	*46		*120	

*These columns do not add up to the total because some projects have multiple funders.

In examining the data (fig 1), it is valuable to remember that we asked each respondent to describe a single project, and that each project involved a varying number of datasets. The projects the respondents described took place over a long period of time, but our focus in the second phase of interviews on researchers who had projects that began earlier this decade gives us a heavy concentration in the early 2010s, with the peak in the years from 2010 to 2013. The data also ranged broadly in size, from modest (less than 100 megabytes) to quite large (hundreds of terabytes). The diversity of time frame and size gives us confidence that despite the limited number of interviews, our work captures the diversity of data stewardship that our informants have experienced.

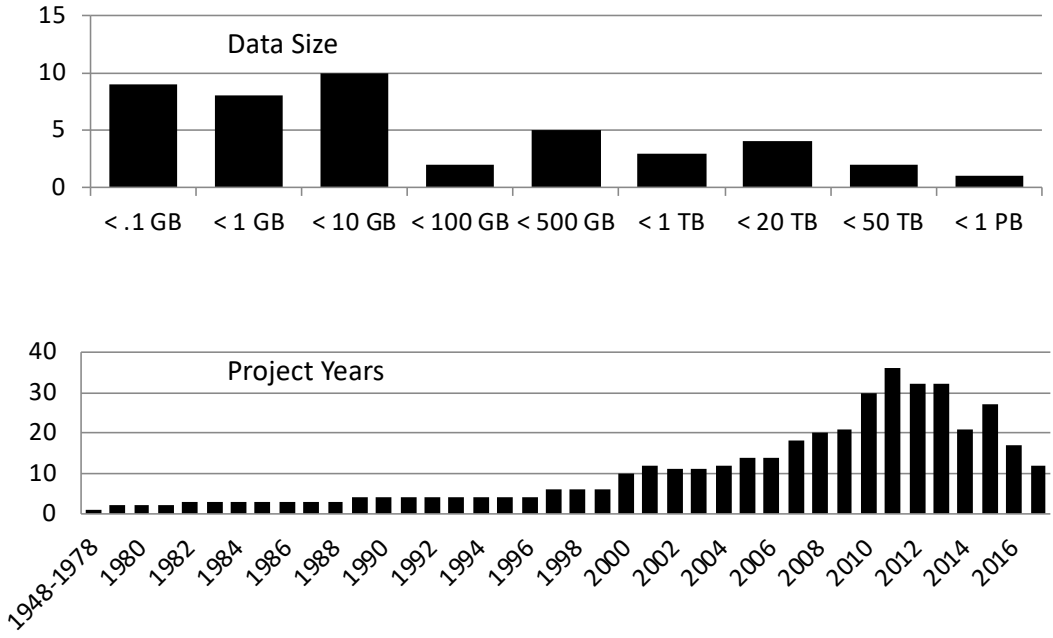

Figure 1: Size and Initial Creation Date of Data Collections as Described by Respondents (number of datasets)

We see key questions about various stewardship gaps as operating across three dimensions, in which one dimension informs us about the value of data, a second dimension explores the risks to data, and a third dimension considers the first two dimensions across time.

We begin with the reasons that researchers believe that their data have value. We asked researchers to agree or disagree with a series of statements about the reasons that their data have value, dividing their responses into agreement, neutral response, and disagreement (fig 2). One of the significant conclusions that we draw from these data is that researchers do not feel that *all* of their data have value, and they feel strongly that their data do not have value for some purposes. It is not surprising that a large proportion of researchers believe that their data have value for their own research, or because they would be difficult to recreate, but it may be surprising -- especially to the data preservation community -- that only half of researchers believe that their data will be useful to their community, and only roughly a third believe that their data will be useful outside their immediate research community.

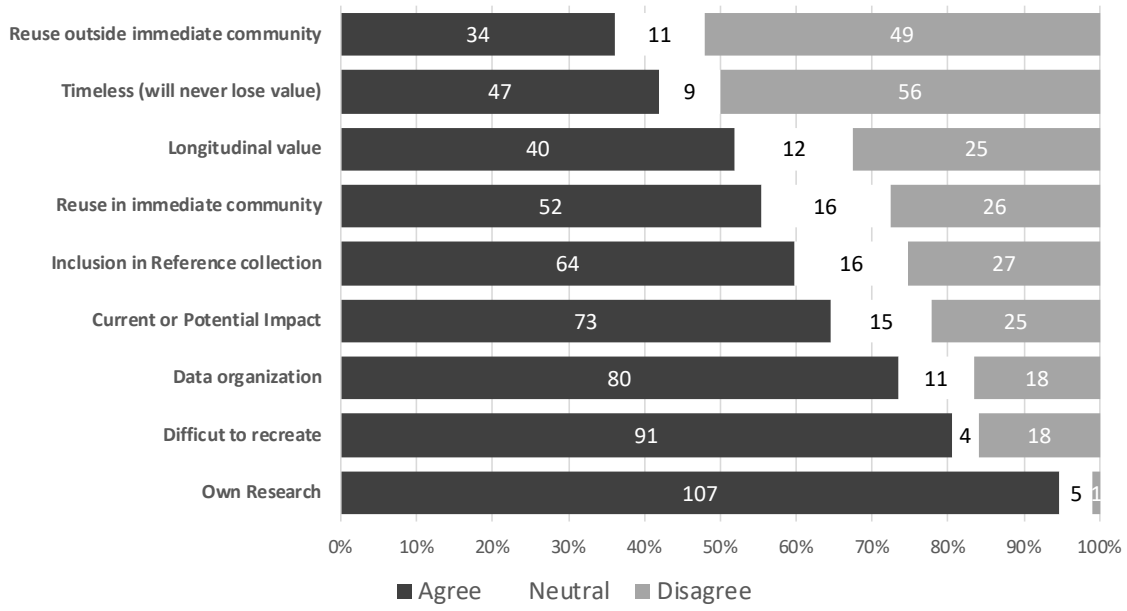

Figure 2: Researcher Attitudes: Extent of Agreement with Type of Value, (number of data sets)

When we add the reasons for value that researchers tell us drive preservation decisions about their data, a different but related story emerges. Here we asked researchers to rank the first, second, and third most important types of value that influenced the decision to preserve data from their project (fig 3). Overall (prioritized by users as first, second, or third) "reuse by others" was the most common, expected use for their “own research” second, “potential reuse” for the researchers themselves, historic purposes, or reproducing results third, and “difficult to recreate” data was fourth. If we limit our analysis to their first order priority, then “reuse by others”, their “own research”, and “longitudinal value” are most important, with “uniqueness” having an interesting place as a strong first order priority, but not as a second or third priority.

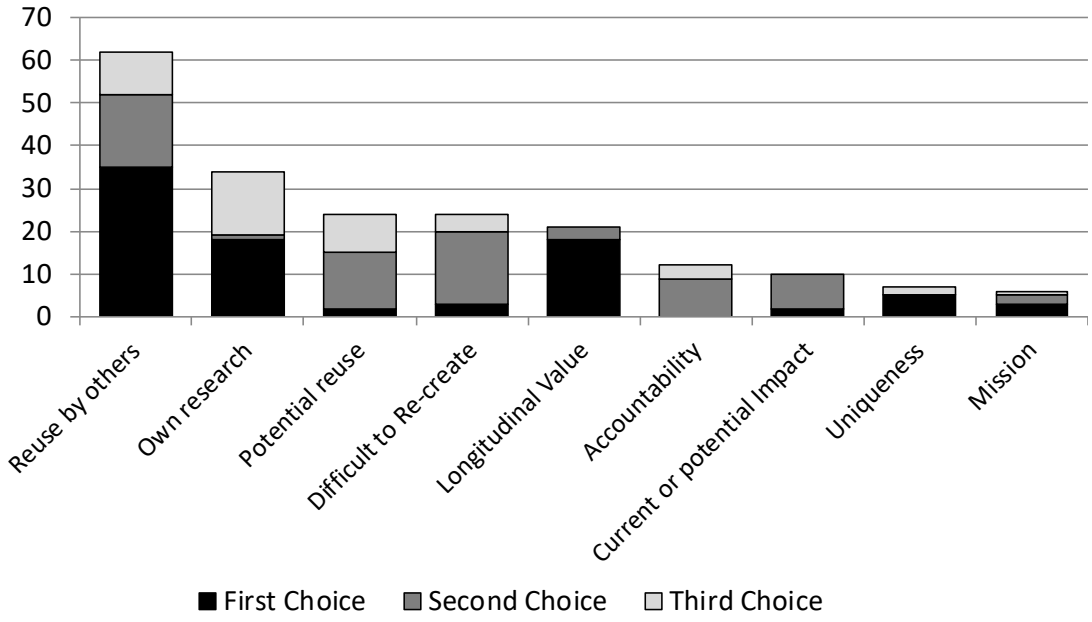

Figure 3: Reasons for Value with Greatest Impact on Preservation Decisions (number of datasets)

The length of time that researchers believe their data have value adds another dimension to the results. Table 6 shows the kinds of commitment that researchers have for preserving their data, and the length of time that they expect their data to have value. The most interesting characteristic of this table is the fact that even several years after their project began, researchers had an intention to preserve 79% of the data sets they described (95 of 120 datasets), but still no solid commitment from a stewardship organization for sustainable preservation. This, despite the fact that researchers expected more than half of the data to have value either indefinitely or for more than ten years, and intended to preserve the data for that length of time. Note that we defined commitment in our interviews as having a written agreement about preservation with a

stewardship organization. This definition sets a high bar, but one that we feel is important because it indicates more than an informal intent to commit.

Table 6: Number of Datasets by Type of Commitment and Term of Value

Type of Commitment	Term of Value					
	Indefinite	> 10 years	<= 10 years	<= 5 years	Undetermined	*Total
Commitment	4	2		1		7
Intention	33	20	13	22	7	95
No Intention	10		4	1		15
Temporary			1	2		3
Unsure				1		1

*This column adds up to 120 because one interview subject indicated that a single dataset had a temporary commitment but also an intention to preserve the data.

It is difficult to over-emphasize the consequences of the finding just articulated. *Researchers value their data and intend for those data to be preserved, but long after their projects have begun, few of them have solid commitments to steward those data.* Put in concrete terms, for only 11 of 120 datasets did researchers have definite plans for stewardship—with some commitments having been confirmed and others not—all involving deposit in a third-party data repository. Another 29 had tentative plans, but many of those lacked the kind of commitment that an expert in data preservation would find sustainable: 13 of the 29 datasets involved backing up their data on a hard drive, while another 9 assumed that their institution would take responsibility for stewarding the data. These are not strong commitments, especially as many institutions are still developing their strategies for institutional repositories.

Our findings of low levels of stewardship planning and preservation commitment are consistent across all durations of time for which researchers believe that their data have value, as

the two panels of Table 7 show. Panel A shows the number of datasets tabulated by the length of time that researchers believe they will have value (columns) and the length of time that they intend to preserve those data. The picture in Panel A appears optimistic at first, because, for the majority of datasets (95 out of 120, excluding the dataset with an undetermined term of intention), the duration of intention to preserve is equal to or greater than the duration that researchers believe the data will have value (along and above the diagonal). Panel B shows the sharp contrast, however, in the number of datasets of any duration of perceived value that have a formal preservation commitment. When we single out written commitments, the picture is much less optimistic, something in keeping with the work of others (12, 13).

Table 7. Length of time that data have value compared with intention and commitment.

A. Length of time that data have value compared to length of time researchers intend for data to be preserved

	Term (length of time) that data have value				
Term (length of time) that researchers have intent to preserve data	**Indefinite**	**> 10 years**	**<= 10 years**	**<= 5 years**	**Undetermined**
Indefinite	31	6	9	18	4
> 10 years	0	11	0	0	2
<= 10 years	10	4	7	1	0
<= 5 years	3	1	1	7	0
Undetermined	3	0	0	1	1

B. Length of time that data have value compared to length of time researchers have a commitment for data to be preserved

	Term (length of time) that data have value				
Term (length of time) that researchers have a commitment to preserve data	**Indefinite**	**> 10 years**	**<= 10 years**	**<= 5 years**	**Undetermined**
Indefinite	2	0	0	1	0
> 10 years	0	1	0	0	0
<= 10 years	0	0	0	1	0
<= 5 years	2	1	0	0	0
Undetermined	0	0	0	Unsure	0

Panel A displays the length of time that researchers intend for their data to be preserved. Panel B displays the length of time that researchers have a commitment for data to be preserved. Cells below the diagonal indicate datasets where the term of the commitment or intention to preserve is longer than the time the researcher believes the data will have value; cells above the diagonal indicate datasets where the term of the commitment or intention is equal to the perceived duration of value; cells on the diagonal indicate datasets where the term of the commitment or intention to preserve is less than the perceived duration of value.

As parties concerned with data stewardship and reuse, we see the lack of commitment to stewardship reported by many of our respondents as dangerous to the long-term sustainability of the research enterprise. On their side, our respondents have their own concerns about their data, which we summarize in Table 8. Researchers expressed concerns about roughly half of the datasets they described (63), with no concern for 18 datasets. They provided less information to us about another 39, making it more difficult to judge their issues.

Table 8: Major Areas of Concern about stewardship sustainability for Interview subjects

Major Concern	Elements of Concern
Understanding	Adequate documentation (metadata) to track, find, access, use data; access to those who worked with the data
Infrastructure	Amount of storage, security, geographic separation
Support	Funding; support from administration
Responsibility	Uncertainty about who will be responsible for the data
Interest	Sustained interest in and attention to the project and data
Technology	External drives will fail; software to access or backup data will no longer work
Management	How the university will manage computer upgrades; how older data is managed over time to be compatible with newly collected data

As we might expect given the topic and the depth of our interviews, there is a complex interaction between the attitudes and knowledge of researchers about stewardship, their concerns for the future of their data, and the value that they see their data holding, now and in the future. One way to look at this is to think about the temporal perspective that researchers hold: where researchers are very confident in the stewardship of their data in the short-term, their concerns have to do with the ability to **understand data** over time (e.g., adequate documentation), and **financial and administrative support** for stewardship and preservation over time. Where researchers are less confident in the short-term, their concerns have to do with **continuity of responsibility**, **technology and infrastructure** (e.g., hardware and software failures and the amount and security of available storage), and the ability to **understand data** over time. These concerns show the interest researchers have in ensuring the stewardship of research, but also the challenges that they face in ensuring that their data can be used in the future. We can also generalize in other ways. Where durations of commitments or intentions match or are greater than durations of value, concerns about understanding data and infrastructure predominate;

where durations of commitments or intentions are less than durations of value, concerns about ongoing responsibility for data predominate. These relationships are shown in Table 9.

Table 9: Number of Studies with each Major Concern by type of value and relationship between term of value and term of commitment.

Major Concern	Term of Commitment or intention Equals Term of Value	Term of Commitment or intention Greater than Term of Value	Term of Commitment or intention Less than Term of Value
Understanding	7	14	1
Infrastructure	7	6	4
Support	4	6	3
Responsibility	0	2	10
Interest	0	6	0
Technology	5	0	0
Management	0	0	5
No concern	8	10	0
No information	25	7	7

Researchers' types of concerns vary according to the durations of commitment or intention and value data have, but what about according to types of value? Table 10 shows the reasons shaping preservation decisions for each term of commitment/intention and term of value combination. The top four reasons in each category are numbered in order.

Table 10: Type of Value with Greatest Impact on Preservation Decisions (rank).

Type of Value	Term of Commitment or Intention Equals Term of Value	Term of Commitment or Intention Greater than Term of Value	Term of Commitment or Intention Less than Term of Value
Reuse	x	x	x
Difficult to re-create	1	4	
Longitudinal	2	x	1
Own research	3	3	2
Uniqueness	4		
Potential reuse	x	2	3
Accountability	x	x	4
Good scholarly practice	x	1	x
Impact	x	x	x
Mission	x	x	x

An "x" in a cell indicates that respondents said that the factor in the left column had an impact on preservation decisions for studies with the terms of commitment/intention and value in each column. The numbers in each column indicate the order of importance for the four most important factors for each set of terms of commitment/intention and value.

We see that decisions driving the preservation of data where the term of commitment or intention is longer than the term of value most frequently have to do with good scholarly practice or the likelihood of continued use by the researcher or others. No researchers expressed the opinion that data in this category were preserved due to the data's uniqueness. Where the term of commitment or intention is equal to the term of value, data are most frequently preserved because they are difficult to re-create or because they are longitudinal in nature. Where the term of commitment or intention is less than the term of value, data are most frequently preserved because they are longitudinal in nature and for the researcher's own use. No data in this last category were preserved due to the difficulty of re-creating them or their uniqueness.

These findings reveal not only systemic risks to data—as evidenced in the relative weakness of preservation commitments and plans, as well as researchers' concerns about data stewardship—but also patterns of risks associated with different types of data value, and different types of data value over different durations of time. The systemic and patterned nature of the relationships between risk and value suggest that systematic policy and procedural interventions could effectively mitigate concerns and ensure adequate commitments to valued data over time. The findings also strongly suggest that success in this area will not be achieved without an accurate assessment and understanding of the reasons data have value and of how long they have value.

An additional finding underscores this point. Table 7 shows a number of datasets that might be characterized as being preserved "too long" (those above the diagonal) and those where the intent to preserve is "not long enough" (those below the diagonal). When we asked researchers about the appropriate length of time after which the ongoing value of their data should be appraised, in nearly all cases where appraisal was deemed worthwhile (more than half of those where a response was recorded), the appraisal term was "just right." That is, if data were appraised at the recommended time, they would be reviewed and either definitively preserved before they were lost, or de-accessioned before they were preserved past their useful life.

Discussion and Conclusions

We have identified aspects of value, aspects of risk, and aspects of value and risk over time that can be identified, measured, and evaluated. Our data and analysis show that stakeholders, stewards, and researchers may have different perspectives about value (e.g., value of data to themselves and reasons for value that drive preservation decisions) and risks (the

relative importance of preservation commitments to data being preserved; the relative strength of different stewardship situations). These groups may also perceive time differently. For example, many researchers do not believe that it is important to think about stewardship until the end of a project or until their retirement despite the long period of time that they may believe the data to have value for reuse (by themselves or others).

Our research leads us to believe that none of these perspectives (of stakeholders, stewards, or researchers) is “right” or “wrong.” Rather, developing effective strategies for stewarding valuable data involves combining information and knowledge from these different sources—researchers’ understanding of value and ability to appraise their own data; stewards’ understanding of infrastructure and data management; stakeholders’ ability to form partnerships and collaborations, allocate resources, and in general foster environments that promote good data management, sharing, and disposition practices.

The conclusions we draw suggest that a large fraction of research data created with federal and foundation R&D support is "at risk." We can go further in describing these at risk data, and divide them into high risk data (intent but no commitment), medium risk data (intent but a term of commitment that is less than the term of value) and low risk data (where the commitment term is equal to or greater than the term of value). We believe that the implications of data being at risk are particularly high when the length of the preservation commitment or intention is less than the amount of time that the data will have value.

These conclusions lead us to focus on the four categories of stewardship gaps identified during this research. Two of these categories are among the stewardship gaps discussed in this paper: ***commitment*** and ***responsibility***. The third is the notion of the ***value*** of data. The last is the role of ***time***, more specifically the amount of time that data have value and the amount of time

for which there are preservation intentions and commitments. Put simply, for any given level of data value, and for any specific amount of time for which we assess value and commitment, research sponsors and research-conducting institutions appear to be generally insufficiently committed to stewarding data, at least from the perspective of researchers generating those data. (Note that these researchers and their institutions are often required by "open data" policies to ensure its access, and therefore stewardship). Coupled with an insufficient commitment, we also find a shortage of responsibility.

Closing the commitment and responsibility gaps should be the first order of business for those who are obligated to think about the long-term value of data, starting with research sponsors and research-supporting institutions, but also including research data repositories and researchers themselves. The starting point must be a clear set of policy directives from government agencies and other research sponsors, in partnership with the institutions that they fund, designed to close the gaps that we find between intention and commitment. This will require a transformation of policy, and an acceptance of responsibility, with financial and human resource consequences. Given the way that we asked our questions, it is difficult to assess how much more we can ask of researchers. Virtually all researchers voiced an intention to have their data preserved.

Our emphasis on two of the gaps we identified should not leave the other gaps unaccounted for. They, too, are important, and demand a response from responsible parties. We especially want to emphasize the importance of the gaps we call culture, knowledge, and action, because these are the areas where the research community needs to stand up, consider their roles, and work to improve our sustainable access to data. In all of these cases we call on researchers, their professional organizations, and the institutions at which they work to build a program of

education and activity that will narrow and ultimately close these gaps. Education sounds easy, but changing culture, improving knowledge, and spurring action is always challenging.

Acknowledgements

We are grateful to 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.

References

1. McCain KW. Mandating Sharing Journal Policies in the Natural Sciences. Science Communication. 1995;16(4):403-31.

2. Noor MAF, Zimmerman KJ, Teeter KC. Data Sharing: How Much Doesn't Get Submitted to GenBank? PLoS Biology. 2006;4(7):e228.

3. Piwowar HA, Chapman WW, editors. A review of journal policies for sharing research data. Conference on Electronic Publishing; 2008 2008/06//. Toronto, Canada.

4. Piwowar HA, Chapman WW. Public sharing of research datasets: a pilot study of associations. J Informetr. 2010;4(2):148-56.

5. Waller M, Sharpe R. Mind the gap: Assessing digital preservation needs in the UK. York, United Kingdom: Digital Preservation Coalition; 2006 2006.

6. Thaesis, van der Hoeven J. Insight Report: PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe. 2010 2010/04//.

7. Addis M. Estimating Research Data Volumes in UK HEI. 2015 2015.

8. Wicherts JM, Bakker M, Molenaar D. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLOS ONE. 2011;6(11):e26828.

9. Vines Timothy H, Albert Arianne YK, Andrew Rose L, Débarre F, Bock Dan G, Franklin Michelle T, et al. The Availability of Research Data Declines Rapidly with Article Age. Current Biology. 2014;24(1):94-7.

10. Vines TH, Andrew RL, Bock DG, Franklin MT, Gilbert KJ, Kane NC, et al. Mandated data archiving greatly improves access to research data. The FASEB Journal. 2013;27(4):1304-8.

11. Pepe A, Goodman A, Muench A, Crosas M, Erdmann C. 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. 2014;9(8):e104798.

12. Read KB, Sheehan JR, Huerta MF, Knecht LS, Mork JG, Humphreys BL, et al. Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. PLOS ONE. 2015;10(7):e0132735.

13. Pienta AM, Alter GC, Lyle JA. The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. The Organisation, Economics and Policy of Scientific Research; Turin, Italy 2010.

14. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol. 2015;13(11):e1002295.

15. Federer L, Lu Y-L, Joubert D, Welsh J, Brandys B. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. PLoS ONE. 2015;10(6).

16. Perry C. Archiving of publicly funded research data: A survey of Canadian researchers. Government Information Quarterly. 2008;25(1):133-48.

17. Tenopir C, Allard S, Douglass K, Aydinoglu A, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE. 2011;6(6):e21101.

18. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS ONE. 2015;10(8):e0134826.

19. Akers K, Doty J. Disciplinary differences in faculty research data management practices and perspectives. International Journal of Digital Curation. 2013;8(2):5-26.

20. Cragin M, Palmer C, Carlson J, Witt M. Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences. 2010;368(1926).

21. Averkamp S, Gu X, Rogers B. Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs.

22. Guindon A. Research Data Management at Concordia University: A Survey of Current Practices. Feliciter. 2014;60(2):15-7.

23. Scaramozzino J, Ramírez M, McGaughey K. A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University. College & Research Libraries. 2012;73(4):349-65.

24. University of North Carolina Chapel H. Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership. Chapel Hill, North Carolina: University of North Carolina Chapel Hill; 2012.

25. Westra B. Data Services for the Sciences: A Needs Assessment. Ariadne. 2010(64).

26. Marcus C, Ball S, Deslserone L, Hribar A, Loftus W. Understanding Research Behaviors, Information Resources, and Service Needs of Scientists and Graduate Students: A Study by the University of Minnesota Libraries. Minneapolis: University of Minnesota Libraries; 2007.

27. Vice Chancellor for Research's Data Management Task Force. Research Data Management at the University of Colorado Boulder: Recommendatins in Support of Fostering 21st Century Research Excellence. Boulder: University of Colorado Boulder; 2012.

28. Johnston LR, editor User-needs assessment of the research cyberinfrastructure for the 21st century. International Association of Scientific and Technological University Libraries, 31st Annual Conference; 2010; West Lafayette, Indiana: International Association of Scientific and Technological University Libraries.

29. York JJ, Gutmann MP, Berman F. What Do We Know about the Stewardship Gap. Data Science Journal. 2018;17.

30. York JJ, Gutmann MP, Berman F. Will Today's Data Be Here Tomorrow? Measuring the Stewardship Gap. iPres 2016 13th international conference on digital preservation; Bern, Switzerland2016. p. 102-11.

31. Fecher B, Friesike S, Hebing M. What Drives Academic Data Sharing? PLOS ONE. 2015;10(2):e0118053.

32. National Science Board. Science & Engineering Indicators 2016. Arlington, VA: National Science Foundation; 2016.

Supporting Information

S1 File. Appendix. Stewardship Gap Project Questions.

S2 Table. Researcher Discipline and Research Data Domain.