The International Journal of Digital Curation
Issue 2, Volume 6 | 2011

"You made it, you take care of it"
Data Management as Personal Information Management

Kathleen Fear,

School of Information, University of Michigan

Abstract

This study explores how researchers at a major Midwestern university are managing their data, as well as the factors that have shaped their practices and those that motivate or inhibit changes to that practice. A combination of survey (n=363) and interview data (n=15) yielded both qualitative and quantitative results bearing on my central research question: In what types of data management activities do researchers at this institution engage? Corollary to that, I also explored the following questions: What do researchers feel could be improved about their data management practices? Which services might be of interest to them? How do they feel those services could most effectively be implemented?

In this paper, I situate researchers' data management practices within a theory of personal information management. I present a view of data management and preservation needs from researchers' perspectives across a range of domains. Additionally, I discuss the implications that understanding research data management as personal information management has for introducing services to support and improve data management practice.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Researchers produce an ever increasing amount of data in their daily work, and there is more and more pressure from funders and publishers to make research data available for other researchers to use. The benefits of preserving and reusing data are significant, but making data preservable and reusable is not a simple task. The extra work required to manage data can be a burden to researchers, especially when it is not a part of their established workflow and when it does not have a measurable, direct benefit for them. Universities have a vested interest in protecting the data their researchers produce, and thus have an opportunity to provide services that will support those researchers in data management. In order to do so, it is important to understand what researchers are doing with their data and how they think about them. Understanding what researchers think of preservation and what they feel to be their most important needs for support in managing their data are critical to incorporating preservation planning into scientific work.

This study explores how researchers at a major Midwestern university are managing their data, as well as the factors that have shaped their practices and those that motivate or inhibit changes to that practice. A combination of survey (n=363) and interview data (n=15) yielded both qualitative and quantitative results bearing on my central research question: In what types of data management activities do researchers at this institution engage? Corollary to that, I also explored the following questions: What do researchers feel could be improved about their data management practices? Which services might be of interest to them? How do they feel those services could most effectively be implemented?

In this paper, I situate researchers’ data management practices as a facet of personal information management. I present a view of data management and preservation needs from researchers’ perspectives across a range of domains. Additionally, I discuss the implications that understanding research data management as personal information management has for introducing services to support and improve data management practice.

Background

Chris Anderson, writing in Wired Magazine, has designated the present time as the “Petabyte Age.” All disciplines are producing digital data in increasingly massive amounts. While high-energy physics and astrophysics have led the way in the “data deluge,” the rest of the sciences and the humanities are not far behind (Anderson, 2008). These data are an invaluable resource for researchers. The ability to reuse data has transformed many fields of study and even created new ones. However, data are rarely standardized across different projects, either within a research group or across fields; frequently data is poorly documented and understandable only to those who created it, and associated information (such as field conditions and instrument calibrations) may not be stored or recorded anywhere at all (Borgman, 2008). In a recent survey of high-energy physicists, about 45% felt that results of their current experiments could have been improved by access to past data, and 40% thought that important historical data had already been lost (Holzner et al., 2009).

Beagrie et al. (2008) identified four major benefits to data preservation. First, data are expensive to generate or recreate, and they may be unique (as in the case of environmental data, which record a snapshot of conditions that cannot be re-measured). Preserving data also gives other researchers the opportunity to validate results or to re-purpose the data in new ways. Readily available data can serve a similar purpose to more traditional publications: promoting the work of the researcher and their home institution, and reinforcing their scholarly reputation. Finally, data preservation is economically sensible: well-preserved data requires less time to search out and convert into a usable form, thus contributing to lower costs and higher productivity for scholars who reuse the preserved data.

Research funding is increasingly contingent on the ability to produce a data management and sharing plan. In the United States, the National Science Foundation (NSF) and the National Institutes of Health (NIH), among others, require grant applications to describe their plan for maintaining and sharing their data during and after the course of the project. However, these mandates are often vague (Borgman, 2008) and unevenly enforced (Noor et al., 2006; Ochsner et al., 2008). Furthermore, researchers who do not receive funding from national funding agencies are typically subject to no such mandates (Lynch, 2008). Funding mandates are not a guarantee of data preservation, even if they are an incentive. Many researchers do not know how to manage their data or are unaware of any services available to help them (Jones et al., 2008; Lynch, 2008). Additional support is necessary to ensure that researchers are able to create and carry out a data management plan.

The NSF's 2005 report, "Long-Lived Data Collections" provided a definition of data that has informed much of the work in the area. This report defined data as: "any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." generated or collected through "observation, computation, or experiment" (National Science Board and National Science Foundation, 2005). This definition includes raw and processed data collected from instruments or through field observation, interview recordings or transcripts, simulations and software used to generate models. However, it excludes, for example, manuscripts written from an experiment, metadata recorded in a laboratory notebook or documents such as signed consent forms.

In 2010, the NSF announced that, beginning January 18, 2011, all proposals will be required to include a data management plan. The NSF, however, has refrained from providing a universal definition of data; the specifics of the data to be documented in the data management plan "will be determined by the community of interest through the process of peer review and program management."[1] This flexibility is important, given that the kind of data generated across disciplines can vary greatly. Further, the ability to share data is complicated by other factors. Kaye et al.'s editorial, while calling for increased data sharing in genomics, acknowledged the difficulties in carrying out that sharing: researchers must "simultaneously fulfil the requirements of funding bodies, honour their obligations to study participants and protect their own interests and careers" (Kaye et al., 2009).

[1] NSF Data Management & Sharing Frequently Asked Questions: http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp.

As the problems of data sharing and preservation move to the foreground, it becomes critical that institutions provide services that support researchers in data management. However, determining which services to provide and in what manner is not a simple proposition.

Literature Review

Cragin et al. (2010) conducted a study of the norms and practices around data sharing in several disciplines, and examined the implications of promoting the use of institutional repositories for data preservation and sharing. They found that there were no well established norms in any of the fields examined, that few scientists had shared data beyond their own collaborators, and that the data that was seen as the most “shareable” tended to be that which was easiest to share, rather than that which might have the highest value to other researchers. In this same vein, Martinez-Uribe and MacDonald (2009) argued that user engagement is the key to successful research data curation, and explored whether the idea of open data is an appropriate tool for engaging users. They found that the idea of Open Access is not sufficient to engage users, and attribute users’ lack of engagement with Open Access repositories to the fact that researchers’ goals are not necessarily aligned with those of repository developers. This work suggests that a major component of enabling data sharing is an understanding of what is important to researchers; if data sharing is seen as less of a priority than other work, or if the benefits of data sharing are not seen as meaningful, less effort will be devoted to curating and making data available.

Improving infrastructure for data sharing is one possible way of making data preservation easier, but several studies have demonstrated that the capabilities of databases are limited. Hine conducted a study of the development of a mouse genome mapping resource, finding that databases in and of themselves do not impose order, standardization or structure on a discipline; rather a database “is an emergent structure that needs to be embedded in an appropriate set of work practices” (Hine, 2006). Properly embedded in work practices, though, successful databases can serve a greater role than simply infrastructure. Cragin and Shankar make the case that scientific data collections (SDC) are sites of Distributed Collective Practice (DCP). SDCs are “constitutive of new kinds of science”; contain a wide variety of data types and associated objects; involve many individuals with a range of skills; consist of a diverse technical infrastructure; and include policies and procedures for sharing, curating, preserving and maintaining the data and the SDC itself (Cragin & Shankar, 2006). This approach takes into account the complex interrelationships between the people that produce and share data, their collaborative practices and the technology in which SDCs are embedded.

The influence of work practices is just as important in the data management of an individual scientist or local collaboration as it is in large scale collaborations or databases. A curatorial approach is difficult if not impossible to carry out if the complexity of the scientific work and the embeddedness of data management practices within that work is not fully appreciated. In her ethnographic study of a biological science laboratory, Shankar observed that scientists use documentation as a way of externalizing memory about experiments and data. She noted that this act can create a personal bond between individuals and they records they create, and that this personal bond “dominates the way individuals and institutions of science conceptualize record keeping practices” (Shankar, 2006). In a later paper, Shankar further emphasized the relationship between scientists and the records they create. Scientists engage in selection and integration of data, synthesize multiple data sources and create annotations in service of creating "a document that is, paradoxically, wholly personal and yet intrinsically professional" - the final published paper (Shankar, 2007).

The extant literature suggests that understanding what researchers perceive the role of data and data management in their work to be is critical to developing appropriate strategies for supporting preservation.

This paper explores the interaction of scientific work practices and data management, framing data management as a facet of personal information management. Personal information management (PIM) deals with information that individuals use to complete everyday tasks. Typically, PIM has focused on information objects like email, files, bookmarks and other sources commonly encountered in an office environment, both paper-based and digital. Researchers deal with all of these kinds of information, but they also use another major source: the data they produce or obtain.

As my subjects spoke about their data, they moved beyond discussing data in the sense defined by the NSF report. The scientists in my study spoke of their data as one of the many things they used to get their jobs done; they did not strongly distinguish the experimental, observational or computational data they collected from other information they worked with, such as laboratory notebooks, manuscript drafts, and emails between collaborators. This paper suggests that bringing data into the fold of PIM can shed light on new and better ways to approach data management. Furthermore, this paper will show that researchers interact with their data as personal information and will explore the implications this orientation toward data has for supporting data management.

Methodology

This study uses mixed methods to examine data management practices and needs at a Midwestern university. First, a survey provided a broad but relatively shallow view; interviews that followed the survey examined issues in more depth and allowed participants to share more detail on their practices.

This study took place at a major Midwestern university, and it included researchers who had conducted funded or human subjects research in the last four years, including faculty and research staff. The university is a large, public institution with growing research expenditures. Research is conducted across the nineteen academic schools and colleges which make up the main campus, in addition to interdisciplinary work between different departments and research carried out in University-affiliated research institutions. Research sponsors include major funders, such as the NSF, NIH and Department of Energy.

Lists of eligible subjects were obtained from the University's Research Division and from the Institutional Review Board (IRB). Together, these lists comprised all researchers who applied for grant funding, plus all those who had studies approved by the Institutional Review Board in the last four years. The University has three campuses and a decision was made to remove research staff not affiliated with the main campus, as well as duplicate names and students, resulting in a population of 2,947 researchers.

This sampling strategy potentially excluded many Humanities researchers, who may be less likely to work with human subjects or rely on external funding. While researchers in the Humanities certainly have data, the primary focus of this study was on the experiences of STEM researchers. To the extent that Humanities researchers were captured in our sample, their responses were included in the analysis, but the survey – and by extension, the interview protocol – was geared primarily toward researchers in science, medicine and engineering.

The lists we received contained only researcher name, email and departmental affiliation. Departmental affiliations were sorted into one of five categories, based on the kinds of methods predominant in the field: medicine, physical science and engineering, social science, humanities, and other. The "other" category contained subjects whose departmental affiliation was unknown and could not be discovered, subjects whose affiliation was with administrative departments rather than content areas, and subjects affiliated with departments or schools that typically employ a wide range of methods (the School of Information, for example).

The survey was developed and administered in collaboration with researchers from the Inter-university Consortium for Political and Social Research (ICPSR). We conducted a survey with the goal of getting a broad view of the range of data management activities at the University. The survey was designed to be very brief: 20 questions, which required only 15 minutes to complete. The survey was based on interview and survey instruments used in other studies of data management (Henty et al., 2008; Martinez-Uribe, 2008).

The survey was structured in four sections. The first was an introduction reiterating important details from the recruiting email. The second section was a background section that asked respondents to identify their school/college and departmental affiliations, along with whether they typically provide or use data management or sharing plans, and whether their data are subject to the Health Information Portability and Accountability Act, which governs the security and privacy of health data in the U.S. The third focused on current data management and sharing practices, and the final section asked respondents to rate a list of possible services to support data management. The survey also included space for respondents to provide contact information if they were willing to be contacted for interviews.

The interview protocol was developed based on the survey instrument. Our goal in the interviews was to gain a better and more detailed picture of data management than the survey provided, so the questions were designed to probe based on the answers given in the survey. The protocol was divided into four sections: background and research overview, current data management practice, current data sharing practice, and evaluation of services. Over the course of the interviews, the interview protocol was refined in response to issues raised by interviewees, which are described in more detail in the discussion section.

An initial email was sent to all 2,947 eligible subjects containing an explanation of the project and a link to the survey. The email was endorsed by the Provost and Vice-President for Research, and was sent via an official ICPSR email address. The survey was administered through Surveymonkey.com.

Of the 2,947 emails sent, 115 were undeliverable. This was because our sample reached back four years, some researchers were no longer affiliated with the University and their email addresses were not active. A further 14 subjects opted out of the survey and any further emails. In the first week of the survey, there were 215 responses. One reminder was sent a week after the initial email, and an additional 165 responses were received over the next week. The survey was intended to be open for ten days, but because subjects were still accessing the survey at the end of that time, the survey window was extended to two weeks and closed after 24 hours of inactivity.

At the end of the survey, participants were asked if they would be willing to be contacted to discuss their results in further detail. Fifteen interviews were conducted, distributed proportionally across the subject domains: seven in medicine, five in physical science/engineering, two in social science and one from the humanities. After interviews with two individuals from the "Other" group, those subjects were reassigned to one of the other categories, based on the kind of data they use and produce (both subjects were added to the physical science/engineering group).

All interviews were conducted in person or by telephone by one researcher and lasted between 20 and 30 minutes. The interviews were then transcribed and then analyzed using NVivo[2]. The transcripts were initially coded according to a codeset based on the interview protocol. The original codeset was not developed with personal information management in mind; instead, those concepts were added to the codeset as the themes emerged.

Findings

The overall response rate was 13.5% (380 responses from 2,816 successfully delivered emails). Of those responses, the completion rate was 94.2% (358 out of 380). Responses from three subjects were removed from the analysis because those individuals were not affiliated with the university. Another 14 responses were removed because subjects indicated that they did not have or use data, or because they did not answer any questions beyond their name. Thus, the total usable responses numbered 363.

Breaking down the respondents by subject area, Medicine had the highest number of responses (173, or 46.5% of the total), followed by Physical Science and Engineering (80, and 23.0%). Social Science was next, with 72 responses, which made up 19.4% of the total responses. "Other" and Humanities trailed, with 33 responses (9.2%) in "Other" and five responses (1.9%) in Humanities. The response rate varied slightly across categories. Social Science had the highest response rate, 17.9%, followed by Medicine and "Other" (12.8% and 11.4% respectively). Physical science and engineering was somewhat lower, at 9.9%, and the Humanities response was quite low (5.1%).[3]

[2] NVivo (QSR International, http://www.qsrinternational.com/products_nvivo.aspx) is a qualitative data analysis software used for creating and managing codesets and coding documents.

[3] The naming convention for subjects is as follows: interview subjects are identified by their disciplinary area (SS = Social Science, MED = Medicine, PS = Physical Science and Engineering, HUM = Humanities) and a number, while survey subjects are identified with an S, followed by their response number and their disciplinary area.

The large response in Social Science is likely due to a disproportionately enthusiastic response by individuals in departments which were associated with the sponsor of the survey (50.0% response rate within that group). Removing those respondents from the group yields a 13.4% response rate for Social Science. The Humanities response is low, as expected. Other than these issues, no systematic differences were found between those who responded and those who did not.

There were 148 interview volunteers, the distribution of which reflected the demographics of the survey respondents. Roughly half the volunteers (49.3%) were in Medicine (including medical research, nursing, and dentistry). Another quarter (22.3%) were in the Physical Sciences and Engineering, 12.8% were in Social Science, 2.7% were in the Humanities, and 7.4% were initially uncategorized.

The survey revealed a varied data environment at the university. Nearly every subject indicated that he or she dealt with data in some shape or form. The few exceptions included theoretical mathematicians (math is often considered a humanities) and some other humanities scholars. The ways in which these researchers interacted with their data differed greatly from discipline to discipline. There is enormous variety in the types and amounts of data handled by different researchers, as well as in the expected lifetimes, frequency of sharing, and formality of management of the data. However, important similarities across all respondents were revealed, particularly in their orientation toward their data as one piece of the work they do. These findings address the activities researchers engage in, their preferences for the kinds of services they would like to be available to them and how they could be implemented. Understanding these details, as well as the way they fit in with the daily work of researchers, is critical to the ability to implement services that will be accepted and useful.

Is Data Personal Information?

Can data be considered personal information? Jones described six ways in which information can be personal:

1. Information that is controlled by (owned by) me;
2. Information that is about me;
3. Information that is directed toward me;
4. Information that is sent (posted, provided) by me;
5. Information that is (already) experienced by me;
6. Information that is relevant (useful) to me (Jones, 2008).

These categories reflect several different ways (described in more detail below) that subjects in this study described relating to their data.

Any given information object need only satisfy one of the conditions above in order to be considered personal information. Data clearly meet at least three of Jones' descriptions:

Controlled by (owned by) me.

Jones describes this kind of information as "the information a person keeps, directly or indirectly … for personal use" (Jones, 2008). Personal use, in this case, means accomplishing some kind of task, whether a personal one (like planning a vacation) or one that is part of an individual's job (like scheduling a meeting or writing a report). The critical piece is the perception of control over the information. In the survey, respondents were asked to select the individual or entity responsible for managing their data; three quarters (n=258) selected themselves. In the interviews, my subjects frequently expressed a feeling of control over their data. SS01 described himself as "protective" of his data and indicated that he felt a strong responsibility to ensuring its safety and appropriate use.

Jones places "owned by" in parentheses because control and ownership do not always go hand-in-hand. Legal rights can complicate control and ownership – the person who holds the rights for the information is not necessarily the one who controls it. After describing his protectiveness toward his data, SS01 went on to say, "It's interesting because it's the university's property. If you get a grant from the federal government, the university actually owns the data." Despite the fact that SS01 does not actually own his data (and knows it), he still feels control over it and a personal responsibility towards it.

Sent (posted, provided) by me.

Jones includes emails, published reports and articles, and personal web pages as examples of this kind of personal information, which suggests that "generated by me" is an appropriate characterization of this category (Jones, 2008). As MED01 succinctly put it: "You made it, you take care of it."

Relevant (useful) to me.

This is the broadest of Jones' categories, and includes information from the two categories above as well as information that is known to be relevant or useful, but has not yet been generated or acquired by the individual. This point is described in more detail below, but this category is important because my subjects often talked about their data management beginning even prior to the data collection process. Even before the data exist, they are known to be relevant to the scientist and are thus personal information.

Data as Part of a Researcher's Personal Space of Information

Jones defines a personal space of information (PSI) as information that is personal, together with information tools, objects, and constructs used to manage information (Jones, 2008). Data, as personal information, are tightly connected with other information and information objects scientists use in the process of writing a paper. The "connectedness" of data to other information in scientists' work is clear in how they talk about the raw data: they are nearly useless without a lot of other information. As MED01 describes, "that file full of numbers means nothing unless you know about the experiment that built the file." This is true across fields, as demonstrated by HUM01's claim that "you need to know the structure of the experiment and the data, the design structure, in order to make sense of these things [data]."

Viewing data as a part of a researcher’s PSI, rather than a discrete unit, contextualizes some survey questions that proved difficult for subjects to answer. Respondents had particular trouble estimating the total volume of their data. Thirty-eight skipped this question entirely. Of those that answered (n=325), 54 (14.9%) answered “I don’t know” or, in interviews, indicated that the amount of data they had varied too much to estimate (SS01). SS01 noted that different studies he has been involved with produced very different types and amounts of data, which made it challenging for him to try to figure out how much he had in total.

Within each category, “less than 1GB” was the most common answer, ranging from 45.8% in the Physical Sciences and Engineering category to 60% in the Humanities. Physical Science and Engineering reported the highest percentage of data over 1TB (22.2%), and while the Humanities reported the highest percentage of data over 1GB (40.0%), the sample size in that group is too small to draw concrete conclusions. Although it was expected that the Physical Sciences and Engineering category would have the largest volume of data, the number of researchers with smaller amounts of data across all groups was surprising. Interviewees in the Physical Sciences and Engineering and Medicine groups indicated that their data is often very small, with the exception of researchers in fields that make use of massive datasets like physics and astronomy. PS01 explained that while he might have many data points from a single experiment, each one is on the order of megabytes per file; he estimated that his lab had generated under 100GB in all their work.

Respondents also found it a challenge to estimate the lifespan of their data, or how long they would have value to them or to others (fewer than five years, five to nine years, ten years or more, or “I don’t know”).

The most common answer (128 responses, or 35.3%) was “ten years or more.” Nearly a fifth (n=70, 19.3%) answered “I don’t know.” In interviews, I asked subjects to elaborate on how they made their estimate. They moved fluidly between discussing grant proposals, raw data, intermediate products of analysis – such as processed data files – manuscript drafts and the final publication, often noting that these different documents are stored in the same place and managed in the same way. MED02’s narrative is both typical and illustrative:

> “[O]n the front end, it’s writing the grant and on the front it’s some of the data and usually it’s sort of written with collaborators. I usually serve as the storehouse on the R drive - it’s a college server, it’s where the budgets and so forth are stored as we put them together and the protocols are stored. Sharing of the writing parts is usually done via email to my collaborators wherever they are in the country. Then usually almost all submissions nowadays are electronic, so via internet, email, they can submit it that way. Then we do a trial, and so data collected on patients will also be stored on our college server. The assay work for the drugs usually I don’t do, somebody else does somewhere out in the country. The’'ll send it to me, like I just got some today. Email. Which I will get and I will store on our server and all analysis is stored on our college server as well. And manuscripts will be written on that server, shared again via email and submitted electronically. The raw data and all versions of manuscripts are stored on the college server as well.” (MED02)

This subject locates the starting point of his data's lifespan not at the moment they are collected but rather when the study is conceptualized and the grant proposal is drawn up. He makes it clear that the complete story of the data includes other documents as well: the grant proposal, emails among his collaborators, and different versions of the manuscript.

Managing Data Over the Short Term

During the active part of data's lifetime, while data are being collected, analyzed and written up, scientists described a number of tools they used to manage their data. Although some used programs they developed themselves or that their students had come up with (PS05, MED01), others noted that the tools they use (or want) are the same as those they use to manage other files, namely Sharepoint (MED02) and Sugarsync (MED01).

While some researchers, like S61MED, felt that they were completely capable of managing their data ("All my issues re data are within my skill set and were easily solved"), others described a conflict between scientists who do data management as one part of their job and individuals who are primarily technical staff, responsible for computing support. MED03 described a "scary" situation: many faculty do not use centralized file storage for data, which she feels means that the data are not backed up. She attributed this to an attitude among some researchers that they could do backup more cheaply than the institution offers by purchasing their own hard drives. But "they don't do it every day because it's not part of their expertise. And then something that was really cheap is not free" because they can easily lose their data.

MED05 expanded on the problem. He felt that researchers "don't know what it takes" to do good data management, and they apply for "nowhere near enough" funding to provide the level of service that they want. He noted that this is especially problematic when a scientist has submitted a data management plan with a grant but doesn't receive enough funding to carry out that plan.

MED04 offered a view from the other side of the problem:

> "The health system has some servers, but it's insanely expensive. When you can buy 1TB from Staples, it doesn't make sense that 1GB costs $3000 or whatever it is. Who could get a grant for that? It does require some extra cost to train, to do the secure storage that they do. It does cost more money, but not that much." (MED04)

Subjects were also troubled by the sense that when they had to troubleshoot data management challenges, especially for large amounts of data, they were reinventing the wheel. MED07 noted that he and his group tended to create their own resources for managing data and that on a later project, "we started working with people who had done basically exactly the same thing we had." The duplication of effort in this area can be a point of frustration.

Additionally, researchers are sometimes stymied by a lack of access to or knowledge of resources available at the university. MED04 described a data management problem she had run into when working with an unusually large dataset:

“We had this huge dataset. […] It was too large for us to crunch numbers on. I’m sure there was some way to get a huge computer, but we ended up having to chop it up and do it one piece at a time.” When researchers must troubleshoot problems on their own, they sometimes end up with a less than ideal fix. However, MED04 was willing to make do with a “good-enough” solution because it allowed her project to move forward; the extra time and effort it would have taken to track down the optimal solution was not worth it.

Data are managed for scientists’ own use and in service of an immediate need. Subjects expressed the view that data management was in their own best interest, rather than being concerned necessarily with the longevity of data for the sake of sharing or reusing it. The primary concern is getting manuscripts done and, in the case of graduate students, completing their dissertation (PS01). MED01 described his awareness from the beginning of his career that carelessness with data could have a negative impact: “I didn’t want to repeat [experiments], so I had to take care of it.” None of my subjects had received any kind of training in data management, instead learning how to do it on the fly. (“I learn, badly, as I go,” PS02 put it bleakly).

Current graduate students are similarly left on their own. PS01 expects his students to figure things out themselves (“I hope they have enough common sense to kind of organize it”), though he does set a day once a year where everyone must sit down and back up their data. One subject noted that he will step in if a graduate student is doing something particularly complex: “If I have a graduate student that does something like that, I take them by the hand and walk them through the process myself,” (HUM01). As SS01 noted, the only training he received in data management was in the context of conducting research generally: “There was no special attention being paid to “here’s a database”, it’s like, “here’s a study, we’re going to analyze it” so you put it in some form that’s analyzable.”

Teaching and learning data management is bound up with teaching and learning to do research, and as a result, the way data management activities are carried out vary as much as individual research styles do. Although my subjects agreed on the range of activities encompassed in data management (developing naming conventions and organizational structures, file backup, etc.), in practice, data management reflects individual’s organizational style, which can be problematic when others try to interpret that data. As PS02 put it, “ultimately it’s an organizational thing, and I’m organizationally challenged.” Somewhat more charitably, MED01 noted that “everybody has their own style […] so it can be tough to decipher their system if they had a system.”

Despite the idiosyncratic nature of data management practices, close to half, 160 individuals (44.1%), indicated that they do typically provide a data management plan when applying for funding. 185 (51.0%) said that they did not do so, and only 16 (4.4%) answered “I don’t know.” Breaking the responses down by category reveals differences across fields.

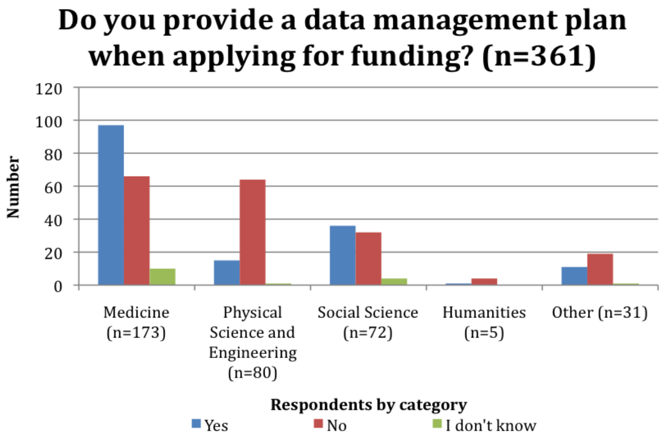

Figure 1. Frequency of providing data management plans.

In the Medicine category, 97 respondents (56.1%) typically provide a data management plan, while in the Physical Science and Engineering category, only 15 (18.8%) do so. Half (36 responses, 50.0%) of Social Science respondents provide a plan. These kinds of plans seem to be seen as just one more hoop to jump through. SS02 indicated that she has “standard language” that her graduate students are instructed to drop into applications; MED03 described the plans she submits as “generic.” SS01, though, noted that his plans do change significantly when the type of data changes from study to study. He also indicated that because the length of the applications is restricted, and he did not feel like a data management plan was worth the space it would take up, if a plan were not required, one would not be included. Subjects agreed that these plans were typically descriptive rather than prescriptive, and the plans allow enough leeway that no major changes to practice are needed to comply with them. Thus, having a plan in place for a project does not necessarily mean that data management practice is more standardized than in a project without a plan.

While most researchers feel that their data management practice works for them, nearly all expressed difficulty with sharing data. A majority (280, or 77.1%) said that they felt their data do have some value for other researchers either at the university or outside of it. A greater proportion of subjects thought their data would be useful to researchers outside the university (266, or 73.3%) rather than inside (226; 62.3%). Just 162 respondents, however, actually share data with other researchers at the university, and 178 said they shared with researchers outside.

Close to two thirds (238; 65.6%) of respondents share their data: 195 (53.7%) share with other researchers at their institution, and 200 (55.1%) share outside of it. However, the interviews made it clear that these numbers reflect sharing in the sense of collaboration, not sharing by making data publicly available. As PS04 described it, “data sharing and work sharing is all kind of wrapped up in the same bundle.” MED07 answered in the survey that he shared with researchers both at the institution and elsewhere, but clarified in his interview. “We only share with collaborators. We don’t just hand it over.” Interviewees shared with colleagues at other institutions (PS01) both domestic (PS02) and international (SS02), and within the institution (MED02), as well as with students in their department (SS01), but in no case did an interviewee indicate that they shared their own research data with someone who was not already a project member, although three (SS01, SS02 and PS02) discussed future plans to make data available.

The picture of sharing as a work process is bolstered by respondents' reports of how they share. Although a third (123, 33.9%) share via a repository of some kind, which implies that at least some of the data is widely accessible, the largest proportion (178, 49.0%) share via physical media like hard drives or CD/DVDs, create shared drives, or use email, among other tools that limit sharing to known individuals.

Interview subjects discussed their difficulties with making data publicly available. HUM01 raised the issue of the amount of time it would take him to make the data understandable to someone else. Sharing raw data that is divorced from its context in the production of a manuscript removes much of its meaning and usefulness. PS05 pointed out the difficulty of gleaning information from raw data: "interesting results are not easily available; just sharing data cubes is no good to anyone else." He further explained, "sometimes people ask for not original dataset but processed data, which is almost a final product. Nobody wants original data." MED02 noted that her data would only be valuable to other researchers doing very similar work to hers, who could use her data to fill gaps in their own existing data.

Somewhat less than half of the respondents provide a data sharing plan when applying for funding: 41.3% (150 responses) answered "Yes," while 52.9% (192 responses) answered "No," and 19 (5.2%) answered "I don't know."

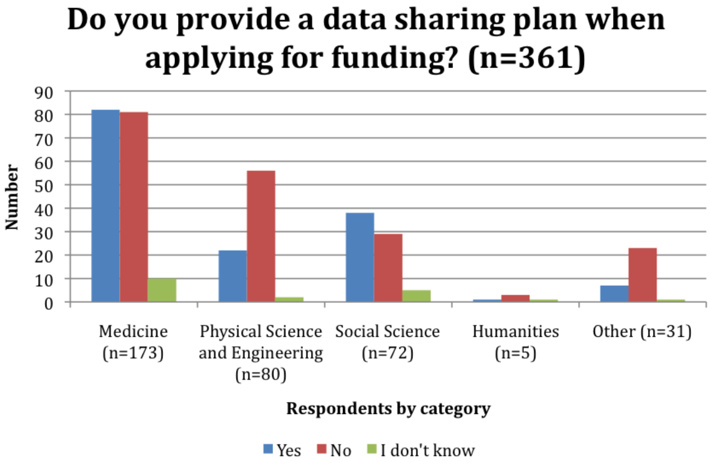

Figure 2. Frequency of providing data sharing plan.

Providing a data sharing plan does appear to be correlated with sharing. Of the 330 respondents who answered both the question about providing a data sharing plan and the questions about sharing, 142 (43.0%) provide a data sharing plan and 188 (57.0%) do not. Among those who provide a data sharing plan, 75.4% (107) reported that they do actually share data; among those who do not, only 64.4% (121) share data. This difference is significant ($p < 0.05$). However, this does not necessarily mean that requiring a plan increases the likelihood that someone will share data. It is possible that the difference reflects the fact that individuals who are collaborating on a project may be more likely to spell out in a grant how that collaboration will work. MED04 noted that some of the data she works with is very tightly controlled, and as a result, when applying for funding, she must specify how the data will be handled within her research group.

Managing Data Over the Long Term

Some interview respondents explained that they felt publishing a paper meant the data was sufficiently shared: “you have your data, you publish it and this is the way you share it” (PS01). The final product, the published manuscript, embeds elements of all the information objects in a scientist’s PSI, and in that context, the data (or some portion) becomes valuable for sharing. PS01 noted that it was rare for anyone to contact him to ask for more data after he published a paper, which he took as an indicator of the relative uselessness of the data beyond what are published.

After data are published, they enter the final phase of their lifetime and are either discarded or preserved. Some data obsolesce, especially that of subjects in the Medical group. Other data simply lose theoretical interest. PS05 noted that his data, while they would remain technically accurate over time, would stop being used eventually because better models could be developed. “It’s like running an old car,” he explained. “It works, but you could get a better one.” However, just because data becomes obsolete or uninteresting does not mean they are discarded; many researchers indicated an impulse to save data “just in case” (MED02, MED06, MED07). No subject indicated that they ever discarded all of their data, and generally only identifiable data (MED06, MED07) were discarded. PS04 noted that she sometimes worked with datasets that were owned by an external entity that specified that she destroy her copies after completing her analysis, but other than that, “data may live on my computer indefinitely.” Although subjects indicated that they did want to retain their data, describing themselves as “conservative” (PS01) or “paranoid” (MED01), preservation planning was generally not more advanced than leaving data on a server that was backed up. No additional clean-up is done with retained data, so saved data can be difficult to interpret, especially once the original researcher graduates, retires or moves to a new institution.

SS02 was particularly concerned with this problem:

> “In fact we’re looking into this right now for someone who’s recently retired and is now hospitalized and has an office full of primary data that no one understands but him and one former graduate student. There is another issue of graduate students […] who’ve done fieldwork and then don’t end up writing a dissertation, which is where the data would be accessible. […] Again, since archaeology is destructive, once you’ve done this, it’s really unethical to not make the data accessible.” (SS02)

Her main concern is with data that does not result in publications. Had the students finished their dissertation, or had the retired professor completed analysis on his data before becoming ill, SS02 felt that the publication would have sufficiently provided for preservation and accessibility of the data.

At the end of the survey, respondents were given a list of potential services to support data management and asked to rate each one on a scale of one to five, with one being least helpful and five being most.

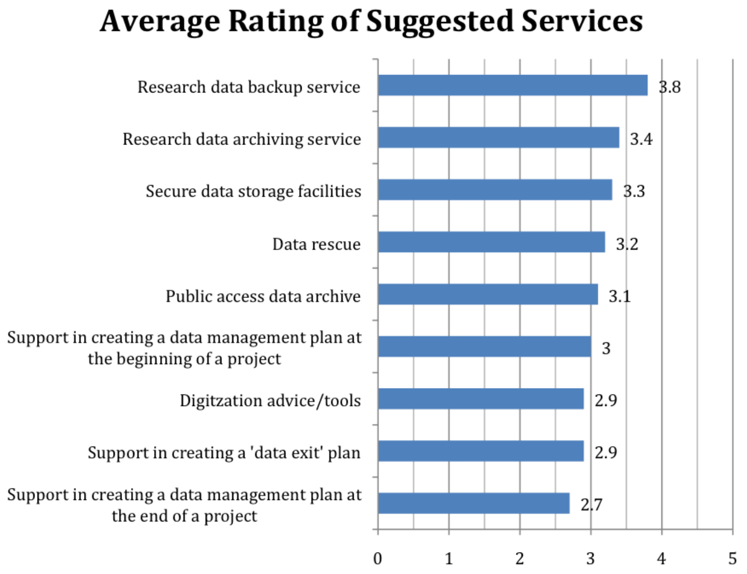

Figure 3. Average rating (1-5) of the helpfulness of data management support services.

Across all categories, a research data backup service was the most popular service, with the exception of the Social Science group, which rated a research data archiving service and secure data storage facilities higher.

Different preferences emerged among subsets of respondents:

- Individuals who typically submitted a data management plan to their funders rated support in creating a data management plan at the beginning of a project higher, giving it a 3.4 overall.
- Respondents who indicated that their data were subject to HIPAA or other confidentiality restrictions placed a high value on secure data storage facilities, ranking it at 3.9.
- Respondents who had larger amounts of data rated a research data backup service especially highly: individuals who had > 1GB of data gave the service a 4.0, and individuals who had more than 1TB ranked it at 4.2.

These preferences were confirmed in interviews, and some additional service interests emerged. Researchers across fields expressed an interest in having training available for both graduate students and faculty that could teach them how to manage their data, how to share it, how to create data management plans, etc. They suggested that training opportunities should be offered in such a way that researchers could take advantage of them as needed, and that consultants should be available to help with data management.

Nearly every participant noted at some point that their data management practice would be better if only it were not so much work. One noted that consent forms sometimes piled up on her desk because the storage area is far away and taking them there is "too much to do every day" (MED02). Another commented that he would not be opposed to sharing his data, but "it's not as well organized as it could be, and certainly not organized well enough that it could be useful for somebody else. The main reason is just that it would just take so much more of my time to do that. I just don't have that amount of time" (HUM01).

Counter Examples: When is Data Not Personal Information?

Three of the individuals[4] I spoke with were not just responsible for the data they produce as individual researchers. They also managed large data collections that are resources for other researchers. Their experience managing that data provides an interesting counterpoint. These datasets fall outside of the scope of personal information management. The information is not under their control (in Jones' sense, meaning the data are not kept by these researchers for their own use), it is not created by them, and although they may use data from it, the collection as a whole is not a part of their day-to-day work as researchers. Correspondingly, these researchers express a somewhat different orientation to these data than to the data they actively produce and use.

The only subjects in this study who expressed a need for significant changes to their data management practices and who were enthusiastic about intervention from the institution were those who manage large data collections, either museum collections or research data centers. In all three cases, these individuals were much more concerned with providing access to the data and with the idea that the data should stand on its own, which is a sharp contrast to the attitude of researchers describing their own data. This suggests that even if the data managed in the context of large-scale, public data collections is the same as that managed by individual researchers, the needs of the people in charge of those collections might be different. The orientation of a researcher toward their data (i.e., "These data are for me and my work" vs. "These data are not mine and not part of my own work") is an important determinant of the kinds of services and support that they might need. A personal information management framework is appropriate for active, individual collections of data, but it may not fit larger-scale, public or widely distributed data collections. Conversely, frameworks that work well for large, shared collections may miss the particular issues and concerns of researchers who manage data locally, on a small scale, and for themselves only.

Discussion

The findings confirm the idea that data management is one part of a spectrum of activities researchers engage in to accomplish their work. As such, the activities they carry out related to data management are not necessarily different from the activities they carry out to manage other information objects they use. This understanding of data management clarifies the reasons behind researchers' preferences for services (and further, their overall reluctance to introduce change into their practices). Data management is difficult to separate from the rest of a researchers' work, and is part of embedded routines that involve other information objects and the production of knowledge. Any change to data management activities potentially has a serious impact on the rest of a researcher's work.

[4] These researchers will not be referred to individually by subject number, as this may make them easily identifiable.

What is Data Management?

In the initial interviews I conducted, I ran into some defensiveness on the part of subjects regarding the concept of data management and whether what they did with their data really was data management. One subject felt that what he did was not really data management because his main purpose was making sure that he himself did not lose access to the data:

> “[Data management] makes it sound more structured than what it really is. […] I end up with hundreds of WAV files and I usually record directly on to my laptop but make multiple backups. So I’ll show up back in the United States with a copy on my hard drive then two or three backup copies on CDs or DVDs or something like that. […] So for myself, I keep very detailed notes about what I did on which data and how I decided which portions to analyze and what to analyze, where I saved what and what the naming convention for the different files were. […] I keep backup copies in my office and at home. So I have multiple CDs and they’re labeled well, so I know if five years from now if someone asks me a question or questions my results or whatever, at least in principle I could come back and retrace my steps.” (HUM01)

This kind of discomfort aligns with Jones’ (2007) observation that in many PIM studies, subjects express unease with their information management and can be self-deprecating about what might appear to be disorganization to someone unfamiliar with their system.

As indicated above in the methodology section, in response to subjects’ qualifying their remarks and general discomfort with answering questions about their practices, I changed the protocol so that the interview began with a broad definition of data management, essentially encompassing any activities that had to do with collecting, analyzing, publishing or preserving data. After that point, participants became more open. In later interviews, I asked participants to describe what they thought the term “data management” meant, whether they felt that they engaged in data management, and whether there was anything they did with their data that did not count as data management. Subjects gave definitions of data management that aligned with my own definition: a broad range of activities that covered the lifespan of the data, from conceptualization to creation or collection through the analysis process to caring for it once a project was completed. All participants felt that data management – according to that definition – was, in fact, something they did (or had done, if they had moved into more of an administrative role than a research one), and that only a few activities – such as sending data to a statistician for help in analysis – did not fall under the umbrella of the term.

An additional source of discomfort may have been due to the fact that we were clear that our results were going to be presented to the Provost and the Vice President for Research at the institution, who had signed the recruitment letter for the survey. We did this to increase the odds of participation, but it may have unintentionally suggested to participants that they were going to be evaluated on their practice. In the survey, seven individuals responded despite indicating that they did not have any data, and had never had any, and never intended to have any, but as one of those participants noted, "I thought I was more or less required to participate in the survey" (S126PS). The effect may have carried over into the interviews. However, this was only a problem in a handful of interviews, and despite their nerves, participants gave valuable answers.

The finding that my subjects initially felt "data management" was too formal a term to apply to what they do with their data echoes the JISC *Incremental* study's finding that it is important to connect with researchers using an appropriate vocabulary (Freiman et al., 2010). However, viewing this finding in light of a PIM framework, it also suggests an important conceptual difference between the way data curators and librarians think about data management and the way researchers think about it. "How do you manage your data?" elicited nervous stares, whilst "What do you do with your data?" prompted rich narratives that revealed connections between data, grant proposals, notes and memos, and manuscripts. Again, this reinforces the idea that data is linked with the other information that makes up a researchers' PSI. Consequently, it is difficult to separate data management activities from the activities researchers carry out to manage all the rest of the information they deal with. Data management is not necessarily a formalized process, but rather actions taken in response to a researcher's current information needs and work goals.

Additionally, data management activities vary over the lifetime of data. Returning to MED02's description of the lifespan of his data, his narrative illustrates another commonality among the researchers I spoke with. In the interviews, it became clear that there are distinct phases in the life of data: first, the kind of data and the methodology are identified, and grants written; second, data are collected, sometimes from multiple sources, and then analyzed; and third, the manuscript is written and reviewed, at which point the data comes to the end of their active life and are either preserved or discarded.

This aligns with Williams et al.'s (2009) three-stage model of PIM. In the first stage, individuals acquire information either by actively acquiring or creating it or by passively receiving it. In the second stage (short-term information management) individuals make the decision to retain or discard information they acquired in the first stage, and perform document and task management activities as they use and reuse the information. In the final stage (long-term information management) individuals decide, once a need is fulfilled, whether to keep information or discard it, and if they decide to keep it, how to do so.

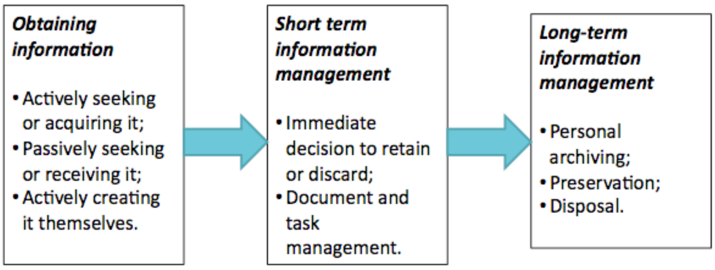
Obtaining information
• Actively seeking or acquiring it;
• Passively seeking or receiving it;
• Actively creating it themselves.
Short term information management
• Immediate decision to retain or discard;
• Document and task management.
Long-term information management
• Personal archiving;
• Preservation;
• Disposal.

Figure 4. Figure compiled based on Williams et al.'s (*op cit.*) three-stage PIM model.

Dealing with data is a part of a researcher's livelihood, and it is not surprising that tenured or tenure-track faculty have worked out data management practices that work for them. As indicated in the results above, many of the researchers I spoke with are satisfied with their data management practices and are wary of making changes or taking on additional work, even while they recognize where there might be room for improvement. Data management is a means to an end – for the subjects I spoke with, uniformly a published manuscript, but in some fields this might be a software product, a publicly available database or other products – not something that is accomplished for its own sake, and so, it cannot take precedence over the other activities researchers engage in.

It is clear that researchers do not think of data management as a separate activity from the rest of their work, and in supporting data management, it is important to remember that it is closely tied to other activities and subservient to researcher's major goal: to produce publishable results. At that point, the data become dormant, for example sitting on a server or computer (PS01, HUM01) or on a shelf in the lab (MED01). Although from the perspective of the researcher, this is the end of the line for the data, from a data curator perspective, this is an opportunity for data professionals to play a part in ongoing curation and management. Treloar and Harboe-Ree (2008) describe this handoff point as the "publication curation boundary." While the data are actively a part of a researcher's work, they are locally controlled and part of that researcher's personal information management process. By contrast, once they are in the publication domain researchers are less engaged with data curation, the responsibility for which can pass to data curators.

Services to Support Data Management

The overall picture of researchers' preferences suggests that they are primarily interested in infrastructural support. In particular, they are interested in support for functions that they do not have time to carry out and that are not part of their day-to-day work, like backup and long term archiving.

The response to the suggested services was, for the most part, lukewarm. It is important to keep in mind, though, that what researchers are currently doing more or less works for them. Because data management is embedded in the work of producing a manuscript, most researchers have developed the expertise that they need to accomplish their main goals. While they recognize that changes to their practice could better enable long term preservation or data sharing, they express concern that "top-down" initiatives to address those problems will compromise their ability to do their job:

> "My major frustration with [this institution's] "services", whether they be IRBs, UCUCA, offices, committees or whatever designed to "facilitate" some aspect of university life, is that the service's modus operandi typically involves the faculty/investigator doing all the work, and the service then looking at the finished product and saying what cannot be done or what is wrong with the proposal, thus sending the faculty/investigator back to re-write or re-think a plan. We already have sufficient peer-review going on; we don't need another layer of this. Services, if they are implemented, should be sharing more of the up-front work load, providing boiler-plate language and assistance required of the outside peer-review mechanisms (NIH, HIIPA *(sic)*, etc.)" (S360MED)

These results suggest that understanding data management practices and how they vary between fields, and even among individual researchers, is critical to implementing any kind of effective service. In particular, initiatives to improve data management across an institution should aim to support researchers in performing activities they feel they do not have time for, not to change practice researchers feel is working adequately for them. Critical to this is challenging assumptions about where the most challenging data management problems reside and what they look like: although "Big Data" sciences have received a lot of attention for the amount of data they handle, this study reinforces the point that other fields are facing challenges as well.

PIM and Data Management

Bringing data into the fold of "personal information" holds promise both for studying data management and personal information management. In particular, thinking about data management as part of the activity of producing scholarly work may also provide inroads into a problem brought up by one survey respondent. S357HUM noted that humanities researchers often "find ourselves excluded from the conversation" by studies like this one that are primarily geared toward science data. While they are involved in what, from the outside, are seen as data management activities, many humanities researchers do not consider themselves to work with "data." Moving the conversation away from "data management" and toward "working with information en route to creating a publication" might allow for a more inclusive discussion of data management.

Including data in personal information management opens up a new area for PIM researchers. Aside from Marshall's (2008) study of scholarly archiving, which focused primarily on archiving manuscripts and characterized datasets as "not archival," there has been little attention paid in the PIM literature to data. However, the results of this paper suggest that expanding explorations of PIM and testing PIM tools not just in a scholarly environment but with the full range of information researchers' work with will be fruitful.

Limitations

Graduate students often take on the burden of data management for a laboratory or research group, and they have far less experience with data management than established researchers do. This group was beyond the bounds of the sample in this study, but further work focusing on this group could be enlightening, as it is possible that graduate students may have a very different view of how well their data management practices work and what services would be beneficial. In the future, it would be useful to conduct interviews with graduate students to get a sense of how different their perceptions may be. If significant differences arise, conducting another survey focused specifically on graduate students could be enlightening.

Additionally, there is unavoidable bias in having individuals volunteer for interviews or fill out surveys. People who are willing to talk about data management are those who are most likely to have something to say about data management. Further, the sample was primarily comprised of researchers in quantitative fields; researchers in the qualitative sciences may have had a more difficult time answering the questions, and the use of the term "data" may have been objectionable to humanities researchers.

Conclusion

Data management is strongly connected with researchers’ daily work creating and publishing manuscripts, and because of this, there is no bright line between data management and what is more often considered personal information management. Researchers engage in a range of data management activities that vary over the course of the data’s lifetime, which is bound up with the other information and documents researchers use to produce their work. In the course of a scientist’s work, raw data are collected, sliced up, analyzed, summarized and condensed, with annotations and notes and files generated all along the way. Pieces of all these documents are compiled into a manuscript, which itself goes through multiple versions before the final version is submitted (and then revised). Data management is part of a continuum of processes, with a grant proposal at one end and a manuscript or other final product at the other and which can include many different processes, all of which tend to blur together as researchers move from document to document and activity to activity. Separating data management from other research activities is confusing to researchers and counterproductive. Bringing data management into the fold of personal information management will aid archivists and curators in understanding how to support data management as part of a larger work process.

Acknowledgements

I would like to thank George Alter, Myron Gutman and Marci Schalk at ICPSR for their work on the development and administration of the survey, as well as their feedback on the interview protocol. I would also like to thank Elizabeth Yakel for her support throughout this project, and Ann Zimmerman for her feedback on this paper.

This work was supported in part by funding from the University of Michigan’s Office of the Vice President for Research. Additionally, this material is based upon work supported by the National Science Foundation under Grant No. 090362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine, 16*(7). Retrieved July 5, 2011, from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). Keeping Research Data Safe. JISC. Retrieved July 5, 2011, from http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx.

Borgman, C.L. (2008). Supporting the “Scholarship” in E-Scholarship. *EDUCAUSE Review, 43*(6). Retrieved July 5, 2011, from http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/SupportingtheScholarshipinESch/163260.

Cragin, M.H., & Shankar, K. (2006). Scientific data collections and distributed collective practice. *Computer Supported Cooperative Work, 15*(2-3), 185-204. doi:10.1007/s10606-006-9018-z

Cragin, M,H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368*, 4023-4038. doi:10.1098/rsta.2010.0165

Freiman, L., Ward, C., Jones, S., Molloy, L., & Snow, K. (2010). Incremental: Scoping Study Report & Implementation Plan. Retrieved July 5, 2011, from http://www.lib.cam.ac.uk/preservation/incremental/Incremental_Scoping_Report_062010.pdf.

Henty, M., Weaver, B., Bradbury, S., & Porter, S. (2008). Investigating data management practices in Australian universities. Retrieved July 5, 2011, from http://hdl.handle.net/1885/47627.

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science, 36*(2), 269-298. doi:10.1177/0306312706054047

Holzner, A., Igo-Kemenes, P., Gjovik, U., & Mele, S. (2009). Data preservation, reuse and (open) access in high-energy physics. Digital Preservation Europe. Retrieved July 5, 2011, from http://www.digitalpreservationeurope.eu/publications/briefs/dp_in_high_energy_physics.pdf.

Jones, W. (2007). Personal information management. *Annual Review of Information Science and Technology, 41*(1), 453-504. doi:10.1002/aris.2007.1440410117.

Jones, W. (2008). Keeping found things found: the study and practice of personal information management. Morgan Kaufmann Series on Interactive Technologies. Amsterdam: Morgan Kaufmann Publishers.

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics: Re-shaping scientific practice. *Nature Reviews Genetics, 10*, 331-335. doi:10.1038/nrg2573

Lynch, C. (2008). The institutional challenges of cyberinfrastructure and e-research. *EDUCAUSE Review, 43*(6). Retrieved July 5, 2011, from http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264.

Martinez-Uribe, L., & Macdonald, S. (2009). User engagement in research data curation. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), *Research and advanced technology for digital libraries: 13th European Conference, ECDL 2009* (Lecture Notes in Computer Science, 5714, pp. 309-314). Berlin: Springer. doi:10.1007/978-3-642-04346-8_30

Marshall, C.C. (2008). From writing and analysis to the repository: Taking the scholars’ perspective on scholarly archiving. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. Pittsburgh, PA: ACM. doi:10.1145/1378889.1378930

National Science Board and National Science Foundation. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century* (Report No. NSB-05-40). Washington, DC: National Science Foundation. Retrieved July 5, 2011, from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

Martinez-Uribe, L. (2008). Findings of the scoping study and Research Data Management Workshop. University of Oxford. Retrieved July 5, 2011, from http://www.ict.ox.ac.uk/odit/projects/digitalrepository/findings.xml.

Noor, M.A., Zimmerman, K.J., & Teeter, K.C. (2006). Data sharing: How much doesn’t get submitted to GenBank? *PLoS Biology, 4*(7). doi:10.1371/journal.pbio.0040228

Ochsner, S.A., Steffen, D.L., Stoeckert, C.J., & McKenna, N.J. (2008). Much room for improvement in deposition rates of expression microarray datasets. *Nature Methods, 5*, 991. doi:10.1038/nmeth1208-991

Shankar, K. (2006). Recordkeeping in the production of scientific knowledge: An ethnographic study. *Archival Science, 4*(3-4), 367-382. doi:10.1007/s10502-005-2600-1

Shankar, K. (2007). Order from chaos: The poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology, 58*(10), 1457-1466. doi:10.1002/asi.20625.

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., Chiang, K., et al. (2008). Digital research data curation: Overview of issues, current activities, and opportunities for the Cornell University Library. Ithaca, NY: Cornell University Library. Retrieved July 5, 2011, from http://hdl.handle.net/1813/10903.

Treloar, A. & Harboe-Ree, C. (2008). Data management and the curation continuum: How the Monash experience is informing repository relationships. In *Proceedings of VALA 2008, Melbourne, Australia*. Retrieved July 5, 2011 from http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf.

Walters, T.O. (2009). Data curation program development in U.S. universities: The Georgia Institute of Technology example. *International Journal of Digital Curation, 4*(3). Retrieved July 5, 2011, from http://www.ijdc.net/index.php/ijdc/article/view/136.

Williams, P., Leighton, J., & Rowland, I. (2009). The personal curation of digital objects: A lifecycle approach. *Aslib Proceedings, 61*(4). doi:10.1108/00012530910973767