JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Data Management Practices Across an Institution: Survey and Report

Cunera M. Buys, Pamela L. Shaw

Buys, C. M., & Shaw, P. L. (2015). Data Management Practices Across an Institution: Survey and Report. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1225. http://dx.doi.org/10.7710/2162-3309.1225

External Data or Supplements:

Buys, Cunera M.; Shaw, Pamela, L., 2015, “Northwestern University Data Management Survey”, http://dx.doi.org/10.7910/DVN/ZMZ5RD, Harvard Dataverse.

ISSN 2162-3309 10.7710/2162-3309.1225

RESEARCH

Data Management Practices Across an Institution: Survey and Report

Cunera M. Buys
E-Science Librarian, Northwestern University
Pamela L. Shaw
Biosciences & Bioinformatics Librarian, Northwestern University

INTRODUCTION Data management is becoming increasingly important to researchers in all fields. The E-Science Working Group designed a survey to investigate how researchers at Northwestern University currently manage data and to help determine their future needs regarding data management. **METHODS** A 21-question survey was distributed to approximately 12,940 faculty, graduate students, postdoctoral candidates, and selected research-affiliated staff at Northwestern's Evanston and Chicago Campuses. Survey questions solicited information regarding types and size of data, current and future needs for data storage, data retention and data sharing, what researchers are doing (or not doing) regarding data management planning, and types of training or assistance needed. There were 831 responses and 788 respondents completed the survey, for a response rate of approximately 6.4%. **RESULTS** Survey results indicate investigators need both short and long term storage and preservation solutions. However, 31% of respondents did not know how much storage they will require. This means that establishing a correctly sized research storage service will be difficult. Additionally, research data is stored on local hard drives, departmental servers or equipment hard drives. These types of storage solutions limit data sharing and long term preservation. Data sharing tends to occur within a research group or with collaborators prior to publication, expanding to more public availability after publication. Survey responses also indicate a need to provide increased consulting and support services, most notably for data management planning, awareness of regulatory requirements, and use of research software.

Received: 02/26/2015 Accepted: 06/05/2015

Correspondence: Cunera M. Buys, Mudd Library, 2233 Tech Drive, Evanston, Illinois 60208, c-buys@northwestern.edu

IMPLICATIONS FOR PRACTICE

1. While it is difficult to extend results of a single institution's survey to an entire community of practice, this survey's results indicate that researchers at Northwestern University struggle with the same issues regarding data management as their peers: issues with long term storage, data organization and management, knowledge of data management plans, and need for consultation and instruction.
2. The U.S. federal government is moving toward enacting requirements for storage and sharing results of federally funded data with the public. Researchers may not be aware of these future requirements. Library-based data management experts can serve as sources of authority on newly enacted policies and requirements.
3. Different levels of data management knowledge exist in a single institution: faculty may have a greater stake in good data management, but the individuals who are managing the data (staff and students) may not have as great an understanding of institutional practices or general practices of good data management. The library and its partners can provide guidance on these practices.

INTRODUCTION

Researchers and other individuals supported by United States federal funding agencies are being placed under pressure to provide evidence of data management practices: including practices for sharing data, protecting sensitive data, and storage or retention of data. The federal government has issued a number of mandates, recommendations, and policies regarding treatment of research data and publications that arise from federally funded projects. Activity surrounding data management practices, especially in academic institutions, is increasing so rapidly that researchers, research offices and university administrators struggle to keep up with the pace of federal recommendations. University libraries are working to place themselves in positions of support and authority in matters of data management. This paper describes the results of a survey of digital data management practices of researchers across all disciplines at a single university. The results of this survey are discussed in the context of the knowledge and awareness of data management at the institution and how the library can respond to improve data management practices.

Federal Mandates and Response from Academic Communities

Since at least 2011, the federal government has advocated public access to peer reviewed publications and digital data resulting from federally funded research. (White House Office of Science and Technology Policy, 2011a; White House Office of Science and Technology Policy, 2011b; White House Office of Science and Technology Policy, 2012). In February 2013, the White House Office of Science and Technology Policy (OSTP) directed federal funding agencies to develop and implement public access plans for both scientific publications and digital scientific data (White House Office of Science and Technology Policy, 2013). On August 4, 2014, the U.S. Department of Energy announced its public access plan (U.S. Department of Energy, 2014). As of June 8, 2015, 16 agencies have released their preliminary public access plans. A group of academic library based data specialists created and maintain a crowd-sourced Google table summarizing the funding agencies' responses. This Google table also contains links to each agency's responses (Whitmire et al., 2015).

While the National Institutes of Health (NIH) has had a Data Sharing Policy for grants in amounts of $500,000 or more since 2003 (National Institutes of Health, 2003), and the National Science Foundation (NSF) required a Data Management Plan (DMP) with each NSF grant application starting in January 2011 (National Science Foundation, 2011), the OSTP directive suggests a more coordinated effort across federal funding agencies to establish policies on open access to research data. This places a burden on researchers to manage, preserve, and share digital research data.

Academic institutions' libraries, computing centers and offices for research have closely followed these proposed mandates for handling research data. Required sharing and retention of federally funded research data will have a great impact on the workflows and practices of researchers and will place a burden of policy definition and computing infrastructure on institutions employing these researchers.

This research study was designed to gather information about management practices of digitally stored data at Northwestern University. The perception among many individuals may be that "data" are primarily generated by practitioners of science, but the term data, when applied to digital objects, can be used for many products of intellectual or scholarly productivity. The simple definition of data becomes less simple when the concept is defined by individuals from arts, humanities, social sciences, mathematics, biological, or physical sciences—or indeed if the term data is used at all. This study describes a survey of digital data management practices across all disciplines at a private university.

The E-Science Working Group (ESWG), which includes representatives from Northwestern University Libraries, Galter Health Sciences Library, Northwestern University Information Technology, Weinberg College of Arts and Sciences, and the Office for Research conducted the study. The survey's goal was to guide the development of training modules, consultation services, tools, and platforms for digital data management and sharing.

LITERATURE REVIEW

The topic of data management and library involvement is popular in current library literature, and more libraries have developed or are developing programs to become involved in data management practices. Much of the literature describing surveys of data management practices and programs in libraries discusses surveys of librarians themselves. The Association of Research Libraries (ARL) conducted a web survey of its members, in attempt to determine the extent of involvement of libraries in e-science or data management (Soehner, Steeves, & Ward, 2010). The ARL survey results suggest that many libraries are involved in data management practices, or are surveying their users in order to develop services for data curation and storage.

Similar surveys of librarian data management involvement are described by Tenopir, Sandusky, Allard, and Birch (2013; 2014) in surveys of U.S. and Canadian academic librarians, by Antell (2014) in a survey of science librarians at institutions affiliated with the Association of Research Libraries and by authors from the Information School at the University of Sheffield in a survey of university librarians in the United Kingdom (Cox & Pinfield, 2014). Pinfield and Cox also conducted a series of interviews with librarians who engage in data management within their institutions (Pinfield, Cox, & Smith, 2014). They identified 7 "drivers" to research data management (RDM) in libraries including storage and security of data and 12 influencing factors on successful RDM in libraries, such as roles, incentives, and skills. They concluded that librarians acknowledge that RDM support is important in libraries but exists on a variety of levels at different institutions, and the combination of drivers and influencing factors present at each institution affects the success of an RDM program at the institution's library.

It is also common to find library case studies on establishing a library based data management program, as libraries describe the lessons learned from starting data services and extend their experiences to offer advice on implementing library based programs (Ball, 2013; Charbonneau, 2013; Henderson & Knott, 2015; Johnston & Jeffryes, 2014).

It is rarer in library literature to find published reports from librarians at academic institutions who have surveyed and interviewed their users to identify institutional needs for data management solutions. Scaramozzino published results of a survey of primarily math and science faculty in 2012 (Scaramozzino, Ramirez, & McGaughey, 2012). This paper investigated data preservation, sharing, and educational needs of faculty. The report described issues in long-term data storage practices and needs for education on data management among faculty. An interesting detail in this paper was that while 65% of faculty considered it important to share data, less than half of those respondents reported that they "always" or "frequently" shared their data openly, despite their belief in the importance of sharing.

Additionally, in 2012, Steinhart published results of a survey of NSF principal investigators (Steinhart, Chen, Arguillas, Dietrich, & Kramer, 2012). This survey was conducted to determine the preparedness of researchers to meet NSF data management plan requirements. The study found that researchers produce a wide variety of data types and sizes, but the majority of respondents create no metadata or do not use metadata standards. The paper concluded that researchers were uncertain about how to meet the NSF DMP requirements.

Averkamp, Gu, and Rogers (2014) published a report on the University of Iowa's data management needs survey. This survey was sent to all faculty and staff directly involved in research. Like the Steinhart survey, researchers generated data in a wide variety of formats. Storage was also a concern among University of Iowa researchers.

Purdue University Libraries have taken a prominent role in data management with their Data Curation Profiles Toolkit, created with the University of Illinois Graduate School of Library and Information Science (Witt, Carlson, Brandt, & Cragin, 2009). Authors from Purdue libraries interviewed faculty and surveyed graduate students to discover students' levels of readiness to manage research data during their careers. In their report, they described the concept of data information literacy, which encompasses aspects of data competency such as metadata, file versioning, ethics, basic database skills, and other skills necessary for the responsible conduct of research and management of digital data (Carlson, Fosmire, Miller, & Nelson, 2011).

Other countries' universities and institutions are equally concerned with issues of good data practice. Oxford University published results of a survey of their institution in a blog post (Wilson, 2013). The authors surveyed investigators from all disciplines (humanities, mathematics, physical sciences and life sciences, medical sciences, social sciences). This survey's results highlighted diversity in data sharing attitudes and a "disappointing" awareness of the university's existing data management infrastructure.

Other authors surveyed the academic community at large to investigate common practices and difficulties in research data management. In 2011, Tenopir surveyed data storage and management needs across several academic institutions and found many investigators are satisfied with short-term data storage and management practices, but are less satisfied with long-term data storage options (Tenopir et al., 2011). This study also revealed that researchers do not believe their institutions provide adequate funds, resources, or instruction on good data management practices. Additionally, there were differences in data sharing or reuse among different academic disciplines, suggesting multiple data cultures within a single institution.

In her 2012 paper, Christine Borgman analyzed four major reasons for data sharing and the challenges associated with them. Data sharing may differ by research community (e.g. the biology community will have different challenges to data sharing than the astronomy or sociology communities). Borgman noted that researchers' reasons for not sharing data are becoming better understood. Barriers to data sharing include the amount of work needed to get data into a form that can be shared, the time and expense involved in sharing and curating data, technological challenges, and the fact that researchers cannot imagine who might use their data.

Each published survey and report adds to the level of knowledge of data management practices and enables comparison across academic institutions. This report describes a survey of data management practices distributed to all academic departments at a research institution and describes differences in responses between schools of the university. Difficulties in surveying such a vast diversity of disciplines will be discussed.

METHODS

Survey Design

Many of the questions used in this survey were borrowed with permission from other university libraries. The majority of questions were adapted from data storage and management surveys from Florida State University and Carnegie Mellon University, who shared the surveys in response to a direct request by the ESWG. Qualtrics software was used to conduct the survey (Qualtrics, LLC, n.d.). Further analysis of the results such as filtering on aspects of the responses to a particular question ("drill down") and cross tabulations were also performed in Qualtics. Text responses to open questions or to questions with an "Other" response were analyzed for frequency of themes using the ATLAS.ti text analysis software package (Scientific Software Development GmbH, n.d.).

This survey contained an introduction and 21 questions. While demographic data on school, department, and appointment status were collected, no names were collected. After review, the university's Institutional Review Board categorized the survey as "Not human subjects research." Survey questions were mostly multiple choice, with an "Other" option (allowing respondents to enter a text answer) provided for many questions. Questions also covered a variety of topics including type and size of data, data storage, data management, willingness to share data, and types of additional assistance or training desired. While the ESWG recognizes that there are many forms of data, the survey was concerned with digital data only.

Survey Population and Distribution

Northwestern University is a private not-for-profit institution composed of two United States campuses and one campus in Qatar, with approximately 21,000 students enrolled full- and part-time. The university comprises an undergraduate college of arts and sciences and several graduate schools, including a medical school and school of law. Northwestern's Carnegie Classification identifies the university as a large 4-year or above level school, with an undergraduate program classified as arts and sciences plus professions, and a graduate program classified as comprehensive doctoral with medical programs. The enrollment profile is majority graduate/professional, and the university is classified as a research institution with very high research activity (Carnegie Foundation for the Advancement of Teaching, n.d.). The university library is an institutional member of the Association for Research Libraries.

The survey was distributed to the United States campuses only. Northwestern's main campus is in Evanston, Illinois and is home to the college of arts and sciences, the graduate school and schools of engineering, communications, business, and most academic departments. The second campus, located in Chicago, Illinois, houses the medical school, law school, and a school of continuing studies.

The survey was conducted between January 15, 2014 and February 17, 2014. The survey URL was sent via Northwestern's bulk email system to all faculty, graduate students, postdoctoral candidates, and selected research-affiliated staff at Northwestern's Evanston and Chicago Campuses (approximately 12,940 recipients). Two reminder emails were sent during the course of the survey.

The survey attempted to obtain as much information as possible about researchers' digital data management practices from a wide variety of subject areas, not only in sciences, but in humanities and social sciences as well.

Survey Responses

Response rate was approximately 6.4% (833 responses with 788 respondents completing the survey). Respondents were allowed to skip questions causing variances in the number of responses for each question. Additionally, five questions allowed for the selection of multiple answers.

RESULTS

Demographics

Respondents were asked to identify their school affiliation and appointment status. Results of school affiliation are shown in Table 1, and appointment status is shown in Figure 1.

School Affiliation	%
School of Medicine	38%
College of Arts & Sciences	24%
School of Engineering & Applied Science	14%
Other	8%
School of Communication	6%
School of Management	3%
School of Education & Social Policy	2%
School of Continuing Studies	2%
School of Journalism, Media, Integrated Marketing Communications	1%
School of Law	1%
School of Music	1%

Table 1. School Affiliation (Percentage of respondents by school)

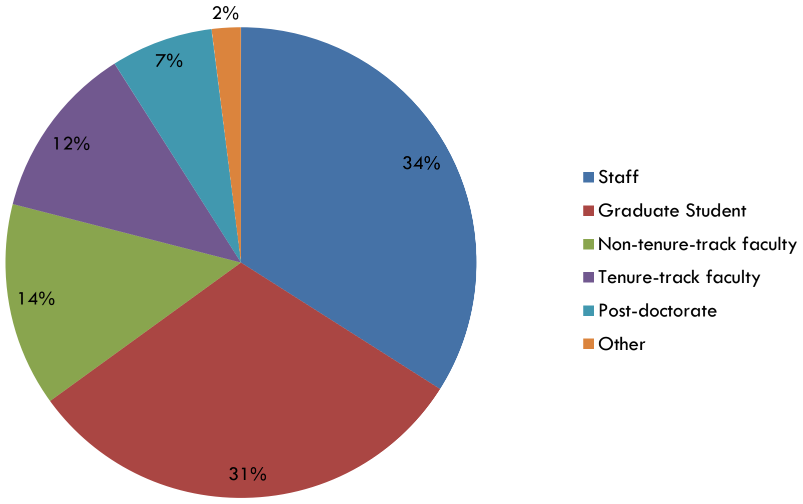

Figure 1. Appointment Status

When cross-tabulated with university school affiliation, the largest group of respondents in all schools was graduate students except for the medical school, where staff was the largest responding group.

Additionally, the survey asked respondents to provide their departmental affiliation (643 responses). Analysis shows that respondents were affiliated with 159 departments. Eighteen departments had more than 10 respondents. Seventy-six departments had 2-9 respondents and 65 departments had only 1 respondent. Departments/programs with more than 10 responses are depicted in Table 2.

Department	Number of responses	Percentage of total responses
Chemistry	45	7%
Preventive Medicine	22	3.42%
Materials Science & Engineering	21	3.27%
Chemical & Biological Engineering	20	3.11%
Psychology	17	2.64%
Psychiatry	14	2.18%
Communication Sciences and Disorders	13	2.02%
Medical Social Sciences	13	2.02%
Physics & Astronomy	13	2.02%
Biomedical Engineering	12	1.87%
Electrical Engineering and Computer Science	12	1.87%
Obstetrics and Gynecology (division of Dept. of Medicine)	12	1.87%
Earth and Planetary Sciences	11	1.71%
Mechanical Engineering	11	1.71%
Pediatrics	11	1.71%
Sociology	11	1.71%

Table 2. Top departments (Departments with more than 10 respondents each)

Types and Size of Data

The survey shows that research data comes in many different forms. Most common data types include spreadsheets (68%), structured data (e.g. csv, xml) (58%), text (74%), and images (52%). Respondents were allowed to choose more than one answer for this question.

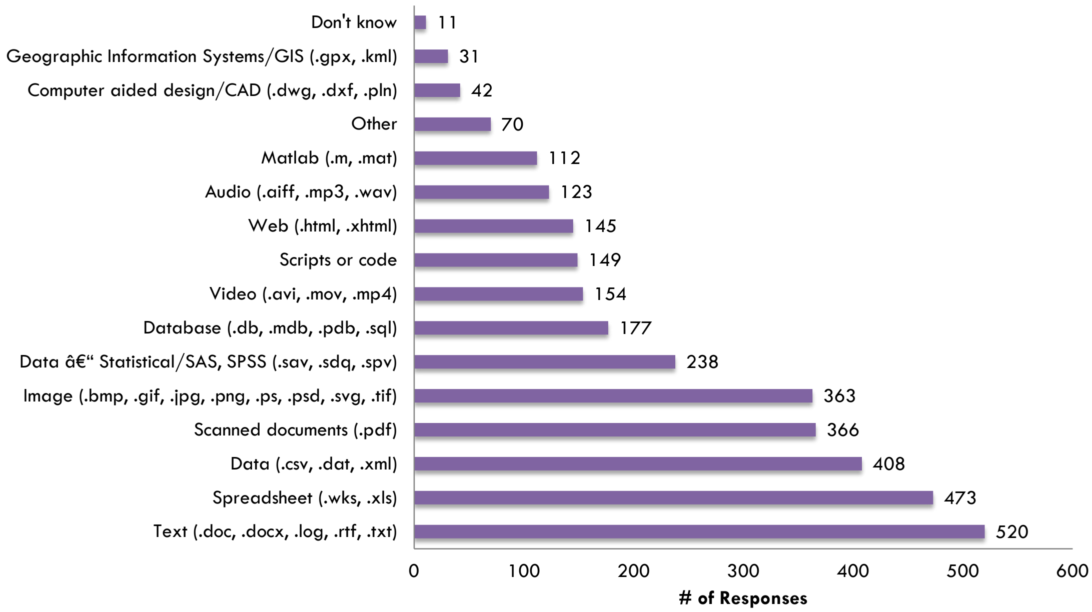

Figure 2. Type/Format of Data

Ten percent of respondents selected "Other" data types. These included crystallography data, mathematics, a custom format used by the lab, historical archives, various types of experimental measurements (EEGs, NMR spectroscopy, seismic data, medical image data, and genetic sequencing data files). Four respondents stated that they had no data.

Further analysis to determine the types of files most common for each school of the university was conducted. Since respondents were allowed to select more than one data type, percentages of responses for each school were recalculated to a maximum total of 100% per school to create a relative proportional display of data types per school. These results are shown in Figure 3 (following page).

All departments rely on text, images, and spreadsheets. The school of medicine is the only school that utilizes spreadsheet data more than text data and uses database (.db, .mdb, .pdb, and .sql) data proportionately more than any other school. The school of law uses fewer data types than any other schools, relying primarily on documents and scanned documents. Least-used document types are geographic information systems (GIS) data and computer aided design (CAD) data. The largest schools (college of arts and sciences, medical school, and school of engineering) employ all types of data choices in the survey. Somewhat surprising is that the school of education uses audio data files proportionately slightly more than the school of music.

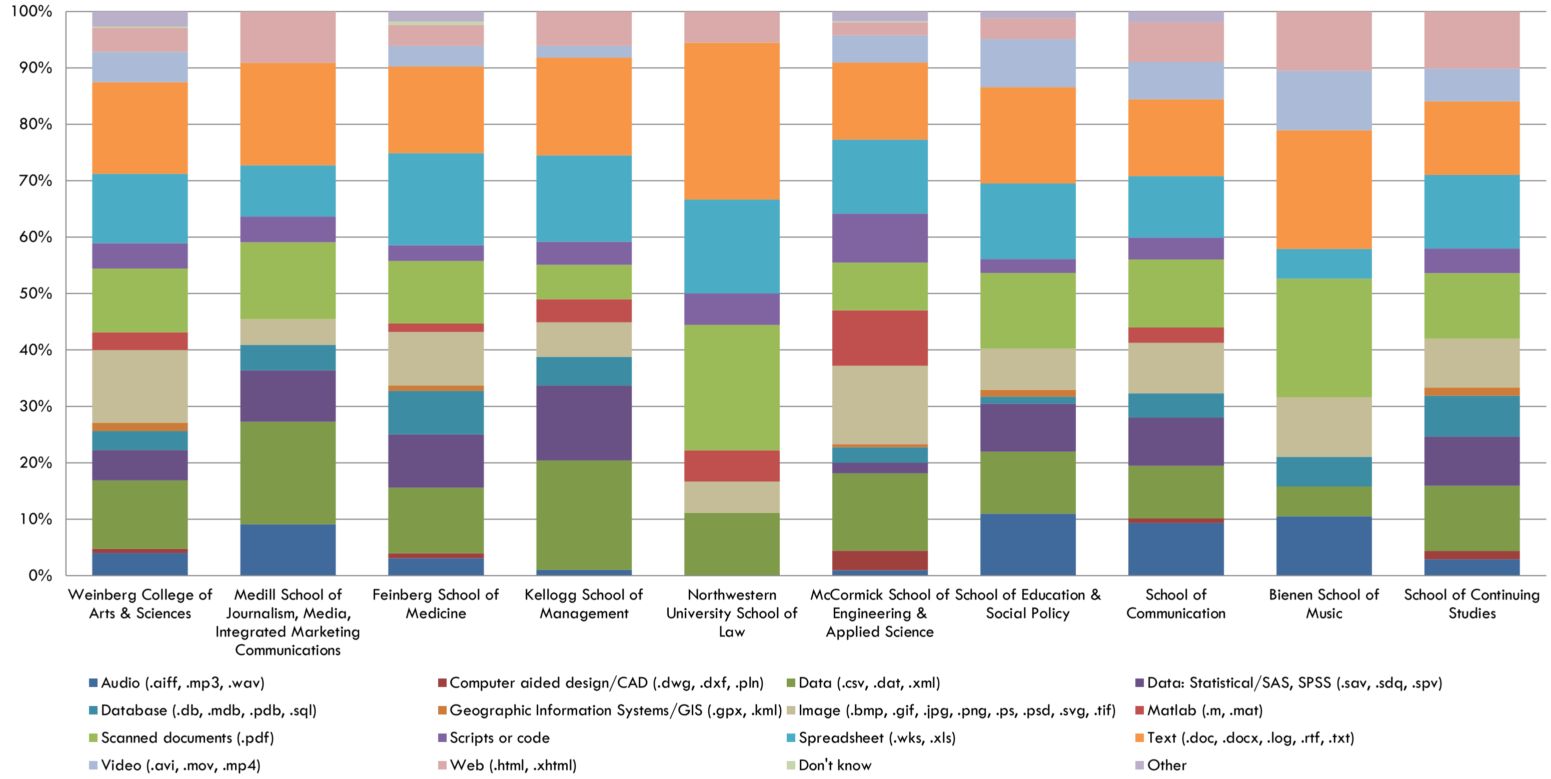

Figure 3. Total numbers of data types used by each school of the university were adjusted to a total percentage of 100% to create an adjusted relative percentage representation of data type usage by school.

Data Storage and Retention

In order to better understand researchers' data storage needs, respondents were asked where they currently store their data and how long they plan to store their data. Respondents could choose more than one storage solution. Results show that researchers store data in a variety of different ways. Sixty-six percent use computer hard drives, 47% use external hard drives, 50% use departmental or school servers, 38% store data on the instrument that generated the data, and 27% use flash drives. Additionally, 31% use cloud-based storage services. When asked to name their cloud-based storage (180 written responses) Dropbox (Dropbox, Inc., 2007) was the most popular choice (63%). Only 6% of the respondents use external data repositories. Written responses to the "other" choice for this question included a wide variety of personal or laboratory backup servers.

A further analysis by school affiliation showed that most schools' top storage medium is "computer hard drive" except for the schools of medicine and communication, both of whom selected "departmental or school server" as their top storage location. The college of arts and sciences is the top user of cloud-based storage solutions, but this choice still ranks fourth for the college, after computer hard drives, external hard drives, and "hard drive of the instrument which generates the data."

When asked how long they expected to store data, "indefinitely" was the most common response for both raw and published data. Many respondents also selected 5-10 years by, largely attributed to publisher or funding agency requirements. Very few respondents keep data for less than one year. Responses are depicted in Figure 4 (following page).

Retention preferences were analyzed by school. Some trends that emerged were:

- The college of arts and sciences prefers "indefinitely" for ALL data types
- All schools prefer "indefinitely" for published data, except the law school, which prefers 1-5 years
- The school of medicine prefers 5-10 years for all data types except published data
- The school of engineering prefers 1-5 years for all data types except published data
- "Indefinitely" is the preferential choice for raw data only for two schools: the college of arts and sciences, and the school of management.

Written comments across all schools suggest that data are perceived as relevant for long periods of time or indefinitely. Keeping raw data / source material was useful because researchers may potentially use it for future / new studies (77 responses), utilize it for

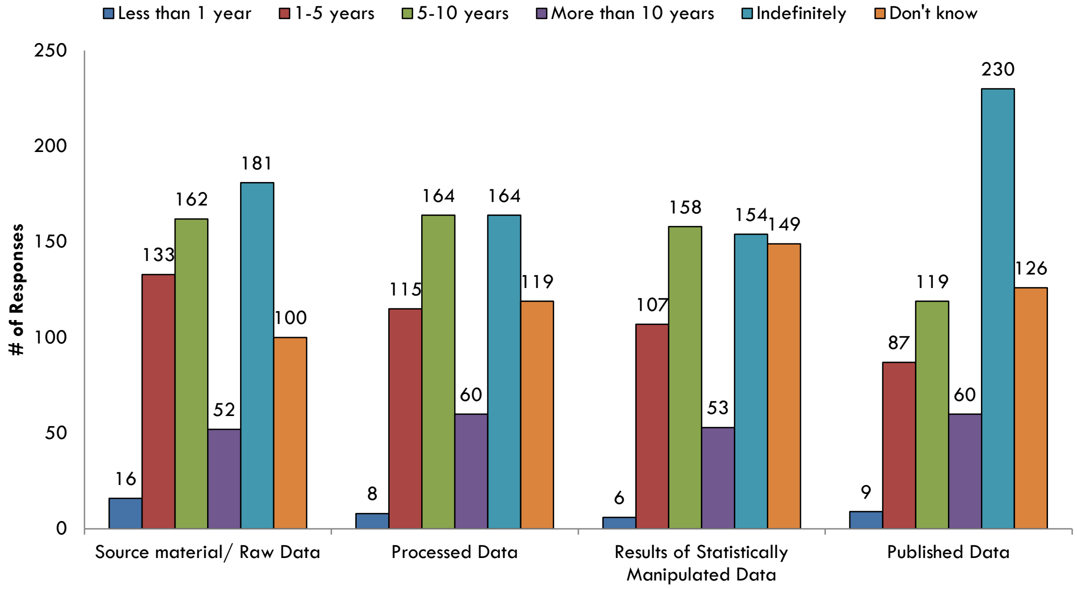

Figure 4. How long are data stored?

longitudinal studies (9 responses) or share it with colleagues (6 responses). Data were also seen as valuable for replicating study results (10 responses), responding to challenges of published results, or because data had been gathered from human or animal subjects and were difficult or costly to replicate. A few responders simply stated that it is good scientific practice to retain data (4 responses).

When asked how much new or additional storage will be needed for their research, 66% indicated they would need additional storage. The amount of storage estimated is depicted in Table 3. The most common responses were1-500 gigabytes and “don’t know.”

Future data storage needs	%
Less than one Gigabyte	9%
1 - 500 Gigabytes (GB)	31%
500 - 1000 GB	11%
1 - 500 terabytes (TB)	14%
500 - 1000 TB	1%
>1 petabyte (PB)	0%
Don’t know	31%
Other (please specify)	1%

Table 3. Future Data Storage Needs. (Respondents were asked to estimate the amount of storage they would require for future projects.)

Responses indicating “don’t know” were further analyzed by appointment affiliation. This analysis showed that 47% of these respondents were staff members and that this was the most common answer from staff. Thirty-two percent of non-tenure track faculty responded "don't know" while 31% indicated 1-500 GB as their anticipated storage needs. Tenure track faculty, post-doctorates, and graduate students selected 1-500 GB more often than any other response.

Data Sharing

Sixty percent of respondents indicated that they share or plan to share their data. Seventeen percent were unwilling to share their data and 23% did not know.

Willingness to share data with various audiences is depicted in Figure 5. Not surprisingly, overall willingness to share with colleagues in their field (both inside and outside the institution) and data sharing with the public at large increased after publication of research results. Before publication, the majority of respondents would share with their colleagues within their research group alone.

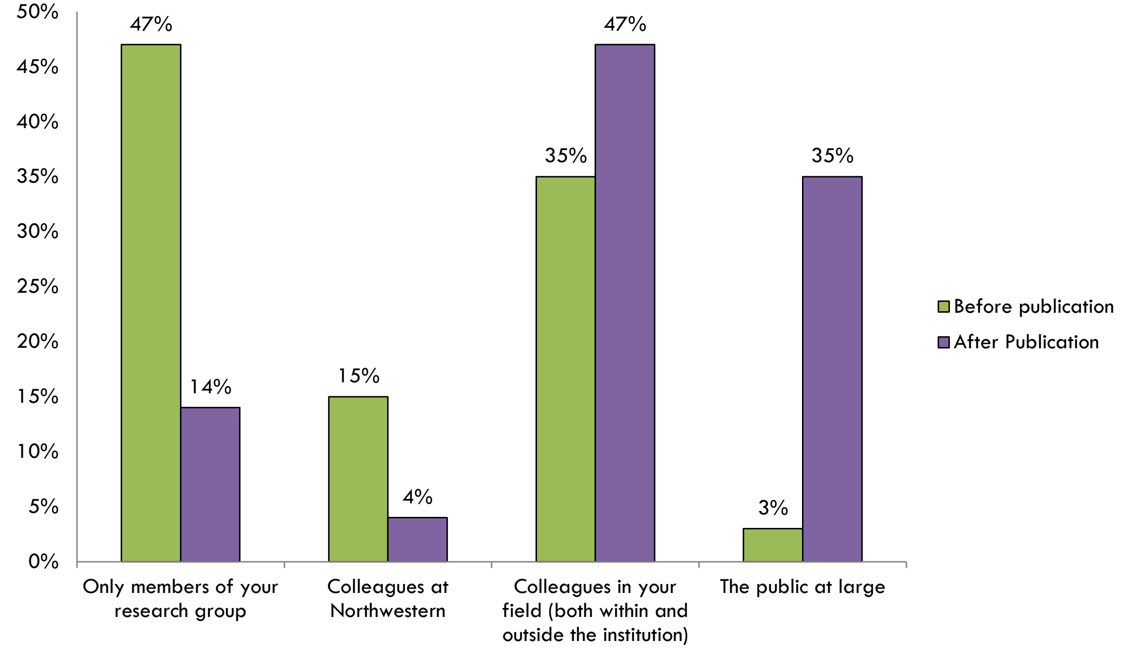

Table 5. Data Sharing Before and After Publication

Personal choice (55%) was the primary reason for data sharing, followed by "required by funder" (22%), "recommended by funder" (7%), and "other" (17%). Other reasons for sharing data included collaboration or sharing with colleagues, advancing science, requirements by the journal in which they publish, and belief(s) that publishing results is sufficient for sharing.

The most common method of sharing data was by personal request (41%). Other methods were through supplemental materials (16%), through a shared site with restricted access (15%), via a discipline-specific public repository (9%), and a university-managed public repository (7%). "Other" responses include sharing through a laboratory server or a website. Privacy or protection of subjects was the main reason for not sharing data (37%). Other reasons why respondents would not share data were protection of intellectual property rights (21%) and that others would not be interested in the data (19%).

Data Management Plans

45% of the respondents indicated that they had data management plans (DMPs), 33% did not and 22% did not know if they had a plan. Main reasons for having DMPs were that one was required by the Institutional Review Board (51%) or required by their funding agency (47%).

The primary reasons for not having a DMP were lack of information about DMPs (58%) or that DMPs were not necessary (42%). A breakdown by appointment status shows that a "lack of information about data management plans" existed in all employment types from tenure-track faculty down to staff and students.

Training and Assistance

Respondents were asked to choose services that would be useful in managing research data. Respondents could select more than one answer. Top responses include long term data access and preservation (63%), services for data storage and backup during active projects (60%), information regarding data best practices (58%), information about developing data management plans or other data policies (52%), assistance with data sharing/management requirements of funding agencies, and tools for sharing research (48%).

DISCUSSION

Results of this survey suggest that researchers need assistance with data storage and preservation as well as data management.

The different types and sizes of data indicated by the respondents show that a universal data storage/preservation solution will be difficult. Estimation of university-wide storage solutions is also difficult as 31% of respondents did not know the size of their data storage needs. Additionally, many respondents indicated that they planned to keep their data indefinitely. Any institutional storage solution will need to accommodate many data types and uncertain storage capacity needs over long periods of time. The university lacks a long-term storage solution for large data, but has short term storage available for data from active experiments. Comments indicated a lack of understanding of current university resources for data storage and requested a university-wide single storage solution for data.

Many researchers also store data on personal or laboratory computers, laboratory equipment, and USB drives. These storage solutions increase the risk of data loss since computers and equipment can fail, and USB drives can be lost. Storage on these types of systems also limits the ability to share data. Thus, there appears to be a need for educating researchers on best practices for data storage and backup.

The fact that very few respondents reported usage of external data repositories is disappointing. This presents another opportunity for education of users on the wide variety of open data repositories that are available for them to store and share data.

Several written comments regarding duration of storage indicated that respondents do not know funding agencies' requirements for data retention. This provided the impetus for the library to provide a clear set of funders' mandated data retention policies that are linked from the library's data management web guide.

The fact that researchers wish to retain their data indefinitely or for long periods of time also indicates the value of this data for reuse or further evaluation by the researchers. However, long-term storage of data is problematic for researchers due to the size of the data and the lack of stable storage solutions.

Survey results show that researchers' attitudes toward data sharing are diverse. Sixty percent of respondents indicated they were willing to share data. This level of general willingness to share data was somewhat surprising and suggests that the university research community is more open to sharing than was originally expected. Whether this "willingness" is personal choice or to comply with funder or institutional requirements is more difficult to determine, although 55% of responders cited personal choice as their reason for sharing, making choice the highest motive. Only 47% were willing to share outside of their research group or colleagues prior to publication. More researchers appear willing to share outside their research groups after publication.

Sharing before and after publication, when analyzed by school, provides some interesting insights. Respondents from the schools of music and journalism are highly likely to share with the public after publication. This is due to a professional obligation or necessity in these fields. More surprising is that only 34% of respondents from the school of medicine indicated willingness to share with the public at large after publication, though 51% are willing to share with colleagues in their fields within and beyond the university. This percentage of willingness to share with the public is quite low, considering the fact that the NIH has a data sharing policy. However, this policy does not currently mandate data sharing for awards less than $500,000 per year, except for human genomic data, which must be shared regardless of funding level. It is possible that many researchers in the medical school do not have grants that are eligible for these policies and are therefore not required to share their data with the public. Expanded NIH data sharing requirements are likely to be announced in the future; so NIH-funded researchers will need greater support and education from the libraries and the ESWG to meet these new requirements.

Some researchers felt that publication of data in papers was sufficient for data sharing. This suggests a lack of awareness or understanding on federally mandated data sharing and indicates a need for some education in this area. This education would provide researchers with a better understanding regarding the reasons for and the federal funders' requirements on data sharing. Published manuscripts rarely provide any level of data description or sufficient supplemental data that allows for replication of results, so this presents an opportunity to educate users in the importance of providing access to data in support of published findings as requested by federal mandates.

It is perhaps disappointing that the majority of respondents are willing to share their data only by request, but the survey did not ask respondents if their data were "clean" and organized well enough to share. Therefore, it is possible that users are not willing to share data because it is not in a form that can be understood by anyone besides themselves and their research partners. In that case, sharing by individual request provides the only way to prepare data to the specifications of the individual requester. Another reason that respondents may share only by request is that suitable data repositories are lacking for their types of data. Public repositories exist for protein, DNA, RNA, microarray, and small molecule data. Enterprise data warehouses exist for electronic health records data, and there are increasing numbers of repositories for medical imaging and social sciences data. Many other disciplines lack public repositories for their data, making it more difficult to share data openly. This provides another opportunity for the libraries and the ESWG to present general open data sharing options to users.

That the majority of respondents either did not have a data management plan or did not know if they had one also indicates a need for training and assistance in this area. Comments indicated confusion about data management best practices. Confusion regarding data management can also arise because data management can be different in each laboratory or for different projects in the same laboratory. Some respondents thought tools to track data provenance and workflow would be helpful.

When asked about their training and/or assistance needs, data storage and backup during active research and long term access and preservation of data were the top answers. Other training needs included best practices for data management, information about or assistance with DMPs, and assistance with finding agency requirements. These results will help inform future training sessions for researchers.

An open request for additional comments elicited 76 responses on various topics of concern or interest to respondents. The most prevalent topic (32% of the comments) was the need for a comprehensive university-wide policy on data management and storage. Twenty-one percent of comments stressed a need for sustainable storage options. Other comments addressed respondents' concerns for their research groups' lack of data organization and management practices (10.5% of comments). Data security or privacy (11.8% of comments) was also important. All of these comments provide useful information in guiding the ESWG's plan for user education and consultation in data management practices. While the university does not have long-term data storage in place to assist researchers in storing and sharing data, users can be guided to available data repositories and offered suggestions for creating multiple storage and backup locations for their data.

Some comments from respondents in the humanities indicated that the responders did not view their scholarly output as "data," and they did not think the survey applied to their fields of study. Mathematicians commented that they did not collect or retain data, because their discipline is more theoretical than data-driven.

Limitations of the Survey

Limitations of the survey were revealed in the analysis of the survey results. Wording of the survey questions was not always inclusive of all disciplines. For example, one question asked respondents to estimate how much storage would be required for future "grants." Since not all disciplines rely on grants for funding, the question should have used the word "projects." This error was revealed when some responders commented that they did not have grants and did not supply any other answer to this question.

Users were asked to identify their department or center affiliation in a text box written response, instead of a supplied drop-down selection of responses. This resulted in numerous variations of department or center names and led to difficulty in deciphering acronyms and departmental name variations. Lack of a specific value response selection meant that correlation analysis of responses by department name was impossible within the Qualtrics software package. Any attemps at departmental correlations were performed manually, tracking individual responses according to departmental identification throughout each responder's answers.

Departmental response was uneven, and some departments had little or no representation. It was hoped that the delivery of the survey email signed by the dean of the university library, the vice president of information technology, and the vice president of the office for research, would result in good departmental response across all schools and disciplines, but this hope was not realized.

Increased Opportunities for Collaboration

One unexpected result of the survey was an increase in opportunities for the library to collaborate with and/or invitations to participate in events with other campus groups, such as university IT. The survey's wide distribution gained the attention of leaders in research computing and data science at the university's institute for clinical and translational sciences. There are currently several ongoing projects with the library's campus partners, including a review of NSF data management plans, a revised and updated Data Management Guide, development of data management training sessions, and a discussion on developing data services.

CONCLUSION

Participants' comments on the survey were generally positive. Some respondents were grateful for being asked for their input on current and future data storage requirements and practices. However, some humanists and social scientists provided comments that this survey did not apply to them, even though the ESWG tried to make it clear that the survey applied to all disciplines. This suggests a potential disciplinary difference in the definition of digital or electronic data. It is possible that humanists and social scientists do not view scholarly output such as text documents and recorded oral histories as "data."

Responses to the survey, especially written responses to open questions, suggest that Northwestern University is similar to institutions surveyed in Tenopir's study (Tenopir et al., 2011). Storage and management of data, especially over the long-term, is an issue of concern for researchers, and no single solution may fit the needs of all disciplines. Also, the survey shows that there is a need for assistance and education regarding data management across all user groups at the institution.

Results of the survey were shared with heads of university libraries, information technology, offices for research and graduate education, and with school administrators to guide the ESWG and other interested offices to formulate future goals regarding data management and infrastructure. The report of the survey has been made available to the university community by a link on the library's website for any interested parties to access. The survey data have been archived so that additional interrogations, such as school-specific cross-tabulations, can be performed as needed.

Responses to the survey were dominated by graduate students (31% of all responses) and staff (34% of all responses). These two groups most likely directly interact with most types of data on a daily basis (especially research data in the sciences). However, staff responses to estimated data storage needs suggest that, while they work with data regularly, they may not have a good view of the "big picture" of data management over a long term. Faculty, whose grants or appointments may rely on good data management practices, may not be aware of how data are managed in their research groups and may not be sharing data "best practices" with students and staff. The library and its partners in university IT and the office for research can address these potential gaps in data practice by tailoring information modules and sessions to the needs of each user group. Intervention and instruction on data management is important for students early in their graduate experience, especially if students are not receiving this training from their mentors. The university's E-Science Librarian responded to this need and created a primer on data management for incoming graduate students offered during orientation week at the university.

The survey shows that staff are interested in learning more about data management and that they may not be getting the appropriate guidance and support from faculty or principal investigators. The ESWG is looking for ways to expand data management education beyond faculty and graduate students to include interested staff.

Interest in library support for data management practice is high in academic libraries. A Google search for "data management" in libraries at domains with an ".edu" URL will return thousands of results on data management guides, data management plan tools, data management workshops, and consultation services in dozens of unique university and college libraries across the entire United States (*Google search performed on libraries and data management in .edu domains*, 2015). Academic librarians are developing skills and resources to support data management best practices among their communities. Librarians are also building communities of peer support and education to share resources and strategies for reaching their users through portals such as the Association of Research Libraries' E-Science Institute (Association of Research Libraries, n.d.) and Digital Libraries Federation's E-Research Network (Digital Library Federation, n.d.). The greatest challenge facing many librarians is raising awareness among academic researchers that the library can provide guidance in data management.

This survey indicates a need for data management training and education regarding federal data sharing mandates at all levels, especially among staff and graduate students. As a result, the ESWG has embarked on a path of user outreach, scheduling talks at the university's annual Computational Research Day and library-based seminar series. The ESWG is also developing targeted marketing and planning information sessions for research administrators and department faculty meetings. These efforts are led by an E-Science Librarian, whose role is dedicated to tracking news and policies on data management and who is developing online resources in data practices, as well as a "Data Management 101" seminar to present to user groups at the university. The graduate school and university IT department have also established a private and secure university instance of the Box storage platform (Box, Inc., 2005), which provides graduate students as well as Northwestern faculty and staff a more stable and secure cloud storage solution than Dropbox (Dropbox, Inc., 2007).

The survey has served to open doors with offices of research and core facilities, increasing the libraries' pool of partners in data management efforts. Our survey has shown that researchers at Northwestern University are very similar to their peers at other institutions who have been interviewed or surveyed regarding data practices: they have a wide range of competencies in data management and are hungry for support in managing their data, as well as options for long term storage. Few libraries have published data surveys of their users across the academic spectrum. Many concentrate on faculty in the sciences or on librarians involved in data management practices. In this regard this survey is relatively unique. The broad net cast by this user survey reveals some shortcomings of the survey itself, which may help other institutions design and implement their own surveys to better capture disciplinary differences in responses.

The academic research and scholarly community is in need of guidance and information on data management, and, while challenges still exist, the library and librarians are preparing to meet these challenges and provide high quality support and resources for management of digital data.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work done by the E-Science Working Group members on the survey and report: Elizabeth Adams, Chris Comerford, Carol Doyle, Randy Janzen, Michael Klein, Deborah Rose-Lefmann, Harriet Lightman, Joseph Paris, Claire Stewart.

REFERENCES

Association of Research Libraries. (n.d.). *E-Science Institute*. Retrieved from http://www.arl.org/focus-areas/e-research/e-science-institute

Averkamp, S., Gu, X., & Rogers, B. (2014, February 28). *Data management at the University of Iowa: A University Libraries report on campus research data needs*. Retrieved from http://ir.uiowa.edu/lib_pubs/153

Ball, J. (2013). Research data management for libraries: Getting started. *Insights*, *26*(3), 256-260. http://dx.doi.org/10.1629/2048-7754.70

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6),1059-1078. http://dx.doi.org/10.1002/asi.22634

Box, Inc. (2005). Box [computer software]. Calif.: Los Altos. Retrieved from https://www.box.com/

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. *Portal: Libraries and the Academy*, *11*(2), 629-657. http://dx.doi.org/10.1353/pla.2011.0022

Carnegie Foundation for the Advancement of Teaching. (n.d.). *The Carnegie Classification of Institutions of Higher Learning*. Retrieved on June 8, 2015 from http://carnegieclassifications.iu.edu

Charbonneau, D. H. (2013). Strategies for data management engagement. *Medical Reference Services Quarterly*, *32*(3), 365-374. http://dx.doi.org/10.1080/02763869.2013.807089

Digital Library Federation. (n.d.). *E-Research Network*. Retrieved from http://www.diglib.org/groups/e-research-network/

Dropbox, Inc. (2007). Dropbox [computer software]. Calif: San Francisco. Retrieved from https://www.dropbox.com/

Google search performed on libraries and data management in .edu domains. (2015, February 20). Retrieved from https://www.google.com/?gws_rd=ssl - q=%22data+management%22+(library+OR+libraries)+site:.edu

Henderson, M. E., & Knott, T. L. (2015). Starting a research data management program based in a university library. *Medical Reference Services Quarterly*, *34*(1), 47-59. http://dx.doi.org/10.1080/02763869.2015.986783

Johnston, L., & Jeffryes, J. (2014). Steal this idea: A library instructors' guide to educating students in data management skills. *College and Research Libraries News*, *75*(8), 431-434.

National Institutes of Health. (2003). *NIH data sharing policy*. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/

National Science Foundation. (2011). *NSF data management plan requirements*. Retrieved from http://www.nsf.gov/eng/general/dmp.jsp

Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS One*, *9*(12). http://dx.doi.org/10.1371/journal.pone.0114734

Qualtrics, LLC. (n.d.). Qualtrics [computer software]. Utah: Provo.

Scaramozzino, J. M., Ramirez, M. L., & McGaughey, K. J. (2012). A study of faculty data curation behaviors and attitudes at a teaching-centered university. *College & Research Libraries*, *73*(4), 349-365. http://dx.doi.org/10.5860/crl-255

Scientific Software Development GmbH. (n.d.). ATLAS.ti [computer software]. Germany: Berlin.

Soehner, C., Steeves, C., & Ward, J. (2010). *E-science and data support services: A study of ARL member institutions*. Washington, DC: Association of Research Libraries. Retrieved from http://www.arl.org/storage/documents/publications/escience-report-2010.pdf

Steinhart, G., Chen, E., Arguillas, F., Dietrich, D., & Kramer, S. (2012). Prepared to plan? A snapshot of researcher readiness to address data management planning requirements. *Journal of eScience Librarianship*, *1*(2), Article 1. http://dx.doi.org/10.7191/jeslib.2012.1008

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One*, *6*(6), e21101. http://dx.doi.org/10.1371/journal.pone.0021101

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2013). Academic librarians and research data services: Preparation and attitudes. *IFLA Journal*, *39*(1), 70-78. http://dx.doi.org/10.1177/0340035212473089

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library and Information Science Research*, *36*(2), 84-90. http://dx.doi.org/10.1016/j.lisr.2013.11.003

U.S. Department of Energy. (2014). *DOE public access plan*. Retrieved from http://www.energy.gov/downloads/doe-public-access-plan

White House Office of Science and Technology Policy. (2011a). *Request for information: Public access to digital data resulting from federally funded scientific research*. Retrieved from https://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific

White House Office of Science and Technology Policy. (2011b). *Request for information: Public access to peer-reviewed scholarly publications resulting from federally funded research*. Retrieved from https://www.federalregister.gov/articles/2011/11/04/2011-28623/request-for-information-public-access-to-peer-reviewed-scholarly-publications-resulting-from

White House Office of Science and Technology Policy. (2012). *Obama Administration unveils "Big Data" initiative: Announces $200 million in new R&D investments*. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

White House Office of Science and Technology Policy. (2013). *Increasing access to the results of federally funded scientific research*. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Whitmire, A., Briney, K., Nurnberger, A., Henderson, M., Atwood, T., Janz, M., ... Zilinski, L. (2015). A table summarizing the Federal public access policies resulting from the US Office of Science and Technology Policy memorandum of February 2013. *figshare*. http://dx.doi.org/10.6084/m9.figshare.1372041. Retrieved June 8, 2015.

Wilson, J. (2013). *University of Oxford Research Data Management Survey 2013: The results*. Retrieved from http://blogs.it.ox.ac.uk/damaro/2013/01/03/university-of-oxford-research-data-management-survey-2012-the-results/

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, *4*(3), 93-103. http://dx.doi.org/10.2218/ijdc.v4i3.117