Report on Data Management Survey, Northwestern University

May 22, 2014

Submitted by the E-Science Working Group:
Cunera Buys (co-chair)
Pamela Shaw (co-chair)
Elizabeth Adams
Chris Comerford
Carol Doyle
Randy Janzen
Michael Klein
Deborah Rose-Lefmann
Harriet Lightman
Joseph Paris
Claire Stewart

Report on Data Management Survey, Northwestern University

Executive Summary

Northwestern researchers have expressed increasing interest in data, data management and data storage due in part to Federal mandates requiring data sharing and preservation. The E-Science Working group designed a survey in 2013 to investigate how researchers at Northwestern University manage data and to help determine researchers' needs or requirements for data storage, preservation and sharing.

The 21 question survey looked at several areas involving data and data management, including the types and size of data generated by researchers, 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.

The survey results indicate a need for assistance with data management as well as long term data storage and preservation solutions. Key findings include:

- 31% of respondents did not know how much storage they will require, highlighting the difficulty of establishing a correctly sized research storage service.
- Much of Northwestern's research data is stored on local hard drives, departmental servers or equipment hard drives. 31% of respondents use web-based storage services, most notably Dropbox.
- 60% of the respondents indicated that they share or plan to share their research data. Sharing tends to occur within a research group or with collaborators prior to publication, expanding to more public availability after publication.
- Responses on data management plans and support needs 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.

The survey results will be used to inform potential services or education opportunities as well as to assist stakeholders in defining future goals regarding data management and infrastructure. Further analysis of the results is also possible upon request.

Introduction

This research study was designed to gather information about management practices of digitally stored research data at Northwestern University. The survey was conducted by the E-Science Working Group (ESWG), which includes representatives from the Northwestern University Libraries, Galter Health Sciences Library, Northwestern University Information Technology, Weinberg College of Arts and Sciences and the Office for Research. The purpose of the survey ~~is~~ was to understand how

researchers manage data and what are their concerns and preferences in order to determine data storage needs and data management services required or desired by the research communities across Northwestern University. This report summarizes the findings from this survey.

Background

In November 2011, the White House Office of Science and Technology Policy (OSTP) released two requests for information (RFI) and public consultation: one on public access to peer-reviewed publications resulting from federally funded research [1] and one on public access to digital data resulting from federally funded research [2]. These requests for information and public comment initiated a series of OSTP announcements and memoranda concerning advancing research in and sharing of "big data" and data resulting from federally funded research projects. In spring 2012, the OSTP released an announcement on the "Big Data Initiative": an investment of $200 million sponsored by six federal agencies to advance the ability to "extract knowledge and insights from large and complex collections of digital data" [3].

Following the collection of responses to the RFIs, in February 2013, the White House released a policy memorandum that directed agencies and departments to develop and implement public access plans within 2-3 years [4]. While the National Institutes of Health (NIH) has had a Data Sharing Policy for grants in amounts of $500,000 or more since 2003 [5], and the National Science Foundation (NSF) mandated the inclusion of a Data Management Plan (DMP) with each NSF grant application ~~since~~ starting in January 2011 [6], the OSTP directive of 2013 suggests a more coordinated effort across federal funding agencies to establish policies on the open access to research data, placing a burden of responsibility upon researchers to manage, preserve and share their digital research data.

The need for data storage and management services in colleges and universities is a topic of great interest in academic libraries and computing centers, as academic institutions work to develop training programs and computing resources to support their federally funded faculty in complying with policies on responsible management of digital research data. A number of institutions have begun surveying their users and the academic community at large to identify institutional needs for data management solutions. A widespread study by Tenopir et.al in 2011, which surveyed data storage and management needs across many academic institutions, found many investigators are satisfied with their short-term data storage and management practices, but are less satisfied with long-term data storage options [7]. The Tenopir study also discovered that researchers do not believe their institutions provide adequate funds, resources or instruction on good data management practices. Tenopir found differences in data sharing or reuse among different academic disciplines, suggesting multiple data cultures within a single institution.

Other academic institutions, mainly through their libraries, have surveyed their faculty in an attempt to determine attitudes and needs for data storage and management at their institutions [8, 9]. Most recently, Oxford University published results of such a survey in a blog post, which highlighted the Oxford University Policy on Research Data Management [10]. The Association for 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 [11]. Results of the ARL survey suggest that many libraries

are involved in such data management practices, or are surveying their users in order to develop services for data curation and storage.

These studies and related activities reinforce the importance of our current survey to understand Northwestern University researchers' needs for data storage and management services. This in turn informed the questions asked in the survey, notably those about potential data curation and management services provided by NU Libraries, NUIT and other units. The collection of demographic information will help us determine what types of different data management needs exist across the many disciplines at Northwestern.

Survey Design

Many of the questions used in our survey were borrowed from other university libraries, who shared their survey instruments with the E-Science Working Group by request. The majority of the 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. The software used for this survey was Qualtrics.

This survey consists of an introduction and 21 questions. In addition to demographic information, questions covered a variety of topics including the type and size of data, data storage, data management, willingness to share data and types of additional assistance or training desired.

Questions with an "Other" option allowed respondents to enter a text answer.

Survey Distribution

This survey was sent to all faculty, graduate students, postdoctoral candidates and selected staff at both Northwestern's Evanston and Chicago campuses. A survey link was by email sent to approximately 12,940 Northwestern email addresses.

The survey URL was distributed by Northwestern's bulk mail system. The survey opened January 15, 2014 and closed February 17, 2014. Two reminder emails were sent during the course of the survey.

Survey Responses

There were 831 responses and 788 respondents completed the survey, for a response rate of approximately 6.4%. Respondents were allowed to skip questions, which accounts for the variance in the number of responses for each question. Additionally, 5 questions allowed respondents to select more than one answer.

Findings

The survey questions covered these general areas.

- Demographics
- Types and size of data
- Data storage
- Data retention
- Data sharing
- Data management planning
- Training or assistance needed

The full responses to the survey with commentary can be found in Appendix A.

Demographics

The Feinberg School of Medicine was the largest respondent population, accounting for 38% of all responses, followed by the Weinberg College of Arts and Sciences (24%) and the McCormick School of Engineering (14%). Additionally, preliminary analysis shows that respondents were affiliated with 159 different departments. The Department of Chemistry had the highest number of respondents (7%) followed by Preventive Medicine (3.42%), Materials Science & Engineering (3.27%) and Chemical & Biological Engineering (3.11%).

When analyzed by appointment type, the highest number of responses were from staff (34%), followed by 31% graduate students, 28% faculty (14% non-tenured faculty and 12% tenure track faculty) and 7% post-doctorates. Graduate students were the largest group of respondents in all schools except for Feinberg, where the largest responding group was staff.

Answer	Tenure-track faculty	Non-tenure-track faculty	Post-doctorate	Graduate Student	Staff	Other
Weinberg College of Arts & Sciences	29	14	23	84	24	1
Medill School of Journalism, Media, Integrated Marketing Communications	4	0	0	0	1	0
Feinberg School of Medicine	20	64	17	16	155	6
Kellogg School of Management	5	3	3	5	9	0
Northwestern University School of Law	1	2	0	4	0	0
McCormick School of Engineering & Applied Science	16	4	6	76	3	0
School of Education & Social Policy	0	2	2	6	6	1
School of Communication	14	0	2	20	10	0
Bienen School of Music	0	0	0	5	0	0
School of Continuing Studies	0	4	0	0	4	3
Other	2	5	1	12	37	0
Total	91	98	54	228	249	11

Types and size of data

The most common data types were spreadsheets, structured data (e.g. csv, xml, xls), text and images. 31% of respondents thought they would need 1-500 Gigabytes (GB) of new or additional data storage. 31% did not know how much storage they will require. Of these nearly half (47%) of staff respondents did not know how much data storage would be needed for future projects. Furthermore, the staff response accounted for 49% of *all* "Don't know" responses. Non-tenure track faculty responded "Don't know" (32%) slightly more frequently than their next common response: 1-500 Gigabytes (31%). Tenure track faculty, post-doctorates and graduate students selected 1-500 GB more often than any other response.

The variety of data types and the size of data indicate that any data storage solution will be complicated and challenging, since it will have to accommodate many different data types and unclear storage capacity needs.

Data Storage

Researchers store data in a variety of ways, often employing multiple different types of storage. Computer hard drives are used by 66%, external hard drives are used by 47% and departmental or school servers are used by 50% of respondents. Cloud-based storage services are also used by 31% of respondents, with Dropbox named as the most popular Web-based storage choice. Relatively few researchers used NU storage services or external subject specific data repositories.

Comments regarding Northwestern's Vault storage show that Vault has had some use among the research community. However, comments indicate that the Vault interface is not as simple to use as some alternatives such as Box.net, there is a desire for a secure storage system for regulated data and that Vault does not fit all needs. This feedback can help guide selection of other data storage platform options for the University.

Some respondents' comments referred to the REDCap data acquisition platform. REDCap was developed for use by clinical and translational science award institutions (CTSAs) for data acquisition and surveying in biomedical research but its use is growing rapidly across the institution. However, REDCap is not meant as a storage solution. This leaves users unsure as to where they should archive data collected through REDCap. The same issue applies to data collected through other survey instruments such as Qualtrics: once surveys are closed, users may not be aware of how or where to archive the survey data for continued access and use.

Other comments indicate that cost models are an important factor in deciding to use a particular service. Some respondents believe that the burden of long term data storage and preservation should be funded by the University. This could point towards the lack of funds for handling long term preservation of research data. The large use of Dropbox and similar services and some comments indicate that an easy to use cloud based system that allows for convenient collaboration might be a good solution. It is possible that box.net, when implemented by the University, may meet some but not all of these needs.

Use of personal or lab computers, laboratory equipment and USB drives to store data increases the risk of data loss. Data can be lost if computers or equipment fail or are upgraded. USB drives are easily lost. The fact that some researchers are storing data in this manner indicates a need for education on best practices for data storage and backup. Additionally, storage of data on a closed system limits data sharing.

Data Retention

Many respondents commented that raw data are valuable for replicating results or pursuing new research questions, thus are retained indefinitely. Published data are also kept indefinitely by most respondents. A time span of 5-10 years was also selected by many respondents, with some citing funding agency or publisher requirements for data retention periods. Very few researchers keep data for less than one year.

Tenured faculty selected “Indefinitely” most often over all types of data collectively. Published data were also preferred to be retained indefinitely by this group. Post-doctorates selected “Indefinitely” for published data, but preferred the period of 5-10 years for raw data, processed data and results of statistically manipulated data. Non-tenured faculty also selected the time span of 5-10 years most often for all categories except for published data, in which they slightly preferred “Indefinitely”. Staff selected “Don’t know” most often over all totaled responses across all categories of data. By category, for raw data, the time spans of 5-10 years and “Indefinitely” were chosen slightly more often by staff. Graduate students selected “Indefinitely” most often for totaled responses across all categories as well, and also in most categories separately, except “Results of statistically manipulated data”, in which they slightly preferred “Don’t know”.

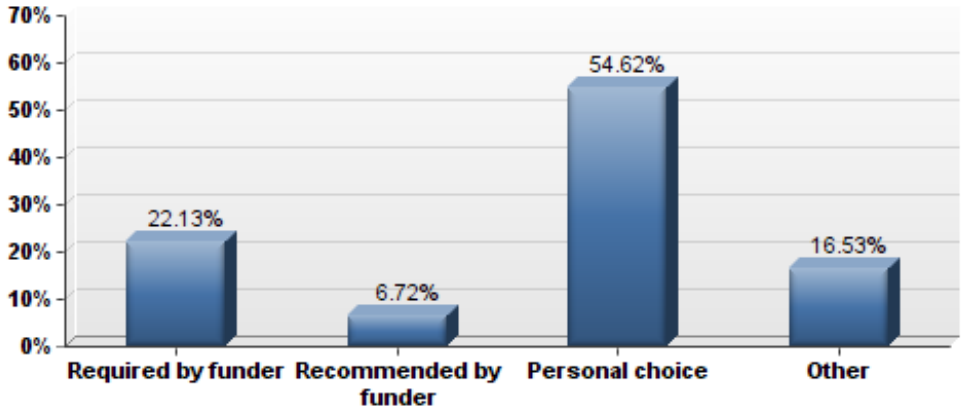
The predilection for saving many types of data indefinitely emphasizes the value of these data for use in new studies or for longitudinal comparison studies. However, combined with size of data and researchers’ concerns for lack of stable storage solutions, the storage and protection of data for indefinite periods of time presents a challenge to researchers.

Data Sharing

60% of the respondents indicated that they share or plan to share their research data. 17% stated they would not share their data and 23% did not know.

Most respondents were willing to share their data with only members of their research group or colleagues in their field before publication. Not surprisingly, more researchers are willing to share their data outside their research groups or outside the University after publication. Personal choice was the top reason why researchers would share data, followed by funder requirements.

Figure: Graph of respondents' reasons for sharing data after publication of results



The majority of those who do not share data cited privacy or protection of subjects as the main reason not to share data. Other top reasons for not sharing data were protection of intellectual property rights and the belief that others would not be interested in the data.

Few respondents indicated that they share data via a university managed repository or a discipline specific repository. Most (41%) indicated that they share data by personal request only. The relatively high percentage of sharing with colleagues inside and outside Northwestern is notable for its potential impact in designing research storage services.

Comments suggesting that publishing papers is sufficient for data sharing, that data is specific to the laboratory, that there is no need to share since there have been no requests, or that sharing is just not done in their field suggest a need for more education regarding what constitutes data and why it is important to share data beyond publications. Researchers may not be aware of proposed federal requirements for data sharing, so early intervention and education on these requirements would benefit researchers in planning for data sharing in the future.

Data Management Planning

Most respondents indicated that they had data management plans (DMPs) because they were required by IRB or a funding agency. Respondents without data management plans indicated that they had a lack of information regarding DMPs or felt that they were not necessary.

Some comments indicated confusion on what were the best practices for data management in their research areas. Additionally data management was different in each lab or even for each project in the same lab (depending of the funding agency) which resulted in confusion around how to manage data. Tools to track data provenance and workflow were also suggested.

These responses indicate that more training and assistance on data management plans and data management in general for all appointment types would be useful.

Training and Assistance

Respondents were asked to choose services that might be useful to them. Data storage and backup during active research projects and long term data access and preservation had the highest number of responses. This indicates a need for both short and long term storage and preservation on campus.

Information regarding data management best practices, information about developing a DMP or other data policies and assistance with funding agency requirements were also seen as useful. This information can be used to inform future training and/or education in this area.

Comments also indicate that some respondents would also like the University to provide access to software such as NVIVO or Atlas TI, electronic lab notebooks, and/or a flexible, free and secure web based management system because not everyone has access to a departmental or school based server.

Further steps

The survey asked whether respondents would be willing to participate in a longer in-depth interview regarding their research data. The "Yes" respondents were asked to send an email to e-research@northwestern.edu so that they could be contacted for a further in depth interview. Of the 213 yes respondents, only 2 sent a follow up email. The ESWG will need to determine a means of identifying those participants that were willing to discuss data in more detail. Once the ESWG has identified willing participants, it plans to use the Data Curation Profiles Toolkit developed by Purdue University to conduct interviews.

The ESWG also plans to use the results of this survey to inform potential services and to build a corpus of education modules and materials accessible to all users, addressing many levels of data management throughout the data lifecycle.

Additionally, the ESWG will analyze NSF Data Management plans submitted by Northwestern researchers. The ESWG hopes to use this information to further identify the needs of researchers regarding data management, storage and preservation.

Conclusions

The responses to this survey were generally positive. Some humanists and social scientists felt that this survey did not apply to them, even though the ESWG tried to make it clear that the survey applied to all disciplines.

In the survey, a question soliciting open comments received a number of varied responses. The largest percentage of comments indicated respondents' desire for comprehensive University-level policies and guidelines on data management. This can prove difficult, due to the diversity of types and size of data and discipline-specific requirements for data management and sharing. Responses to the survey,

especially written responses to open questions, suggest that Northwestern University is similar to institutions studied by Tenopir in her 2011 study [7]. 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 Northwestern.

Additionally, the results can be employed by stakeholders in defining future goals regarding data management and infrastructure. The report can be shared with heads of university libraries, NUIT, offices for research and graduate education and with school administrators, to guide the ESWG and other interested offices to formulate services in these areas. The survey data will be archived so that additional interrogations, such as school-specific cross-tabulations, can be performed as needed.

References

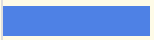
1. White House Office of Science and Technology Policy. Public Access to Peer-Reviewed Scholarly Publications Resulting from Federally Funded Research. Available from: https://www.federalregister.gov/articles/2011/11/04/2011-28623/request-for-information-public-access-to-peer-reviewed-scholarly-publications-resulting-from#p-2
2. White House Office of Science and Technology Policy. *Public Access to Digital Data Resulting from Federally Funded Scientific Research*. Available from: https://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific#p-8
3. White House Office of Science and Technology Policy. *Big Data Initiative*. Available from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
4. White House Office of Science and Technology Policy. *Increasing Access to the Results of Federally Funded Scientific Research*. Available from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
5. National Institutes of Health. *NIH Data Sharing Policy*. Available from: http://grants.nih.gov/grants/policy/data_sharing/
6. National Science Foundation. *NSF Data Management Plan Requirements*. Available from: http://www.nsf.gov/eng/general/dmp.jsp
7. Tenopir, C., et al., *Data Sharing by Scientists: Practices and Perceptions.* PLoS One, 2011. **6**(6): p. e21101.
8. Scaramozzino, J.M., M.L. Ramirez, and K.J. McGaughey, *A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University.* College & Research Libraries, 2012. **73**(4): p. 349-365.
9. Carlson, J., et al., *Determining Data Information Literacy Needs: A Study of Students and Research Faculty.* portal: Libraries & the Academy, 2011. **11**(2): p. 629-657.
10. Wilson, J., *University of Oxford Research Data Management Survey 2012 : The Results*, in *DaMaRO: Data Management Rollout at Oxford* 2012, Oxford University.

11. Soehner, C., C. Steeves, and J. Ward, *e-Science and data support services: a survey of ARL members*, in *International Association of Scientific and Technological University Libraries, 31st Annual Conference* 2010: Purdue University, West Lafayette, Indiana.

Appendix A (Survey results with commentary)

The survey results provide an opportunity to provide response sorting ("drill-down") by a number of different concepts. For this report, drill-downs were applied by school affiliation as well as appointment affiliation, but a large number of different analyses can be performed.

Please indicate your University affiliation:

Answer		Response	%
Weinberg College of Arts & Sciences		175	24%
Medill School of Journalism, Media, Integrated Marketing Communications		5	1%
Feinberg School of Medicine		279	38%
Kellogg School of Management		25	3%
Northwestern University School of Law		7	1%
McCormick School of Engineering & Applied Science		105	14%
School of Education & Social Policy		17	2%
School of Communication		46	6%
Bienen School of Music		5	1%
School of Continuing Studies		11	2%
Other		57	8%
Total		732	100%

732 respondents answered this question.

The Feinberg School of Medicine was the largest respondent population, accounting for 38% of all responses (279 responders); followed by the Weinberg College of Arts and Sciences, accounting for 24% of responses (175 responders); the McCormick School of Engineering, 14% (105 responders); with the rest of the options composing the remainder. 8% selected "Other", totaling 52 respondents who selected this option. Of these responses, 14 wrote in "The Graduate School", 10 wrote in "NUIT", 8 wrote in "Office for Research", 3 "Alumni Relations & Development", 2 from NU-Qatar, 2 from the Robert H. Lurie Comprehensive Cancer Center. The remainder of the unique write in responses included individuals with joint appointments, student health, student affairs, University Center and the Block Museum.

Please indicate your Department/Program or other institutional affiliation:

There were 643 responses to this question in the form of a text answer from respondents. Preliminary analysis shows that respondents were affiliated with 159 departments. 18 departments had more than 10 respondents. 76 departments had 2-9 respondents and 65 departments had only 1 respondent.

The departments/ programs with the highest number of responses were:
Chemistry 7.00% (45 respondents)
Preventive Medicine 3.42% (22 respondents)
Materials Science & Engineering 3.27% (21 responses)
Chemical & Biological Engineering 3.11% (20 responses)
Psychology 2.64% (17 responses)
Psychiatry 2.18% (14 respondents)
Communication Sciences and Disorders 2.02% (13 respondents)
Medical Social Science, Medicine, Neurology and 2.02% (13 respondents)
Physics & Astronomy 2.02% (13 respondents)
Biomedical Engineering 1.87% (12 respondents)
Electrical Engineering and Computer Science 1.87% (12 respondents)
Obstetrics and Gynecology 1.87% (12 respondents)
Earth and Planetary Sciences 1.71 % (11 responses).
Mechanical Engineering Department 1.71 % (11 responses).
Pediatrics 1.71 % (11 responses).
Sociology 1.71 % (11 responses).

Please indicate your affiliation to the university:

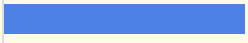
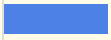
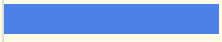
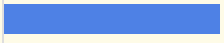
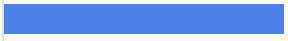
Answer		Response	%
Tenure-track faculty		91	12%
Non-tenure-track faculty		99	14%
Post-doctorate		54	7%
Graduate Student		228	31%
Staff		249	34%
Other		11	2%
Total		732	100%

Of the 732 respondents to this question, the highest number of responses were from staff (34%), followed by 31% graduate students, 14% non-tenured faculty and 12% tenure track faculty (for a total of 28% faculty) and 7% post-doctorates.

Of the responders who indicated “Other”, the most common write-in response was adjunct faculty or instructor (5 of 9 written responses). Other responses included “undergrad”, “RIC”, “Contributed services” and one respondent who identified as a “student, too”.

When cross tabulated with University Affiliation (School; Question 2), graduate students were the largest group of respondents in all schools except for Feinberg, where the largest responding group was staff.

Choose all of the following formats that best describe your research data (examples of specific file extensions are included):

Answer		Response	%
Audio (.aiff, .mp3, .wav)		123	18%
Computer aided design/CAD (.dwg, .dxf, .pln)		42	6%
Data (.csv, .dat, .xml)		408	58%
Data – Statistical/SAS, SPSS (.sav, .sdq, .spv)		238	34%
Database (.db, .mdb, .pdb, .sql)		177	25%
Geographic Information Systems/GIS (.gpx, .kml)		31	4%
Image (.bmp, .gif, .jpg, .png, .ps, .psd, .svg, .tif)		363	52%
Matlab (.m, .mat)		112	16%
Scanned documents (.pdf)		366	52%
Scripts or code		149	21%
Spreadsheet (.wks, .xls)		473	68%
Text (.doc, .docx, .log, .rtf, .txt)		520	74%
Video (.avi, .mov, .mp4)		154	22%
Web (.html, .xhtml)		145	21%
Don’t know		11	2%
Other		70	10%

Participants were asked to choose among formats that best described their research. They were allowed to choose more than one format and there were 698 responses. The most common data types were spreadsheets (473 responses or 68%), text (520 responses or 74%), PDFs (366 responses or 52%), data (408 responses or 58%) and images (363 responses or 52%).

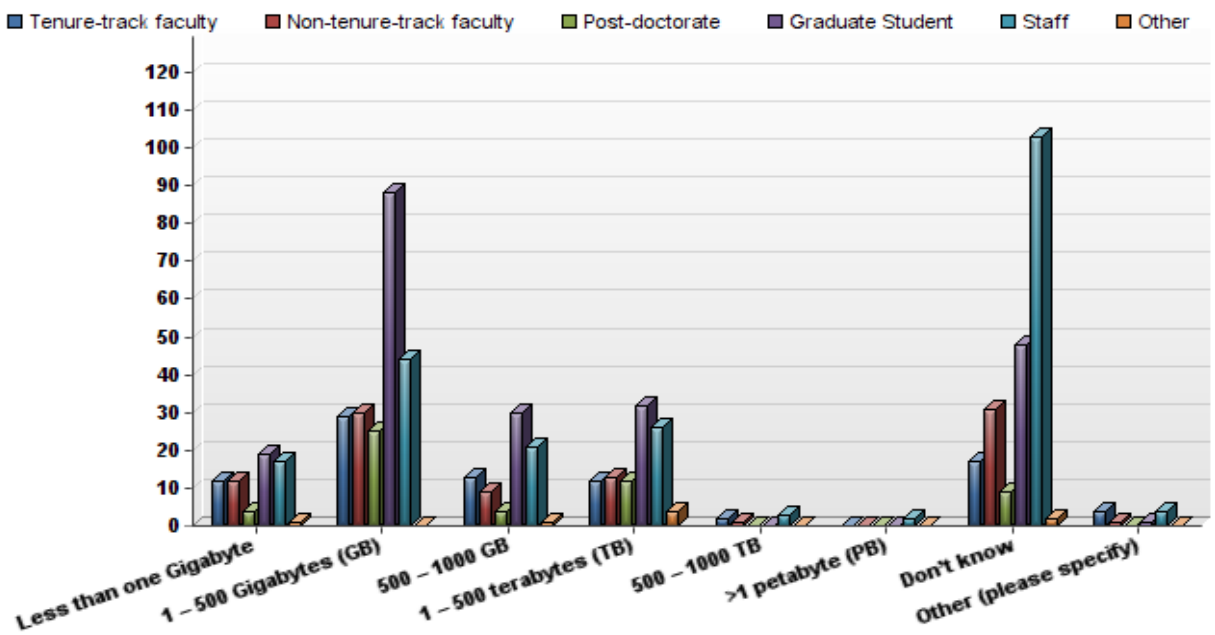
Seventy respondents (10%) selected “Other” data types. Other data types identified include crystallography data, mathematics, a custom format used by the lab, historical archives, various types of experimental measurements such as EEGs, NMR spectroscopy, seismic data and medical image data and genetic sequencing data files. Four responded that they had no data.

How much new or additional data storage do you estimate is required to meet the demands of each new grant over the lifecycle of the grant and for whatever duration you will be required to keep the data?

Answer		Response	%
Less than one Gigabyte		65	9%
1 – 500 Gigabytes (GB)		216	31%
500 – 1000 GB		78	11%
1 – 500 terabytes (TB)		99	14%
500 – 1000 TB		6	1%
>1 petabyte (PB)		2	0%
Don’t know		210	31%
Other (please specify)		10	1%
Total		686	100%

66% of those who answered this question indicated that they would need additional data storage. 216 respondents (31%) believed they would need an additional 1-500 gigabytes, 99 (14%) 1-500 terabytes and 78 (11%) 500-1000 gigabytes. 210 (31%) indicated that they did not know how much data storage they would need and 10 (1%) choose “Other”. Of these, 4 indicated that they did not have grants, 3 had no data and 1 indicated that some of the data was stored on a remote server. Extremely large storage needs were selected by very few respondents.

Because the response “Don’t know” was selected by 31% of respondents, responses to this question were separated by appointment affiliation to determine if knowledge of future data storage requirements differed among faculty, staff and students. It was discovered that nearly half of staff respondents did not know how much data storage would be needed for future projects. 47% of staff responded “Don’t know”, making this the most common response among staff. Furthermore, the staff response accounted for 49% of *all* “Don’t know” responses. Non-tenure track faculty responded “Don’t know” (32%) slightly more their next common response: 1-500 Gigabytes (31%). Tenure track faculty, post-doctorates and graduate students selected 1-500 GB more often than any other response.



Indicate where your data are currently stored (choose all that apply):

Answer		Response	%
Hard drive of the instrument which generates the data		259	38%
PC hard drive		451	66%
External hard drive		318	47%
Departmental/School Server		340	50%
University storage service (e.g. Vault)		95	14%
CD/DVD		63	9%
USB flash drives		186	27%
Internet-based storage (e.g., cloud or grid storage). Please specify provider.		217	32%
External data repository (e.g. Protein Data Bank)		38	6%
Don't know		4	1%
Other (please specify)		51	7%

681 people responded to this question. Most respondents selected more than one option, with computer hard drives and departmental or school servers as the most common responses. Internet based services were also fairly popular (217 responses: 32%).

For option number 8, respondents were asked to name providers if they used internet-based cloud or grid storage solutions. There were 180 written-in responses to this option. Dropbox was the most popular response, with 100 respondents naming it as their sole online storage provider and 13 others naming Dropbox in combination with other online or cloud options, making Dropbox the choice of 63% of respondents who specified an online storage solution. Other common choices were Google (Drive/Docs or Gmail), accounting for 31 responses (17%), Box.net (9 responses; 5%), CrashPlan (9 responses; 5%), REDCap (4 responses; 2%), Amazon Web Services (4 responses; 2%). A variety of other entities made up the rest of the responses.

Fifty respondents selected “Other” storage solutions as response number 11. Most common responses were lab or research group servers or clusters (12 responses; 24% of responses), RAID or Network Attached Storage (NAS) (6 responses; 12%). A variety of other solutions composed the rest of these responses, including paper storage files and servers at partner research institutions.

Identify the minimum number of years you need to preserve your data (in general):

Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don’t know	Total Responses	Mean
Source material/ Raw Data	16	133	162	52	181	100	644	3.85
Processed Data	8	115	164	60	164	119	630	3.97
Results of Statistically Manipulated Data	6	107	158	53	154	149	627	4.10
Published Data	9	87	119	60	230	126	631	4.26

By appointment / status this question was answered by:

- 91 tenured faculty
- 99 non-tenured faculty
- 54 post-doctorates
- 228 graduate students
- 249 staff
- 11 “Other”

"Indefinitely" was the most common response for published and raw data when totaled across all responder groups. The time span of 1-5 years was selected least often by all responder groups, accounting for 7% or less of all responses for totals for all types of data.

Tenured faculty selected "Indefinitely" most often over all types of data collectively, 92% more than any other time span. Published data were preferred to be retained indefinitely 2.4 times more than any other length of time by this group. Post-doctorates selected "Indefinitely" two times more than any other length of time for published data, but "Indefinitely" only accounted for 27% of answers for all data types among post-doctorates. Post docs preferred the period of 5-10 years for raw data, processed data and results of statistically manipulated data. Non-tenured faculty also selected the time span of 5-10 years most often for all categories except for published data, in which they slightly preferred "Indefinitely". Staff selected "Don't know" most often over all totaled responses across all categories of data. By category, for raw data, the time spans of 5-10 years and "Indefinitely" were chosen slightly more often by staff. Graduate students selected "Indefinitely" most often for totaled responses across all categories as well, and also in most categories separately, except "Results of statistically manipulated data", in which they slightly preferred "Don't know".

The following tables show the responses to this question sorted by appointment.

Tenure-track faculty						
Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don't know
Source material/ Raw Data	3	11	23	10	35	4
Processed Data	1	9	20	10	39	5
Results of Statistically Manipulated Data	0	8	17	9	33	15
Published Data	0	10	18	8	43	5
Non-tenure-track faculty						
Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don't know
Source material/ Raw Data	0	17	35	9	23	6
Processed Data	0	11	34	11	21	9
Results of Statistically Manipulated Data	1	11	33	11	21	9
Published Data	1	13	26	10	30	8
Post-doctorate						
Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don't know
Source material/ Raw Data	1	9	13	7	17	5
Processed Data	2	10	15	9	10	5
Results of Statistically Manipulated Data	0	14	14	8	8	6
Published Data	1	10	8	6	20	5

Graduate Student						
Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don't know
Source material/ Raw Data	7	54	37	9	58	38
Processed Data	3	43	50	10	51	43
Results of Statistically Manipulated Data	2	37	45	9	52	55
Published Data	3	25	25	18	86	44
Staff						
Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don't know
Source material/ Raw Data	5	39	51	17	48	46
Processed Data	2	40	43	19	42	56
Results of Statistically Manipulated Data	3	36	45	16	39	63
Published Data	4	28	39	17	50	63
Other						
Question	Less than 1 year	1-5 years	5-10 years	More than 10 years	Indefinitely	Don't know
Source material/ Raw Data	0	3	3	0	0	1
Processed Data	0	2	2	1	1	1
Results of Statistically Manipulated Data	0	1	4	0	1	1
Published Data	0	1	3	1	1	1

Please explain why you plan to keep the data for this amount of time, or explain what types of projects you are pursuing that require different lengths of time to retain data (example: NIH grants require data to be retained for 5 years, NSF require data to be retained for a minimum of 3 years after reporting, etc.):

This question solicited a text answer from respondents. 309 responses were written in.

As a funding body, the NIH was the most-named agency by respondents when providing explanations for their data retention span. Fifty respondents cited NIH requirements for data retention as their reason for keeping data. Eighteen respondents named NSF requirements as their reason for retaining data. Some respondents also cited that journals or publishers required them to retain data upon which their publications were based (7 responses).

Most comments suggested that data are perceived as relevant for long periods of time or indefinitely. Written responses referenced keeping raw data / source material because researchers may potentially use it for future / new studies (77 responses), utilize it for longitudinal studies (9 responses) or share it with colleagues (6 responses). Data were seen as valuable for replicating study results (10 responses),

responding to challenges of published results or because they had been gathered from human or animal subjects who are difficult or costly to replicate. Some responders simply stated that is it good scientific practice to retain data (4 responses).

Do you have a data management plan or policy?

Answer		Response	%
Yes		285	45%
No		213	33%
Don't know		142	22%
Total		640	100%

If you do have a data management plan or policy, indicate the reasons why (choose all that apply):

Answer		Response	%
Yes		285	45%
No		213	33%
Don't know		142	22%
Total		640	100%

There were 297 responses to this question.

71 written comments were supplied for "Other". These included multiple comments relating to data management plans being best practice, good organization, or good sense (14 comments). At least 29 comments were made relating that it was the user's own personal preference or directive to create a plan. Other comments related to protecting privacy of subjects, concerns over loss of data, or requirements by other outside institutions or partners.

If you do not have a data management plan or policy, indicate the reasons why not (choose all that apply):

Answer		Response	%
Lack of information about data management plans		162	58%
Not necessary		116	42%
Other (please specify)		24	9%

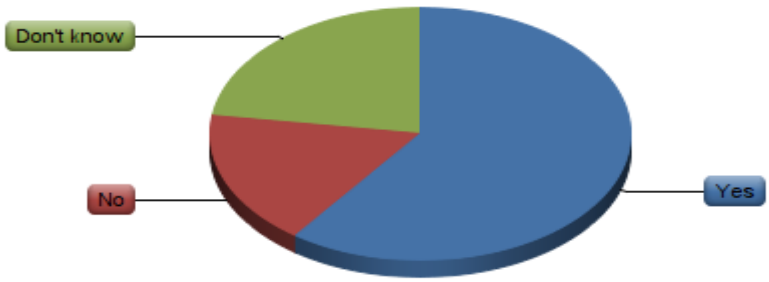
There were 279 total responses to this question.
Cross tabulation with Question 3 (appointment or status) shows that there is a need for education on data management plans for all respondent types.

The breakdown of those that indicated a “Lack of information about data management plans” is as follows:

- 13 Tenure track faculty (38%)
- 23 Non-Tenured-track faculty (58%)
- 14 Post-doctorate (58%)
- 75 Graduate students (57%)
- 35 Staff (51%)
- 2 Other (34%)

For the 24 written responses to “Other”, more than one responder mentioned that creating a plan had not been considered or discussed (4 responses); creating a plan takes too much time or work and thus has not been a priority (5 responses). Some respondents indicated they have a plan in development but it is not finished, or that a single plan could not define all their data, because they are dependent on the nature of the projects (5 responses). Other single responses mentioned lack of storage, servers or trained personnel and the need for an institutional policy on data management.

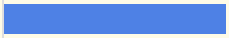
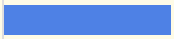
Please indicate if you share or plan to share your research data:



Answer		Response	%
Yes		384	60%
No		110	17%
Don't know		145	23%
Total		639	100%

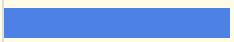
Sixty percent of the respondents said they plan to share their data while 17% said they would not.

If you do share or plan to share your data, with whom is it shared before publication of any manuscripts arising from the research?

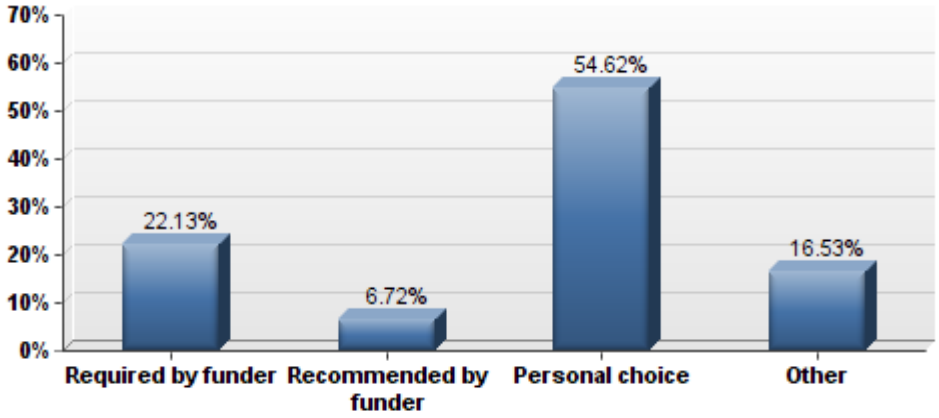
Answer		Response	%
Only members of your research group		174	47%
Colleagues at Northwestern		57	15%
Colleagues in your field (both within and outside the institution)		131	35%
The public at large		11	3%
Total		373	100%

Of the 373 respondents who will share their data prior to publication, 47% said they would share only with members of their research group, 35% indicated with colleagues in their field (both inside and outside the institution) and 15% will share with colleagues at Northwestern. The relatively high percentage of sharing with colleagues inside and outside Northwestern is notable for its potential impact on designing research storage services.

If you do share or plan to share your data, with whom is it shared after publication of any manuscripts arising from the research?

Answer		Response	%
Only members of your research group		50	14%
Colleagues at Northwestern		14	4%
Colleagues in your field (both within and outside the institution)		175	47%
The public at large		130	35%
Total		369	100%

If you do share or plan to share your data, please indicate why:



Answer		Response	%
Required by funder		79	22%
Recommended by funder		24	7%
Personal choice		195	55%
Other		59	17%
Total		357	100%

Of the 59 respondents who indicated “Other”, 53 wrote in responses. Of these responses, 19 mentioned collaboration or sharing with colleagues (36% of written responses); several responders indicated that sharing data was good for the advancement of science or public knowledge (12 responses; 23%); reasons for sharing were a combination of the choices offered by the survey question (5 responses; 9%); sharing is decided by the department, PI, or research group (4 responses; 7.5%); some wrote that simply publishing results is sharing by default (3 responses; 6%); sharing is required by journal in which they publish (2 responses); sharing is standard in their field (2 responses), and a variety of other single responses.

If you share or plan to share your data, how do you share it?

Answer		Response	%
Upload to a university-managed public repository		26	7%
Upload to a federal or discipline-specific public repository (such as GenBank, etc.)		31	9%
Shared site with restricted access		53	15%
Include datasets as part of Supplemental Materials files submitted to journal publisher's site upon publication		57	16%
Share by personal request only		149	41%
Other		47	13%
Total		363	100%

Of the 47 respondents that chose "other", 44 wrote responses. Seven wrote that they share data either through a laboratory server or a web site. Two respondents use Northwestern's Vault and 2 were "not sure". Five indicated that they used more than one method described by the choices. Additional responses include by personal request, publication or YouTube, commercially hosted web servers, by email attachment, subject specific repositories, and systems built on top of university managed servers.

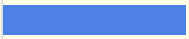
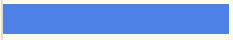
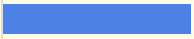
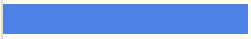
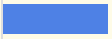
If you do NOT share or plan to share your data, please indicate why not:

Answer		Response	%
Privacy or protection of subjects		64	37%
To protect my intellectual property rights		36	21%
I don't know where to share it		9	5%
No repository exists for my type of data		14	8%
I don't think others would be interested in the data		34	19%
Other		18	10%
Total		175	100%

There were 18 "Other" responses with 17 written responses. Reasons for not sharing data varied and included patent rights, prohibition from sharing by licensing agreements, non-ownership of data, subject

confidentiality, data represents potential projects or inconclusive experiments, it is not done in their field of study, no one has asked for the data and there is no reason to share the data.

Which of the following services might be useful in regard to the management of research data? (choose all that apply)

Answer		Response	%
Assistance with data sharing and/or data management requirements of funding agencies		281	47%
Information about developing a formal data management plan or other data policies		313	52%
Information regarding data management best practices		346	58%
Assistance with selecting data to preserve for the long-term		211	35%
Tools for sharing research data during active research projects (short- or middle-term storage)		288	48%
Data storage and backup during active research projects (short- or middle-term storage)		359	60%
Long-term data access and preservation		375	63%
Assistance applying metadata to research data		126	21%
Assistance finding and accessing data resources		161	27%
Information about citing data resources		120	20%
None of the above		42	7%
Other (please specify):		19	3%

Respondents were allowed to select more than one answer. The total number of responses was 598. Respondents had the most interest in services for data storage and backup during research and long term access and preservation. This indicates a need for both short and long term storage and

preservation on campus. There also appears to be high interest in assistance with Federal requirements, data management plans and data management best practices.

The “Other” option was chosen by 19 (3%) of the respondents. The responses were quite varied. One respondent indicated interest in access to good electronic lab notebook software. Others suggested some sort of cloud based data storage. One respondent stated that they used Dropbox for collaborative work and would like to see Northwestern have a comparable system. This person also felt that Vault was not convenient for collaborations. Others wished for assistance with database development, long-term storage of video data and one respondent stated “technical skills” would be useful. Another respondent suggested that a Northwestern provide university-wide qualitative data management software like NVivo or Atlas TI.

Please provide any additional comments regarding research data management, potential data curation services, or this survey:

There were 76 written responses to this request for additional comments. 25 responses indicated concerns or need for Northwestern university-wide solutions or policies in the area of data management, making this concept the most commonly-addressed issue among the comments. Also prevalent were mentions of needed sustainable solutions for backup or storage of data (16 written comments); comments about sharing research data with colleagues or collaborators (10 comments); concerns about data security (6 comments) or privacy (3 comments). Several comments referred to the adequacy (or inadequacy), or comprehensiveness of current policies on data management in their labs, research groups or disciplines (8 comments). Some users indicated that assistance with data management would be appreciated (5 comments). Some users referred to specific solutions: electronic lab notebook solutions were requested by 4 commenters; a university-wide subscription to Dropbox (or similar) was requested by 4 comments. Specific references to Northwestern’s Vault storage were made (5 comments): one user stated that they “love” Vault; another commented that “mounted Vault drives are slow, and the total space is still limited. More investment to make Vault a truly LUXURY offering would set the University's data management apart for small independent labs”; one respondent felt that Vault as a shared solution has created a situation in the past that violated FERPA laws on protection of human subjects’ data; two respondents commented that Vault was not user-friendly. Three written comments requested data storage to specifically support the needs of students during their studies at the university. Many responders referred to several of the above concepts in a single response.

Sample comments include:

“At present, it seems that everyone comes up with their own individualized plan, if at all. It would be nice to have something more centrally organized or at least suggested. (I suspect many people who could use a plan don't have one and don't even know how to think about one.)”

“This is a major problem in our research field. Thank you for looking into it!”

“I am one of the de facto data management admins of my research group. It was a challenge to set up, and we're still in the process of doing so. I would have appreciated more information from NUIT or other

Northwestern sources. If we had more information easily available about what the best ways to do this were, it would have taken much less time for us to reinvent the wheel, so to speak."

"We are part of a large collaboration (~100 people), and while long-term storage of our raw and reconstructed data is handled by a national lab, we are responsible for the management of the derived data sets created in the process of analyzing data for a publishable paper, so additional resources and/or support for long-term management of these data and codes would be helpful. It would also be helpful to have tools to track data provenance and workflows, to assist in the reproducibility [*sic*] of results."

"Need FISMA and HIPAA compliant web-based data management system with participant portal, e-mailed surveys, flexible reporting, statistical interface which is broader and more flexible than REDCap and essentially free to the investigator."

"Improved interfaces for investigators not well versed in technology would greatly enhance our ability to use all that wonderful information stored in the EDW and other places."

"It would be great to have a central NU repository where research data can be posted and cited!"

"This is a much needed service. I and everyone in my lab are completely clueless to these things despite our best efforts of trying to figure out best practices. Having assistance from the NU Library would be fantastic... both in an *easy* way to store and share data, helping with DOIs, copyright licenses, and other things that don't even come to our minds would be *very* helpful."

Would you be willing to participate in a follow-up interview regarding research data management?

Answer		Response	%
Yes		213	35%
No		400	65%
Total		613	100%

The "Yes" respondents were asked to send an email to a specific email address so that they could be contacted for a further in depth interview. Of the 213 yes respondents, only 2 sent a follow up email. The ESWG will need to determine a means of identifying those participants that were willing to discuss data in more detail.

Once the ESWG has identified willing participants, it plans to use the Data Curation Profiles Toolkit developed by Purdue University to conduct interviews.