Variability in academic research data management practices: implications for data services development from a faculty survey

Citation	Whitmire, A. L., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: implications for data services development from a faculty survey. Program, 49(4), 382-407. doi:10.1108/PROG-02-2015-0017
DOI	10.1108/PROG-02-2015-0017
Publisher	Emerald Group Publishing Limited
Version	Version of Record
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Oregon State
UNIVERSITY OSU

Variability in academic research data management practices
Implications for data services development from a faculty survey

Amanda L. Whitmire and Michael Boock
Oregon State University, Corvallis, Oregon, USA, and
Shan C. Sutton
University of Arizona, Tuscon, Arizona, USA

Abstract

Purpose – The purpose of this paper is to demonstrate how knowledge of local research data management (RDM) practices critically informs the progressive development of research data services (RDS) after basic services have already been established.

Design/methodology/approach – An online survey was distributed via e-mail to all university faculty in the fall of 2013, and was left open for just over one month. The authors sent two reminder e-mails before closing the survey. Survey data were downloaded from Qualtrics survey software and analyzed in R.

Findings – In this paper, the authors reviewed a subset of survey findings that included data types, volume, and storage locations, RDM roles and responsibilities, and metadata practices. The authors found that Oregon State University (OSU) researchers are generating a wide variety of data types, and that practices vary between colleges. The authors discovered that faculty are not utilizing campus-wide storage infrastructure, and are maintaining their own storage servers in surprising numbers. Faculty-level research assistants perform the majority of data-related tasks at OSU, with the exception of data sharing, which is primarily handled by the professorial ranks. The authors found that many faculty on campus are creating metadata, but that there is a need to provide support in how to discover and create standardized metadata.

Originality/value – This paper presents a novel example of how to efficiently move from establishing basic RDM services to providing more focussed services that meet specific local needs. It provides an approach for others to follow when tackling the difficult question of, "What next?" with regard to providing academic RDS.

Keywords Research data services, Data management, Academic libraries, Metadata, Survey, Data sharing

Paper type Case study

1. Introduction

The increasing ease and speed with which researchers can collect large, complex data sets is outpacing their development of the knowledge and skills that are necessary to properly manage them. These skills are crucial to ensuring data quality, integrity, shareability, discoverability, and reuse over time. As funding agencies steadily enact mandates for the submission of data management or sharing plans with proposals, investigators will be held accountable to them (Holdren, 2013). Similar expectations for data accessibility are emerging from some journal publishers, such as PLOS.

Academic libraries are increasingly sources of infrastructure and research support in the area of data stewardship (Akers and Doty, 2013; and references therein), and directly assessing researchers' data needs through the use of surveys is a common tactic employed during the process of developing services (Akers and Doty, 2013;

Averkamp *et al.*, 2014; Marchionini, 2012; Rolando *et al.*, 2013; Scaramozzino *et al.*, 2012; Steinhart *et al.*, 2012; Tenopir *et al.*, 2011). For example, Akers and Doty (2013) used results from a campus survey to make the decision not to expand institutional repository functionality to include preservation and sharing of data sets. Averkamp *et al.* (2014) discovered widespread dissatisfaction with the lack of both centralized data storage and university-supported cloud storage, and shared these concerns with the research services arm of the university information technology group. Survey results gathered by a Provost's Task Force on the Stewardship of Digital Research Data at the University of North Carolina (UNC) at Chapel Hill revealed that more often than not, researchers were relying upon themselves to store data and were using "less desirable practices for data storage" (Marchionini, 2012). They also found that less than 25 percent of survey respondents were aware of certain data management support services that were available. Based on direct feedback from faculty, the Task Force was able to make strong recommendations to the UNC campus administration regarding establishing or expanding cyberinfrastructure and data support services. Several surveys have found that creating metadata is something that researchers struggle with, and that they often use non-standardized methods to document their data or fail to document their data at all (Rolando *et al.*, 2013; Steinhart *et al.*, 2012; Tenopir *et al.*, 2011). The proposed solution to this challenge largely involves training for researchers (e.g. Rolando *et al.*, 2013), but site-specific survey data can also elucidate the extent to which researchers would be receptive to such training if it were developed. For example, Steinhart *et al.* (2012) found that, "nearly two-thirds of respondents reported they would not use a metadata service, whether fee-based or free of charge." In that case, despite the fact that researchers need training, developing a metadata service would likely be a wasted effort. While the results of faculty surveys often reveal common themes, there is no substitute for having an understanding of local research practices when investing in the development of research support services.

This case study reviews the history of data services development at Oregon State University (OSU), and describes how recent faculty survey results are being used to further refine these services. An online survey was distributed to all OSU faculty during the fall of 2013. The survey covered several aspects of research data management (RDM), ranging from characterizing the data that faculty generate, to asking what RDM tasks they struggle with, and what their opinions are regarding who should pay for data services and infrastructure. In this case study paper, we focus on five areas of the survey that generated surprising or particularly important results, and discuss how we will use or have used these discoveries to modify or develop our existing research data services (RDS). First, we discuss the types of data that faculty in different colleges are generating, and review the possible implications for targeting outreach and training. Then we discuss the volume of data that faculty report they are generating, and how this informs planning for future data storage and sharing infrastructure. One of the most important aspects of practice variation among faculty is where they store their data, and we present some unexpected results in this area. As much of the support that our data services group provides occurs one-on-one with researchers, it is critical to understand to whom we should target for assistance. We asked the faculty to describe who performs the majority of RDM tasks in their research endeavors, and now have a better understanding of who to reach out to when we develop new services or products. Lastly, we review current practices on campus for creating metadata, and discuss how we may try to address gaps in this area.

1.1 Timeline of data services development at OSU

OSU Libraries has investigated and engaged in the provision of data services on a limited basis for some time. Historically, OSU Libraries' data services have focussed in two areas: aggregating and visualizing Oregon natural resources-related geospatial data; and building data repository services. In 2000, OSU Libraries partnered with the OSU College of Forestry, College of Science, USDA Forest Service Lab in Corvallis, and the Northwest Alliance for Computational Science and Engineering to create Virtual Oregon, a data archive and portal for "environmental and other place-based data on Oregon and associated areas" (Keon *et al.*, 2002). Virtual Oregon was discontinued due to lack of funding, but was soon replaced with Oregon Explorer (http://oregonexplorer.info/), a series of web portals that include data archiving as well as data visualization tools pertaining to Oregon natural resources.

In addition to portal development to make specific types of data available to the OSU and wider communities, OSU Libraries have worked with faculty and staff on campus in a variety of ways to better ascertain campus needs regarding data. In 2006, meetings were held by members of the OSU community to discuss issues relating to the management and curation of research data across campus, and the feasibility of establishing a spatial data repository for OSU. Underpinning these conversations was the recognition that increasingly large volumes of data were being produced across campus, with no way of knowing what was stored where, by whom and how it was organized. The series of meetings served to gather knowledge about the different kinds of research data that were being produced at the university and potential avenues for sharing information about best practices. The library was an active participant in these meetings, and one result was that the ScholarsArchive@OSU institutional repository was deemed to be an appropriate repository for static data sets smaller than two gigabytes (Avery *et al.*, 2010).

In 2010, OSU Libraries invited faculty from across the university to two lunch meetings at which attendees were asked a series of questions about their data and the libraries' potential role in relation to those data. At this point, the ScholarsArchive@OSU institutional repository, built on the DSpace platform and managed by the libraries, housed a variety of spatial data sets from faculty involved in the 2006 data meetings, as well as a small number of data sets associated with student theses and dissertations. One outcome of these meetings was that the libraries decided to focus on research data associated with theses and dissertations "as a way for the libraries to learn how to do the work involved in curating data" (Boock and Chadwell, 2011).

Although the data services that OSU Libraries currently provides are informed by this history of engagement with OSU faculty, the library still lacked sufficient staffing and critical details that it deemed necessary to provide targeted services and support that would meet campus researcher needs. In 2012, a Data Management Specialist position was established in OSU Libraries to provide leadership in formalizing and expanding the organization's data services. One of the position's initial roles was to participate in the 2012 ARL/DLF/Duraspace E-Science Institute (E-Science Institute, 2012) as part of a small team of librarians and a member of the university's Information Services (IS) department, in order to produce a strategic agenda for RDS at OSU. The agenda provided a roadmap for the development of services in four primary areas: planning and consultation services, access and preservation infrastructure, data management training, and open data consortia and collaborations. It also identified the campus survey whose results are discussed in this paper as the best way to further discern campus needs, and direct an expansion of library and technology support services pertaining to the university's research data (Sutton *et al.*, 2013).

2. Methods

The purpose of the 2013 survey was to improve our understanding of practices, opinions, concerns, and needs regarding data at OSU. We endeavored to understand the nature of the data sets that OSU researchers are generating, and how they are being managed. As such, we distributed the survey to faculty across all ranks from professorial (assistant, associate, and full) to support faculty (faculty research assistants (FRAs) and research associates) and post-doctoral researchers. Survey questions generally fell within the following areas that represent primary issues in data stewardship: data stewardship policies, roles and responsibilities; data characteristics and short-term management practices; data management services and support; data management funding; research data standards and documentation; data sharing; and long-term preservation. The web-based survey was developed using Qualtrics software, referring to the survey from Marchionini *et al.* (Marchionini, 2012) as a starting point. We obtained significant constructive feedback from the OSU Survey Research Center to refine aspects of the survey structure, flow, and question design. The survey was distributed to all OSU faculty members via e-mail addresses that were obtained from the Office of Human Resources (HR). The HR database query resulted in 2,562 e-mail addresses. Data were then downloaded from Qualtrics and analyzed in the software program R (Whitmire, 2015).

3. Results

3.1 Survey Response

The survey was open from October 31-December 5, 2013. After the survey was deployed, it became evident that 528 emeritus faculty were inadvertently included in the e-mail list. While 25 of 39 responses in the "other" category of the faculty rank question actually self-identified as emeritus (via write-in response), we excluded all "other" answers from the results and from our response rate calculation. In total, 572 surveys were started; 443 surveys were completed. There were no required questions, so response rates for each question vary. In total, 76 e-mails bounced or failed to deliver. Therefore, a response rate of 20.6 percent was estimated based on how we treated "other" faculty responses. Excluding all "other" faculty ranks from responses (numerator) and emeritus and bounced e-mail addresses from denominator, we find:

$$\frac{443-39}{2,562-528-76} = 20.6\%$$

We utilized the "Anonymize Response" feature in the Survey Termination section of the Qualtrics Survey Flow to disassociate responses from the individual survey link and scrub the IP address. This effectively de-identified the survey results.

Faculty from every college and unit responded to the survey (Table I), and response rates were generally greater than 20 percent (Table II). Response rates varied across the ranks, ranging from 12 percent for full professors ($n = 97$) to 50 percent for instructors/other/unknown ranks ($n = 39$; Table III).

3.2 Data types, volume, and storage locations

The most common data types that OSU researchers produce are quantitative data (e.g. spreadsheets, delimited text, SPSS, XML; 90.6 percent of total responses), digital images (80.1 percent), and non-digital (handwritten) text (74.9 percent; Figure 1). As expected, differences in the most common data types are evident across colleges.

College or unit	Prof.	Assoc. prof.	Asst. prof.	SFRA/ FRA	Res. assoc.	Post doc	Total
Agricultural Sciences (Agr)	30	18	16	41	3	14	122
Business (Bus)	2	3	4	0	0	0	9
Earth, Ocean and Atmos. Sci. (CEOAS)	17	9	5	16	2	2	51
Education (Edu)	3	3	3	0	1	1	11
Engineering (Engr)	6	7	10	5	1	2	31
Forestry (For)	3	3	7	17	1	4	35
Liberal Arts (LibArt)	12	8	12	0	0	0	32
Pharmacy (Pharm)	2	4	4	0	0	0	10
Public Health and Human Sci. (PHHS)	4	8	7	6	1	2	28
Science (Sci)	12	5	4	11	3	8	43
Veterinary Medicine (Vet Med)	2	3	6	3	0	2	16
University Libraries (Lib)	1	4	4	0	0	0	9
Other	3	0	0	3	0	1	7
Total	97	75	82	102	12	36	404

Notes: These response numbers do not include responses from faculty who responded as "other" to the question regarding their rank ($n = 39$). The college and unit abbreviations used in figures are shown in parentheses. Ranks are Professor (Prof.), Associate Professor (Assoc. Prof.), Assistant Professor (Asst. Prof.), Senior Faculty Research Assistant and Faculty Research Assistant (SFRA/ FRA), Research Associate (Res. Assoc.; not including post-docs), post-doctoral researchers (all types, including Research Associate, Fellow, etc.) and other (affiliations include Research Centers and Institutes, student affairs and academic programs such as the Graduate School, Extension; ranks include instructors, courtesy faculty, support faculty affiliated with a Research Center, etc.)

Table I.
Number of completed responses from each college or unit, by rank

College or unit	Contacts	Responses	Response rate (%)
Agricultural Sciences	743	122	16
Business	71	9	13
Earth, Ocean and Atmos. Sci.	251	51	20
Education	35	11	31
Engineering	263	31	12
Forestry	177	35	20
Liberal Arts	206	32	16
Pharmacy	59	10	17
Public Health and Human Sci.	152	28	18
Science	281	43	15
Veterinary Medicine	76	16	21
University libraries	31	9	29
Other	217	7	3
Total	2,562	404	16

Notes: The number of contacts shown includes emeritus faculty who were inadvertently contacted, and bounced e-mails. As such, these response rates shown are slightly lower than the estimated response rate for the survey as a whole. These response numbers do not include responses from faculty who responded as "other" to the question regarding their rank ($n = 39$)

Table II.
Response rates by college or unit

Position/rank	Contacts	Responses	Response rate (%)
Professor	826	97	12
Associate professor	495	75	15
Assistant professor	453	82	18
Research associate/fellow	297	48	16
Faculty research Assistant	284	61	21
Senior faculty research assistant	129	41	32
Instructor/other/unknown	78	39	50
Total	2,562	443	17

Notes: The number of contacts shown includes emeritus faculty who were inadvertently contacted, and bounced e-mails. As such, these response rates shown are slightly lower than the estimated response rate for the survey as a whole. These response numbers include responses from faculty who responded as "other" to the question regarding their rank, but these responses were removed from the survey analysis
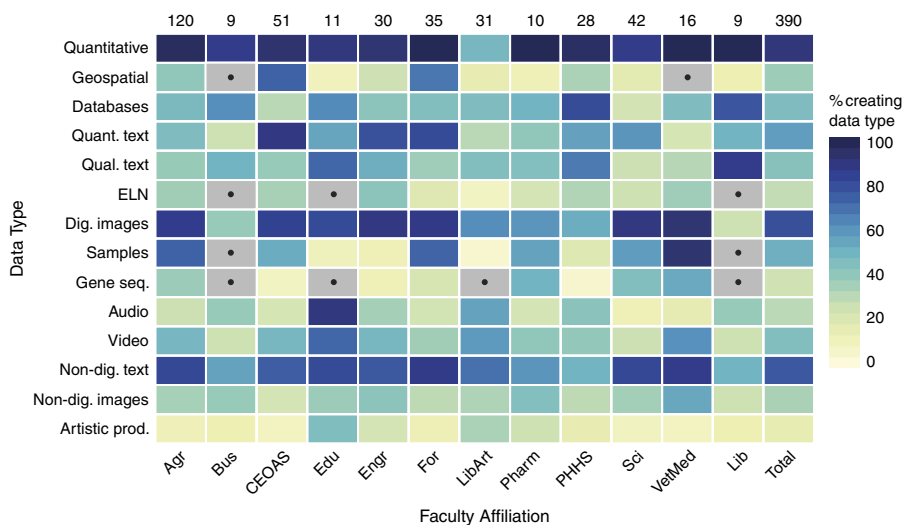
Table III.
Survey response
rates by rank

Figure 1.
Responses to the
question, "Please
indicate whether or
not you generate
each of the following
data format(s) as a
part of your research
process. Select Yes
or No for each"

Notes: Color scale indicates what percentage of respondents in each college or unit selected "Yes" for each data type. Light gray with a bullet indicates zero "Yes" responses. The number above each column shows the total number of faculty responses for that college/unit

For example, a higher percentage of researchers within the Colleges of Earth, Ocean and Atmospheric Sciences (CEOAS) and Forestry (For) produce geospatial data than in other colleges, while qualitative text (e.g. an interview transcript) is more prevalent in Education (Edu), Public Health and Human Sciences (PHHS), and Liberal Arts (LibArt).

When asked about how much data they are producing, OSU faculty report that for a "typical" research project, they generate less than 100 GB in most cases ($n = 186$; Figure 2). Again, depending on their discipline, some researchers produce much more and some much less. There were no responses in the ranges from 100 TB-1 PB or > 1 PB. In total, 15 percent of respondents indicated that they did not know how much data they

typically produce. When asked about the largest amount of data they have produced for a single project, only three respondents answered in the 100 TB-1 PB range (Colleges of Agriculture ($n = 2$) and Liberal Arts ($n = 1$)), and one in the > 1 PB range (CEOAS; data not shown).

Overall, OSU faculty report storing short-term data (data less than five years old) most often on personal computers (PC; 85 percent) and external storage devices (83 percent; faculty could report storing data in multiple locations; Figure 3). In several colleges and departments, faculty report storing data on servers held within their research group, in most cases despite the fact that their college or department offers replicated, network server-based storage as a service. Colleges with high numbers of respondents using their own research group servers include CEOAS (75 percent report having their own server), Engineering (53 percent), Science (58 percent), and Vet Med (56 percent).

College and departmental servers are also well utilized by faculty for storing short-term data, especially in Agriculture (62 percent), Business (75 percent),
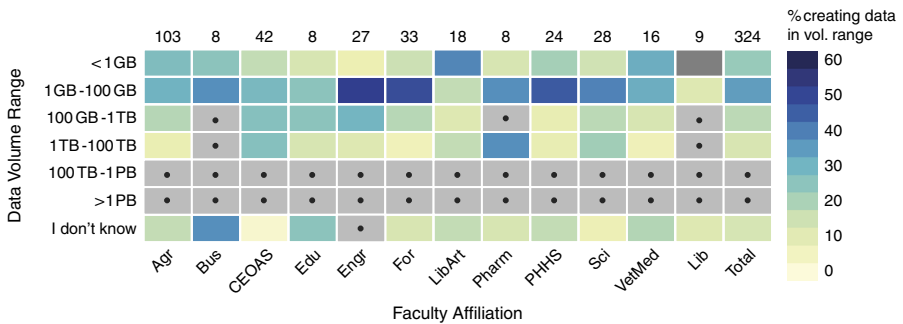
Figure 2.
Responses to the question, "What has been the typical amount of digital data for a single project you have worked on in the past 5 years?"

Notes: Color scale indicates what percentage of respondents in each college or unit selected the given data volume range. Light gray with a bullet indicates zero responses. The number above each column shows the number of faculty responses in college/unit. The percent of library faculty responses is off-scale at 78 percent in the <1GB range (dark gray)
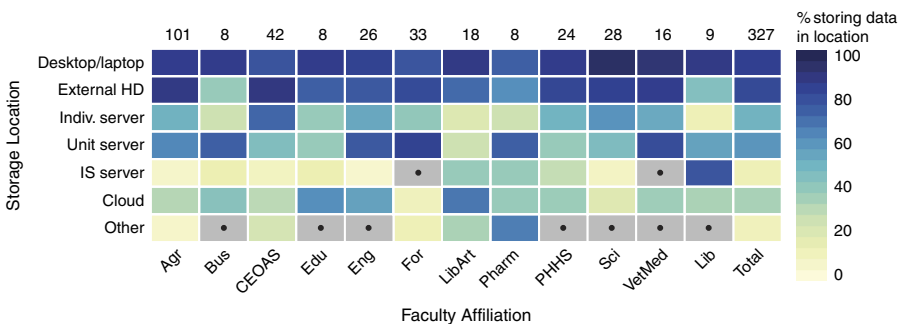
Figure 3.
Responses to the question, "Thinking about data you've generated in the last five years (short-term data), please indicate where you store and/or backup these data. Select Yes or No for each"

Notes: Color scale is the percent of faculty that responded "Yes," where the total responses include "Yes," "No," and "I don"t know.' Light gray with a bullet indicates zero "Yes" responses. The number above each column shows the total number of faculty responses (Y+N+IDK) for that storage location and college/unit

Engineering (77 percent), Forestry (83 percent), Pharmacy (75 percent), and Vet Med (82 percent). Cloud-based storage options are utilized at rates between 9 percent (Forestry) and 63 percent (Education). It is interesting to notice that campus-wide server-based storage infrastructure, offered though IS, is not heavily utilized. Only 11 percent of respondents indicated that they store data with IS, and most were in units that also reported producing smaller data sets. It is important to note that a given faculty member may employ different data storage options at different times, so the use of one method does not entirely preclude the utilization of others.

3.3 Data management tasks and roles
With the exceptions of data analysis, sharing, and disposal, the survey results indicate that FRAs handle the majority of data management tasks (Figure 4). At OSU, personnel in research support positions, such as laboratory technicians and research assistants, are distinguished from administrative staff in that they have non-tenure track faculty status (as opposed to "classified staff" status). As such, research personnel are known as "FRAs," or FRAs, and they are almost exclusively supported on "soft money" by research grants. In the case of researchers in less data-intensive colleges (e.g. Liberal Arts or Business) however, principle investigators (PIs) handle the majority of these tasks themselves (college-level data not shown). Graduate students are almost never responsible for data sharing outside of the research group, nor are they typically involved in data archiving or data disposal. While less involved than research assistants, faculty reported that graduate students do participate in data collection, metadata creation, quality control, and analysis. The only data management tasks for which faculty reported involvement by IS were data backup and archiving. The professorial ranks handle the majority of data sharing.

3.4 Metadata practices
The proportion of faculty who report that they create metadata varies widely by college. Only 19 percent of Veterinary Medicine faculty create metadata, while
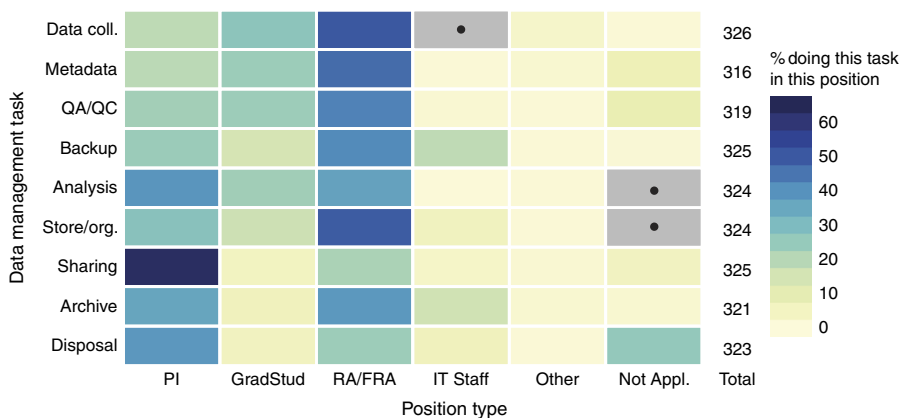
Figure 4.
Responses to the question, "Who performs the majority of each of the following digital data management tasks associated with your research?"

Notes: For each task, respondents could choose one position type only. Color indicates what percent of each task is being conducted by the given position type. Light gray with a bullet indicates zero responses. The numbers in the "Total" column show the total number of responses for each task. Note that the color scale range is 0 to 65 percent

88 percent of Earth, Ocean, and Atmospheric Sciences faculty do (Table IV). Response rates to the question were also highly variable. In total, 94 percent of Forestry faculty responded to this question, while 56 percent of Liberal Arts faculty did. Only faculty who responded "Yes" to the first metadata question were asked about which metadata standard they are currently using. Faculty who are creating metadata are overwhelmingly using a schema that has been standardized within their research group (Table V). Interestingly, 17-19 percent of faculty do not know if they are using the metadata standards listed in the survey question.

4. Discussion

4.1 Nature and volume of data produced at OSU
The survey results clearly demonstrate that faculty are generating a wide variety of data types. This was not unexpected, but it's helpful to see "who" (faculty in which colleges) is generating "what" (data types) as we consider adding support services or training. For example, a high percentage (> 80 percent) of faculty in the colleges

College or unit	Yes	No	Total	% Yes	Survey responses	Response rate (%)
Agricultural Sciences	41	62	103	40	122	84
Business	3	5	8	38	9	89
Earth, Ocean and Atmos. Sci.	36	5	41	88	51	80
Education	5	3	8	63	11	73
Engineering	19	8	27	70	31	87
Forestry	24	9	33	73	35	94
Liberal Arts	6	12	18	33	32	56
Pharmacy	3	5	8	38	10	80
Public Health and Human Sci.	12	11	23	52	28	82
Science	15	13	28	54	43	65
Veterinary Medicine	3	13	16	19	16	100
University libraries	4	5	9	44	9	100
All units	171	151	322	53	397	81

Table IV.
Responses to the question, "Do you generate metadata? For example, do you currently document or describe your data, create code books, data dictionaries, 'README' files, etc.?"

Note: For comparison, the two "Responses" columns on the right show the number of respondents from each college for the survey as a whole, and the subsequent within-survey response rate for this question

	Yes	No	I don't know	Total
DC (Dublin Care)	5	121	25	151
DwC (Darwin Core)	3	124	25	152
DDI (Data Documentation Initiative)	3	121	26	150
DIF (Directory Interchange Format)	3	120	26	149
EML (Ecological Metadata Language)	16	112	24	152
FGDC (Federal Geographic Data Committee)	17	106	27	150
ISO 19115 (Geographic Information)	14	105	28	147
OGIS (Open GIS)	4	118	25	147
Metadata standardized within my lab	90	55	14	159
Other (specify):	18	21	13	52

Table V.
Responses to the question, "Please indicate which metadata standard you currently use to describe your data. Select Yes or No for each"

Notes: Only respondents who answered, "Yes" to the question regarding metadata creation were prompted to answer this question

of Earth, Ocean and Atmospheric Sciences, Engineering, and Forestry are creating digital text for quantitative research (Figure 1), which was described in the survey as, "software scripts and codes, descriptive information/metadata." We have been considering offering Software Carpentry (SC; http://software-carpentry.org/) workshops, which would provide instruction on topics like version control, programming languages (R, Python, and Matlab), and using databases. The survey results indicate that we would want to target our outreach and marketing for these workshops to those three colleges, due to the prevalence of software code being generated by their constituents. Results also indicate that faculty, graduate students, and research assistants are all involved in the analysis phase of the research lifecycle (Figure 4), so we would have to consider either targeting the SC workshop to an audience with a broad level of experience (beginning coder to very experienced), or creating separate workshops for students and faculty.

In seven of the 12 colleges surveyed, more than 80 percent of faculty report that they are creating digital image data, and in three of the colleges more than 50 percent of faculty are (Figure 1). Despite the fact that several campus surveys have found the same results with respect to the prevalence of digital image data (Averkamp *et al.*, 2014; Marchionini, 2012; Rolando *et al.*, 2013; Steinhart *et al.*, 2012), this was an interesting and unexpected finding that has broad implications for data storage and backup, file organization and naming, metadata, and data sharing and preservation. The fact that so many digital images are being created on campus points to a potential need for providing specialized support materials (e.g. Cornell University Library, R.D., 2000; Jisc, 2015) or a workshop on best practices for managing digital images. This observation also generates more questions: what are they taking pictures of? How are they being analyzed? With increasing funder and publisher mandates for data sharing, how will researchers share them? These questions point to a need to further engage with faculty on the topic of the use of digital images for research, perhaps via a series of Data Curation Profiles (Carlson, 2013; Witt *et al.*, 2009) dedicated to the issue.

A topic closely related to the types of data that faculty are generating, is the volume of data being generated. While the topic of "big data" has been grabbing headlines, research funding (Zgorski, 2012), and even its own journals (e.g. Elsevier's *Big Data Research* and Springer's *Journal of Big Data*), the large majority of researchers (75 percent at OSU) are still creating what we consider to be "regular data" (Figure 2), which we arbitrarily define as being less than one terabyte in size. Only 40 researchers out of 324 respondents (12.3 percent) report that they are producing "typical" data sets in the 1-100 terabyte range, and none report creating anything larger than that under usual circumstances (three researchers report that the largest data set they have created is 100 terabytes-1 petabyte in size, and one reported their largest data set was >1 petabyte; data not shown). The main implication of this finding is that meeting the data storage needs of our faculty is likely to be a tractable challenge.

Faculty at OSU already have access to 30 gigabytes of free Google Drive cloud storage, and will soon also each have access to 1 TB of free storage with Microsoft's OneDrive cloud storage service. While not universally ideal, cloud-based, vendor-provided data storage has some advantages over using laptops, external hard drives, and individually maintained servers. Most notably, cloud storage is replicated and secure, and assuming an Internet connection is available, can be accessed from anywhere. Unlike using laptops and external drives, data stored in the cloud is not at risk of physical theft or accidental damage. Drive and OneDrive also offer variable access permissions at the file and folder level, so that researchers involved in collaborative work can more easily share data within the project. A drawback of using cloud-based

storage on our campus is the potential for upload/download bottlenecks due to the limited speed of our connection with the internet. Researchers who need to move large data sets around would be better off working on local servers, which have much faster transfer rates within the campus network. Ultimately, there will be plenty of space available for most researchers to store their short-term data, but how well faculty are managing those data (e.g. using a thoughtful and consistent file-naming convention) remains an open question. Perhaps the more interesting question is whether or not faculty will take advantage of these storage options as they become available. Results from the data storage location question reveal some unexpected and disconcerting habits.

4.2 Data storage habits

Perhaps the most surprising discovery revealed by this survey is the fact that a large percentage of faculty are managing their own data servers. We expected to see that faculty are storing short-term data (less than five years old) on desktop and laptop computers and external storage drives, and that is born out in the results (Figure 3). In addition to central data storage options that are available through IS, several colleges on campus have their own computing support services for data storage and backup. In light of this, we did not expect faculty to be maintaining their own servers in any appreciable number. However, as noted in the Results section, significant proportions of faculty in CEOAS (75 percent), Engineering (53 percent), Science (58 percent), and Vet Med (56 percent) report that they are storing data on servers that they maintain themselves, despite the fact that replicated, networked storage is available within their college. Likewise, only 11 percent of faculty store data with IS. This indicates that either the centralized (college and university level) cyberinfrastructure resources do not currently meet faculty needs in this area, or that faculty are unaware of their data storage options. A sample of write-in responses to the survey shed some light on how faculty view this problem:

> Having reliable, scalable, and relatively inexpensive short term data storage is critical for our work. Our current model requires us to buy lab specific equipment that degrades over time and that is not completely backed up. It would be fantastic to have a central repository for data that can use economies of scale to increase reliability and redundancy and allow us to focus on analyzing the data rather than managing it.

> Some college level services (data storage and backup) should be available at the university level in a visible OSU data center.

> We constantly struggle with adequate, secure, backed-up disk space for our projects. Central data storage (on the order of 10s to 100s of TBs), provided by the college/university, would be a big help.

Given that centralized data storage services do exist with IS, a critical issue to explore further is the degree to which faculty's low use of IS options is attributable to a lack of awareness vs shortcomings in the options themselves, so those centralized services may be improved to enable wider adoption.

Between PCs, external hard drives and personal servers, this level of ad hoc, do-it-yourself data storage exposes a significant proportion of the data produced at OSU to serious risks. How much of the data stored in these locations is backed up in multiple locations (i.e. replication)? Are faculty aware of the life expectancy of PC and external hard drives and their rates of failure? Are researchers adequately prepared

to manage their own storage servers effectively? Why do faculty rely on themselves for data storage at such high percentages? Is self-reliance in this area a personal preference, or is it a lack of high-quality, affordable centralized infrastructure (either within the college or the university)?

Our library has an effective, valuable working relationship with IS, and the results of this survey have sparked a substantive conversation about possible causes of and remedies for the lack of uptake of centralized cyberinfrastructure by the OSU community. IS is currently working on implementing an advanced "object storage" cyberinfrastructure system, which is expected to effectively eliminate storage volume limits and significantly reduce costs for campus users. We will be working with them to advertise this option and provide support for how to take advantage of its features once this option becomes available. As faculty shift over to a new storage system, this will also provide opportunities to have conversations with them on topics such as file-naming conventions and folder organization.

4.3 Targeting outreach and services

One of the biggest challenges that we have faced in developing RDS at OSU Libraries has been our lack of visibility on campus. For example, only 13 percent of faculty reported that they are aware of our services related to developing or reviewing data management plans (data not shown). In light of this, we believed that it would be beneficial to better understand who on campus, in terms of their position, is handling which data-related tasks. The hope was that a better understanding of who is doing what would enable us to focus outreach efforts on the appropriate audience. This would make better use of our limited RDS resources and improve our chances of service uptake.

FRAs perform the majority of several RDM tasks, including data collection, metadata creation, quality control, and data backup, storage, and organization (Figure 4). They share about equally in data analysis and data archive tasks with those in the professorial ranks. The conclusion we can draw from these results is clear: we need to be reaching out to research assistants with support and training in all aspects of RDM best practices. FRAs play a large role in data storage and organization, with 49.4 percent of respondents indicating that FRAs are primarily responsible for this activity. As we collaborate with IS in building out new storage infrastructure, we need to make the professorial ranks aware of the resource, but teach FRAs how to use it.

Of all RDM tasks, data sharing had the highest percentage of responses clustered in a single rank, at 63.1 percent. In this case, professors were predominantly responsible for the sharing of their research. Given that professors are the ones serving as PI on the grants that support much of the research performed at OSU, and that they oversee the projects and the work, it is no surprise that they act as gatekeepers to the products of their work. As federal mandates for the sharing of research results continue to expand across agencies, and become more rigorously audited (Holdren, 2013), we can play an important role in helping PIs stay up to date with the sharing requirements. It will also be important for us to be aware of the growing number of options for archiving and sharing research data, so that we can help PIs discover and utilize them effectively. These options currently range from Federal, discipline-specific archives (e.g. the National Center for Biotechnology Information, or the National Ocean Data Center), to private, discipline-agnostic sharing platforms (such as figshare) or repositories that exist to support the sharing of data sets associated with publications (like Dryad).

Another data sharing option that is becoming increasingly available for academic researchers is using an institutional repository. OSU Libraries has just begun the development of a new platform for our institutional repository (IR; currently on DSpace). The IR will be built on a Hydra/Fedora repository backend (http://projecthydra.org), which will enable a much more nuanced and robust data model. It will also allow for much more expansive and flexible metadata, and explicitly define relationships between objects in the repository. For example, we will be able to assign researcher IDs to data set creators (e.g. ORCID) and we will be able to retain the folder and file structure of data deposits. While the transition to a new repository system should be invisible to campus users (with the exception of a significantly improved user interface), we will be able to encourage depositors to use RDM best practices by supporting an expanded range of metadata schemas, preserving folder and file structure, and adding functional links between data sets and related content (both inside and outside of the IR). While the programmers work to develop and refine the IR platform, our data specialists will need to invest significant effort toward developing outreach and training materials for PIs and FRAs (since FRAs share in the work of data archiving). We will also need to be prepared to spend time offering workshops and guest lectures for broad audiences on features of the new IR and how to take advantage of them.

4.4 Metadata support is needed

The survey results regarding metadata practices are promising, but also provide a potential area for engagement with faculty and the development of training exercises. With 397 (or 81 percent) of survey respondents answering the question of whether or not they create metadata, 53 percent report that they do (Table IV). This agrees well with the results of an international survey on researcher data management practices, which found that 54 percent of researchers were creating metadata (Tenopir *et al.*, 2011). Within the group that reported creating metadata, Tenopir *et al.* (2011) found that a combined 78.2 percent of the researchers were either not using a metadata standard, or were using a standard devised within their lab. Likewise, we found that a total of 74.5 percent of respondents to our survey were either using a standard within their group (56.9 percent) or were not using one at all (17.6 percent; Table V, including "Other (specify)" responses not shown).

It is encouraging that nearly half of OSU researchers report that they create metadata. However, the extent to which researchers are not using standard metadata schemas is an area where we can improve data stewardship on campus. Data sets that have metadata that conforms to a standard will be more interoperable with other data sets, more discoverable (by machines and by humans), and are likely to be more thoroughly documented compared to those that have an ad hoc schema. Since so many researchers are already creating metadata, it's not likely that we would get much traction with an introductory metadata workshop (unless perhaps, we geared it toward early-stage graduate students). There appears to be more of a need for training in how to implement specific metadata standards, ideally using an available tool to do so. For example, Ecological Metadata Language (EML) and FGDC/ISO 19115 were among the most commonly selected schemas among faculty who are using a standard (Table V). It would make sense, then, to develop a workshop to train faculty in metadata creation under each of those standards, using existing tools for doing so (e.g. Morpho, in the case of EML). Survey respondents were also asked about how important it was for OSU to invest in providing certain data services, including guidance in how to use metadata standards. A combined total of 58 percent of faculty rated this type of guidance as moderately or

very important (Figure 5), which indicates that there may be a sizeable pool of faculty who would be receptive to training in this area and that development of this training should be strongly considered.

We also know from the survey results that faculty rate creating metadata as one of their most difficult tasks (data not shown), and that research assistants are most commonly performing data documentation tasks (Figure 4). This implies that we should be tailoring the content of metadata training, and how we do outreach, for the FRAs. This is a particularly good example of how valuable local survey results can be in helping to determine how to most effectively invest limited RDS resources and time.

5. Conclusions

The primary goal of launching a data stewardship survey was to characterize the RDM practices of OSU faculty, and subsequently determine where expanded RDS efforts could be most effectively applied. In this paper, we focussed on five aspects of the survey: the types and volume of data being generated by OSU faculty; their data storage habits; the roles and responsibilities for various data-related tasks; and metadata. We had a response rate of just over 20 percent, with almost 450 faculty completing the survey. After excluding results from faculty of unknown rank, we had data from 404 completed surveys (though no survey questions were required, so response rates vary by question).

We found that OSU researchers are generating a wide variety of data types, and that practices vary between colleges. We were surprised to discover that such a large percentage of researchers on campus are generating digital images as a part of their research (80.1 percent). We are motivated to further engage with faculty on this topic in order to better understand their habits and how we can support them. We also discovered that faculty are largely not availing themselves of centralized cyberinfrastructure resources. Instead, even faculty in colleges that have computing support are often going so far as to maintain their own data storage servers. In several cases, faculty who are maintaining their own servers are also generating data at a higher volume. This level of ad hoc storage exposes a significant portion of the OSU research data corpus to significant risk of loss. This finding provides impetus for library collaboration with IS and the campus administration on how we can increase the utilization of centralized, replicated data storage options.

At OSU, faculty-level research assistants perform the majority of data-related tasks, including data collection, metadata creation, quality control, and data storage, organization, and backup. They share data analysis and data archiving responsibilities with PIs, while PIs play the most significant role in data sharing. These observations provide clear direction regarding who we should be targeting outreach, training and
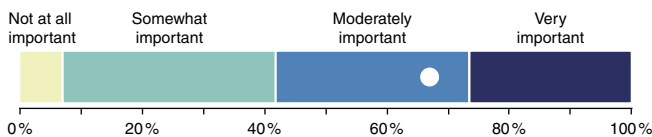
Notes: $n = 303$. The white dot shows the mean response.
"How important do you think it is for OSU to spend resources on providing the following services [...] Guidance on how to use appropriate metadata standards?"

Figure 5.
Categorical responses to the question about the importance of providing metadata support

support for given RDM tasks. Since faculty report that they currently struggle with how and where to share data (results not shown), we now understand that we need to provide support directly to PIs in this area.

Finally, we reviewed the metadata habits of our faculty. While we are buoyed by the discovery that such a large percentage of faculty are creating metadata (53 percent of respondents to that question, which had an 81 percent response rate itself), we see room for improvement in the area of increasing the use of standardized metadata schemas. Since so many faculty are already creating metadata, we believe that a more advanced workshop that would enable attendees to learn how to create standardized metadata pertinent to their disciplines is warranted. There is also likely a place for an introductory metadata workshop for early-stage graduate students (and open-minded faculty), given that over half of OSU faculty are not creating metadata (when those who did not answer the question are included).

Overall, the results of a campus-wide faculty survey on research data stewardship has provided us with significant insight into local RDM practices. We see several areas where we can develop targeted services and training, and areas where we would like to delve more deeply into what, how and why researchers do what they do with their data. The purpose of this case study, which shares abbreviated results from that survey, was to provide other academic libraries and/or RDS personnel with a few examples of the value and utility of conducting such a survey. To the extent that it agrees with the findings of other RDM practices surveys, we also believe that some of these results may be generalizable to the wider academic community (e.g. PIs are likely to be the best point of engagement for offering data sharing services, are often employing their own servers for data storage, and are somewhat unlikely to be employing standard metadata schema). It is almost certainly true in most places that university faculty produce an impressively diverse corpus of data types, formats, and sizes, and that these data sets are stored in a myriad of locations, from ideal to less so. The results of our survey, taken in context with the results of other such surveys, point to a ubiquitous need for thoughtfully planned academic RDS that are simultaneously broad in scope and strategically focussed on addressing specific local needs. While it is possible to generalize about the common challenges that researchers face with respect to RDM, it is also undoubtedly true that in the endeavor to address those challenges, the devil is in the details.

Acknowledgments

References

Akers, K.G. and Doty, J. (2013), "Disciplinary differences in faculty research data management practices and perspectives", *Int. J. Digit. Curation*, Vol. 8 No. 2, pp. 5-26. doi: 10.2218/ijdc.v8i2.263.

Averkamp, S., Gu, X. and Rogers, B. (2014), *Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs*, Univ. Iowa Libr. Staff Publ, Iowa City.

Avery, B.E., Chau, M., Vondracek, R. and Wirth, A.A. (2010), "OSU libraries and research dataset curation: a beginning", working paper, Oregon State University Libraries, Corvallis.

Boock, M. and Chadwell, F.A. (2011), "Steps toward implementation of data curation services", Oregon State University Libraries, Corvallis.

Carlson, J. (2013), "Opportunities and barriers for librarians in exploring data: observations from the data curation profile workshops", J. EScience Librariansh, Vol. 2 No. 2, available at: http://dx.doi.org/10.7191/jeslib.2013.1042

Cornell University Library, R.D. (2000), "Digital Imaging Tutorial – Contents", Mov. Theory Pract. Digit. Imaging Tutor, available at: www.library.cornell.edu/preservation/tutorial/contents.html (accessed 2 June, 2015).

E-Science Institute (2012), "Home Page", available at: http://duraspace.org/e-science-institute (accessed 10 August, 2012).

Holdren, J.P. (2013), "Memorandum for the heads of executive departments and agencies: Expanding public access to the results of federally funded research", Executive Office of the President, Office of Science and Technology Policy, Washington, DC, February 22, available at: www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Jisc (2015), "Systems for managing digital media collections", Jisc Digit. Media, available at: www.jiscdigitalmedia.ac.uk/guide/systems-for-managing-digital-media-collections/ (accessed 2 June, 2015).

Keon, D., Pancake, C. and Wright, D. (2002), "Virtual oregon: seamless access to distributed environmental information", Proceedings of the 2Nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '02, ACM, New York, NY, pp. 387-387.

Marchionini, G. (2012), Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership, University of North Carolina, Chapel Hill, NC.

Rolando, L., Doty, C., Hagenmaier, W., Valk, A. and Parham, S.W. (2013), "Institutional readiness for data stewardship: findings and recommendations from the research data assessment", technical report, Georgia Institute of Technology, Atlanta.

Scaramozzino, J.M., Ramírez, M.L. and McGaughey, K.J. (2012), "A study of faculty data curation behaviors and attitudes at a teaching-centered university", Coll. Res. Libr, Vol. 73 No. 4, pp. 349-365.

Steinhart, G., Chen, E., Arguillas, F., Dietrich, D. and Kramer, S. (2012), "Prepared to plan? A snapshot of researcher readiness to address data management planning requirements", J. EScience Librariansh, Vol. 1 No. 2. doi: 10.7191/jeslib.2012.1008.

Sutton, S., Barber, D. and Whitmire, A.L. (2013), Oregon State University Libraries and Press Strategic Agenda for Research Data Services, Oregon State University, Corvallis.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. and Frame, M. (2011), "Data sharing by scientists: practices and perceptions", PLoS ONE, Vol. 6, No. 6, e21101. doi: 10.1371/journal.pone.0021101.

Whitmire, A.L. (2015), "Data and code from: variability in academic research data management practices: implications for data services development from a faculty survey", Oregon State University Libraries, Corvallis, available at: http://dx.doi.org/10.7267/N9J1012R

Witt, M., Carlson, J., Brandt, D.S. and Cragin, M.H. (2009), "Constructing data curation profiles", Int. J. Digit. Curation, Vol. 4 No. 3, pp. 93-103. doi: 10.2218/ijdc.v4i3.117.

Zgorski, L.-J. (2012), "NSF leads federal efforts in big data, press release 12-060", Natl. Sci. found, available at: www.nsf.gov/news/news_summ.jsp?cntn_id=123607&org=NSF&from=news (accessed 2 June, 2015).

Appendix 1. Survey instrument

INTRODUCTION. Thank you for participating in the Center for Digital Scholarship & Services Survey on Research Data Stewardship at Oregon State University. Your responses will help us better understand the data landscape at OSU: how much data are being created and in what types and formats, and how faculty are managing them. **Results from this research survey will contribute to our efforts to build better support and services for research data stewardship on campus.**

This survey covers topics including funding agency and publisher mandates regarding data, perceptions of data ownership, funding support and services for research data management, and current researcher practices. These topics are relevant to all researchers at OSU, and your participation may benefit you and the wider OSU research community by enabling informed, targeted expansion of services to meet current needs.

Your participation is voluntary; you may skip questions or end the survey at any time. After the conclusion of the survey, your name and e-mail address will not be associated with your responses in any way. Results from the survey will be reported in aggregate by such factors as rank and college appointment, and the data set and analysis will be shared in a publicly accessible repository and via conference proceedings and publications. It is theoretically possible that your identity may be ascertained by pairing your rank and department, but there is no risk associated with answering the survey.

The survey will take approximately 10-15 minutes to complete. You may close the survey and return to it at any time using the survey link you received in the e-mail invitation.

The security and confidentiality of information collected from you online cannot be guaranteed. Confidentiality will be kept to the extent permitted by the technology being used. Information collected online can be intercepted, corrupted, lost, destroyed, arrive late or incomplete, or contain viruses.

If you have any questions or concerns, please contact the Principal Investigator of this research, OSU Libraries' Data Management Specialist, Dr. Amanda Whitmire, at amanda. whitmire@oregonstate.edu or 541-737-3133. If you have questions about your rights as a survey participant, please contact the Oregon State University Institutional Review Board (IRB) by e-mail at IRB@oregonstate.edu and refer to study number 5790 (Survey on Research Data Stewardship at Oregon State University).

Please select 'I Agree' and click the [>>] button below to provide consent and begin the survey.

❍ I Agree
❍ I Do Not Agree

If I Do Not Agree Is Selected, Then Skip To End of Survey

DATA. OSU Libraries' Research Data Services acknowledges that "[m]any variables govern what constitutes 'data' and the management of data, and each discipline has its own culture regarding data." Examples of data could include sensor readings, laboratory or field observations, physical samples, novel processing scripts or code, databases, interview transcripts, and documentation (metadata). Thus, where it appears in this survey, "data" may be referred to simply as "information generated during your research process" with respect to your research discipline.

If you have other views regarding ways in which you or the particular research community to which you belong defines "data," please briefly share them below.

[TEXT BOX]

Q1. Please select your current position and/or rank.

- ❍ Professor (including Extension, Senior Research and Clinical)
- ❍ Associate Professor (including Extension, Senior Research and Clinical)
- ❍ Assistant Professor (including Extension, Senior Research and Clinical)
- ❍ Faculty Research Assistant
- ❍ Senior Faculty Research Assistant
- ❍ Research Associate
- ❍ Post Doc (all types, including Research Associate, Fellow, etc.)
- ❍ Other (Specify): _____

Q2. Please select your primary unit of affiliation. NOTE: if your primary work unit is in a research center, institute, or with the Extension Service or an Agricultural Experiment Station, please select the college that most closely represents the type of research that you do.

- ❍ College of Agricultural Sciences
- ❍ College of Business
- ❍ College of Earth, Ocean and Atmospheric Sciences
- ❍ College of Education
- ❍ College of Engineering
- ❍ College of Forestry
- ❍ College of Liberal Arts
- ❍ College of Pharmacy
- ❍ College of Public Health and Human Sciences
- ❍ College of Science
- ❍ College of Veterinary Medicine
- ❍ University Libraries
- ❍ Other (Specify): _____

Q3. Please indicate whether or not you generate each of the following data format(s) as part of your research process. If you are unsure, please select "Other" and provide a brief description of your data format(s). Select Yes or No for each.

	Yes	No
Quantitative, tabular data (e.g. spreadsheets, delimited text, SPSS, XML)	❍	❍
Geospatial data (e.g. vector and raster data, shapefiles, geodatabases)	❍	❍
Digital databases (e.g. surveys, census data, government statistics)	❍	❍
Digital text for quantitative research (e.g. software scripts and codes, descriptive information/metadata)	❍	❍
Digital text for qualitative research (e.g. survey transcripts, code books, metadata)	❍	❍
Electronic laboratory notebook	❍	❍
Digital images or digital copies of images	❍	❍
Biological, organic, or inorganic samples or specimens	❍	❍
Digital gene sequences or similar digital renditions of biological, organic, or inorganic samples or specimens	❍	❍
Audio recordings	❍	❍
Video recordings	❍	❍
Non-digital text (e.g. hand-written notes, sketches, figures, paper laboratory notebooks)	❍	❍
Non-digital images	❍	❍
Artistic products (14)	❍	❍
Other (Specify): (15)	❍	❍

Q4. What is your opinion about the ownership of original research data produced in the course of research conducted at, or under the auspices of, OSU (i.e., who owns the data)?

- ❍ The researcher owns the data
- ❍ The University owns the data
- ❍ The publisher of any research based on the data owns the data
- ❍ The funding agency owns the data
- ❍ The public owns the data
- ❍ Other (Specify): _____
- ❍ No opinion

NOTE: For the remainder of the survey, any mention of "data" refers specifically to digital data, or data recorded or stored electronically as part of OSU research. Data management for research involves the organization, cleaning, analysis, storage, sharing, disposal, and/or preservation of research data at or under the auspices of OSU.

Q5. Do you engage in research activities at OSU that include digital data management?

- ❍ Yes
- ❍ No

If No Is Selected, Then Skip To End of Survey

Q6. Who performs the majority of each of the following digital data management tasks associated with your research?

	PI/Director	Graduate Students	Researchers/ Research Staff	IT Staff	Other	Not Applicable
Data collection	❍	❍	❍	❍	❍	❍
Data documentation (metadata)	❍	❍	❍	❍	❍	❍
Data cleaning (QA/QC)	❍	❍	❍	❍	❍	❍
Backing up data	❍	❍	❍	❍	❍	❍
Data analysis	❍	❍	❍	❍	❍	❍
Data storage and organization	❍	❍	❍	❍	❍	❍
Data sharing outside of your research group	❍	❍	❍	❍	❍	❍
Data archive or long-term storage	❍	❍	❍	❍	❍	❍
Data disposal/destruction	❍	❍	❍	❍	❍	❍

Q7. Please indicate which of the following people or groups typically provide technical support or assistance with your OSU research data. Select Yes or No for each.

	Yes	No
Personnel within my research group	❍	❍
Designated computer support or research computing personnel in my college, department, school or unit (e.g. COSINe in the College of Science)	❍	❍
Designated computer support or research computing personnel at a center or institute where I have a formal affiliation (e.g. LPI, HMSC, CGRB)	❍	❍
OSU Information Services (IS) or Central Networking (CN)	❍	❍
University Libraries	❍	❍
Personnel at a non-OSU organization	❍	❍
Other (Specify):	❍	❍

Q8. How easy or difficult is each of the following activities with regard to managing your OSU research data?

	Easy	Somewhat Easy	Neutral	Somewhat Difficult	Difficult	Not Applicable
Storing data short-term (5 years or less)	○	○	○	○	○	○
Storing data long-term (more than 5 years)	○	○	○	○	○	○
Backing up data	○	○	○	○	○	○
Analyzing/manipulating data	○	○	○	○	○	○
Finding my own data	○	○	○	○	○	○
Finding data produced by other researchers	○	○	○	○	○	○
Accessing data produced by other researchers	○	○	○	○	○	○
Ensuring data are secure	○	○	○	○	○	○
Ensuring data are trustworthy	○	○	○	○	○	○
Making data accessible to other researchers	○	○	○	○	○	○
Controlling access to data	○	○	○	○	○	○
Tracking updates to data (i.e., versioning)	○	○	○	○	○	○
Describing the data to be more usable at a later time or by others (i.e., creating metadata, code books, etc.)	○	○	○	○	○	○
Protecting intellectual property rights	○	○	○	○	○	○
Ensuring appropriate professional credit/citation is given to data sets I generated	○	○	○	○	○	○
Other (Specify):	○	○	○	○	○	○

Q9. How often do you generate original data as part of your research?

○   Always
○   Most of the time
○   Sometimes
○   Rarely
○   Never

Q10. How often does your research include the analysis of data collected by others (also referred to as secondary data)?

○   Always
○   Most of the time
○   Sometimes
○   Rarely
○   Never

If Never Is Selected, Then Skip To What has been the typical amount of d...

Q11. From which of the following types of sources do you acquire secondary data? Select all that apply.

❑   Public data sources
❑   Non-public/restricted data sources
❑   Directly from other researchers
❑   Other (Specify): _____

Q12. What has been the typical amount of digital data for a single project you have worked on in the past 5 years?

❍ 1 GB (gigabyte) or less
❍ More than 1 GB but less than 100 GB
❍ More than 100 GB but less than 1 TB (terabyte)
❍ More than 1 TB but less than 100 TB
❍ More than 100 TB but less than 1 PB (petabyte)
❍ More than 1 PB
❍ I don't know

Q13. What is the largest amount of digital data for a single project you have worked on in the past 5 years?

❍ 1 GB (gigabyte) or less
❍ More than 1 GB but less than 100 GB
❍ More than 100 GB but less than 1 TB (terabyte)
❍ More than 1 TB but less than 100 TB
❍ More than 100 TB but less than 1 PB (petabyte)
❍ More than 1 PB
❍ I don't know

Q14. Thinking about the data you've generated within the last five years (short-term data), please indicate where you store and/or backup these data. Select Yes or No for each.

	Yes	No	I don't know
Desktop or laptop computer hard drive	❍	❍	❍
External hard drive or media (USB, disk, magnetic tape, etc.)	❍	❍	❍
Local server (within your research group)	❍	❍	❍
College, department, school, institute or unit network server	❍	❍	❍
Central campus network server (IS/CN)	❍	❍	❍
Cloud or web-based storage service	❍	❍	❍
Other (Specify):	❍	❍	❍

Q15. Do you generate metadata? For example, do you currently document or describe your data, create code books, data dictionaries, "README" files, etc.?

❍ Yes
❍ No
If No Is Selected, Then Skip To End of Block

Q16. Please indicate which metadata standard you currently use to describe your data. Click Yes or No for each.

	Yes	No	I don't know
DC (Dublin Care)	❍	❍	❍
DwC (Darwin Core)	❍	❍	❍
DDI (Data Documentation Initiative)	❍	❍	❍
DIF (Directory Interchange Format)	❍	❍	❍
EML (Ecological Metadata Language)	❍	❍	❍
FGDC (Federal Geographic Data Committee)	❍	❍	❍
ISO 19115 (Geographic Information - Metadata)	❍	❍	❍
OGIS (Open GIS)	❍	❍	❍
Metadata standardized within my lab	❍	❍	❍
Other (Specify):	❍	❍	❍

Q17. How often do you share your data with others (outside of your research team) using the following methods?

	Never	Rarely	Sometimes	Often	Always	I don't know
Data are made publicly available on the principal investigator's website, or a project-specific website	❍	❍	❍	❍	❍	❍
Data are made available through my organization's website (college, department, school, unit, institute, etc.)	❍	❍	❍	❍	❍	❍
Data are made available through OSU's digital repository, ScholarsArchive@OSU	❍	❍	❍	❍	❍	❍
Data are made publicly available through a regional, national or global network or repository, including agency repositories (e.g. LTER, ICPSR, NODC, PDB, GenBank, DataONE)	❍	❍	❍	❍	❍	❍
Data are made publicly available on a third-party platform (e.g. Dryad, figshare, Pangaea)	❍	❍	❍	❍	❍	❍
Data are not made publicly available, but I respond to individual requests from others	❍	❍	❍	❍	❍	❍
Data are not made publicly available beyond the members of the research team	❍	❍	❍	❍	❍	❍
Other (Specify):	❍	❍	❍	❍	❍	❍

Q18. Please indicate how much each of the following limits the sharing of your data outside of your research team?

	Does Not Limit Sharing	Limits Sharing Somewhat	Very Much Limits Sharing	Not Applicable
Ability to maintain confidentiality of research participants	❍	❍	❍	❍
A lack of funding	❍	❍	❍	❍
A lack of standards (e.g. data or metadata format)	❍	❍	❍	❍
An opinion that people don't need the data	❍	❍	❍	❍
An insufficient time to make data available	❍	❍	❍	❍
A lack of mechanism to share the data	❍	❍	❍	❍
An opinion that the data shouldn't be made available	❍	❍	❍	❍
A sponsor requirement regarding sharing	❍	❍	❍	❍
A lack of rights to make the data public	❍	❍	❍	❍
Ability to gain appropriate intellectual property rights protection	❍	❍	❍	❍
The potential for data to be misinterpreted or misused by others	❍	❍	❍	❍
A lack of mechanism to receive citation or credit once data are public	❍	❍	❍	❍
Other (Specify):	❍	❍	❍	❍

Q19. Please indicate which of the following strategies you use to protect your data from corruption or loss. Click Yes or No for each.

	Yes	No	I don't know
I make multiple copies in multiple locations (onsite + offsite copies)	○	○	○
Copies of datasets are saved on a disk, USB drive, tape, or computer hard drive	○	○	○
Copies of datasets are saved on a local server (within your research group)	○	○	○
Copies of datasets are saved on a college, department, institute or center server	○	○	○
Copies of datasets are saved on a central campus server (IS/CN)	○	○	○
Copies of datasets are saved on a web-based or cloud server	○	○	○
Copies of datasets are stored in a data repository or archives	○	○	○
Backup files are automatically generated	○	○	○
Backup files are manually generated	○	○	○
Access to files is restricted	○	○	○
Other (Specify):	○	○	○

Q20. Please indicate what typically happens to your data after the research project has concluded. Click Yes or No for each.

	Yes	No	I don't know
The data are destroyed	○	○	○
The data are stored in a discipline/domain-specific public data repository or archives (e.g. LTER, ICPSR, NODC, PDB, GenBank, DataONE)	○	○	○
The data are stored in an institution-based public repository (e.g. ScholarsArchive@OSU) or archives	○	○	○
The data are stored with the publisher of the research based on the data	○	○	○
The data are stored on a disk, USB drive, tape or computer hard drive	○	○	○
The data are stored on a local server (within your research group)	○	○	○
The data are stored on a college, department, institute or center server	○	○	○
The data are stored on a central campus server (IS/CN)	○	○	○
Other (Specify):	○	○	○

Answer If Whattypically happens to your data after research projec... The data are stored in a discipline/domain-specific public data repository or archives (e.g. LTER, ICPSR, NODC, PDB, GenBank, DataONE) - Yes Is Selected

Q21. Please enter the name(s) of the discipline/domain-specific repository or archives that store(s) your data.

[TEXT BOX]

Q22. Please indicate which of the following services for the OSU community that you are aware of.

	Aware	Not Aware
Advanced computing options (e.g., distributed, network-, or cluster-computing; supercomputer-class machines; data visualization)	❍	❍
Statistical and other data analysis support	❍	❍
Short-term data storage/preservation (5 years or less)	❍	❍
Long-term data storage/preservation (more than 5 years)	❍	❍
Acquiring unique identifiers for datasets (e.g. DOI, ARK)	❍	❍
Data security support	❍	❍
Guidance on depositing data into repositories or archives	❍	❍
Guidance on how to use appropriate metadata standards	❍	❍
Guidance on writing a data management plan	❍	❍

Q23. How important do you think it is for OSU to spend resources on providing the following services?

	Not at all Important	Somewhat Important	Moderately Important	Very Important
Advanced computing options (e.g.,distributed, network-, or cluster-computing; supercomputer-class machines; data visualization support)	❍	❍	❍	❍
Statistical and other data analysis support	❍	❍	❍	❍
Short-term data storage/preservation (5 years or less)	❍	❍	❍	❍
Long-term data storage/preservation (more than 5 years)	❍	❍	❍	❍
Data security support	❍	❍	❍	❍
Guidance on depositing data into repositories or archives	❍	❍	❍	❍
Guidance on how to use appropriate metadata standards	❍	❍	❍	❍
Guidance on writing a data management plan	❍	❍	❍	❍
Other (Specify):	❍	❍	❍	❍

Q24. Do you archive or store datasets for 5 years or more?

❍ Yes
❍ No

If No Is Selected, Then Skip To In your opinion, where should the fun...

Q25. Thinking about long-term data storage (5 years or more), are adequate funding and resources available to your department, center, lab, or research group to support this?

❍ Yes
❍ No
❍ I don't know

Q26. Thinking about long-term data storage (5 years or more), from which of the following sources have you or your research group received funding to support such storage? Select all that apply.

❑ External funds
❑ University funds
❑ Other (Specify): _____
❑ I/we have not received funds for long-term data storage
❑ I don't know

Q27. In your opinion, where should the funding come from to cover the costs of data management (organization, cleaning, analysis, storage, sharing, disposal, and/or preservation) for research SUPPORTED by grants, contracts, or other external sources of funding? Select all that apply.

❑   It should be folded into the direct costs of those grants and contracts by the individual researcher or research team
❑   It should be paid for by the University from overhead/F&A funds it receives from grants and contracts
❑   It should be paid for by the University from other sources of funds
❑   Other (Specify): _____
❑   No opinion

Q28. In your opinion, where should the funding come from to cover the costs of data management (organization, cleaning, analysis, storage, sharing, disposal, and/or preservation) for research NOT SUPPORTED by grants, contracts, or other external sources of funding? Select all that apply.

❑   It should be paid for by the University from overhead/F&A funds it receives from grants and contracts
❑   It should be paid for by the University from other sources of funds
❑   It should be paid for by the individual researcher or research team with other funds
❑   Other (Specify): _____
❑   No opinion

Q29. Please provide any thoughts or comments related to research data management, including suggestions for services or support that might improve the quality or effectiveness of your research process or data stewardship activities.

[TEXT BOX]

END. You have reached the end of the survey.

By clicking on the advance button [>>] below, your survey will be finalized, and you will NOT be able to return to the survey to review or edit your responses. If you would like to review or edit your responses at a later time, please close your browser window now. You can return to your survey at any time using the survey link you received in the email invitation.

<END OF SURVEY>

End of survey flow >> redirect to the OSU Libraries Center for Digital Scholarship & Services' Research Data Services web page: http://guides.library.oregonstate.edu/research-data-services

Corresponding author
Dr Amanda L. Whitmire can be contacted at: amanda.whitmire@oregonstate.edu