9-1-2015

Boise State Data Management Needs Report

Michelle Armstrong
Boise State University

Megan Davis
Boise State University

Margie Ruppel
Boise State University

EXECUTIVE SUMMARY

To ensure compliance and to achieve greater value for the research sponsored by the University and outside funders, researchers need to take an active part in the management of the data they produce. Albertsons Library, in coordination with University units such as the Office of Sponsored Programs and the Office of Information Technology, is well-positioned to play an integral role in educating researchers and providing support on issues related to data management.

Albertsons Library established a working group in early 2014 dedicated to learning about data management issues and becoming involved in the University's research data management activities. In order to better serve the University community in this regard, the Albertsons Library Core Data Management Team worked with librarian liaisons to interview campus researchers on their current knowledge of and practices related to data management. Responses were grouped into three major topic areas for analysis: Data Characteristics, Data Storage, and Data Discoverability and Access. Researchers reported a variety of practices, some demonstrating effective management of their researcher data, while others described significant gaps in their current practices.

Based on this review, eleven recommendations were developed.

- Educate researchers on the why, how, and benefits of sharing data.
- Regardless of discipline, encourage researchers to utilize metadata schema and data description standards throughout the research lifecycle.
- Provide basic descriptive metadata templates that can be incorporated into research processes or through data gathering instruments.
- Establish university-wide metadata services sufficient to provide both consultation and direct support services.
- Provide FAQs and other informational materials explaining data ownership policies.
- Partner with other campus grant/data management groups to ensure consistent language and guidance is provided to researchers.
- Educate researchers about proprietary hardware and software limitations and methods for ensuring long-term data access.
- Increase researcher awareness of best practices regarding data storage and backup procedures, including availability and proper use of OIT storage space.
- Promote better utilization of additional OIT research services, including consultations.
- Educate researchers about effective practices for naming and organizing files, version control, documentation, and other data management tools.
- Make researchers aware of safe and ethical methods for sharing de-identified information and provide support to transform sensitive data into public-use datasets.

BACKGROUND

As a growing metropolitan research university, Boise State University faculty members, researchers, and staff are generating ever-growing amounts of data. Management of this data is becoming increasingly challenging, at both a technical level and a policy level (Pinfield, Cox, & Smith, 2014). Whyte and Teds of the Digital Curation Centre define research data management as "the organisation of data, from its entry to the research cycle through to the dissemination and

archiving of valuable results." (2011). At an institutional level, this may include issues and activities related to data storage, security, preservation, compliance, quality, sharing, and jurisdiction (Pinfield, Cox, & Smith, 2014).

While major research funders like the National Science Foundation and the National Institutes of Health require applicants to provide a plan to manage their data, they provide little support to those not accustomed to organizing, storing, and preserving data in a systematic way. Universities are therefore responsible for coordinating efforts through policy development and the evolving research infrastructure. As academic libraries have long played an expert role in both the preservation of documents and access to the published literature, they are well-positioned to expand into the data management arena. In 2012, the academic library community identified data curation as one of the top trends (Tenopir, Birch, & Allard). In addition, a survey conducted in 2013 by the Association of Research Libraries of their members identified two emerging service areas: assisting with data management planning at the grant proposal stage and providing support for the archiving of research data (Fearon, Jr., Gunia, Pralle, Lake, & Sallans, p. 11).

Albertsons Library established a working group in early 2014 dedicated to learning about data management issues and becoming involved in the University's research data management activities. In order to better serve the University community in this regard, the Albertsons Library Core Data Management Team interviewed campus researchers on their current knowledge of and practices related to data management.

METHODOLOGY

The Core Data Management Team established three primary objectives for the interviews with campus researchers:

- Develop relationships with faculty regarding their data management activities
- Document current processes for collecting, storing, and sharing data
- Identify gaps in data management services or supports

Based on Purdue University's "Conducting a Data Interview" worksheet (Witt & Carlson, 2007) and the Australian National Data Service's "Research Data Interviews" methodology (n.d.), the Library's Core Data Management Team created the following interview protocol.

Faculty library liaisons were responsible for conducting the interviews which involved three basic steps: Prepare, Interview, Share. The Core Team provided instructions for each of these steps during a data management workshop and through a video explaining the process. Support materials including background research worksheets, sample email invitations, and interview questions were also provided *(See Appendix A).* During the interviews, the Core Team provided key support and facilitation by serving as interview partners and data management experts for more advanced questions, coordinating interviews, and compiling interview notes.

From June to October 2014, library liaisons conducted 30 interviews with faculty members, researchers, and staff from 25 different departments. Interviewees were selected by the liaisons according to topics of interest and the availability of researchers' schedules. Interviews ranged from

approximately 30 minutes to an hour. Once the interview was completed, the liaison and interview partner compiled and shared their notes with the Core Team.

The Core Team grouped the responses into three major topic areas: Data Characteristics, Data Storage, and Data Discoverability and Access. Interviewees were also given the opportunity to discuss other research needs and concerns. Once compiled, each member of the Core Team was assigned a topic area and summarized the responses, identifying common responses and providing numerical tallies where appropriate. These summaries were then discussed among the Core Team members and further refined. The final summaries were analyzed for trends in current data management practices, comparison of current practices versus best practices, and implications for Boise State if current practices continue. Based on this analysis, the Core Team developed a list of recommendations for the university.

DATA CHARACTERISTICS

Types of research and data

Boise State University researchers conduct original research using quantitative, qualitative, and mixed-methods. Analysis of the interview responses show this research falls into three general categories: research which requires specific instruments to collect data, human subjects research, and data for secondary analysis. The research and resulting data collected vary greatly.

Form and format

Boise State University researchers use a plethora of applications to create, store, and analyze their data, which results in many forms. In some cases, two or three software programs are used for different actions on the same set of data. The data is numerical, statistical, textual, audio, images, photographs, and sometimes paper. Other researchers mentioned data formats such as video, MyBoiseState data, business analytics data, MSA (metropolitan statistical area) data, and market research data. The most common types of applications researchers use to store and analyze this data are spreadsheets, statistical software, word processing, survey software, database software, and satellite navigation systems.

Given the variety of data forms and formats used, as well as software applications, Boise State researchers should consider using stable, non-proprietary software and hardware in order to make the data accessible to future researchers. The *DataONE Best Practices Primer* states, "File formats should ideally be non-proprietary (e.g. .txt or .csv files rather than .docx or .xls), so that they are stable and can be read well into the future. Consider the longevity of hardware when backing up data" (p. 5). If stable, non-proprietary software and hardware formats are not used by Boise State researchers, they risk its future usability because not all users have access to proprietary software such as Microsoft Office or the specific hardware needed to access the data.

Organization and management

The *DataONE Best Practices Primer* lists two main reasons for researchers to establish a data management plan at the beginning of their research project: the researcher will spend less time on

data management and more time on research, and it will be easier for the researchers, collaborators, and future users to find, use, and analyze the data (Strasser, et al, n.d.). To explore this, Boise State researchers were asked about their organization and management practices. In response, Boise State researchers provided examples of how they organize and manage their data during the data collection phase. These questions included naming conventions, organization and management, and file version control.

Date, topic, researcher initials, notebook numbers, and version numbers are the most common file naming conventions the researchers mentioned. Two researchers added that they use lab notebooks to keep a log of each file name. Researchers also discussed how they organized the individual files such as by project, student/grade/school/date, and folders within data sets. For example, one researcher stated that each project gets an individual folder and related "Read Me" file as he progresses through his research project which are then organized using ArcCatalog. Another researcher lamented the fact that she often has to reorganize files initially set up by graduate students who are not as skilled in these tasks.

Several researchers mentioned working with graduate assistants to support the overall management of the research project. Students were described as helping update project manuals, creating and maintaining relational databases, creating additional datasets from the researcher's original data, and contributing files to shared Google drives or a team Dropbox account.

Another important topic highlighted during the interviews were methods for maintaining version control. While the Boise State researchers interviewed mentioned some good practices in this category, a few researchers described problematic habits such as the lack of any formal version control. One researcher said she recognizes the need for formal version control, but doesn't do it. To enforce version control, other researchers listed assigning one part of the data to each team member and using Dropbox for "idiot level version control." If researchers do not formally plan for version control, they risk losing data, as well as valuable research time.

Sensitive information

A majority of the Boise State researchers we interviewed were conducting research that included sensitive information, such as names, grades, mental health diagnoses, email addresses, patent issues, HIV status, and locations of sacred Native American sites. While the researchers detailed their methods for de-identifying sensitive information, it is possible that some researchers and grantors may attach certain restrictions on public access to the raw data. Researchers need to be made aware of safe and ethical methods for sharing de-identified information.

One goal of effective data management practices is to make the raw data available for future researchers in a data repository or other publicly-accessible location. Best practices are in place to "design public-use datasets that maintain the confidentiality of respondents and are of maximum utility for all users" (*ICPSR Guide to Social Science Data Preparation and Archiving*, p. 39). Boise State researchers were not asked to give an opinion on making sensitive data available to other researchers, but given the amount of human subjects research occurring on campus, we recommend that they be made aware of how their data can be transformed into public-use data sets.

DATA STORAGE

Expected data set lifespan

When asked about the length of time researchers needed to store their data, responses varied from "not sure" to less than 3 years to forever. For those who did specify particular time periods for storage, reasons included requirements from funding agencies and personal research goals including long-term data comparisons. Although only two of the respondents stated that data storage periods were dependent upon grant requirements, it is likely that an increasing number of grant funding agencies will require a minimum time period for data retention that researchers will need to prepare for in their data management plan. Whether the expected storage length of a project is short or long, planning for storage is a crucial part of the data management process.

Without taking these issues into account during the planning process, data can fall through the cracks due to outdated technology or personnel issues. It might also be noted that not every piece of data may need to be included when research is being preserved for a significant length of time. According to DataONE, it is critical to consider the entire research process when determining which data products should be preserved, as well as the cost of preserving that particular type of data (DataONE, n.d.a).

Current data storage practices

Data storage locations for researchers included desktop and laptop hard drives, external hard drives, shared department or college drives and servers, USB drives, cloud storage applications, paper hard copies, software applications, and university-wide services like Globus and the server provided by Boise State's Office of Information Technology (OIT). The preferred locations were desktop and laptop hard drives, shared department or college drives and servers, and cloud stor-age options.

More than half of the respondents stated that they stored data in multiple places, yet it was not always clear whether this meant that the same data was backed up in alternate locations or that unique data files were stored in different places. Without clarity on this issue, it is difficult to gauge the security of Boise State research data. Documenting where each file is located and on which type of storage device will greatly assist all members of the research team who might need to access the data.

One possible secure storage location is freely available to researchers through OIT, yet few of those interviewed mentioned using this space. Wherever and however data is stored, security related to the method of storage is of paramount importance. Boise State researchers should consider where desktop and laptop computers are kept and who has access to them as well as the advantages and disadvantages of using local and/or cloud storage.

DataONE's best practices site recommends creating a backup policy that applies to data being collected and/or created while the project is ongoing (DataONE, n.d.a). Boise State researchers should create such a policy, including information about who is responsible for data storage and backups, and where and when backups take place. This policy should be reviewed periodically and updated when necessary. Backups should be automated when possible, but checked in person on a

regular basis. At the very least, several copies of the data should be available in non-proprietary, standard formats in multiple places, preferably including one off-site location.

Data set size and growth

A reasonable estimation of the size of a researcher's data set should be done at the beginning of the research process so plans can be made for appropriate storage methods and locations. Very few Boise State researchers interviewed had a clear idea of the size of their files and data sets. When asked about file size, many seemed to focus on the number of records in a data set or the number of study subjects involved in the project. Those who were more unambiguous in their responses had files ranging from a few megabytes to multiple terabytes of data. It is important to recognize that different types of files will likely have unique storage requirements. As stated above in the summary of data characteristics, a majority of the file types from these specific researchers are numerical and textual, which trend toward smaller file sizes. Images, audio, and video files are likely to require much more storage space.

Physical data storage will have space limitations as will cloud storage in many cases. Physical storage methods (desktop and laptop computers, external hard drives, and departmental and university servers) all require space in an office, lab, or IT server. If increased physical storage is needed during or after the project, that requires additional space and money. Although a growing number of researchers envision cloud storage as a solution to physical space and fiscal obstacles, that is not always the case. Not all cloud storage providers behave the same way in regards to storage space and security. For example, cloud storage may be free only up to a certain total amount (e.g., 5 GB) before users are charged for additional storage space. Critically, space for backups as well as original data files must be considered in storage and security planning regardless of where data is stored.

DATA DISCOVERABILITY AND ACCESS

Use of metadata

Metadata is often referred to as data about data. It describes the data and provides valuable context needed for the future use of that data. According to the *DataONE Best Practices Primer* (Strasser, et al, p. 5.), "Without a thorough description of the context of the data file, the context in which the data were collected, the measurements that were made, and the quality of the data, it is unlikely that the data can be easily discovered, understood, or effectively used."

Despite the importance of metadata, there was generally a lack of awareness or clear understanding among the researchers when responding to questions about metadata. Researchers stated that they didn't fully understand what was meant by the term "metadata" or that the idea of a discipline-specific metadata schema was not common in their field, indicating that there were no standards or they were not sure if any had been established.

The few researchers who provided specific answers to the questions about metadata had either received grants from federal funders requiring data management plans which included sections on metadata, or they were in a field that commonly used established metadata standards. When asked

whether or not they needed assistance with creating metadata, responses were fairly evenly split between "yes" and "no". A few of the faculty who said they did need support also noted that it would be valuable to teach their students about metadata.

As previously noted, this lack of attention regarding metadata has implications for long-term usability of Boise State's data. Metadata helps current researchers manage their data during the lifecycle of a research project and assists in ensuring quality control of that data. As more of Boise State's data is deposited into external repositories, application of appropriate metadata schema will increase visibility of that research.

Discoverability and potential uses of the data

In addition to secondary research, making one's data both discoverable and accessible allows for the replication and verification of that research. As large, institutional or discipline-specific repositories develop, the ability to discover, use, and obtain additional gains from the data will become increasingly easier. Additionally, techniques to cite and apply permanent identifiers to datasets are evolving. These changes are providing opportunities for use and recognition of one's research that have not previously existed.

In contrast to the above benefits, when asked about potential uses of the data being collected, most researchers provided only a few potential ways their data could be utilized. Interviewees stated that the data they collected could be used for publishing, training students, conducting secondary research, or in direct application by practitioners or other members of the community. Of these various uses, secondary research was the most frequently mentioned type of use and included both further study by the interviewee or other researchers in their discipline. Several researchers were either not sure how their data could be useful or believed that they would be the only ones who could utilize it.

When asked about making their data discoverable and accessible, a large portion of the respondents mentioned releasing their data through their publishing activities. However, unless a paper is a formal data paper or the publisher has required authors to share the supporting data, this approach is not typically considered a form of data sharing and would not meet most funder's requirements. Another large portion of respondents stated that they either were not going to share their data or were not sure how they would do so. Only one researcher stated that they regularly submit their data to a specific, external archive.

Researchers were also asked if they knew whether their discipline used a particular repository. Although several researchers knew of and had worked with specific repositories or archiving services, nine researchers said there were no repositories for their specific research or discipline, and nine researchers said they weren't sure or didn't know of any services. Of those who were aware of specific repositories, only three indicated during the interview that they used the repository for their data. Because so few researchers had worked with specific repositories, there was no significant discussion regarding the cost of data repositories.

Overall, the Boise State researchers interviewed did not emphasize data sharing and discovery as a primary research activity. This perspective is in contrast to the data sharing policies established by

federal funders and the known benefits of making data appropriately accessible. By becoming familiar with methods and services for permanent repository deposit, Boise State researchers can take advantage of these benefits.

Data ownership and related policies

When asked about ownership of the data, respondents stated that they own their data or the entity/group with which they are working with owns the data. A few respondents indicated that Boise State is the owner of the data being produced. When asked clarifying follow-up questions, several of the respondents changed their answers and stated that they were unsure who actually owns the data being produced during their research.

As the emphasis on data as a research asset continues to increase, issues of ownership and intellectual property will also increase. Obligations to funders or a researcher's affiliate universities, either for the Boise State researcher or their off-campus colleagues, can influence how the data is actively managed during the project. Requirements may include how it is handled once the project is completed, who can access that data, and even how data resulting from secondary analysis can be used or disseminated. Based on the responses given and the on-going development of data ownership and intellectual property policies, there is a need for awareness in this specific area.

Data management plans

As discussed in the previous "Organization and Management" section, utilization of data management plans allows researchers to spend more time on their actual research. To understand current practices in this area, researchers were asked about the policies they referred to when creating data management plans, resources they would find useful when creating a data management plan, and if they had used the DMPTool (www.dmptool.org).

Responses to the question of data management policies used fell into four groups: IRB policies, funding agency policies, another organization's policies, or no specific policy at all. Several researchers saw an overlap between IRB requirements and data management plans required by funding agencies. When asked about useful resources for data management planning, researchers listed templates and direct support, such as individual consultations.

The majority of researchers interviewed had never heard of the DMPTool and of the few who had, they had never used it to create a data management plan. As one of the data management planning resources available to Boise State researchers, the DMPTool provides templates, including relevant questions and prompts for each section, designed to comply with federal and foundation funder policies.

As more and more federal funders require data management plans, Boise State researchers are expected to need appropriate support materials. Additionally, to capitalize on the benefits of spending more time actually researching, researchers are encouraged to think of data management as a practice that occurs throughout the lifecycle of the project.

CONCLUSION

In order to achieve greater value for the research sponsored by the University and outside funders, researchers need to take an active part in the management of the data they produce. Not only does the management of data need to happen, but it needs to happen at the right time "to enable the survival of the data, and maximize the initial investment made in its creation or collection" (Pryor 2012, p. 18). In contrast, poor data management can lead to loss of research data, retraction of published scholarly works, and denial of future funding. In many institutions, the university library is seen as an important partner in the management of this data. Taking on this partner role can be seen as a "natural extension" of the library's mission to provide access to research data once it has been published in the scholarly literature (Lewis 2010). Albertsons Library's Core Data Management Team, in coordination with University units like the Division of Research, the Office of Sponsored Programs, and the Office of Information Technology, is well-positioned to play an integral role in educating researchers and providing support on issues related to data management. It is in the best interest of everyone at the University to ensure that data produced at the institution is made discoverable, accessible, and reusable long into the future.

RECOMMENDATIONS

Based on this analysis the following recommendations are provided for Boise State University.

- Educate researchers on the why, how, and benefits of sharing data.
- Regardless of discipline, encourage researchers to utilize metadata schema and data description standards throughout the research lifecycle.
- Provide basic descriptive metadata templates that can be incorporated into research processes or through data gathering instruments.
- Establish university-wide metadata services sufficient to provide both consultation and direct support services.
- Provide FAQs and other informational materials explaining data ownership policies.
- Partner with other campus grant/data management groups to ensure consistent language and guidance is provided to researchers.
- Educate researchers about proprietary hardware and software limitations and methods for ensuring long-term data access.
- Increase researcher awareness of best practices regarding data storage and backup procedures, including availability and proper use of OIT storage space.
- Promote better utilization of additional OIT research services, including consultations.
- Educate researchers about effective practices for naming and organizing files, version control, documentation, and other data management tools.
- Make researchers aware of safe and ethical methods for sharing de-identified information and provide support to transform sensitive data into public-use datasets.

REFERENCES

Australia National Data Service (n.d.) *Research data interviews*. Retrieved from http://ands.org.au/datamanagement/data-interviews.html

DataONE. (n.d.a). Create and document a backup data policy. *DataONE: Best Practices.* Retrieved from: https://www.dataone.org/best-practices/create-and-document-data-backup-policy

DataONE. (n.d.b). Decide what data to preserve. *DataONE: Best Practices.* Retrieved from: https://www.dataone.org/best-practices/decide-what-data-preserve

Fearon, Jr., D., Gunia, B., Pralle, B.E., Lake, S., & Sallans, A.L. (2013). *Research data management services.* Washington, DC: Association of Research Libraries.

Lewis, M. (2010). Libraries and the management of research data. In S. McKnight (Ed.), *Envisioning future academic library services: Initiatives, ideas and challenges* (pp. 145-168). London: Facet Publishing.

Pinfield, S., Cox, A.M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS One*, *9*(12), 1-28. http://dx.doi.org/10.1371/journal.pone.0114734

Pryor, G. (2012). *Managing research data*. London: Facet Publishing.

Strasser, C., Cook, R., Michener, W. & Budden, A. (n.d.). *DataONE primer on data management: what you always wanted to know**. Retrieved from https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Tenopir, C., Birch, B., & Allard, S. (2012). *Academic libraries and research data services: Current practices and plans for the future, an ACRL white paper*. Retrieved from: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

Whyte, A.,& Tedds, J. (2011) *Making the case for research data management*. Edinburgh: Digital Curation Centre. Available: http://www.dcc.ac.uk/webfm_send/487.

Witt, M. & Carlson, J. (2007). *Conducting a data interview*. Retrieved from http://docs.lib.purdue.edu/lib_research/81

APPENDIX A - DATA MANAGEMENT NEEDS INTERVIEW QUESTIONS

The purpose of this interview is to help the Library, Office of Information Technology and the Office of Sponsored Programs document your current process for collecting, storing and sharing data and to identify gaps in service or support that would assist you with your data management needs. The information you share will be compiled with other interview notes and summarized in a final report which will be released to the university.

Interviewee Name:

- Interviewee Department:
- Date of interview:

Research Overview:

- Would you tell us about your research?

Data Characteristics:

- What data do you collect (nature, scope, scale)?
- What form and format are the data in?
- How will the data be organized (version control, naming conventions, etc.)
- Does the dataset include any sensitive information?

Data Storage:

- What is the expected lifespan of the dataset?
- Where do you currently store your data?
- How large is the dataset, and what is its rate of growth?

Data Discoverability and Access:

- Are there domain-specific metadata standards for your data?
- Do you need assistance with metadata development?
- How could the data be used, reused and repurposed?
- Who are the potential audiences for the data?
- Who owns the data?
- How will your data be made accessible?
- How will your data be made discoverable?
- Does your discipline use a particular repository?
- What is the cost of utilizing the repository?
- How are you currently funding the cost?

Institutional Level Information: Needs and Resources

- What are your particular pain points with regard to data management?
- What policies do you refer to when developing your data management plan?(i.e., IP policy)
- What resources would you find most helpful when developing a data management plan?
- Have you used the DMPTool to create a data management plan? If so, what has your experience been like?

Other comments/suggestions: