An Integrated Data Management Plan Instructional Program

William H. Mischo, University of Illinois, Urbana-Champaign

William Mischo is Head, Grainger Engineering Library Information Center and Professor, University Library at the University of Illinois at Urbana Champaign (UIUC). He has been a Principal Investigator on a number of digital library grants from the National Science Foundation (NSF), including the National Ethics Portal grant, several National Science Digital Library (NSDL) grants, and the Digital Library Initiative I grant. He has also received an Institute of Museum and Library Services (IMLS) National Leadership Grant, and several Andrew Mellon Foundation grants. Bill has published some 70 articles and conference papers in the field of library and information science and has presented at more than 75 national and international conferences, including at ALA, SLA, the NSDL Annual meeting, Internet Librarian International, LITA National, and ASEE annuals. He served on the NSDL Policy Committee from 2003 to 2006. In 2001, Bill received the Homer I. Bernhardt Distinguished Service Award from the American Society for Engineering Education Engineering Libraries Division and he was the recipient of the 2009 Frederick G. Kilgour Award for Research in Library and Information Technology from the American Library Association and OCLC. In 2015, he was elected as an AAAS Fellow.

Ms. Christie A. Wiley, University of Illinois, Urbana-Champaign

Engineering Research Data Services Librarian

Prof. Mary C. Schlembach, University of Illinois, Urbana-Champaign

Mary C. Schlembach is Chemistry & Physical Sciences Librarian and Associate Professor, University Library at the University of Illinois at Urbana-Champaign (UIUC). She has been on a number of digital library grants from the National Science Foundation (NSF), including the National Ethics Portal grant, and the Digital Library Initiative I grant. She has published many articles and conference papers in the field of library and information science and has presented at national and international conferences, including at ALA, SLA, Internet Librarian International, QQML, and ASEE annuals.

Heidi J. Imker, University of Illinois, Urbana-Champaign

An Integrated Data Management Instructional Program

Christie A. Wiley
William H. Mischo
Mary C. Schlembach
Heidi J. Imker
Grainger Engineering Library Information Center, University of Illinois at Urbana-Champaign

ABSTRACT

Much has been written about researcher's data management and data sharing practices and needs. The published studies show that researchers have an awareness of the data sharing mandates and policies of federal grant agencies and journal publishers and there is a growing acceptance of the intrinsic value of data sharing albeit with some concerns and caveats. However, establishing an effective and consistent data management service presents challenges for libraries, given the known disciplinary differences in data management needs and the fact that faculty have not yet significantly changed their data management practices to conform to federal agency and publisher mandates. After conducting in-depth interviews with twenty-one engineering and atmospheric science faculty at the University of Illinois at Urbana-Champaign, it became clear that scientists and engineers view the research lifecycle as a holistic endeavor and treat data as one of many necessary elements in the scholarly communication workflow. The generation, usage, storage, and sharing of data are part of the integrated scholarly workflow, and are not necessarily wholly separate processes.

Building on these interviews, the authors have developed an instructional and training program that better focuses on integrating data management activities focusing on research and scholarly communication processes. The goal of our project was to examine data management practices in the context of researcher scholarly workflow needs and behaviors and develop and implement an instructional program that addresses researcher data needs. The development and assessment of this program is underway.

INTRODUCTION

In response to federal grant agency and publisher mandates for data sharing, science and technology libraries have become actively involved in designing and implementing programs to meet the data management needs of researchers (Tenopir et al, 2015; Samuels et al, 2015). Along with these new opportunities comes challenges in establishing effective programs and services that support the data workflow needs of researchers. Libraries have an opportunity to develop instructional programs that allow better collaboration with researchers and are consistent with the data management behaviors of researchers and students.

A large number of studies have looked at researcher's data management and data sharing practices and needs. The literature shows that researchers have a broad awareness and growing acceptance of data-sharing mandates from federal agencies and journal publishers (Van Tuyl and

Michalek, 2015; Whitmire et al, 2015)). However, they also have concerns regarding the value of data-sharing (Borgman, 2015). Several studies have revealed that disciplinary differences in data management requirements are significant (Weller and Monroe-Gulick, 2014; Akers and Doty, 2013; Kim and Stanton, 2016)). This is a major concern in designing a one-size-fits-all data management scheme for researchers. It is also clear from the literature that faculty have not yet significantly changed their data management practices to conform to federal grant agency and publisher mandates (Whitmire et al, 2015, Diekema et al, 2014). Overall, these observed data management practices present many challenges for libraries when setting up data management services and training programs (Wiley and Mischo, 2016).

In 2016, the lead author interviewed 21 faculty in engineering and atmospheric science at the University of Illinois at Urbana-Champaign to better understand their data management practices and needs within the overarching context of their research workflow. Expanding on the interview questions from the Data Curation Profile Toolkit (http://datacurationprofile.org), open-ended questions were asked about current research projects, funding sources, data types, format, description, disciplinary and/or institutional repository use, data-sharing, scholarly communication processes, awareness of library data management and preservation services, and challenges and struggles researchers have encountered.

The faculty interviewed indicated that they had received a number of communications from the campus and the library regarding support for data management yet were still unsure of what resources were actually being offered or how they could utilize the support. Department heads and group leaders were uncertain about how their faculty and students addressed data management needs and practices. 95% of the interview participants stated that their funding source(s) required them to preserve and share their data and 50% of the participants indicated that they were interested in having a librarian provide instruction on data management practices and options. Three of the researchers had used an online data management plan template provided by the engineering library for National Science Foundation grant requests. Other findings from the interviews: researchers rarely receive requests for their data; they often use internal access mechanisms (departmental web sites, campus computing clusters) for data storage and are unfamiliar with the campus institutional repository; they feel publishers need to take a more active role in data management; and they feel funding agencies are not providing clear guidelines and expectations for data sharing and preservation. The researcher interviews are discussed in greater detail in Wiley and Mischo (2016).

One clear observation from the interviews and from the literature review is that scientists and engineering faculty take a holistic view of the research lifecycle and treat data as one of many elements in the scholarly communication workflow. Data generation, usage, storage, and sharing are an integrated aspect of a larger scholarly workflow, and are not necessarily treated as a separate entity. Acknowledging this fact allowed us to develop instructional and support programs that better integrate data management support into scholarly communication instruction and training.

In addition, researchers also feel that there has been a sea change in the way research is being conducted, that e-journals and e-conferences, grid and cloud computing, simulation software, and data analytics tools have changed the face of research.

Based on the outcomes and lessons learned from this investigation, it was determined that our researchers regard their data management lifecycle as a dynamic process and as one element, albeit a key element, in their scholarly workflow. Researchers, for the most part, have the fundamentals of this workflow in mind, but do not necessarily have it explicitly outlined. This is particularly critical as scientific researchers often rely on graduate students and/or post-docs for day-to-day management of laboratory studies and data recordkeeping. In developing a data management instructional program, libraries take on the responsibilities of orienting graduate students and other personnel on basic data management skills.

INSTRUCTIONAL PROGRAM

It is clear that changing e-research technologies and methodologies have led to rapid changes in scholarly communication models and the support services offered by libraries (Carlson, et al, 2011; Borgman et al, 2015). This changing knowledge creation environment presents libraries with an opportunity to become more integrally involved in the research workflow and scholarly communication lifecycle (Tenopir, et al, 2015).

A number of instructional and training programs for engineering faculty and researchers have been developed and implemented (Carlson et al, 2011; Johnston and Jeffrys, 2014). Many of these programs are built around data management plan assistance (Samuels at al, 2015; Wang and Fong, 2015; Nelson, 2015). Zilinski et al (2014) have developed a program directed at undergraduate STEM students. Many libraries have established research data services and have developed data management presentations (Tenopir et al, 2015). We have borrowed from several of these. Our focus has been on developing a data management and curation instructional program that can be integrated into our presentations on other aspects of scholarly communication. It is natural to discuss dataset and article institutional repository preparation and deposit procedures together. Likewise open access and open data topics are closely related as are topics such as the importance of the literature review, peer review processes, and the selection of appropriate journal article and dataset publication and deposit venues for achieving optimum exposure of the researcher's work. In addition, we have found that it is advantageous to tie together data management discussions with instruction on citation and impact measures, publication metrics, and altmetrics.

As a precursor to the instructional session, the librarian fills out a pre-data management engagement checklist. This helps the instructor to better address the broader scholarly communication needs and familiarize themselves with the research focus of the particular group. The checklist is comprised of the following tasks and questions:

Pre-Data Management Engagement Worksheet

ABOUT THE FACULTY MEMBER

Faculty Member's Name Primary Department Rank (Assistant, Associate, Professor, Specialized)

Group website

What is the main topic of the faculty member's research?

About how many undergraduates/graduate students/postdocs work with this faculty member?

Can you tell if the faculty member is involved in any large centers or consortia or other major projects?

ABOUT THE FACULTY MEMBER's Publications

List three recent publications from this faculty member

Find the faculty member(s) H-Indexes and publication graphs in Scopus

What kind of data is contained within each of the publications?

Do they say if they published their data anywhere?

Find the data-related policies for publishers (be sure to check if the author guidelines are unique to specific journals for the publisher)?

Paste the links here and pull out any good quotes that indicate obligations

ABOUT THE FACULTY MEMBER's Grants

What funders were listed in the publications?

What funders are listed elsewhere (e.g. NIH Reporter, NSF Award search)?

What are the data policies for those funders?

The data management instructional program core features are based on the familiar six progressive stages. To address the known data workflow disciplinary differences, the presentations are tailored in terms of focus on specific data types, formats, descriptions and metadata. There are also differences with funding sources, repository types, and scholarly communication processes depending on discipline culture. We typically begin presentations by defining data, for funding or publication mandates and repository deposit, as "*the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications*". This is consistent with OMB Circular A-110 (OMB Circular, A110). We point out that this definition

precludes: laboratory notebooks, preliminary analyses, drafts of papers, peer review reports, email communications, and, usually, physical objects and lab specimens.

Our presentations also typically include several data management horror-story and retrieval failure scenarios to illustrate the consequences of poor data management.

The instructional sessions begin with a general discussion of the faculty or group's publication patterns and research focus. General principles of knowledge creation, the data lifecycle, publication venue choices, and scholarly metrics are introduced as a means to better integrating data management and the research workflow. The data management sections revolve around previously defined six stages or elements of good data management. For those not familiar with this schema, these six elements are:

PLANNING TOOLS

Our instructional program provides information on the use of the DMPTool (http://www.dmptool.org) and also a custom Data Management Plan (DMP) template that has been developed by Grainger Engineering Library librarians that is designed to guide users and expedite the DMP process. The custom template has been used by engineering faculty in over 150 NSF proposals. We also present a constantly updated chart showing which federal agencies require a DMP (currently NSF (National Science Foundation), DOE (Department of Energy), FDA (Food and Drug Administration), USGS (United States Geological Survey) and soon NIH (National Institutes of Health)) and also which agencies require the deposit of all articles generated with grant support from the agency. This serves to demonstrate the interconnections between datasets and the other outputs generated during the research process.

ORGANIZATION OF DATA

Being able to accurately organize data that is generated and stored during the project workflow is important. The core principles that we try to impart are; identify and distinguish raw vs. processed data; use consistent file naming and file directory structures; define and exercise version control using clarity in file names and date formats. This assists in quickly sorting and finding files and folders related to a specific aspect of the research workflow. We give examples of effective file and folder naming. Again, here we tie the data and metadata to other products of the research process, including grant progress reports, key findings notes, group assignments, article drafts, and laboratory notebooks. Utilizing clear and well documented naming conventions allows the research group to locate, for example, all the graphs produced as part of the data generating process. Being able to sort by date is particularly important.

DOCUMENTATION

Data description through file naming conventions, metadata header files, readme files, and data dictionaries allows other researchers to comprehend and reuse the data of another research group. It also allows a research group to identify and remember the details about the data after a period of time has passed. There are three types of metadata or documentation: descriptive metadata (often connected with a specific subject schema), structural metadata (expressing relationships to

other files), and administrative metadata (including rights and environmental information). These metadata elements are particularly important for data deposited in institutional or disciplinary repositories.

STORAGE AND BACKUP

Our instruction focuses on recommendations regarding number and locations of data copies; schedules for backup; and the robustness and efficacy of storage solutions. We demonstrate the 3-2-1 rule which prescribes: 3 copies of the data; 2 kinds of media; and placing 1 copy in a remote location. Also important are appropriate considerations for data that is sensitive in nature. This includes protected human information under FERPA (Family Educational Rights and Privacy Act), HIPAA (Health Insurance Portability and Accountability Act), and data that is personally identifiable, proprietary, classified, or locally sensitive.

As storage formats evolve, concerns about the feasibility, validity, and awareness of archived data becomes critical. Decisions may have to be made around using software emulation or simulation techniques to either replicate (emulation) or mimic (simulation) older data formats when storing or presenting data in current computing environments. .

ARCHIVING

The longevity and storage requirements for archived data are an active topic of discussion. Several funding agencies place a time limit on data storage requirements. In order to be in compliance with funding agencies, researchers have to consider how, where and for how long to archive their data. Also, as in the mandated open access policies of federal agencies, the question of the embargo period before data becomes publically available must considered.

SHARING

At Illinois, the Library has implemented two data archiving and sharing solutions. Researchers can deposit and store their data in the IDEALS institutional repository when the dataset file size is less than 2 GB, the file(s) are static, and the datasets are fixed or static and not subject to change. Datasets not meeting these criteria can be deposited in the Illinois Data Bank (IDB). The IDB has the ability to mint (create) DOIs for datasets, has a broader deposit capability, and will commit to at least 5 years of online or near-line (mountable tape or disk) storage. The Library is building out additional preservation capabilities in the IDB that conform to the OAIS (Open Archival Information System) reference model.

OUTCOMES AND CONCLUSIONS

The goal of our project was to examine data management practices in the context of researcher scholarly workflow behaviors in order to develop and implement an instructional program directed at faculty groups that addresses researcher data needs. The instructional program is designed to be used in both classroom and research lab environments. The extensive literature review and interviews with 21 researchers in atmospheric science and engineering revealed that researchers have a broad awareness of data-sharing mandates from federal agencies and journal publishers but still have concerns regarding the value of data-sharing. They also showed that that

data management requirements are discipline dependent and that the research data workflow is connected to all aspects of knowledge creation.

From this information, we have developed a data management instructional program that can be integrated into our more overarching presentations on the elements of scholarly communication. The instructional program is built around the six established elements of data management and is enhanced with content aimed at specific discipline and research practice variations. This enables participants to gain skills relevant to all aspects of data management, but tailored to the nuances of their particular needs and funding agency requirements.

A second goal was to increase awareness of the library's programs and role in data management support and services for campus faculty and students. We are buoyed by the early success of having 10 of the 21 interviewees arrange for a library-led presentation to their group regarding data management best practices. We are also investigating new tools such as cloud computing, data visualization and simulation on our visualization wall, collaborative information seeking techniques, and data analytics to provide richer distributed knowledge environments in which we can work collaboratively with researchers. Our hope is that instructional programs will help bridge the gap between researchers' data management practices and the advantages of data sharing as mandated by various funding agencies. Testing and assessment of this instructional program is underway and future modifications are being explored. These results will be part of the conference presentation.

REFERENCES

Akers, K. G., & Doty, J. 2013. Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*. 8(2): 5–26.

Borgman, C., Darch, P.T., Sands, A.E., Pasquetto, I.V., Golshan, M.S., Wallis, J.S. & Traweek, S. 2015. Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries* 16:207-227. DOI: 10.1007/s00799-015-0157-z

Borgman, C. 2015. Data management: One scientist's data as another's noise, *Nature* 520, 157 DOI: 10.1038/520157d

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

Data Curation Profiles Toolkit. 2013. [Internet]. Available from: http://datacurationprofiles.org/

Data Management Plan Tool. 2011. [Internet]. Available from: www.dmptool.org

Diekema, A.R., Wesolek, A., & Walter, C.D. 2014. The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories. *Journal of Academic Librarianship*, 40:322 331.

Johnston, L. & Jeffryes, J. 2014. Data management skills needed by structural engineering students: Case study at the University of Minnesota. *Journal of Professional Issues in Engineering Education and Practice* 140 (2):05013002. DOI: 10.1061/(ASCE)EI.1943-5541.0000154.

Kim, Y., & Stanton, J. 2016. Institutional and individual factors affecting scientists' data sharing behaviors: A multilevel analysis. *Journal of the Association for Information and Technology*, 67(4):776-799.

Nelson, M. S. 2015. Data Management Outreach to Junior Faculty members: A Case Study. Journal of eScience Librarianship, 4(1): e1076, DOI 10.7191/jeslib.2015.1076.

OMB Circular A-110. Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations.

Samuels, S. M., Grochowski, P. F., Lalwani, L. N., & Carlson, J. 2015. Analyzing Data Management Plans: Where Librarians can make a Difference. 122nd ASEE Annual Conference & Exposition, June 14-17, 2015, Paper #12072.

Tenopir, C., Hughes, D., Allard, S., Frame, M., Birch, B., Baird, B., Sandusky, R., Langseth, M. & Lundeen, A. 2015. Research Data Services in Academic Libraries: Data Intensive roles for the future. *Journal of eScience Librarianship* 4(2) e1085. DOI: 10.7191/jeslib.2015.1085

Van Tuyl, S.V., & Michalek, G. 2015. Assessing research data management practices of faculty at Carnegie Mellon University. *Journal of Library and Scholarly Communication, 3(3).*

Wang, M., & Fong, B. L. 2015. Embedded data Librarianship: A Case Study of Providing Data Management Support for a Science Department. *Science and Technology Libraries* 34(3), 228-240, DOI: 10.1080/0194262X.2015.1085348.

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: Electronic library and information systems*, 49(4):382-407.

Wiley, C., Mischo, W. H. 2016. Data Management Practices and Perspectives of Atmospheric Scientists and Engineering Faculty. *Issues in Science and Technology Librarianship* 85(Fall). DOI: 10.5062/F43X84NJ.

Zilinski, L. D., Nelson, M. S., & Van Epps, A. S. 2014. Developing Professional Skills in STEM Students: Data Information Literacy. *Issues in Science and Technology Librarianship*, Summer 2014, DOI: 10.5062/F42V22Z.