eSLib Research Projects: Final Report (Cover Page)

Results of the University Research Data Policy Project & Recommendations

eScience Librarianship Program

School of Information Studies

Syracuse University

Authors: Thea Atwood and Kayleigh Ayn Bohémier

Supervisors: Andreas Kuehn and Jian Qin

Date of Submission: 9 May 2011

Table of Contents

Table of Contents 2
Executive Summary 3
Introduction 4
Methodology 4
Table 1. Randomly Generated Very High Research Activity Universities 5
Data Analysis 5
Figure 1. Coding Frequencies by Coder 6
Findings 6
Figure 2. Top Thirteen Codes. 7
Figure 3. Codes With Fewer Than 15 Occurrences 8
Recommendations 9
References 10

Executive Summary

The project described below offers an overview of an investigation into the data policies of twenty High Research Output Universities. This project was initiated in response to the National Science Foundation's new data management requirements for researchers applying for funding.

Data policies on twenty High Research Output Universities were collected. Two coders created a code book and normalized terms and definitions for use in analysis. All 58 policy documents were analyzed with ATLAS.ti using the aforementioned normalized terms. The final intercoder reliability was 95%.

The findings demonstrated here offer a general overview of the trends in data policies. Most policies focus on the following: (1) access, (2) curation, (3) sharing, (4) restriction and (5) metadata. However, many policies lack a discussion of other, equally important aspects of data management, such as collaboration, data provenance, and scientific workflow. The lack of this discussion may be due to how scientific data has been previously handled or to a perceived lack of need.

It is recommended that the university offers a data policy that covers the key important aspects of data management, including: data access, use, sharing, provenance, metadata, data management, curation, and collaboration. The university should also investigate creating a template that will aid researchers in filling out the Data Management Plan section of their NSF grant proposals. Furthermore, it will be important to offer the infrastructure to support scientists in managing and archiving their data. Finally, establishing a way to reinforce the data management behaviors of scientists will be of importance, as many will be prone to filling out the data management plan, and only minimally fulfilling the obligations they initially outlined. By only minimally fulfilling these obligations, scientists will only minimally add to the use and re-use of their data, and the benefits inherent in a data management plan will fail to manifest.

Introduction

New National Science Foundation (NSF) guidelines declare that researchers wishing to obtain grants in 2011 and beyond must submit a data management plan no more than two pages in length. To promote Syracuse University's adjustment to the new policies, we conducted a content analysis of 20 research colleges and universities to determine the state of data policies across the United States. Our research had two objectives: (1) Determining the emphases of data policies to learn which aspects of data management institutions currently stress and (2) Providing Syracuse University with recommendations to further its development of institutional scientific data policies.

The possible need for an institutional policy concerning scientific research data was raised at the first meeting of the eScience Librarianship cohort in September 2010. In an ideal environment, a basic, uniform research data policy will support grant applicants when they address the new NSF requirements. This content analysis will provide key information surrounding other universities' efforts at creating data policies to the Office of Research. It only surveys publicly-accessible policies, so any information other institutions have internally was omitted from our content analysis. We also limited our content analysis to policies associated with departments, projects, and the institution as a whole, yielding 58 policies from the 20 sampled public and private institutions.

Methodology

Before data collection could begin, it was necessary to determine from which universities to collect data. To ensure that universities were not selected upon pre-determined biases of the data collectors, a random sampling of twenty *Carnegie Classification: Research Universities – Very High Research Activity universities* was generated using SPSS. This way, data was collected from a representative group of universities, instead of from a group of universities that were deemed to be using "best practices" in data policy. The resulting universities determined by the randomly-generated SPSS schema are seen in Table 1, seen below.

We defined an institutional research data policy as a document that included any combination of the following six aspects: data curation, management, use, access, publishing, and sharing. Policies that did not include information on the handling of scientific data were excluded from data collection efforts. To discover research data policy documents, queries were used on both Google and the university's internal search engine. Upon discovery, these documents were saved either in PDF format, or, if in HTML, were copied to a text file and saved for further analysis.

Table 1. Randomly Generated Very High Research Activity Universities

Private Institutions	Public Institutions	
Columbia University in the City of New York	Oregon State University	University Michigan – Ann Arbor
Dartmouth College	SUNY at Albany	University of Minnesota – Twin Cities
Northwestern University	University of California – Irvine	University of Missouri – Columbia
Tulane University of Louisiana	University of California – Santa Barbara	University of Pittsburgh – Main Campus
University of Rochester	University of Colorado at Denver and Health Sciences Center	University of South Carolina – Columbia
Washington University in St. Louis	University of Kansas Main Campus	University of Utah
	University of Massachusetts – Amherst	University of Washington – Seattle Campus

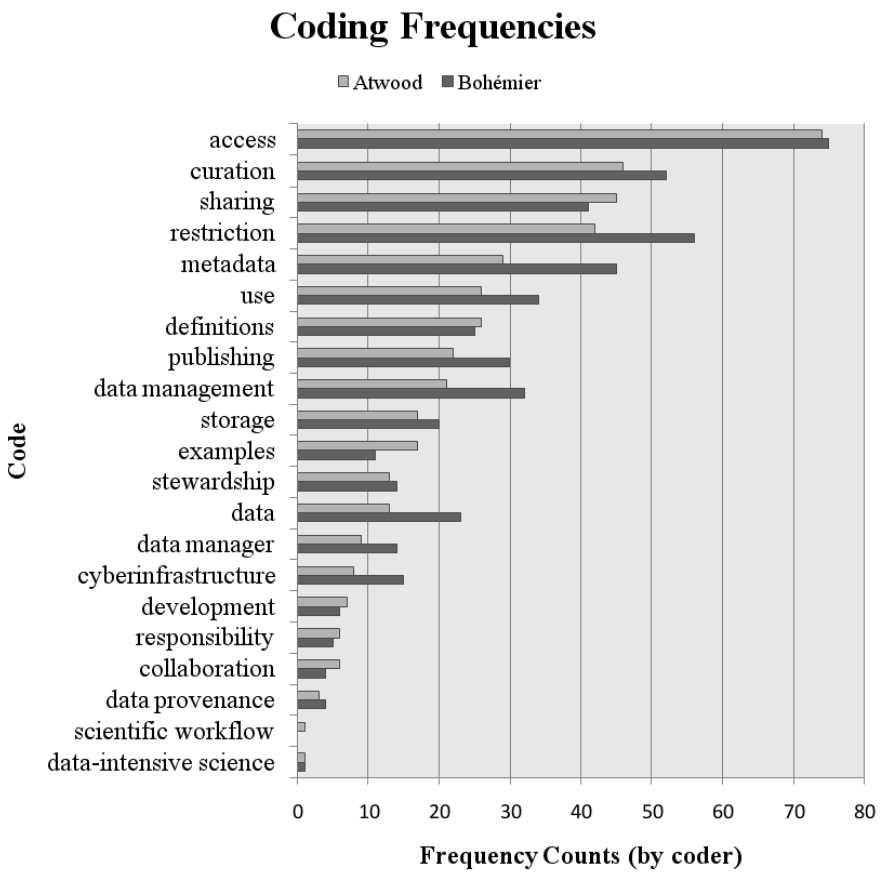
Data Analysis

Retrieved documents were coded in ATLAS.ti 5.2. Prior to coding, it was necessary to normalize terms used to analyze the data policies previously collected. This is because, while many universities use the same terms, they may be using these terms in different ways. For example, one university may use the word *access* to refer to who within the university has clearance to access scientific data, but another university may use the term *access* to determine who is able to use or re-use the data for scientific analysis. Clearly, these two universities are using the same term in two different manners. Furthermore, universities may be using "headings" in their documents to define sections of policies, but as it is unclear as to how each university is using each term, it is imperative to thoroughly read each section of each data policy document. Thus, to facilitate a high degree of intercoder reliability, a code book with key terms and definitions was created.

To select terms used in the code book, we referenced *The Data Deluge* (Marcum & George, 2010). The essays contained within this book offer insight into issues with data and data management, and the challenges libraries face in managing scientific data. With the use of *The Data Deluge*, twenty-one key terms and ideas were identified, and the definition and interpretation of each of these terms normalized between coders. By normalizing the definition of each code between the coders, we would work towards eliminating any discrepancies in term usage, as seen in the example above.

The final intercoder reliability was 95.4%. The difference between the coders was small, and deviations in the code may have been due to the level of granularity to which that individual coders adhered. Figure 1 describes the coding frequencies by coder.

Figure 1. Coding Frequencies by Coder



Findings

Our content analysis showed general trends in the topics individual data policies covered. The content analysis also showed differences in the level of the data policies, or whether the policies covered departments, projects, or the entire institution. We have prepared several graphs and charts showing the distribution of the codes. Using this graphical information, we identified trends and needs for improvement in institutional data policies as a whole.

Figure 2. Top Thirteen Codes.

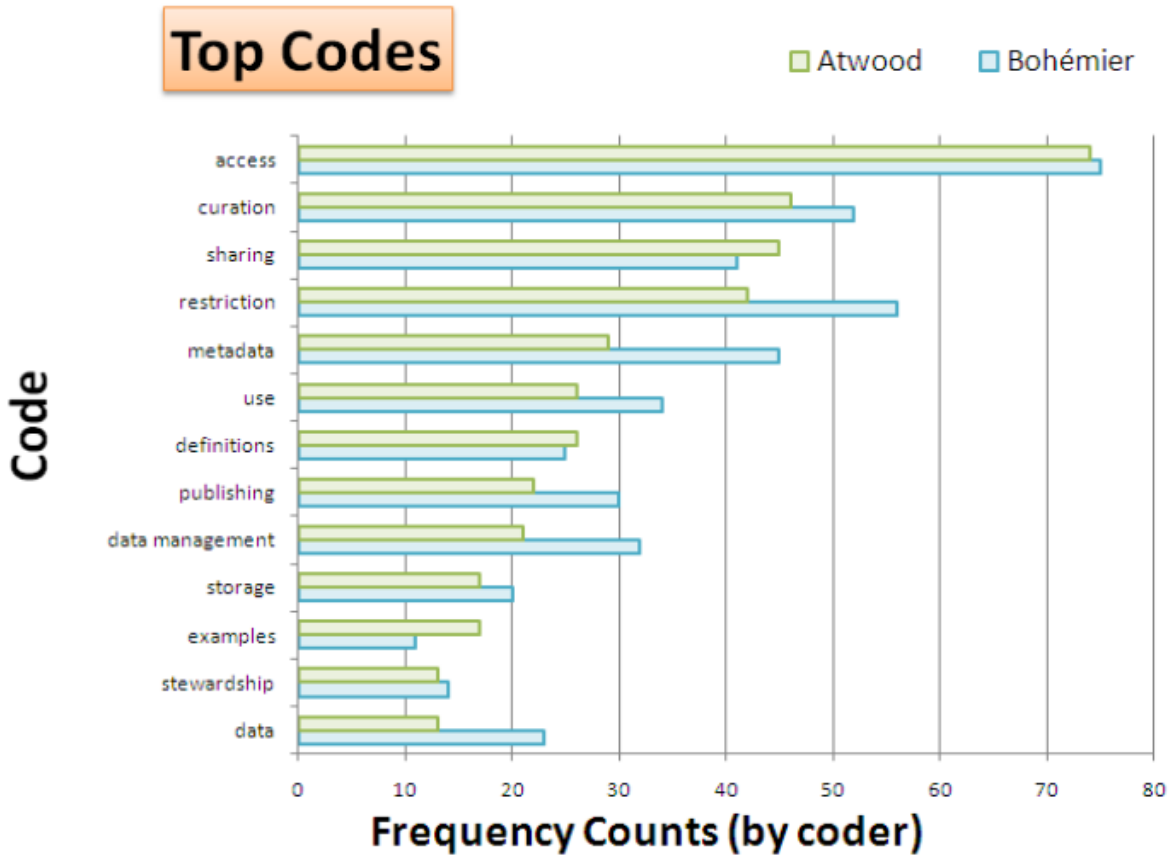
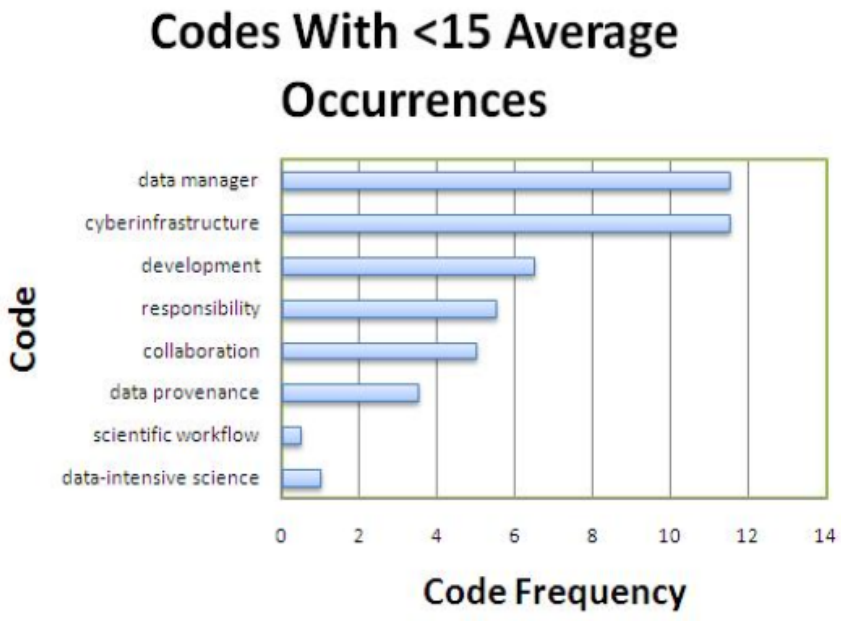


Figure 2 shows a list of the top codes associated with institutional data policies from each coder. Scientific data policies focus primarily on access, curation, sharing, restriction, and metadata. These terms match the type of things people care about in the Internet age: who can access data, what kinds of restrictions exist due to HIPAA and other regulations, and the extent of the description, storage, and maintenance of the data set. Much of this emphasis is backed by statements about the importance of sharing and accessing data. The High Level Expert Group on Scientific Data released a report that argued for "the tremendous value in having the data made seamlessly available, to use, reuse and recombine to support the creation of new knowledge," but also emphasized the need for protections and restrictions on this data to protect it (Wood, Andersson, Bachem, Best, Genova, Lopez et al., 2010, p. 18).

While data policies frequently discuss access, they do not often discuss other important aspects of data policies. Figure 3 shows those codes with less than 15 average occurrences in the data set — among these collaboration, data provenance, and scientific workflow. The reasons for not having this information addressed in the data policies are complex, but often occur because the infrastructure or precedence that support having this information included in policies is insufficient.

Figure 3. Codes With Fewer Than 15 Occurrences



Collaboration stems from both access to the data and the ability to use the data in conjunction alongside others, such as the originator of the data, and to consolidate the distribution of controlled samples when different researchers want to do the same things. Researchers often have little guidance on how to create collaborative partnerships and must instead rely on unspoken disciplinary culture. Bammer (2008) summarizes these issues, revealing that "investigators funded to lead partnerships are often left to learn by trial and error, making success a hit-or-miss affair" (p. 876). Despite not existing in many policies, collaboration is stressed by the High Level Expert Group on Scientific Data during their discussion of the challenges presented by remote collaboration (Wood et al., 2010).

Data provenance information describes where the data comes from and the context in which it was produced, especially those conditions known only to the research team. Having clear data provenance information included in a data set's metadata makes the scientific information more usable to future researchers. Deelman and Chervenak (2008) identify that data provenance information is required "to support both scientific and engineering reproducibility" (p. 5). Hey and Trefethen (2003) describe the challenges to annotating data sets with data provenance information, placing the burden on funding agencies to require creating this metadata because most researchers do not see the value of using part of their grant award to curate their data instead of producing more.

Scientific workflow does not appear in policy documents for very similar reasons. Not only do scientific workflow descriptions often contain very technical information, but the idea of modern scientific workflow has not matured enough for inclusion in scientific data policies. Researchers require better methods for creating scientific workflows, which technology has often missed providing adequately

(Ludäscher, Bowers, McPhillips, & Podhorszki, 2006). Ludäscher et al. also argue that "the most precious resource in e-Science is often neglected: human time" and that acknowledging this requires better methods for scientific workflow (p. 7).

Recommendations

Many universities have yet to establish best practices for research data policies. It is recommended that Syracuse University offer a comprehensive data management policy that covers a variety of areas important for the use and access of information, as well as the storage and curation of data. The university will need to, at a minimum, address the important key topics demonstrated above: data access, data use, data sharing, data provenance, metadata, data management, data curation, and collaboration. The university will need to offer definitions of these terms. As we found in our research, different universities will use the same term to describe different concepts, so eliminating any confusion based upon the university's specific lexicon will be important to cover.

It will be important for the university to offer a Data Management Plan template to help researchers complete the Data Management Plan section of their NSF proposals. The university should also strongly consider offering the technical and personal support needed to fulfill the NSF data management policy requirements. None of the universities we surveyed transparently offered information on how to best approach the NSF Data Management Plan to its researchers, although there is some literature available on how to approach the NSF requirements. Lisa Johnston from the University of Minnesota offers one such overview (Johnston, 2010), and certain university libraries offer guidelines to data management (MIT Libraries, 2010), and other libraries offer services to help scientists manage their data as well as NSF Data Management Plan Templates, such as the University of Virginia's Scientific Data Consulting Group (University of Virginia Library, n.d.).

It may furthermore be worthwhile for Syracuse University to provide a pervasive university-wide data repository, as offering a standard, NSF-compliant area for scientists to deposit their data will facilitate completion of NSF grants, and make those applying for NSF grants more competitive than their non-SU counterparts. While the importance of the cyberinfrastructure supporting data management is not yet highlighted by the literature, this essential component of data management may soon see an exponential increase in importance.

Finally, offering a means to reinforce the data management behaviors of scientists will be important to investigate. There is some literature available that covers how to establish some of these incentives (Barton, Smith & Weaver, 2010; Frigoli, Etgen & Kuhar, 2010) as well as the benefits of establishing such practices (Pascal, 2006). While gaining funding from the NSF does offer some incentive to create a data management plan, there may be little to no incentive to continue to fulfill the obligations outlined by a researcher in their NSF proposal. Making the completion of a proposal, as well as the continued upkeep of that proposal as streamlined as possible will help to ensure that a research team is doing their best to continue to fulfill their data management obligations. Failure to offer a means to demonstrate to

scientists that completely fulfilling their data management obligations may mean that data management will be abandoned in favor of creating more data. This result will compound the issue already at hand - scientists will be creating more data and minimally applying their data management standards, which will yield poorly constructed data management systems which scientists will see little benefit in using.

References

Bammer, G. (2008). Enhancing research collaborations: Three key management challenges. *Research Policy, 37.* 875-887. doi:10.1016/j.respol.2008.03.004

Barton, C., Smith, R. & Weaver, R. (2010). Data practices, policy, and rewards in the information era demand a new paradigm. *Data Science Journal, (9).* pp. IGY95-IGY99.

Deelman, E. & Chervenak, A. (2008). Data management challenges of data-intensive scientific workflows. *Eighth IEEE International Symposium on Cluster Computing and the Grid*, (CCGRID '08). 687-692. doi:10.1109/CCGRID.2008.24

Frigoli, J., Etgen, A.M., & Kuhar, M. (2010). Developing and communicating responsible data management policies to trainees and colleagues. *Science and Engineering Ethics, 1353-3452.* pp. 1-10. doi: 10.1007/s11948-010-9203-9

Hey, T., & Trefethen, A. (2003). E-science and its implications. *Philosophical transactions: mathematical, physical and engineering sciences, 361*(1809, Information, Knowledge and Technology), pp. 1809-1825. doi:10.1098/rsta.2003.1224

Johnston, Lisa (2010). Web Reviews: NSF Data Management Plan: A Primer. *Sci-Tech News, 64*(4). Retrieved from http://jdc.jefferson.edu/scitechnews/vol64/iss4/12

Ludäscher, B., Bowers, S., McPhillips, T., & Podhorszki, N. (2006). Scientific Workflows: More e-Science Mileage from Cyberinfrastructure. *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (e-Science '06).* Retrieved November 7, 2010 from http://ieeexplore.ieee.org.libezproxy2.syr.edu/stamp/stamp.jsp?tp=&arnumber=4031118

Marcum, D.B. & George, G. (Eds.). (2010). *The data deluge: Can libraries cope with e-science?* Santa Barbara, CA: Libraries Unlimited

MIT LIbraries (2010). Data management and publishing. Retrieved May 8, 2011 from http://libraries.mit.edu/guides/subjects/data-management/index.html

Pascal, C.B. (2006). Managing data for integrity: Policies and procedures for ensuring the accuracy and quality of the data in the laboratory. *Science and Engineering Ethics, 12*(1). pp. 23-39.

University of Virginia Library (n.d). Scientific Data Consulting. Retrieved May 8, 2011 from http://www2.lib.virginia.edu/brown/data/

Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D.R., Los, W., Marinucci, M., Romary, L., Van de Sompel, H., Vigen, J., Wittenburg, P., & Giaretta, D. (2010, October). *Riding the wave: How Europe can gain from the rising tide of scientific data*. High Level Expert Group on Scientific Data. European Union.