

Faculty data management practices: A campus-wide census of STEM departments¹

John D'Ignazio

Jian Qin

School of Information Studies

245 Hinds Hall

Syracuse University

Syracuse, NY 13244-1190

jadignaz@syr.edu

jqin@syr.edu

Citation for this paper:

D'Ignazio, J. A. & J. Qin. (2008). Faculty data management practices: a campus-wide census of STEM departments. In: *Proceedings of the American Society for Information Science and Technology, October 24-29, 2008, Columbus, Ohio*. (Poster) <http://sdl.syr.edu/pubs/dignazio2008ASIST.pdf>

¹ Study funded by the U.S. National Science Foundation as part of the Course, Curriculum, and Laboratory Improvement (CCLI) program

Introduction

As scientists' means of communication and information behaviors have evolved over the past several decades due to computing and networking developments, the field of library and information science (LIS) has responded by surveying their changing practices. This study continues this type of research but switches focus away from scientists' use of journal articles, the dominant means by which research libraries have supported their investigations, and instead concentrates on scientists' practices related to data. The view of scientists as increasingly energetic generators, managers, and users of large and growing electronic datasets has lately been recognized and promoted at the societal level by funding agencies, academic societies, and large research centers.

The growth in data production and storage related to science, technology, engineering, and mathematics (STEM) research has been attributed to the development and proliferation of enabling technologies and computer networks associated with cyberinfrastructure advancement, including sensors and sensor networks, high-throughput technologies and instrumentation, automated data acquisition, and computational modeling and simulation. (NSF, 2007) While the technical and societal forces encouraging scientists to produce this born-digital content continue unabated, attention to their resulting burden of managing this content for access and use has lagged behind. LIS-trained practitioners could contribute information management infrastructure to aid scientists but this infrastructure would have to be oriented to the variety of local, idiosyncratic, field-based approaches currently extant. (Borgman, 2007) A campus-wide census of data management practices at a Carnegie Classification Research II institution identifies what the variety of faculty data management practices is across the STEM departments.

Approach and Methods

LIS researchers have studied scientists to understand their information behaviors in a changing scholarly publishing environment. For instance, Brown (1999) polled faculty in four disciplines, Astronomy, Chemistry, Mathematics, and Physics, at her home institution to gather comparative data about their use of and attitudes about electronic versus print journal articles in relation to the academic library. Tenopir and her colleagues (2005) have teased out differences in the rate and manner that scientists in particular disciplines consume online versus print journal articles to conduct their research via multiple, sampled surveys that target a large number of members of the field's dominant academic society, and then comparing the results to other similar studies. Zhang (2001) surveyed authors of articles published in eight scholarly journals covering the LIS field to determine how Internet-based electronic resources were considered during the preparation of their articles.

To help get the most out of computer-based data resulting from the increasing amount of scientific investigation facilitated through cyberinfrastructure, funding agencies in both the U.S. and the U.K. are investing in projects that develop data preservation and management tools and skills. One such effort, the Science Data Literacy (SDL) project at the iSchool at Syracuse University, has been funded by the National Science Foundation to develop a course designed to prepare students with basic knowledge and skills in science data management. In a similar manner as the earlier surveys of scientist information behaviors, the SDL staff as an early part of the project prepared a survey vehicle and conducted a census of the relevant campus faculty. The project team took a pragmatic approach in crossing disciplinary boundaries in order to gather the variety of data management practices in STEM departments across the home institution. Departments were identified that reasonably fell within the rough boundaries provided by the STEM category amalgam; SDL staff erred on the side of including the most

number of researchers likely to be accumulating and working with primary datasets.

In consultation with iSchool faculty practiced at survey construction, SDL staff created a survey vehicle and pilot-tested it on members of the SDL project advisory board who matched the target population. Terminology had to be negotiated and definitions provided to orient concepts from information science to the scholars' research process. For example, we eliminated the word "metadata" from the questionnaire and added an inclusive definition of data from a National Science Board report (2005) in the introduction to the survey: "any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." Demographic questions particular to the discipline-focused and hierarchical environment of the academic community were taken from the Higher Education Research Institute's faculty performance survey (2004). As per Janes' (1999) helpful advice on survey construction, this demographic data-gathering section was placed at the end.

The iterative process of question phrasing and grouping led to a refined instrument designed to capture data management practices. The four-part, web-based survey featured Likert-scale agreement questions related to attitudes, practices, and experience with research data. Two provocative questions about data management in the respondent's discipline were designed not only to motivate participants to express their views, but also to help establish the presupposition of the questionnaire that the respondent was a data producer (Martin, 2006). This rather brute force technique was designed to concentrate the mind of the respondent on their data management practices, but may have had unknown effects on the target population. An initial branching question may have been more effective at weeding out members of the STEM departments who, due to a teaching, practical, or theoretical orientation, did not manage data. In the meat of the survey, branching questions also would have allowed researchers to record more than one management or preservation practice, but the instrument provided for this purpose the ability for the participant to mark multiple options per response where appropriate to simplify the survey design. Textboxes and open-ended questions in each section were meant to elicit a range of practices with data as well as overall reactions provoked by the 25 questions.

Syracuse University, in combination with the symbiotically connected campus of the SUNY College of Environmental Science and Forestry, has over a thousand full-time faculty members; the SDL survey was administered to 362 faculty members from STEM departments at the two institutions. Respondents received a movie ticket coupon as consideration for their time. Possible participants were identified via information posted on school and department websites and then contacted via email solicitation, notification and up to two reminders. Faculty members were given the opportunity to opt out—many did citing current retirement status or lack of involvement in research. As it was a local census using a MySQL-based token response system, anonymity was not an option, so entries are being kept in confidence and analysis is ongoing.

Preliminary Findings

Of those who participated resulting in a 30.7 percent response rate, problems that might affect the census results include a lack of alignment of the respondent with data-producing aspects of STEM research pursuits, such as a solely theoretical orientation in physics and mathematics, or a social orientation in the case of geography and information science. Additionally, faculty who handle different types of data per research project, faculty organized in research groups, and faculty use of research assistants as day-to-day handlers of data are conditions that may have interfered with an accurate and complete perception of data management practices from the

survey responses obtained.

Table 1 indicates some of the variety encountered in data management practices throughout the STEM departments. A value closer to five indicates strong agreement that the researcher responding is a frequent data producer, used data prepared by another researcher, was aware that the digital data may be used by another researcher outside his or her own group, prepared metadata of some kind for the internally produced datasets, and found metadata entries helpful when obtaining data for use from external research groups. All entries could have been the respondent answering as a representative of his or her research group.

Table 2 shows the relation in responses between those researchers who work with a certain size dataset and their perception of the effect of data management practices on their discipline's progress. Researchers who operate with larger datasets appear more confident about their discipline's data management practices.

Table 3 relates actions routinely taken by researchers with their data and with their agreement on the negative impact that inadequate data preservation practices are having on their discipline. It appears that researchers involved in advanced data activity such as calculation and visualization may be slightly more sanguine about their discipline's preservation practices.

Table 1. Relationship between discipline, data production & use, and metadata assignment & use (3 or more responses identified by discipline, 5 point agreement scale, high value indicates more agreement)

Discipline	Frequent Data Producer	Brought in external data for use	Aware of external data use	Metadata entered by research group	Metadata helpful for obtaining external data
7 environmental science	4.71	4.00	4.00	3.57	4.14
5 zoology	4.60	3.60	4.40	4.20	4.20
10 other biological sciences	4.40	3.70	3.40	3.70	3.60
4 bioengineering	4.00	3.00	3.00	4.00	3.75
12 other engineering	3.08	2.58	3.00	3.33	2.83
8 atmospheric science	3.00	2.88	2.75	2.88	3.50
9 oceanography	4.33	3.44	3.00	4.22	4.00
6 physics	5.00	3.17	3.50	3.17	4.33
4 other physical sciences	3.75	3.00	2.50	3.25	3.75
9 experimental psychology	3.78	3.78	3.56	3.00	3.56
5 archaeology	4.80	1.80	3.40	4.40	2.80
3 economics	3.33	2.33	2.67	4.33	4.00
3 political science	5.00	3.33	4.00	5.00	4.33

3 computer science	4.67	3.33	3.67	3.67	3.67
3 library science	4.00	3.00	3.00	4.00	3.67
11 information science	4.36	3.09	4.00	3.18	3.63
7 other	4.71	3.00	3.43	4.00	3.43
111 Total	4.06	3.11	3.32	3.56	3.58

Table 2. Relationship between current data management practice and the size of data files (Selection of data file size X 5 point agreement scale regarding practices limiting disciplinary advancement)

Current practices used to manage data in my field are limiting advances in knowledge	Size of data files								Total
	< 1 MB		> 1 MB and < 100 MB		> 100 MB and < 1 GB		> 1 GB		
	Yes	No	Yes	No	Yes	No	Yes	No	
Strongly disagree	9	10	5	14	2	17	3	16	76
Disagree	13	17	12	18	3	27	6	24	120
Neutral	19	8	15	12	4	23	3	24	108
Agree	12	15	16	11	13	14	5	22	108
Strongly agree	2	6	7	1	1	7	1	7	32
Total	55	56	55	56	23	88	18	93	555

Table 3. Impact of lack of data management strategies on data use and management tasks (Selection of data management action X 5 point agreement scale regarding disciplinary information loss)

Information is being lost in my discipline due to lack of strategies to preserve research data	Action taken in data use and management										Total
	Cleaning		Conversion		Merging		Calculation		Visualization		
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
Strongly disagree	5	11	7	9	6	10	11	5	11	5	80
Disagree	10	22	17	15	14	18	25	7	20	12	160
Neutral	9	13	12	10	10	12	12	10	15	7	110
Agree	16	15	18	13	17	14	27	4	22	9	155

Strongly agree	6	4	9	1	7	3	9	1	7	3	50
Total	46	65	63	48	54	57	84	27	75	36	555

Conclusion

Continuing analysis will show the variations in management and preservation practice, according to faculty affiliation and position. This will allow mapping of local institutional attitudes and behaviors regarding data to those encouraged by major research initiatives at the disciplinary and national level. It is anticipated that results from this survey will help the SDL project team, and the LIS field more generally, understand how to integrate SDL education with STEM departments actively working with science data, as well as develop educational materials at an appropriate level to assist meeting the need for personnel with the skills and interest in science data management and preservation to support effective community and disciplinary use of these digital resources. On a more basic level, it will help provide information on how science and technology researchers obtain and manage their data as part of knowledge production and science communication processes.

References

- Borgman, C. (2007). Data: Input and Output of Scholarship. *Scholarship in the Digital Age*, Cambridge, Mass: MIT Press, 115–147.
- Brown, C. (1999). Information seeking behavior of scientists in the electronic information age: Astronomers, Chemists, Mathematicians, and Physicists. *Journal of the American Society for Information Science*, 50(10), 929–943.
- Cyberinfrastructure Vision for 21st Century Discovery* (2007). Washington, D.C.: Cyberinfrastructure Council, National Science Foundation.
- Faculty Survey (2004). Los Angeles, Calif.: Higher Education Research Institute, University of California, Los Angeles. Retrieved August 07, 2007, from <http://www.gseis.ucla.edu/heri/researchers/instruments/FACULTY/>
- Janes, J. (1999). On Research: Survey Construction. *Library Hi Tech*, 17(3), 321–325.
- Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (2005). Washington, D.C.: National Science Board, National Science Foundation.
- Martin, E. (2006). *Survey Questionnaire Construction*. Research Report Series #2206-13. Washington, D.C.: U.S. Census Bureau.
- Tenopir, C., et al. (2005). Relying on Electronic Journals: Reading Patterns of Astronomers. *Journal of the American Society for Information Science and Technology*, 56(8), 786–802.
- Zhang, Y. (2001). Scholarly Use of Internet-Based Electronic Resources. *Journal of the American Society for Information Science and Technology*, 52(8), 628–654.