Studies of Data Curation for the Long Tail of Science

Carole L. Palmer, Melissa H. Cragin, P. Bryan Heidorn, Linda C. Smith
Graduate School of Library and Information Science
Center for Informatics Research in Science and Scholarship (CIRSS)
University of Illinois at Urbana-Champaign

Abstract

Some sciences have organized data activities around national data centers and "reference" collections (NSB, 2005), and others around disciplinary repositories. The data curation and stewardship needs of sciences that produce and rely on smaller, "research" level data collections are less well coordinated and understood. To better understand the potential of both cross-institutional cooperation and the roles of librarians in curation of research level data collections, the Data Curation Education Program (DCEP) at UIUC is undertaking a series of projects. We have identified several questions for study, including: What attributes of data collections, or of a particular science, make them less likely candidates for resource or reference federations? What are the roles of institutional repositories in managing and maintaining research data collections? What is the impact of accessibility to the long tail of data for the conduct of science? In this poster we focus on the initial study of environmental scientists at UIUC. Overall, our research agenda requires

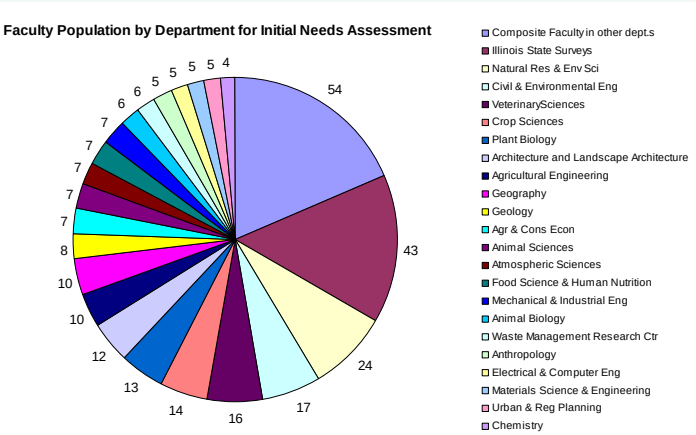
1.) data collection techniques that scale; and

2.) customized instruments that yield integrable, cross-disciplinary data.

Sampling and Scaling Challenges

We have found two fundamental challenges in conducting this study. First, the survey population we selected has proven to be difficult because of its cross-domain make-up. The UIUC Environmental Council Faculty of the Environment is a loosely organized group of approximately 400 people, "dedicated to environmental excellence". This group includes scientists, administrators and community members from all colleges and most departments on campus. The variation in this group required re-evaluation of our target population. Wanting to focus on "data generators," we reduced the list to current researchers and related administrators. Second, we want the survey to scale in application, for use with larger populations or researchers. Yet, we found that a single set of survey questions will not produce reliable data, and possibly invalid data, across this very diverse group of faculty.

Faculty of the Environment – Survey population



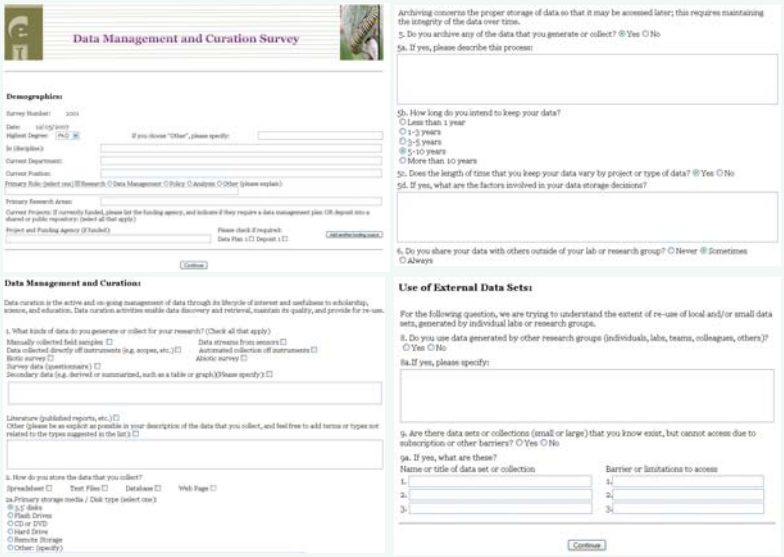
Research Design Refinements

- completing pre-test required personal contacts from PIs and others on the project
- conducting pilot phase of survey "assisted" by interviewing.
- modified survey to better capture data re-use;
- added questions about disciplinary data management training and need for technical assistance

Next Phase in Research Program

Applying refined approach in collaboration with Purdue University Library (D. Scott Brandt, PI) and the UIUC Library on project to investigate "when scientists are willing to share data, and with whom," and how librarians might participate in university eScience initiatives to facilitate curation activities.

Sample pages from the UIUC Environmental Council Survey



Variation of key factors across pre-test participants

	Data management	Data sharing and re-use	Implications
Agricultural Economist - environmental economics	- no perceived problems	- scientist generates composite data sets for each project or series of papers and does not re-use these data	- uncertainties about features of **composite data sets** and related metadata descriptions
Microbiologist - microbial ecology	- full data sets are kept in a database, but several "views" of processed data are kept in spreadsheets.	- some data be made available, but "most data sets cannot be used outside of the context of this lab"	- relationships between disciplinary repositories vs. local IRs - criteria needed for **retention policy**
Sociologist - natural resources sociology	- "grad students know more about where (some) data come from than I do" - data sets growing with need for integrating secondary data; would like training and support with data management and integration	- "more willing to share data with non-academics, extension, public agencies"; - re-visits own data, and compares with new data – "expect to do that a lot"	- role for university libraries and research computing in providing **data services** for data management and curation
Sociologist - water management	- grad students do most of the data management; - quantity of data is a problem.	- sharing dependent on "IRB conditions"; will release data to engineers or water operators - "some parts of a data set become obsolete"	- assessment frameworks for obsolescence, **replication**, and reuse (value)

References

National Science Board (2005). Long-lived digital data collections: Enabling research and education in the 21st century, (http://www.nsf.gov/pubs/2005/nsb0540/).

This work is supported by IMLS grant RE 05-06-0036-06.