Eliciting Faculty Requirements for Research Data Repositories

Michael Witt
Interdisciplinary Research Librarian & Assistant Professor
Purdue Libraries & Distributed Data Curation Center (D2C2)
mwitt@purdue.edu

Open Repositories 2009, Georgia Tech: May 18, 2009
Ten Questions to Begin a Conversation With Your Faculty About Data Curation

1. What is the story of your data?
2. What form and format are the data in?
3. What is the expected lifespan of your data?
4. How could your data be used, reused, and repurposed?
5. How large is your dataset, and what is its rate of growth?
6. Who are potential audiences for your data?
7. Who owns the data?
8. Does the dataset include any sensitive information?
9. What publications or discoveries have resulted from the data?
10. How should the data be made accessible?

Investigating Data Curation Profiles Across Multiple Research Disciplines

Investigators in the Distributed Data Curation Center in the Libraries at Purdue University, and the University of Illinois, Urbana-Champaign will address the question “which researchers are willing to share data, when, with whom, and under what conditions?” The team will produce case studies of researcher data/metadata workflow, data curation profiles describing policies for archiving and making available research data, a matrix to compare parameters across disciplines, system requirements for managing data in a repository, and recommendations for implementing results under diverse systems. The project will describe the roles of librarians and identify the skill sets they need to facilitate scholarly communication and data sharing. Supported by IMLS LG-06-07-0032-07.
Investigators

D. Scott Brandt (PI) – Purdue University
Jacob Carlson – Purdue University
Melissa Cragin – University of Illinois
P. Bryan Heidorn – University of Illinois
Carole Palmer – University of Illinois
Sarah Shreeves – University of Illinois
Michael Witt – Purdue University

Two-year research project began on 11/15/2007.
Project Activities

• Two interviews each with 20 faculty who produce data in a variety of research domains
• Transcription, coding, and analysis (NVivo)
• Creation of “data curation profiles” and wiki
• Developing two case studies in Agronomy and Geology
• Two focus groups with subject-specialist librarians who acted as liaisons

• Distinguish and map needs expressed by faculty to repository functionality
• Assess current capabilities of repository systems and related technologies
• Experiment using institutional repositories for data curation in practical terms
Subjects

**Purdue**
- Biology
- Horticulture
- Civil Engineering
- Electrical & Computer Engineering
- Biochemistry
- Food Science
- Earth & Atmospheric Science
- Agronomy
- Agronomy
- Agronomy

**Illinois**
- Kinesiology
- Atmospheric Sciences
- Speech & Hearing
- Soil Science
- Anthropology
- Anthropology
- Anthropology
- Geology
- Geology
- Geology

**Case Studies**
- Agronomy
- Agronomy
- Agronomy
- Agronomy
- Agronomy
Caveat audiens

- Preliminary findings (project is not yet complete)
- Convenience sample, not statistical
- Exploratory, qualitative study
- The subjects provide much more context and information in the interview transcripts, which are still being coded
Dataset static or dynamic?

- Static: 65%
- Dynamic: 35%

n=20
Data bound by confidentiality or privacy concerns?

- Yes: 79%
- No: 16%
- I don't know: 5%

n=19
Is your manner of organization/description sufficient for another person with similar expertise to be able to understand and properly use the data?
How long to preserve your data?

- 21%: 0 years
- 11%: < 3 years
- 11%: 3-5 years
- 21%: 5-10 years
- 5%: 10-20 years
- 11%: 20-50 years
- 31%: 50-100 years

n=19
With whom would you share your data immediately after the data were generated?
With whom would you share your data after the data were normalized and/or corrected?
With whom would you share your data after the data have been processed for analysis?

- Nobody: 16%
- Collaborators: 53%
- My research center or institution: 21%
- My professional society: 5%
- Anyone: 5%

n=19
With whom would you share your data after the data have been analyzed?

- 42% Anyone
- 26% My research center or institution
- 16% Collaborators
- 11% My professional society
- 5% Nobody

n=19
With whom would you share your data immediately before publication of findings?

- Nobody: 5%
- Collaborators: 16%
- My research center or institution: 21%
- My professional society: 26%
- Anyone: 32%

n=19
With whom would you share your data immediately after publication of findings?
Prioritize your needs for the following types of services

- A secondary storage site for the dataset?
- A secondary storage site for the dataset at a different geographic location?
- Documentation of any changes that were made to the dataset over time?
- The ability to migrate datasets into new formats over time?
- The ability to audit this dataset to ensure its integrity over time?
Prioritize your needs for the following types of services

The ability for researchers within my discipline to easily find this dataset

The ability for researchers outside of my discipline to easily find this dataset

The ability to cite this dataset in my publications

The ability for people to easily discover this dataset using Google

Legend:
- Not a priority
- Low priority
- Medium priority
- High priority
- I don’t know or N/A
Prioritize your needs for the following types of services

- The ability of the repository to provide version control for the data
- The process of submitting this dataset to a repository is automated
- The ability for me to submit this dataset to a repository myself
- The ability to make these data accessible in multiple formats

[Bar charts showing priority levels]
Prioritize your needs for the following types of services

1. The ability to access the data at a mirror site if the main repository is “offline”
2. The ability to see usage statistics of how many people accessed your data
3. The ability to apply standardized metadata from your discipline to the dataset
4. The ability of others to comment on or annotate the dataset
Prioritize your needs for the following types of services

- The ability to restrict access to datasets to authorized individuals
- The ability to support the use of web services APIs
- The ability to connect the dataset to visualization or analytical tools
Summary