Open Repositories 2009 Conference Presentation Proposal - February 20, 2009

Eliciting Faculty Requirements for Research Data Repositories

Michael Witt (mwitt@purdue.edu)
Assistant Professor of Library Science
Distributed Data Curation Center
Purdue University Libraries
504 W. State Street
West Lafayette, IN 47907 USA

In a two-year project sponsored by the Institute of Museum and Library Services, investigators from Purdue University and the University of Illinois, Urbana-Champaign (UIUC) are addressing the question, "Which researchers are willing to share data, when, with whom, and under what conditions?" Extensive, in-depth interviews have been conducted with nineteen faculty researchers in twelve different disciplines about their scientific workflows and range of research outputs (e.g., raw data through published forms) to identify needs for discovery, access, use, and repurposing of their data. The interviews utilized a protocol to elicit responses for cross-discipline comparison and contrast as well as coding to identify categories for deeper analysis. Disciplines covered in the study include:

- Agronomy & Soil Science
- Anthropology
- Biochemistry
- Biology
- Civil Engineering
- Earth & Atmospheric Sciences
- Electrical & Computer Engineering
- Food Science
- Geology
- Horticulture & Plant Science
- Kinesiology
- Speech and Hearing

Three disciplines overlap between Purdue and UIUC, and the interviewers at each institution have conducted interviews with three or more researchers within at least one discipline to generate discipline-specific case studies. The information gathered from the interviews and surveys is being analyzed and a comparative matrix and "data curation profiles" are being developed for each case.

As the study enters its second year, we are ready to share preliminary findings that relate the needs for data curation that were expressed by the subjects to possible functional requirements that a data repository system can provide to meet those needs.

A general overview of the study and its methodology will be presented as background information, and the focus of the presentation will be on the data collection and early findings that relate to the needs/systems requirements expressed for specific, exemplar datasets that were described by each subject. These will be presented as a summary of the survey responses and contextualized by supplementary information provided in the interviews as brief "vignettes" of each dataset and corresponding requirements.

Questions about the exemplar datasets explore the size, quantity, and rate of data growth; data formats and tools; what data faculty researchers are willing to share and at what point in the data lifecycle; embargoes; preservation; organization and description of data; use of standards; and access. Specific needs are prioritized by the subjects, as well.

This is not a statistical study (the sample size is neither large nor randomized) or a comprehensive needs assessment; however, we believe that disseminating our process and early results of discerning user needs into these systems requirements can inform both the developers and implementers of data repositories as well as practitioners who are developing and managing data collections. These "deep dives" provide immediate and valuable insights and can establish an anchor point for more generalized research in the future.

This proposal is for a 20-30 minute presentation to be incorporated into a panel or to complement other presentations in a session. It is being suggested for the "Data, Repositories, and Infrastructure" track, but there is considerable overlap with other conference themes, so it may be also considered for other tracks.