Re: NSF Workshop 9/26-27, Digital Data Curation and Management

Top Issues regarding: Long term preservation, management and curation of scientific data in digital form

Basic theme: Organizational Models

1. **Engaging the Research Community**

Research scientists now face the necessity of a major shift in 'zeitgeist' of how they must think about their data, and their labs in general, if they are to take advantage of the power and promise of the cyber-infrastructure -based digital environment in which we now exist. If research data sets created by individual research labs are to become part of a national or international digital data framework which is "open, extensible, and evolvable," they can no longer function in isolation. For example, they must adopt revised methods of data management which insure data preservation and data sharing. Unless the research community can be brought to understand the significance and the usefulness of such changes, they will not adapt.

2. **Partnerships between research labs and University Library and across Libraries**

Establishing partnerships between research labs and the university library provides a possible new infrastructure which is essential to the overall goal (a new 'system of science' in the digital framework). This partnership fosters an environment where library and research lab do not function in isolation from each other. However, this infrastructure requires: (i) personal investment from both research lab and library personnel; (ii) interpersonal collaboration at an infrastructure level; (iii) university level support; (iv) new middleware for collaboration and coordination; (v) reciprocal adaptation by both the research lab and the library to the information structure of the particular data and materials involved. These needs arise at a repeated but higher order level when inter library exchange is developed.

3. **Establishing Knowledge networks and related ontologies which bridge numerous research centers**

New intermediary infrastructures can potentially help bridge the divides that now exist between individual research labs, between institutions housing these labs, and between lab and university libraries. One such structure is the "Virtual Center" in a knowledge area.

Submitted by B. Lust & J. McCue

Re: NSF Workshop 9/26-27, Digital Data Curation and Management

Basic theme: Sustainable Economic Models

Barbara Lust (Professor, Human Development) and I (Associate University Librarian for Life Sciences) are co-PI's on a project at Cornell that relates to research data and library/laboratory collaboration. The purpose of our NSF Small Grant for Exploratory Research is to test the feasibility of extending the role of large research libraries in supporting value-added services for research data, including access, metadata, outreach, training, and archiving. The exploratory grant was focused on one laboratory (Language Acquisitions Lab); a supplemental award targeted a second research group (Agricultural Ecology Program). In the supplement, we are evaluating whether the conceptual model developed for language acquisition data is applicable to other disciplines. For an overview of the project's goals and accomplishments, see the project website. http://metadata.mannlib.cornell.edu/lilac/

Based on our experience in this planning grant, there are many significant issues to address in considering sustainable economic models, including capacity building, scaling-up, and determining future costs.

1. **Staffing:**
Although the Teragrid is a reality, it will only reach its full potential when it is heavily traveled by a broader spectrum of the research community. It is a significant challenge to build the human capacity to deliver data and associated services in ways that support the research community. We will need skilled programmers to develop the tools for data-driven research and facilitate discovery and access; agile librarians with strong academic backgrounds to curate the collections and support end users; and sympathetic researchers who understand the value of good metadata, best practices, and archival decisions. For example, in our work with Lust's lab, both the metadata librarian and the programmer have linguistics backgrounds; in our AEP grant, our Research Data/Environmental Sciences Librarian, who has a graduate degree in Ecology & Evolutionary Biology, works closely with the 12 co-PI's in the project. Having these specialized backgrounds allows the library to more easily translate the needs of the research lab into services and to understand the curatorial and preservation aspects-of the data.

2. **Scaling-up**:
We are working with two small projects in two labs within a single university, and some close collaborators. How do we scale-up to deal with oceans of data in diverse disciplines bridging multiple institutions? Can we leverage what we learn with one project and apply it to another? Can we mainstream some activities so that specialized staff consult and support staff process? Can we do a better job of capturing data at the point of creation, in formats that can be made accessible, re-purposed, archived, and mined?

3. **Estimating Future Costs:**
It is difficult to determine long-term costs and long-term commitments when the models are still evolving. How do we determine long-term costs when the issues related to long-term availability/preservation of data are still puzzling us? If we develop collaborative repositories, how invested will the individual institutions and individual researchers be in sustaining of a cross-institutional entity? Can the costs be generalized for other institutions and other disciplines? Who is likely to bear the costs associated with research data discovery and preservation—research institutions? granting agencies? Will STM vendors or universities

or new entities offer subscriptions to institutions for services related to research data and will institutions/researchers be willing to pay for those services?

Submitted by J. McCue & B. Lust