New Collaborative Relationships
The Role of Academic Libraries in the Digital Data Universe

Position Statement by Charles Humphrey

Preserving scientific data is a process best characterized by a life cycle model that differentiates the various stages through which research is conducted.[1] The life cycle perspective helps visualize a global representation of this process and helps identify the digital objects produced throughout the full research cycle.[2] Within each stage are outputs in digital format that record or summarize research activities. For example, literature reviews, research prospectuses and grant applications are typical products of the Study Concept and Design stage that later become important sources when documenting data. Some of these products are specific to a particular stage while others are passed between stages. For example, a data file from the Data Processing stage will be passed to a subsequent Analysis stage. The life cycle model helps monitor both the digital objects bound within a stage and those objects that flow across stages.

This type of model also depicts the wide mix of implicit and explicit partnerships that occurs during research, including relationships among researchers, grant agencies, universities, data producers, scientific publishers, libraries, data repositories and others. New scientific research is stimulated by the intellectual capital, resources and infrastructure brought together through such partnerships and much of today's research is shaped by these interdependencies. The big picture from the life cycle model ensures that the combination of relationships within a project is recognized and well described.

Given a life cycle perspective, what are the key issues of long-term preservation, data management and the curation of data in the context of new partnerships, new organizational models and sustainable economic models?

New Partnerships. The preservation of scientific data is dependent on the custodial care of the digital objects produced throughout the research process. The traditional practice of gathering a paper trail of research outputs long after a scientific investigation has been concluded and depositing them with an archive is inapplicable in the digital era. Too much valuable research data are either at high risk of being lost or have been destroyed because of inappropriate practices that were carried over from a time when paper was the dominant medium. In the

[1] A position statement that I wrote for the ARL E-Science Task Force presents an example of a research life cycle model. See "e-Science and the Life Cycle of Research" (2006) available online at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

[2] Life cycle does not refer to life span, which is the time from birth to death. Rather, life cycle is used to describe the processes within an environment under which resources are formed, transformed and re-used. For a brief summary of life cycle models and references to examples in addition to the one previously cited, see the item by Ann Green, "Conceptualizing the Digital Life Cycle" on the **IASSIST Communiqué** at http://iassistblog.org/?p=26

digital era, the challenge is to coordinate among partners the care of research data throughout the life cycle. Digital custodianship requires clearly articulated roles for the care of the digital objects, including which partner has primary responsibility in each stage of the life cycle.

New possibilities exist for librarians to serve as partners in the life cycle of research. Today, data librarians are on staff in many academic research libraries where collections of data are made available through library data services. For the most part, data librarians are not engaged in primary research[3] being conducted on their local campuses. Instead, they mainly support researchers undertaking secondary data analysis. While this is an important service, the potential for data librarians to be much more involved in activities across the life cycle of research remains untapped. For example, data librarians could contribute significantly to the management of metadata throughout a research project.

Research data require high quality metadata for proper preservation and to be of value for re-use. New metadata standards are emerging that facilitate the discovery and repurposing of data.[4] The enhancement of such standards requires participation by all of the partners in the life cycle of research. With the production of comprehensive metadata based on open standards, new partnerships will be needed to develop open-source tools for mining this rich metadata.

Research libraries should provide access both to the body of scientific literature and to the data upon which this literature is based. The publishers of scientific literature and the providers of library data services need to agree upon standard metadata elements that will facilitate the dynamic linking of data with scientific literature. With such metadata in place, new partnerships can be forged to develop the tools for integrating data with literature.

New Organizational Models. Short-term access to data is often best facilitated by keeping the data in close proximity to its origins. However, long-term access is completely dependent upon thorough preservation practices and standards. One challenge we face is to establish a network of organizations with varying levels of responsibility to span the life cycle of research. For example, a local digital repository may take initial responsibility for providing access to research data but the long-term preservation and access becomes the responsibility of a topical (discipline-specific) or general national repository. Coordinating the division of responsibilities across multiple digital repositories is a major

[3] Original data collection is a defining aspect of primary research.

[4] The Data Documentation Initiative (DDI) is an example of such a metadata standard for survey, aggregate and time series data. Version 3 of DDI introduces a metadata model based on the life cycle of data. For further information, see http://www.icpsr.umich.edu/DDI/

organizational task.[5] A model based on a federation of data repositories is one approach that would address the need for strong organizational coordination.[6]

New data repositories of national prominence need to be launched that take on the long-term responsibilities of preserving data and that work closely in coordination with local repositories responsible for short-term access to data.[7] An open consultation is needed to determine how many of these repositories are required and whether their focus should be general or topical.

The emergence of local digital repositories with recognized responsibilities for the care of research products requires a certification process to ensure best practices and to build a level of trust between researchers and the providers of repository services. The work by the joint digital repository certification task force between the Research Libraries Group and the U.S. National Archives and Records Administration has provided a framework for such a system.[8] One or more organizational homes will be needed, however, to implement a certification process.

Sustainable Economic Models. One of today's most serious threats to science is the commodification of research data, which includes the acts of selling research data at a cost in excess of the Bromley guideline[9] and of inappropriately hoarding data under the pretext of intellectual ownership. Science flourishes in an environment of openness where ideas are exchanged, challenged and tested. This principle of openness also applies to the data upon which research findings are based. The replication of research depends on an open exchange of data. The challenge we face as a scientific community is to find ways of preserving and exchanging research data that are not based on the commodification of research data. If we accept the premise that scientific research data must be a public good, how will the services to preserve and provide access to the data be financed?

---

[5] For a recent, comprehensive discussion of the issues of digital repositories and their applications with research data, see Ann Green and Myron Guttmann, "Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives," 2006. Pre-print deposited in: http://deepblue.lib.umich.edu/

[6] An example of a federation of repositories in the social sciences is the Data Preservation Alliance for the Social Sciences (Data-PASS), which is supported by the Library Congress National Digital Information Infrastructure and Preservation Program. For more information, see http://www.icpsr.umich.edu/DATAPASS/

[7] The argument for new national data archives of prominence has been made by James Jacobs and myself in "Preserving Research Data," **Communications of the ACM**, Vol. 47 (9), pp. 27-29.

[8] For further information about the RLG-NARA digital repository certification task force, see http://www.rlg.org/en/page.php?Page_ID=20769

[9] The concept of pricing data at the marginal cost of reproducing a copy of the data is one of the Bromley Principles, which was published by the Committee on Earth and Environmental Sciences, National Science Foundation in "Data Management for global change research policy statements July 1991" **U.S. Global Change Data and Information Management Program Plan**, Washington, 1992, pp. 42-48.