Managing Collections of Highly Dispersed, Hererogeneous Data

S. Hampton, M. Jones, M. Schildhauer, *National Center for Ecological Analysis and Synthesis (NCEAS), UC Santa Barbara*

Libraries have historically played a critical role in the long-term preservation of scholarly works, especially books and other artifacts written in natural language. Curation of scientific data has challenged the capabilities of existing systems because of the unique characteristics of scientific data and the special services which must accompany archiving of scientific data. Unlike books and similar publications, scientific data are generally intended for analysis, modeling, and visualization rather than reading and browsing. Analysis and modeling activities often require the quantitative integration of multiple data sets from dispersed locations that vary tremendously in their structure and semantics, which creates a need for much more detailed metadata than is available in traditional library usage. Depending on the discipline, scientific data can be small and complex, requiring substantial documentation to accurately interpret the data, or large but uniformly structured, requiring less documentation but presenting system scalability issues. These and other fundamental differences in the way one uses scientific data lead to the need for new partnerships that can effectively provide for simultaneous preservation, discovery, access, integration, and analysis of data.

Heterogeneity and dispersion. Dealing with heterogeneity and dispersion in the context of integration, analysis, and modeling is the major key to successfully building data collections. Disciplines that are strongly bounded along lines of similar information (such as genetics and proteomics) can have highly integrated solutions like GenBank or the Protein Data Bank, yet other fields (like Ecology) can vary widely in the types of information that are necessary within the context of a single study (e.g., a single ecological study may require data from population biology, genetics, hydrology, and meteorology). Consequently, one archival solution might not fit all disciplines, unless that solution provides interfaces that enable both breadth of coverage and depth of resolution within any given discipline. For example, traditional metadata systems used in libraries provide metadata that assists with discovery at a coarse-grained level, but understanding heterogeneous data requires detailed metadata that describes the structure, content, and semantics of data and the protocols used to generate the data. Although metadata standards overlap tremendously, discipline-specific extensions must be created to fully understand and utilize data. Data dispersion can play an important role in structuring collections. Local institutions are typically the best curators of scientific data because they best understand the data collection and quality assurance processes. Although specific versions of scientific data sets are static and can be preserved as-is, data users find errors and omissions that are fixed in subsequent versions, requiring an active curatorial system that dynamically links actions of local scientists to regional, national, and global archives. For libraries to provide an effective archival system for science data, they must build semantically rich data infrastructure that allows direct access to heterogeneous data directly from within analysis and modeling systems used by scientists and that allows for curatorial linkages among data systems from local, regional, national, and global scales.

Long-term preservation. Many disciplines, including ecology, lack a mechanism for assuring the long-term preservation of scientific data. Although some nationally-scoped data archives exist (e.g., the NASA DAAC's, the NODC, etc), many of these are federally funded and are subject to the vagaries and cycles associated with public federal funding. These archive centers tend to focus on archiving data without fully dealing with the difficult and expensive aspects of long-term curation, including the creation and maintenance of new data versions, media migration, and software obsolescence. A partnership of libraries and data centers that each contains replicas of scientific data linked to local mechanisms for data curation and update would be far more durable over the long-term than single, centralized data archive systems.

Data sharing. Despite widespread agreement that sharing data is paramount to the scientific method and essential to synthetic advances that span scales and disciplines, institutional and individual sociological barriers to intellectual rights of data use remain a serious problem. Diverse approaches to preserving data that gradually migrate disciplines into more open and unquestioned sharing of data will benefit science but require new partnerships among scientists, data centers, libraries, scholarly societies, and universities. One approach is to provide incentives to data sharing directly to scientists, e.g., as the Kepler scientific workflow system has done by directly linking analysis and modeling capabilities to data archives and sensor networks. New partnerships that promote generic data access interfaces allow us to build integrated systems that scientists can use to access data archives during the course of analysis and modeling, thereby providing an incentive to share data.