

**Position paper for ARL workshop “New Collaborative Relationships:
The Role of Academic Libraries in the Digital Data Universe”**

MacKenzie Smith, MIT Libraries; September 2006.

There are two aspects of scientific and engineering data that relate to academic libraries:

- Data as *primary source material* available for further research and experimentation, using particular datasets or groups of datasets
- Data as part of “*enhanced publications*” that form the basis of modern, digital scholarly communication

Academic research libraries and archives are closely involved with both of these as a part of their mission and expertise. However, broadening the scope of libraries and archives to include digital scientific research data brings big challenges. There are unanswered questions about the

- technical infrastructure, and who will develop and manage it
- collection practices involving decisions about what data will be kept, when, in what form, with what tools, what description
- digital preservation practices (of unknown difficulty and expense)
- legal framework that is necessary to allow this to happen at all

Libraries and archives will probably not be the primary providers of the large-scale storage infrastructure required. Nor will they provide the specialized tools to work with the data (sometime at the level of individual datasets). They will also not provide detailed information about the data (which falls to researchers, or specialists from their societies and publishers). Nor will they provide the legal framework to enable open science. However to achieve economies of scale *across all scientific research domains* and not just create data silos within particular scientific sub-disciplines, there is value in library practices around

- Collection policies and practices (appraisal, selection, weeding, destruction, etc.)
- Data clean-up, normalization, description, and submission to archives
- Collaboration with researchers around scholarly communication practices of the disciplines (e.g. educating students about these practices, or helping researchers find appropriate archives or publications)

It’s unclear whether libraries will provide the technical solutions to long-term digital data preservation. It is certainly within the mission of research libraries and archives to preserve the scholarly record, but the technical challenges and costs involved are large, and libraries will need to invest seriously in this area if they wish to help find solutions.

Finally, for “enhanced publications” that include scientific data as a useful part of networked documents, there are missing standards that academic libraries are well positioned to help define, including

- Ontologies (for complex publications that include data)
- Identifiers for publication parts that work across disciplines
- Consistent description practices for enhanced publications and their parts
- Data structuring conventions
- Interoperability protocols for searching and retrieving data