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

Position Paper, Eric F. Van de Velde, Caltech

Both large and small research projects produce data of historical value worthy of preservation. Large projects must incorporate data preservation as an essential part of their project. Small projects need institutional support to help them implement high-quality data-preservation policies. Funding agencies should create Centers of Excellence in Preservation and encourage peer review of data sets and associated services and software. Computer- and library-science curricula should include data preservation.

The size and specialized nature of the data of large research projects (Caltech-MIT's LIGO, Human Genome Project…) require that data preservation is considered as an essential component of the project. These projects have the responsibility to manage the data, to make it available with appropriate services, and to preserve the data and the associated software and documentation. Often, the desired outcome is a data set that continues to grow indefinitely, supplemented with data from newer, more accurate, observations and experiments. Terminations of projects like these are relatively rare events that should be handled on a case by case basis as part of the closing-down process, which should be supervised and audited by the funding agency.

Small research projects need institutional help with the work required to comply with preservation mandates. Managing scientific data requires scientific know how at an expert level, and the local research library cannot be expected to handle preservation of data of all disciplines. It might be feasible, however, for a library to specialize in the preservation of data in one or two disciplines. For the rest, the library is the agent between local researchers and specialized data archives.

Funding agencies should fund (distributed) Centers of Excellence in Data Preservation, each specializing in a particular discipline. As part of the competitive funding process, interested institutions would develop collaborative organizational networks capable to implement effective preservation of specific data. This approach allows for organic growth, proportional to the actual needs. This approach also builds on the strengths of existing institutions (universities, research laboratories, and their libraries).

Funding agencies should also provide incentives to accomplish more than "just archiving." Data obtained at great effort and expense should be made available as widely as feasible together with supporting services and software. Peer review of data sets and associated services and software would make it easier to consider this kind of work in tenure and promotion. Under suitable conditions, for-profit organizations could provide services using publicly available data, ensuring the use of this data for society's benefit.

Finally, we must ensure that the talent to preserve scientific data will be available. The preferred approach is to provide incentives for computer-science and library-science departments to include suitable disciplines in their curricula.