

## Workshop on Data-Driven Science & Scholarship: Organizations

This position paper was prepared for the NSF-JISC-CNI Workshop on Data-Driven Science and Scholarship, held April 17-19, 2007, in Phoenix, Arizona. Please restrict use of this informal paper for workshop discussions. It is not intended for wide distribution.

### Position Paper, Eric F. Van de Velde, Caltech

Managing the numerous complex data archives being generated by the scholarly community will impose significant and costly demands on our research infrastructure.

It is too early to predict which organization(s) will be able to step up to the task. It could be supercomputer centers, libraries, dedicated research groups, commercial ventures, or brand-new organizations that combine all necessary talents. However, if we accept the motto of modern architecture and industrial design that form follows function, we should focus on developing the functional requirements of a data-archiving infrastructure and let the appropriate organizational forms emerge from those requirements.

Whatever its ultimate form, the infrastructure for long-term data-archiving will require considerable ongoing resources. When putting in place data-archiving policies, we must question their long-term technical as well as economic sustainability. From a purely technical point of view, the minimal requirement is that future data archeologists can decipher the databases we are putting in the archive today. Giving no weight to the future cost of that endeavor, would lead us to minimalistic data-archiving policies. However, if we do give weight to the future cost of the ongoing management and eventual recovery of data, then it would be reasonable to invest some resources up front if that should lead to a more sustainable future. Unfortunately, we have little experience in long-term data archiving spanning many cycles of technology innovation, considering that the widespread use of information technology spans less than 50 years.

In fact, our historical experience with any archives is rather limited to archival libraries and museums. Neither texts nor artifacts are good models for data stored on volatile media with volatile software. Organizationally, however, we can draw a lesson from these traditional archives. Over time, these archives were able to commoditize their collections, which reduced the need for specialists. For example, most of the day-to-day work in a physics library does not require physicists. In early libraries, this would have been unthinkable: early librarians knew their collections by reading them and guided their users based on their knowledge of the discipline. Over time, libraries transformed their information into commodities (books, proceedings, journals...) that could be managed with limited knowledge of the subject discipline. Of course, library science became a professional discipline in the process.

Can a similar transition be achieved for data archives? Can we turn scholarly data into commodities that may be managed by data archivists that do not necessarily have discipline-level knowledge of the archives? Such a transition would relieve scholars from

the enormous task of managing not only existing data archives but also those that will be created over the next decades and centuries. Such commoditization might be one step that leads to a scalable infrastructure to manage data archives that are likely to be inactive for many years, but can be revived to the contemporary scientific and technological environment when future researchers wish to re-use the data.

Commoditizing data is likely to require a significant ongoing international collaboration to develop and agree on suitable standards for a large variety of possible data archives. At first sight, the complexity of the task is overwhelming. On closer analysis, however, there is significant overlap in some areas, and there are many opportunities for concurrent development in other areas.

The overlap is primarily in the traditional metadata (title, abstract, author, institution, date, subjects, references to the literature...) that are necessary for any information object. There are already (too?) many standards and best practices for these traditional metadata fields. In addition, data archives also need specialized metadata that differ from traditional archives, but are rather similar from one data archive to the next: metadata to describe the technology environment in which the data was originally used (computers, manipulation and visualization software, for example). Finally, the data itself needs to be described according to standards: each value must be associated with its dimension and unit. Mathematical relationships between values need to be described. Database structures need to be described. Here, there might be significant differences between the disciplines (compare geographical and genomic databases, for example). Within the same discipline, there are likely to be many different kinds of archives (compare clinical-survey and drug-interaction databases, for example). However, development on discipline-oriented issues can proceed concurrently, though interdisciplinary cooperation when there is common ground (survey databases, for example) will surely benefit everyone.

To ensure that scholars actually use these evolving standards and best practices, adhering to the standards should be the easiest route to creating a database. Fortunately, we have the opportunity to create a valuable service for scholars. Currently, many databases are created in an ad-hoc manner that is labor intensive and duplicative. By developing preconfigured blueprints for a wide variety of possible scholarly databases, bundled together with advanced visualization tools, we can improve the scholar's productivity in creating and managing their databases. When time comes to convert these into long-term data archives, the same software can produce archives according to the latest standards and best practices. One could even anticipate that these data archives could be refreshed to evolving technology and standards by subsequent generations of the software.

This proposed agenda is definitely a tall order. However, once the basic framework is developed, this is a project that many groups can work on simultaneously. For example, it would be reasonable for funding agencies to mandate that larger research projects take on some development of standards, best practices, and software for the management and visualization of data archives in their discipline. The work of these projects should be coordinated and packaged for use by individual researchers.

As this research proceeds, it is likely that various organizations will emerge to participate in the development and management of data archives. In the beginning, special emphasis should be put on support for the individual researcher. Mandates to preserve data are a disproportionate burden on the individual researcher and small research projects. They deserve special consideration and support. We should help individual researchers and small research groups achieve acceptable data preservation with minimal overhead to the researchers.

To coordinate and support the disparate data-archiving activities of researchers, funding agencies should establish (potentially distributed) Centers of Excellence in Data Preservation. Each could specialize in a particular discipline, but all would be required to participate in global standardization efforts. The competitive funding process allows for organic growth of new organizations, created by building on the strengths of existing institutions (universities, research laboratories, and their libraries).

Funding agencies should also provide incentives to accomplish more than archiving. Data obtained at great effort and expense should be made available widely together with supporting services and software. Peer review of data sets and associated services and software should be encouraged. Under appropriate conditions and restrictions, for-profit organizations could provide services using publicly available data, ensuring the widest possible usage of scholarly data for society's benefit.