

NSF* Workshop on
Interfaces to Scientific Data Archives
Pasadena, California, 1998 March 25-27

Organizer
Roy Williams
Center for Advanced Computing Research
California Institute of Technology
roy@caltech.edu

Many scientific endeavors produce large quantities of heterogeneous data that is to be analyzed by loose, distributed collaborations. There is a call for federally-funded data to make itself useful through its availability to those who are not experts in the meaning of the data. Scientific data, in contrast to text or image data, is often useless without sophisticated, customized data-mining and knowledge extraction tools. Given these three conditions, there is an urgent need for software infrastructures to create, maintain, evolve, and federate these active digital libraries of scientific data; infrastructures that consider the newcomer learning to use the system as well as the seasoned professional.

The objective of this workshop was to examine approaches to such active digital libraries through case-studies and tools. The interaction of these illustrates needs for standards and abstraction, identifies similarities, focuses on real-life problems, and thus curbs the excesses of theory. Using small-group discussion, we identified solutions, consensus, and challenges in creating and maintaining active digital libraries of scientific data, for ensuring that the archive is flexible and extensible, that it is as easy as possible to learn how to use, and so that different groups can use each other's development work instead of repeating it. The full report from the workshop is available from the web site below, with findings, recommendations, descriptions of case-studies and tools, and a survey of scientific data archives.

Metadata is an important topic: how to describe data objects, to make catalogs, to form relationships between data objects. An effective way to do this is by using structured documents to describe large binary objects; machines such as search engines and summarizers can then parse the document. Such structure can be provided with the XML language, a rationalized and extensible version of HTML. The Dublin Core metadata standard is an effective and viable way to provide the semantics and structure of these metadata records.

The Web is universally seen as the *lingua franca* for the client who is new, or who does not have special software installed, or who wishes to access the archive from an arbitrary machine. In addition to this important role, the HTTP protocol increasingly provides communications between machines as well as between machines and people. Web servers are also being used as brokers, providing unified access to databases, legacy systems, supercomputers, and data archives.

There were discussions on distributed computing, the major contenders being CORBA and Java RMI; we also contrasted the simplicity of the relational database with the flexibility of the object database. We contrasted text-based interfaces with graphical interfaces: while mature users prefer the speed and flexibility of text, novices prefer a GUI, therefore we should concentrate attention on how a user can reuse what has been learned as a novice when it is time to advance to the text-based world of the mature user. Authentication, security and signatures were important issues, particularly ways to bridge the gap between traditional Unix mechanisms and newer, commercial solutions such as digital certificates. Other topics are: longevity of the archive, who are the librarians, fostering collaboration through interoperation of archives, deep citation, and data-driven computing.

The full report and supporting material may be found at <http://www.cacr.caltech.edu/isda>

*. This workshop was sponsored by the National Science Foundation, under the grant IIS-9803760 (PI: Roy Williams) awarded by the Information and Data Management Program and the Special Projects Program of the Information and Intelligent Systems Division. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the participants, and do not necessarily reflect the views of the National Science Foundation.