Data Management Systems for Digital Libraries

Reagan W. Moore
San Diego Supercomputer Center
August 4, 2000

Digital libraries now have data management requirements that extend beyond the local computing environment. Distributed digital libraries are being created that incorporate collections from multiple autonomous sites. SDSC has developed data management systems to facilitate creation of distributed data collections. The associated infrastructure includes persistent archives for managing technology evolution, data handling systems for collection-based access to data, collection management systems for organizing information catalogs, digital library services for manipulating data sets, and data grids for federating multiple collections. The infrastructure components can be characterized as interoperability systems for digital object management, information management, and knowledge management. Examples of the application of the technology include distributed collections and data grids for astronomical sky surveys, high energy physics data collections, and art image digital libraries.

The creation of international digital libraries will require support for data that are distributed across multiple administration domains and are stored on heterogeneous storage systems. The challenge is to facilitate the organization of these data resources into collections without compromising local control. At the same time, middleware is needed to support uniform access to the data sets, including APIs for direct application discovery and manipulation of the data, command line interfaces for accessing data sets from scripts, and web GUIs for interactive browsing and presentation of data sets.

The development of infrastructure to support international digital libraries must recognize that information repositories and knowledge bases are also needed. One can differentiate between infrastructure components that provide:

- Access to storage systems that hold the digital objects,
- Information repositories that store attributes about the digital objects. The attributes are typically stored as metadata in a catalog or database.
- Knowledge bases that characterize relationships between sets of metadata. An example is rule-based ontology mapping that provides the ability to correlate information stored in multiple metadata catalogs.

A digital library will need to support ingestion of digital objects, querying of metadata catalogs to identify objects of interest, and integration of responses across multiple information repositories. Fortunately, a rapid convergence of information management technology and data handling systems is occurring for the support of data collections. The approach used at the San Diego Supercomputer Center is to organize distributed data sets through creation of a logical collection. The ownership of the data sets is assigned to the collection, and a data handling system is used to create, move, copy, replicate, and read collection data sets. Since all accesses to the collection data sets are done through the data handling system, it then becomes possible to put the data sets under strict

management control, and implement features such as access control lists, usage audit trails, replica management, and persistent identifiers.

Effectively, a distributed collection can be created in which the local resources remain under the control of the local site, but the data sets are managed by the global logical collection. The data handling system serves as an interoperability mechanism for managing storage systems. Instead of directly storing digital objects in an archive or file system, the interposition of a data handling system allows the creation of a collection that spans multiple storage systems. It is then possible to automate the creation of a replica in an archival storage system, cache a copy of a digital object onto a local disk, and support the remote manipulation of the digital object. The creation of data handling systems for collection-based access to data sets makes it possible to automate all data management tasks. Data set handling systems can be characterized as interoperability mechanisms that integrate local data resources into global resources. The interoperability mechanisms include

- inter-domain authentication,
- transparent protocol conversion for access to all storage systems,
- global persistent identifiers that are location and protocol independent,
- replica management for cached and archived copies,
- container technology to optimize archival storage performance and co-locate small data sets, and
- tools for uniform collection management across file systems, databases, and archives.

Data Handling Infrastructure:
The data management infrastructure is based upon technology from multiple communities that are developing archival storage systems, parallel and XML database management systems, digital library services, distributed computing environments, and persistent archives. The combination of these systems is resulting in the ability to describe, manage, access, and build very large distributed scientific data collections.

The ability to manipulate data sets through collection-based access mechanisms enables the federation of data collections and the creation of persistent archives. Federation is enabled by publishing the schema used to organize a collection as an XML DTD. Information discovery can then be done through queries based upon the semi-structured representation of the collection attributes provided by the XML DTD. Distributed queries across multiple collections can be accomplished by mapping between the multiple DTDs, either through use of rules-based ontology mapping, or token-based attribute mapping.

Persistent archives can be enabled by archiving the context that defines both the physical and logical collection organization along with the data sets that comprise the collection. The collection context can then be used to recreate the collection on new database technology through an instantiation program. This makes it possible to migrate a collection forward in time onto new technology. The collection description is instantiated on the new technology, while the data sets remain on the physical storage resource. The collection instantiation program is updated as database technology

evolves, while the archived data remains under the control of the data handling system. As the archive technology evolves, new drivers are added to the data handling system to interoperate with the new data access protocols.

Application:
A collection-based data management system has the following software infrastructure layers:

- Data Grid – for federation of access to multiple data collections and digital libraries
- Digital library - provide services for discovering, manipulating, and presenting data from collections
- Data collection – provide support for extensible, dynamically changing organizations of data sets
- Data handling system – provide persistent IDs for collection-based access to data sets
- Persistent archive – provide collection –based storage of data sets, with the ability to handle evolution of the software infrastructure.

The essential infrastructure component is the data handling system. It is possible to use data handling systems to assemble distributed data collections, integrate digital libraries with archival storage systems, federate multiple collections into a data grid, and create persistent archives.

References:
Moore, R., C. Baru, A. Rajasekar, B. Ludascher, R. Marciano, M. Wan, W. Schroeder, and A. Gupta, "Collection-Based Persistent Digital Archives - Part 1", D-Lib Magazine, March 2000, http://www.dlib.org/

Foster, I., Kesselman, C., "The Grid: Blueprint for a New Computing Infrastructure," Chapter 5, "Data Intensive Computing," Morgan Kaufmann, San Francisco, 1999.

Baru, C., R, Moore, A. Rajasekar, M. Wan,"The SDSC Storage Resource Broker," Proc. CASCON'98 Conference, Nov.30-Dec.3, 1998, Toronto, Canada.

NPACI Data Intensive Computing Environment thurst area, http://www.npaci.edu/DICE/

Moore, R., C. Baru, P. Bourne, M. Ellisman, S. Karin, A. Rajasekar, S. Young, "Information Based Computing," Proceedings of the Workshop on Research Directions for the Next Generation Internet, May, 1997.