DIGITAL LIBRARY
FEDERATION

University of California, Berkeley

Report to the Digital Library Federation
Fall, 2003

Table of Contents

I. Collections, services, and systems

A. Collections

The UCB Library has created or collaborated in the creation of nearly 200 collections of digital content. An inventory of these collections, with brief descriptions and links to the collections themselves, is maintained by the Digital Publishing Group (DPG), a unit of the Library Systems Office. It is available on the DPG Web site at the following URL:
http://dpg.lib.berkeley.edu/scripts/webdb/dpg/collbrw.plx

B. Services

Remote access

The UC Berkeley Library provides a proxy service by means of which members of the UC Berkeley community can access licensed content from off-campus locations. Users authenticate using either their library card number and a PIN, or their CalNet ID and passphrase. CalNet authentication is the campus standard for Web applications and uses the UCB Authentication Web Server, a Kerberos-based service developed and maintained by the Systems and Network Security group. The Library's proxy service is heavily used, handling several million documents per month.
http://proxy.lib.berkeley.edu

C. Systems

GenDL

The UC Berkeley Library began implementing GenDL, its digital library support architecture and system, in conjunction with the DLF sponsored *Making of America II* initiative on which it was the lead institution. This initiative developed the first standard encoding for library digital objects (the MOA2.DTD) against which the library could develop such a suite of digital library tools. Now that the METS schema, which builds on the MOA2 work, has superseded the MOA2.DTD, the UC Berkeley Library has been working to convert its MOA2-based tool suite to produce and present METS objects. It has also been enhancing GenDL in ways that are independent of the core encoding standard.

The GenDL tool suite consists of 4 components: GenDB, for gathering the raw structural, descriptive and administrative metadata pertaining to digital materials, GenX for creating standards-based digital objects from the raw metadata, GenSearch for searching the library's repository of digital objects, and GenView for viewing and navigating the library's digital objects.

As first implemented, GenDB was built on MS Access, and the Library Systems Office produced a separate Access-based GenDB database for each digitizing project and project participant. Over the past 2 years the Library Systems Office has worked to implement a version of GenDB with a centralized database, and an intuitive and highly configurable Web interface. This work is now substantially complete, and the OAC administered California Cultures project is actively using the new version of GenDB. The University of California History Digital Archives, and Environmental Design Archives will begin using this new tool over the next couple of weeks. As Finding Aid support is added, and project management features are improved, WebGenDB is expected to support all UC Berkeley digitizing projects. WebGenDB/GenDB is largely neutral with respect to encoding standards. However it has also been adjusted better to support METS, MODS and MIX output now that these are emerging as the primary standards for target encodings.

GenX, which extracts the raw metadata from the GenDB database to produce standards-based digital objects, previously produced MOA2 objects. A recently revised version of GenX now produces METS objects, with MODS encoded descriptive metadata, and MIX encoded image technical metadata. The new GenX cannot be finalized, however, until appropriate schemas for technical metadata and rights metadata emerge and/or mature.

GenSearch is the least developed component of the Library's tool suite. The Library Systems Office has performed fairly extensive testing of an XML database (Tamino) as a possible support for digital library searching functions. However Tamino's support for XML Schema in general, and METS and MODS in particular, is currently immature and requires that the schemas be rewritten to be loaded. Recent GenSearch research has focused on Cheshire II, and the forthcoming Cheshire III.

GenView has provided library patrons with the ability to view MOA2 objects since the conclusion of the MOA2 initiative. A new version of GenView recently put into production supports METS objects as well as MOA2 objects. Library Systems staff members have also been working substantially to redesign the GenView interface to make it more intuitive and user friendly. It will begin implementing these enhancements in the very near future.

XML Search Tools

With the increase in number of digital collections hosted by the UCB library, providing our patrons with a reliable and efficient search tool has become more important. It has also been essential to find one compatible with the METS standard now used to describe objects' technical, administrative, and descriptive metadata. The SAG XML database Tamino was tested for the purpose, but its inability to handle the full METS schema without modifications prompted us to look at other products, including Cheshire II, an XML search engine developed at UCB's School of Information Management and Information Management and Systems by Ray Larson.

Tests performed to date include the indexing of a million MODS records, which showed a excellent performance for retrievals. The user interface API made it easy to create a web search page. Compatibility with METS was also good. Particularly interesting is Cheshire's support of both Boolean and probabilistic "best match" ranked searching and browsing. Unicode support is not available yet in Cheshire II, but it is expected in the next release, Cheshire III, which is currently under development. Additional information can be found at the Cheshire II Web site.
http://cheshire.berkeley.edu/

II. Projects

Digital Publishing Projects

The Digital Publishing Group (DPG) manages and provides technical and other forms of support for a wide variety of UC Berkeley’s digital publishing projects. The DPG maintains an alphabetical list with brief descriptions of current projects at the following URL.
http://dpg.lib.berkeley.edu/scripts/webdb/dpg/project.plx/

LOCKSS at Berkeley

The UC Berkeley Library is participating as a tester for Stanford University’s LOCKSS system. LOCKSS stands for "Lots of Copies Keep Stuff Safe” and the goal of the project is to create low-cost, persistent digital caches of authoritative versions of Web-delivered content. The LOCKSS software enables institutions to locally collect and archive content, while enforcing the publishers’ access control systems and without harming their business models. The accuracy and completeness of individual LOCKSS caches are assured through a peer-to-peer polling system which is both robust and secure. LOCKSS replicas cooperate to detect and repair preservation failures. LOCKSS is designed to run on inexpensive hardware and to require almost no technical administration. The software has been under development since 1999 and is now distributed as open source. Further information is available at the following URL:
http://lockss.stanford.edu/