The UC Libraries' Digital Preservation Program: aims, overview, and initial priorities

California Digital Library
February 6, 2004

1. Aims and overview

The UC libraries' digital preservation program was established in 2002 to ensure long-term access to selected digital information that supports and results from research and teaching at UC.

To achieve these aims, the program will develop and maintain a standards-based infrastructure comprising a preservation repository and a suite of tools that support the identification, capture, description, organization, presentation, and persistent management of digital scholarly information.[1] In order to enable the successful use of the tools the infrastructure will also include implementation guidelines, support services, and guidance to collection planning, rights management, and other aspects that are key components of the preservation process. Access tools will be available but limited as appropriate to a program established to retain digital information for future generations. The program will not provide "failover" or backup services.

The infrastructure will be used by the CDL to capture and persistently manage digital information in which the UC libraries have a common interest. In addition, campus libraries, acting individually or in groups, may also use the infrastructure to capture and persistently manage digital materials in which they have a local interest.

The infrastructure will be designed to be scaleable so that it can grow to accept increasing volumes of digital information. It will also be extensible so that UC's libraries can cost-effectively manage a growing variety of digital materials including:

- the published record (whether produced and distributed as journal articles, monographs, working or conference papers);
- web-based information content such as the growing volume of Internet-only publications produced by US state and federal governments;
- primary evidence including materials that are "born digital" (e.g. datasets produced by scholars in the course of their research)

[1] The role of standards will be critical. Accordingly, the program will conform to evolving standards including the Reference Model for an Open Archival Information System (OAIS) published by the Consultive Committee for Space Data Systems (CCSDS). The program's credibility as a "trusted" third-party preservation facility will also be crucial in encouraging data owners/distributors to deposit their content with it. Accordingly, the program embraces principles developed initially by CLIR, CNI, and DLF and elaborated in the RLG-OCLC Report, "Trusted Digital Repositories: Attributes and Responsibilities".

- primary evidence including digitally reformatted content (e.g. digital surrogates for analog items held by UC libraries, archives, and museums); and
- online teaching and learning materials (produced by UC scholars for use in instruction).

This document provides an overview of the program's initial 24-month development phase during which the program will:

- design, build, and evaluate core components of the preservation infrastructure;
- gain practical experience preserving a selected array of digital materials;
- gain practical experience of the different uses which the program's infrastructure will be employed by the CDL and campus libraries;
- gain practical experience building partnerships with content producers;
- gain a better understanding of program costs and the business and service models that are best able to meet them; and
- identify and evaluate sustainable funding models for the UC libraries' digital preservation efforts.

2. Initial program priorities

The program's cost and its sustainability are particular concerns that substantially shape the initial developmental phase. Although impossible at this stage to predict, it is clear that the program's costs will far outstrip its annual budget that is derived from centrally managed systemwide preservation funds and amounting to $180,000 *per annum*. For context, service-provider maintenance of and support for the preservation program's computer infrastructure costs a minimum of $125,000 *per annum*. Additional funding for the program has been found in the CDL's general operational budget (the CDL is contributing c.$850,000 in 2003/2004 for the initial development of the preservation repository and hopes to continue subsidizing the program[2]) and in awards from philanthropic and federal funding agencies (to date, the program has benefited from

[2] The design and initial build out of the preservation repository are detailed in the table below. Note that costs not included in the table include those associated with:

- design and development of limited access tools ($250,000)
- practical work preserving selected digital collections (NA)
- hardware and maintenance required to preserve collections selected for ingest during the program's development phase (estimated at c.$700,000)

Table 1. Costs involved in the design and initial build-out of the preservation repository

Resource	**System Design Phase:** ***September 03 – December 03***	**Development Phase:** ***January 04 – August 04***
Staffing	$55,200	$592,100
Training	0	$30,000
Hardware/Storage/Software	0	$170,000
Operational Costs	0	$10,000
Total Phase Costs	$55,200	$802,100
Total Cost for design and development phase	$857,300	

c.$500,000 combined in external grant funding from The Andrew W. Mellon Foundation and the Institute of Museum and Library Services). Although external funding is likely to continue in its importance to the program, operational support will require permanent core funding as appropriate for the management of assets that are distinctive to and the primary archival responsibility of the University of California. Even with these additional revenues, the program's budget is not elastic. It is essential therefore that the program:

- prioritize its efforts;
- pay careful attention to the ongoing commitments it incurs as it adds digital collections to the repository;
- ensure that scarce systemwide funding (including funding derived from the CDL's operational budget) is used to preserve shared or systemwide digital collections; and
- develop an infrastructure that campus libraries can, with additional local investment, use cost-effectively to build and manage archives of local digital content.

Program priorities during the initial development phase reflect these concerns and are described in greater detail below.

2.1. Development of an extensible, distributed digital preservation repository

The preservation program relies on the development of a robust and reliable preservation repository that can be replicated (e.g. at UC campus libraries to cost-effectively enable local preservation efforts and provide appropriately redundant storage) and extended (e.g. to add storage capacity and new functionality as required). To increase confidence, enhance development, and lessen implementation costs, the architecture and source code will be open source and built on known standards. To reduce implementation costs, it will be built by connecting existing system components.[3]

The primary functions supported by the preservation repository are simple: acquire (ingest); persistently manage; store; and provide access. The ingest function combines an incoming object, which may consist of multiple files, together with important metadata, and puts them in a METS (the XML-based Metadata Encoding and Transmission standard) wrapper. Next the object is named in such a way so as to support mechanisms to persistently identify -- for our purposes an ARK identifier is created and bound to the object. An ARK is a special kind of URL that connects users to three things: the named object, its metadata, and the provider's promise about its persistence. The object is then

[3] Generally speaking, the interfaces and components rely on concepts from the OAIS reference model (The Open Archival Information System Reference Model is as an ISO Draft International Standard; see http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html). The repository design uses METS (the Metadata Encoding and Transmission Standard that is becoming widespread amongst digital libraries as a means of recording information about digital objects - http://www.loc.gov/mets) and ARK (Archival Resource Key - Kunze, John A. Towards Electronic Persistence Using ARK Identifiers. *3rd ECDL Workshop on Web Archives Semantic Web.* Trondhiem, Norway, August 21, 2003. http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze) a robust naming scheme necessary to identify objects in the repository . Shibboleth (a project of the Internet2 MACE group – http://shibboleth.internet2.edu/) will likely inform our authentication strategy (as required to ensure appropriate use of archived contents).

pushed into storage. The access function accepts an ARK identifier, and returns the object bound to it. Most of the complexity resides in developing routines capable of ingesting digital objects with different characteristics (electronic journals, scientific databases, PDF files, geographical information systems), enveloping each in a METS wrapper that provides sufficient information about the object and the terms and conditions that surround its use to ensure that it can be located and rendered intelligibly by appropriate end-users. The need to develop and automate low-cost, efficient, and re-usable ingest and data access routines for commonly encountered digital objects underpins collection development priorities.

The design rests on grid of storage servers that may be distributed (e.g. at UC campus libraries). Each grid is composed of "bricks" which are low-cost computers with attached storage; the computer and its storage fits in a small, uniform physical chassis that can be "stacked" (rack- or shelf-mounted) in modest-sized and widely available machine rooms. The bricks in a grid are tied together by a database server that allows them to be viewed as a coherent collection of storage units. The database server is essentially a gateway between a collection of digital objects (as may be developed by the CDL on behalf of the UC libraries, or by a library on behalf of a local constituency) and the repository.

For low-level storage, the repository will rely upon the open-source (to academic, research, government, and non-profit organizations) Storage Resource Broker (SRB) system developed by the San Diego SuperComputer Center (SDSC).[4] The SRB has been selected because it is easily extensible and supports the distributed repository infrastructure described above. The distributed approach not only empowers campus libraries to meet their distinctive preservation requirements at low costs, but also builds in a high degree of geographic replication of archived data content. The CDL or another UC library can configure its SRB grid so that when new data are ingested into one site, this causes an automatic ingest of the same data into one or more other sites.

The design of the preservation repository took place July-December 2003. The initial development phase will take place January-August 2004.

2.2 Design and develop an initial (minimal) suite of access tools

The preservation repository is intended to retain digital content so that it may be rendered intelligibly on future generations of computer hardware, software, and network environments. It is not a “failover” or a backup service and will not provide instantaneous and fully featured individual end-user access to digital information content when access services for that content are no longer operable. The rationale underpinning this approach is purely economic. The cost involved in maintaining the full range of end-user services required for all of the digital materials destined to be stored in the repository is prohibitive.

Access mechanisms built around the repository will be simple. They will include:

[4] See http://www.sdsc.edu/DICE/SRB/Pappres/Pappres.html for recent publications and presentations on SDSC’s Storage Resource Broker.

- an administrator's interface to the tools required to capture, ingest, and populate the preservation repository and, where necessary, to gain authenticated wholesale access to ingested collections (e.g. via ftp); and
- an end-user interface designed to disclose rudimentary information about what collections are stored in the repository and/or in repositories (e.g. at UC campus libraries) with which its contents may be federated.

The design of repository *access* services is currently underway and will proceed through July 2004. Initial development work on the access services is anticipated to begin August 2004.

2.3 Capture, ingest, and preserve selected digital collections

Selected practical uses of the repository infrastructure are crucial to the program's development. They enable the program to:

- gain practical experience with commonly encountered digital materials (e.g. online journals, digital image collections, etc.);
- develop generalizable routines capable of preparing (e.g., wrapping in METS) and ingesting digital objects with different characteristics (electronic journals, scientific databases, PDF files, geographical information systems), and preparing them for ingest into the repository;
- develop an appropriate policy environment and rights framework that will enable the program to manage digital objects persistently while protecting the rights and interests both of data owners and data users;
- gain practical experience of the different uses to which the program's infrastructure may be put (e.g. by the CDL and by the campus libraries); and
- gain a better understanding of program costs and the service and funding models that will work to offset them.

At present, the following collections, listed in priority order, are being considered.

2.3.1. Online commercial journals. In support of faculty research and teaching, the UC libraries have developed very substantial collections of online scholarly journals at an annual cost of more than $20 million. Unlike collections of paper journals that remain on library shelves even after subscriptions to current materials are cancelled, online journals are highly at risk. Although the UC libraries have succeeded in including perpetual access clauses in their subscription licenses, they lack the means of implementing the clauses should the need arise. Even though some progress is evident in the industry, journal publishers are no better prepared than libraries to persistently manage online journal content.

Accordingly, the University Librarians have assigned the highest priority to developing the means of persistently managing online journals. To this end, the CDL is in discussion with some of the largest journal suppliers, notably Kluwer and Wiley, about partnerships through which the libraries will preserve the vendor's online journal publications. Discussions are most advanced with Wiley (publisher of InterScience). At present, the CDL is receiving sample files from Wiley and working out mechanisms for automatically acquiring them for inclusion in the preservation repository. Along a parallel track the CDL is exploring the business and licensing terms under which its preservation efforts would be conducted with Wiley.

Longer term, we recognize that the UC libraries license digital information content from upward of 100 vendors and it will neither be affordable nor practical for them to acquire and persistently manage the sum total of the digital files with which they are supplied. The publisher's readiness to work with the CDL on preserving their content and the scale of their journal database will initially drive what journal collections will be preserved. In the longer term other priorities will be required. For example priority may be given to high-quality scholarly journals that are available in:

- digital files to which UC access is imminently threatened, for example, by the termination of a licensing agreement
- digital files to which access is not imminently threatened but are:
    - produced by publishers that adopt business models that promise reasonable-, low-, or no-cost access;
    - deemed to be at particular risk (for example, because no other responsible entity – which may include the vendor – is making any thoughtful effort to persistently manage them);
    - acquired into the preservation repository as part of a strategic alliance between the UC libraries and the publisher

In light of these considerations, priorities for the preservation of licensed content may be informed by (and may in turn inform) the strategies developed by those responsible for adding commercially-supplied content to the shared digital collections.

2.3.2. Data managed by the CDL and made available through systemwide services. The CDL acquires and/or manages access to an ever-growing volume of digital information that is made available without restriction and without access charges to individual end users through a range of services. With many of these services (the eScholarship Repository, the Online Archive of California, Museums in the Online Archive of California, Counting California), the CDL actively manages the digital files, even as they may reside (and be accessible from) elsewhere. With other services (the public website, the image collections) only a portion of the files are managed by the CDL. The remainder are managed by third parties and integrated virtually (with some federation technique) into the service.

The digital information that the CDL manages is highly diverse with regard both to data and metadata formats and to the level and richness of accompanying documentation.   To preserve the full corpus of material that is managed by the CDL would require a phased development and application of an equally wide range of ingest routines and infrastructure capacity. Preservation repository content will be a subset of the files managed by CDL; all content ingested into the repository must meet minimum requirements (e.g. with regard to data and metadata formats, and access). Given the other pressing demands that will be made on the preservation program's limited financial, technical, and staff resources, the CDL will prioritize its efforts with these materials concentrating primarily on those:

- that have been acquired from third parties for integration into systemwide digital library services in part based on the CDL's commitment to archiving;
- that are essential to the maintenance and integrity of existing CDL services and deemed at risk because no other institution claims responsibility for or ownership over them; and are
- acquired from third parties willing to contribute at-cost services for preservation.

2.3.3. Development and management of special archival collections. From time to time the UC libraries may identify a common or pervasive need to capture and persistently manage third-party digital information that is publicly accessible, essential for their users, and significantly at risk of loss. To date, the UC libraries have identified the web-based materials produced by US state and federal governments as potentially comprising such a collection. Many of these materials are available exclusively online; to date no lasting institution claims primary archival responsibility for them; and they contribute an essential and growing component of the important government document collections that have been developed at 7 of the UC libraries. In time, the libraries may identify other collections of web-based materials in which most, if not quite all, share a common preservation interest.

To develop capacity with these types of materials, the preservation program will develop tools that will enable the UC libraries to identify, capture, describe, organize and ingest into the preservation repository a selected collection or collections of web-based materials, and to provide some minimum access to the collections. Building on work of the Libraries' Government Information Task Force and that of the CDL (funded by The Andrew W. Mellon Foundation and in which the required tool suites were largely defined[5]), the program is likely to focus on web-based materials produced by U.S. state or federal governments, by political parties, or by political movements or political campaigns. Over the long term other priorities will be required. So-called web archiving is expensive and complex and can only be undertaken by the CDL in very specific circumstances, for example where the archived material:

- contributes essentially to systemwide digital library services maintained by the CDL; or
- is deemed as a vital component of the UC libraries' campus collections (for example the CDL's capture of web-sites related to the 2003 California gubernatorial recall election).

2.3.4. Preservation of campus-based digital collections. The infrastructure developed by the preservation program will be available to (indeed, enriched through use by) campus libraries that will use the infrastructure to preserve materials in which they have a distinctive archival interest. Such materials proliferate on the campuses. In addition to the online materials in the shared digital collection, materials in this category include online catalog records, finding aids, and third-party databases that are commercially available and acquired or licensed by campus libraries. . In time, some campus libraries may

[5] *Web-Based Government Information: Evaluating Solutions for Capture, Curation, and Preservation. An Andrew W. Mellon Funded Initiative of the California Digital Library* (November 2003) available from http://www.cdlib.org/programs/Web-based_archiving_mellon_Final.pdf

extend their digital purview by taking archival responsibility for selected collections of web-based materials or for online research and learning materials that are produced by campus faculty.

This section describes the variety of ways in which campus libraries will use tool suites in different combination to meet local digital preservation requirements. It is important to note that the tool suites may be used in any combination. Accordingly, use cases presented below are not mutually exclusive and may in fact be combined as hybrid models. It is also important to realize that the costs associated with each of these different use cases are unknown and accordingly impossible at this stage to compare. Still, given the program's scarce resources, it is anticipated that costs over and above those incurred by the program through its provision of a basic preservation infrastructure, would have to be met by the campus library.

- ***Replication of local archives***. Some campuses may build their own local repositories. UCSD, UCLA, and UCB have expressed interests in this direction. In such cases, the campus would be responsible for building (or outsourcing) and managing its own repository, selecting content for inclusion in it, formatting and documenting content according to likely future needs, developing and implementing appropriate data migration or emulation strategies (to ensure its persistence), and negotiating terms and conditions of access with content owners and/or distributors. The preservation program's repository would, in this case, support replication and redundancy of the campus's persistent collections. The program's preservation repository is being developed with such replication in mind. In a mature operational environment, new material ingested into a campus-based repository would trigger the automatic replication of the new content at the program's repository. Similarly content removed from a campus repository would automatically be removed from the program's repository.

- ***Storage of campus archives***. In some cases, a campus library may wish to use the program's repository as primary filestore for the digital materials that it preserves. Here, too, the campus would be responsible for developing and ensuring persistent management of digital collections and for managing access to them. It would not, however, incur the full costs involved in building and maintaining a preservation repository. Instead, it would use the program's repository as something of a data warehouse. The program's repository would in this case accept campus-supplied content and ensure its integrity. On request the program would return the content as supplied to the depositing campus library. The depositing campus library would take sole responsibility for ensuring that the content is renderable on future generations of computer hardware and software.

- ***Preservation of campus content***. In some cases, campuses may build and manage digital archives that are *preserved* by the program rather than simply replicated or stored there. In such cases, the campus would use the program's repository to store its preserved collections but also to migrate those collections through changing technical regimes. Because the program would in this case inherit responsibility for the content's persistent management and migration, it would necessarily require that the campus-deposited content meet minimum specifications, with regard to format, metadata, and access terms. The campus would be free to determine what acquisition strategies and data processing routines most economically and effectively ensure that its content meets the minimum requirements.

- ***Capture and processing of campus content***. In this scenario, a campus may use data capture and/or ingest routines developed or maintained by the preservation program to build its own archival collections and to prepare them for storage in a preservation repository (whether located at the CDL, the campus, or some third party). The advantage to the campus is that it is able to re-use tried and tested data capture and ingest tools. There may also be disadvantages. Existing tools may require modification to meet the campus's requirements. In some cases, the campuses may be building archives with content for which the program has no adequate data capture or ingest tools, or even experience. For example, the program does not anticipate early involvement with digital film. A campus interested in building a persistent collection of digital film would therefore be unlikely in the near term to benefit from tools developed or supported by the program. In such cases, the campus may develop its own tools or embark on a co-development effort with the program with a view to developing generalizable experience and tools that may benefit other UC libraries in future and the preservation program more generally.

In the program's initial development phase, campus-based initiatives will be selected on the basis of the following criteria:

- Work will help broaden the range of extensible ingest tools and data processing routines. In this regard, the program is likely to prefer campus-based initiatives that extend experience with varying data formats and types of digital materials.
- Work will enable us to learn about, evaluate, and assess the costs involved respectively to the campuses and the CDL in supporting the different kinds of uses that campuses may make of the preservation program.
- Work will involve campuses that are ready and able to co-invest in the essential development effort that it will require.