

Brain Dump Paper #3

**System Architecture Considerations  
for  
Digital Archival Repository Services**

*DRAFT -- 11/4/97*

**Bernie Hurley  
Chief Scientist, UC Berkeley Library**

Campus academic departments, libraries, museums, and archives are engaged in a broad range of projects to produce digital materials for use in research, instruction and for public access. As each projects proceed and the number of digital objects increase, a need is being created for campus supported archival repository services to ensure these electronic collections are preserved for future generations of scholars and students. This paper discusses a proposal to investigate the creation of a digital archival repository that will help libraries better understand the services, technologies and costs related to running an archival repository.

This paper develops a concept for repository services that separates the *access* and *archival* components of standard repositories. *Archival repositories* will hold the *master copy* of each digital object, including the best version of the content and all related metadata. The archival repository will support an *extraction service* that will allow metadata to be queried and digital objects to be selected for export from the repository. Some version of these exported objects will then be loaded into an *access repository*, which is part of an *access system* that will provide the appropriate search, display and navigation features. Note: the archival repository's *extraction service* is actually an specialized access system that allows one to determine what objects are contained in the archive.

Within this framework:

***Archival Repositories Become Manageable***

The function of an archival repository is to store master copies of digital objects in a standard manner that minimizes problems related to the migration of these objects to new technologies over time. The key to realizing this goal is the development of best practices that define the metadata, content and encoding standards used for each class of digital object in the repository. For example, it will be much simpler to migrate

digital books to a new technology if all these books contain the same metadata elements and are encoded in a standard format.

However, mixing access services into an archival repository creates many problems, as different communities require specialized access services to meet their specific needs. For example, access services required to support scholars will vary greatly from those desired by K-12 students. Providing access services from the archival repository becomes impossible, as it cannot scale to provide every access need for every community. Even worse, an expectation is set that the access services provided from an archival repository will be perpetuated over time. Therefore, removing the access component from the archival repository makes these repositories manageable, as their primary mission is to move digital objects forward, through time and new technologies.

### *Access Repositories Become More Flexible*

Communities that have the need and resources will build *access systems* around the access repositories they create. Each access system will be customized to fit the search, display and navigation services required by that community. It may also support special tools required by the community for manipulating the digital objects. The access repository and system can be more flexible, in that it does not have to maintain all the information required in an archival repository. For example, it may not need all the metadata for each object or, it may use only lower resolutions of images. This flexibility can be supported because the access system could query in real-time the archival repository's extraction service with a unique ID for an object. The archival system would then return a full master copy of the object to the access system. In theory, an access system could be written that does not have an access repository, but only indexes and programs that pull master copies of objects as needed from the archival repository. In reality, most access systems will have some form of each object in an access repository, as they will not want to deal with the overhead of processing a master copy of each object in real-time. For example, the overhead of creating a thumbnail from a master copy object for a summary display each time it is needed.

Another important concept is that access repositories and their related systems are transient. They exist only as long as the communities that desire these services are willing to support them. Some access systems, such as ones provided by libraries, will have very long lives. Access systems created by extracting digital objects from archival repositories for research projects may only exist for the length of the project. In both cases, the digital objects will move forward through time and technologies, as part of the archival repository.

This proposal recommends the establishment of a project to investigate the creation of a digital archival repository, as defined by the above framework. In addition, the participants will create at least one prototype access repository and system to test the framework and help the community better understand the services, technologies and costs related to running an archival repository. This project would proceed in three phases.

*1) The Planning Phase*

The planning phase will be used to investigate policies, practices and costs related to the implementation of a prototype archival repository service. This phase will also:

Determine what services an archival repository should provide (e.g., define the extraction service in detail)

Recommend policies that will guide the selection of digital materials to be included in the repository

Identify best practices and standards that are required to support the archival services (e.g., metadata, digitization procedures and standards, digital object naming conventions, etc.)

Develop a distributed archival systems architecture that builds on the best practices and standards and identifies the technologies that are best suited for implementing this architecture (e.g., Java, CORBA, JTIP, etc.)

Provide specifications for at least one access repository and system that will use the services of the digital archive.

Prepare a detailed plan and budget for the creation of a campus archival repository prototype, including recommendations for classes of materials and digital collections to include in the prototype. Although many types of digital objects need repository facilities, classes using digital images will be chosen as the initial type of digital object, as both the Library and MIP have extensive experience with creating digital image databases.

*2) The Prototype Phase*

In this phase, a prototype archival repository will be developed and evaluated. The prototype system will be used as a testbed to:

Evaluate and refine the decisions made during the planning phase in the light of the realities of running an actual system, including customers willingness to provide object for the archival service in a standard form

Develop long-term technical strategies and institutional policies for the establishment, management, and maintenance of repository services.

Gather detailed cost information related to running a production level service

Use the cost information to recommend a fiscal model for a production service

**3) *The Production Phase***

The results of the prototype will be reviewed and a proposal will be submitted for running a campus digital archival service. If the proposal is accepted and funded the production phase will commence.