Preserving Digital Materials

Final Report of the Digital Preservation and Archive Committee

Submitted to SOPAG on October 18, 2001

Membership

Howard Besser, UCLA

Curtis Fornadley, UCLA

Anne Gilliland-Swetland, UCLA

Sal Guerena, UCSB

Bernie Hurley, UCB (Chair)

Richard Marciano, UCSD Supercomputer Center

Reagan Moore, UCSD Supercomputer Center

Barclay Ogden, UCB

David Walker, CDL

Brad Westbrook, UCSD

EXECUTIVE SUMMARY AND MAJOR RECOMMENDATIONS

The long-term retention of digital library materials represents an urgent problem for the University of California Libraries. Therefore, the Digital Preservation and Archiving Committee (DPAC) recommends that the UC Libraries:

1) Establish a centralized UC Library preservation repository service that conforms to the OAIS[1] Reference Model, is administered by the CDL, and provides the following services:

a) ***Education and Outreach –*** promotes the importance of digital preservation, explains policies and procedures that govern the responsibilities of the preservation repository and the libraries using its services, and provides expert consultation and training on digital preservation issues. Centralizing this service will limit the need to develop highly specialized, digital preservation expertise at every UC library.

b) ***Ingest, Data Storage & Dissemination*** – creates *submission and dissemination agreements* that define library and preservation repository responsibilities for depositing, storing and returning digital materials from the repository. Initial submissions to the repository should be limited to the following materials submitted by UC Libraries: EADs, materials created to the CDL Digital Objects Standard (CDL-DOs)[2] and MARC records.

c) ***Digital Object Integrity and Transformative Migration*** – creates policies, procedures, tools and technologies that ensure the *physical and intellectual integrity* of preserved materials. Physical integrity ensures digital objects (i.e., their bits) are not inadvertently or maliciously altered. In the case of a transformative migration, when a digital object's bits must be changed, the preservation repository works with the depositing library to retain all the essential information needed to ensure the continued intellectual value of the object.

d) ***Discovery and Dissemination*** – allows for the identification and retrieval of repository objects. The most basic discovery and dissemination service returns objects that are requested by their unique ID. An advanced *real-time object export* service would allow "access systems" that directly serve end-users the ability to retrieve objects from the preservation repository and therefore, avoid the cost of storing these locally. The DPAC does not recommend that the preservation repository directly serve end-users.

e) ***Data Rescue*** – offers a recharge service to help libraries convert legacy, non-standard digital objects to standards accepted by the repository.

---

[1] The Open Archival Information System (OAIS) Reference Model is currently being reviewed as an ISO Draft International Standard.

[2] CDL Digital Objects (CDL-DOs) contain descriptive, administrative and structural metadata, as well as the actual content (digitized images, text, audio, and video). Content types in CDL-DOs created to date are primarily digitized images and/or text.

2) Form a governance/management structure comprised of major stakeholders to establish policy and guide the CDL's administration of the preservation repository.

3) Fund the preservation repository. The recurring salary cost required to provide the services listed above is estimated at $227K. The approximate hardware and software startup expense required to build an offline (i.e., tape based) preservation repository system is $377K. If the repository were to support "access systems" through the *real-time object export service*, it would need to be configured as a "highly available" online system that would have an estimated additional startup cost of $87K, or a total startup expense of $464K. See the section titled, "Advanced Discovery and Dissemination Service" for a detailed explanation of the real-time object export service.

4) Strive to license digital materials for which preservation rights can be secured, and that the university administration and libraries work to ensure that intellectual property legislation does not impede the preservation of digital materials (e.g., restrict the right to make exact, transformed or derivative copies needed for preservation).

5) Create a UC Library "Preservation Repository Implementation Team" to realize any DPAC recommendations accepted by the University Librarians. If the above recommendations are approved, the DPAC suggests creating an "implementation team" to aid in establishing the governance structure and assist the CDL in developing the preservation repository services. This implementation team could follow the successful model established by the *CDL Request Implementation Team*.

THE NEED TO ADDRESS THE LONG–TERM PRESERVATION OF DIGITAL MATERIALS

The primary motivations to address digital preservation issues are itemized in the Digital Preservation and Archiving Committee's (DPAC's) charge:

> "The long-term retention of digital library materials represents an urgent problem for the University of California. Faculty are hesitant to embrace publishing in "electronic only" formats, as they require more assurance that their scholarship will endure through time. In addition, UC libraries need to create an environment where these digital materials become part of their permanent collections, thus ensuring they are available for use by future generations of scholars. Finally, the University must be concerned with protecting its significant and growing investment in these digital assets."

The DPAC concurs with the above statement and notes that its theme is a concern over the ***longevity and integrity*** of digital materials: the faculties' need to publish an accurate and longstanding record of their work; the librarians' increasing addition of digital materials to their permanent collections and the institutional requirement to protect the long-term fiscal investment made in digital collections.

The most cited risk to maintaining the longevity and integrity of digital materials is the concern over migrating data to new technologies and accompanying data formats. The motivation for adopting new technologies and formats will vary from capturing cost efficiencies to providing better services. Some of these migrations are relatively simple; for example, moving data to new storage media where the "bit streams" that represent the content do not change. Other data migrations will be much more complex, such as moving digital content to new technologies that employ new storage formats (i.e., the bit streams *do* change) In this case, the challenge is to retain the essential information stored in the original object that is necessary to ensure its continued intellectual integrity. Some procedures that will minimize the risk of losing information are listed in Appendix A.

Another serious risk to the longevity and integrity of digital materials is the lack of policies, procedures and best practices for creating, maintaining and preserving digital objects. In many organizations, digital materials are treated as ephemeral resources, with little thought given to their existence beyond posting them to a website. The UC libraries are taking a leadership role in addressing this risk factor and have implemented well-reasoned and practical approaches to building digital libraries objects. The creation and adoption of UC standards such as the Encoded Archival Description (EAD), the CDL Digital Object Standard and the CDL Digitization Standard are helping to ensure that metadata and digital content produced by the University libraries are created and encoded to quality levels that are likely to be preserved.

Protecting the longevity and integrity of digital library materials is an urgent need and major challenge for the University of California Libraries. ***Therefore, the DPAC recommends the UC Libraries create a digital preservation program that focuses on the***

formation of a "preservation repository." The following sections of this report define a preservation repository and itemize the services it provides.

DEFINING A UC LIBRARY PRESERVATION REPOSITORY

FORMAL DEFINITION

The DPAC defines a preservation repository, as a trusted, neutral, shared service where the libraries can deposit digital materials to ensure their long-term integrity and accessibility[3].

SCOPE OF MATERIALS TO BE PRESERVED

The recommended scope for the initial repository implementation is CDL Digital Objects (CDL-DOs), EADs and MARC records from the University libraries. More specifically, the DPAC suggests building the initial preservation repository from:

- **Online Archive of California (OAC) Content**

The OAC provides access to its materials through MARC collection level record that points to their EAD collection description (finding aid), which in turn can point to digital objects included in that collection. A preservation repository would have to ingest these MARC records, EADs and digital object, while preserving their relationships (i.e., links) to each other. Adding OAC content to the preservation repository will help to determine how best to implement a linking structure within the repository. Using OAC content will also help to determine how these CDL-DOs can be harvested for inclusion in the repository.

- **CDL Digital Objects in Addition to OAC-DO Types**

There is a growing number of CDL-DOs that are being created outside of the OAC Program (e.g., topographic maps, Chinese stone rubbings). The preservation repository's adoption of the CDL-DO standard will provide added direction and incentive for libraries to create their objects to this standard, which also simplifies the long-term management and preservation of digital objects.

- **MARC Records Submitted from UC Libraries**

MARC records held in Melvyl do not contain detailed holdings information, such as volume/copy numbers, barcodes used for circulation, and item level notes. Therefore, Melvyl cannot be considered as a place to preserve this information, which is so critical to library operations. Campus libraries may wish to send copies of MARC records that include detailed holdings information to the preservation repository to ensure the investment made in recording this data is not lost.

---

[3] As the DPAC was finalizing its work, the RLG/OCLC Working Group released its report titled, Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources (http://www.rlg.org/longterm/attributes01.pdf). The DPAC believes that this thoughtful report provides a broad framework that supports the preservation repository definition and services as conceived by the DPAC.

- **Experimental Data Formats for Digital Content**

CDL-DOs contain descriptive, administrative and structural metadata, as well as the actual content. The data formats that represent content in existing CDL-DOs are images and text (e.g., digitized photographs, transcribed oral histories). ***The DPAC recommends that the preservation repository program be used as a testbed to experiment with other content formats as well, including those "born digital."*** The goal of these experiments will be to help understand how the preservation repository services can be extended to collections that use these data formats.

Examples of other formats include digital word-processed documents, library websites, digitized audio or video, numeric data, and electronic books and journals. Adding these to the repository can be accomplished by creating CDL-DOs, which would wrap the descriptive, administrative, and some of the structural metadata around these content formats. Determining the range of formats to use for preservation purposes represents the experimental nature of the testbed. For example, a word-processed document could be stored within a CDL-DO in its native, proprietary form, as a PDF document and also as a series of TIFF images. Storing the content in more than one standard format, especially in popular formats like PDF and TIFF, helps to ensure the longevity of the document.

WHY A PRESERVATION REPOSITORY AND NOT AN "ARCHIVE"

The use of the term "archive" can cause confusion, as it has multiple meanings. For example, Webster's Third New International Dictionary defines an archive as:

- "A place in which public or institutional records (as minutes, correspondence, reports, accounts) are systematically preserved;"
- "A repository for any documents or other materials, esp. of historical value (as diaries, photographs, personal correspondence);"
- "Any repository of collections, esp. of information."

The digital library community uses the last and broader definition of an archive when it talks of creating archives of digital materials, as is evidenced by the terminology used in the OAIS Reference Model. Library special collections departments may consider themselves archives under the second definition. Finally, archivists may prefer to use the first and specific definition of an archive.

Adding to the definitional confusion is the fact that "archiving" is a series of actions that includes selecting, processing, storing and providing access to materials. The DPAC envisions that a digital preservation repository implements only a subset of these archival functions, rather than all archival functions. The digital repository is a trusted, neutral, shared service where communities can deposit their digital materials to ensure their long-term preservation. Therefore, the DPAC has decided to use the phrase "preservation repository," in place of "archive."

WHY A PRESERVATION REPOSITORY IS NOT A "BACK-UP"

Backing-up data is not the same as preserving digital materials. A back-up is a copy of a computer's proprietary file system, not a replication of the standardized digital objects it initially imported. It is important to understand that most "access systems" do not store a standard form of a digital object. For example, the OAC access system disassembles EADs and stores their data in a series of proprietary database tables and indexes, which are spread out across the computer's file system. The EAD's cannot be recreated and exported from this system.

Certainly, back-ups are critically important for library computing systems in order to recover from a catastrophic failure (i.e., loss of data). They are optimized to restore lost data by restoring the computer's proprietary file system in the shortest time possible. For this reason, a preservation repository itself would be backed-up.

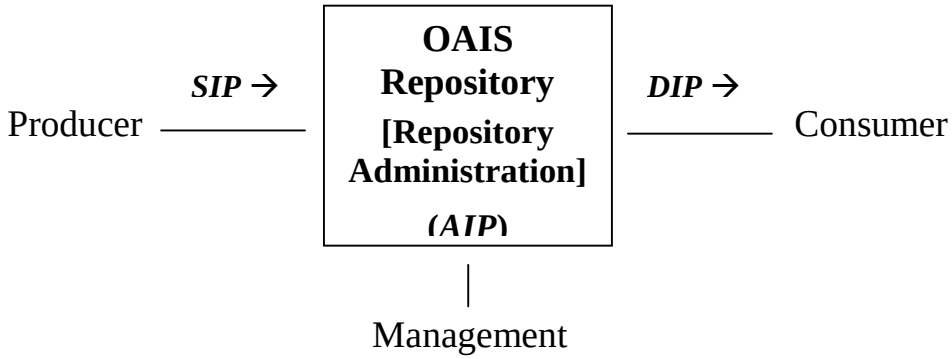
However, back-ups have no intellectual intent, context or consistency, and are worthless outside of the file system for which they are designed, whereas preserved digital objects are able to stand alone outside of any given file system. For example, a digital object encoded to the CDL Digital Object Standard and placed in the preservation repository is a self-contained unit that includes the digital content, its metadata (descriptive, administrative and structural), as well as the relationships among all these elements. A preservation repository also strives to capture the relationships of the digital objects to one another, and to preserve these relationships over time.

PRESERVATION REPOSITORY'S RELATIONSHIP TO THE OAIS REFERENCE MODEL

The Open Archival Information System (OAIS) reference model (http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf) provides a useful framework for identifying the governance, functional and metadata requirements that any implementation of a preservation repository must address. This model is widely used within the national and international digital preservation community to discuss digital preservation issues and is on track to become an ISO standard.

The DPAC recommends that any implementation of a preservation repository for UC libraries be compliant with the OAIS Reference Model.

The basic OAIS model is diagramed as:



The DPAC has adopted the OAIS terminology, as it has precise meaning and is being used internationally. The most notable exception to this guideline is that the OAIS term "archive," has been renamed to "preservation repository," for reasons explained earlier. A more detailed diagram of the OAIS model can be found in Appendix B, or at: http://ssdoo.gsfc.nasa.gov/nost/isoas/dads/OAISOverview.html

OAIS ROLES AND RESPONSIBILITIES

Role of the Producer

Given the scope of the DPAC's charge, the producers will be UC libraries, including its special collections and archival operations. Note, the producer is the organization that submits digital materials to the archive, which is not necessarily the same person or organization that created the materials. Librarians may select, catalog and deposit their electronic books, journals and any other materials into the preservation repository. Curators may select, validate, process and deposit their archival collections. It is the responsibility of the "producer community" to set standards, guidelines and best practices for content creation (including "born digital" materials), metadata capture, and digital object encoding for the materials being deposited. So far, UC libraries have these necessary standards, which include MARC, Encoded Archival Description (EAD) and the CDL Digital Object Standard. It is also the producer's responsibility to maintain the materials in the repository, updating the content and metadata as necessary.

Role of the Consumer

The consumers of the materials in the preservation repository are again, the UC libraries. The consumers are the original producers, who will need to retrieve materials from the repository for occasional maintenance, disaster recovery, etc. A consumer's library access system may also retrieve digital materials for its patrons' use, as described in the section titled *Discovery and Dissemination Services* (see below).

Role of the Preservation Repository Administration

The repository administration is responsible for day-to-day operation of the service. It is required to negotiate *"submission agreements"* with producers to define how and when materials will be deposited, as well as *"dissemination agreements"* with consumers to determine how materials can be discovered and extracted from the repository. The administration also has the responsibility to negotiate digital migration plans with producers, as they become necessary over time. ***The DPAC recommends that the CDL undertake the role of preservation repository administration.***

Role of Management

Management sets overall policy for the preservation repository service and is responsible for securing funding. It must resolve high-level conflicts between producers, consumers and the preservation repository administration. Within the UC system, the Regional Library Facility Boards provide a model for a preservation repository

governance structure. ***The DPAC recommends a preservation repository management membership roster that reflects the preservation repository's stakeholders***. For example, it could include two University Librarians, one or two faculty members, and one representative each from the Collection Development Committee (CDC), Library Technology Advisory Group (LTAG), and the preservation repository administration.

THE SIP, AIP AND DIP

The submission information package (SIP) is the format and encoding standard used by the producer to send digital materials to the preservation repository and is documented in the submission agreement. The first SIPs that would be used for UC libraries will be MARC, EAD and CDL-DO standards.

The AIP is the archival information package and represents the format used to store the materials in the preservation repository. For now, this can also be MARC, EAD and CDL digital object standard definition, extended to include any additional information required by the repository (e.g., a link from the digital object to the submission agreement it was entered under).

Finally, the DIP is the dissemination information package and represents the format and encoding used to send materials extracted from the preservation repository to consumers. As the DPAC is not recommending a public interface, the main consumers are the producers. Again, MARC, EAD and CDL Digital object definitions could be used.

PRESERVATION REPOSITORY SERVICES

INGEST, STORAGE AND DISSEMINATION SERVICES

The preservation repository administration should negotiate "submission agreements" with producers to define how, when and under what terms materials will be deposited, as well as "dissemination agreements" with consumers to determine how and under what terms the materials can be searched and extracted from the repository. The process of negotiating these agreements is extremely useful in that it requires that all parties address their expectations, roles and responsibilities. Equally important, the agreements formalize and document this information for future reference, as informal agreements will be forgotten over time.

DPAC's recommendation that the consumers of preservation repository services be limited to producers simplifies the dissemination agreement. That is, the producers will deposit MARC, EAD and CDL-DOs and, upon request, have these same formats returned. The submission agreement is more complicated; the DPAC has developed a submission agreement template (see Appendix C) that can be used and expanded upon by the preservation repository administration. All materials deposited should have references back to the submission and dissemination agreements under which they were deposited.

The DPAC recommends that only materials covered by a pre-negotiated submission and dissemination agreement be deposited in the preservation repository.

DIGITAL OBJECT INTEGRITY SERVICES

A primary function of a preservation repository is to ensure the physical and intellectual integrity of deposited materials.

Physical and Linking Integrity

Physical integrity services ensure the digital objects (i.e., their bits) are not inadvertently or maliciously altered. There are technologies available to support these services, such as check-sums and digital signatures. Special attention must be given to system security to ensure that unauthorized intruders don't maliciously change data.

Other serious challenges to integrity are cases involving relationships within objects or between objects (e.g., "links," such as URL's, URN's, etc.). Some examples of these challenges follow.

The first case is where links are within an object. For example, the CDL-DO standard has a *file inventory* section that tracks all the different files that make up the digital content. If a digitized book has 100 pages, there could be 100 TIFF files that represent the master images for all pages. These files can be embedded in the object, or pointed to via links. In the latter case, digital object integrity would require depositing the separate content files in the repository.

A more complicated case is where links are used to represent relationships between objects in the preservation repository. For example, an EADs container list points to CDL digital objects that make up the collection being described by the EAD. As it would not be reasonable to try to embed all the CDL objects within the EAD, original links between the EAD and CDL objects would have to be maintained within the repository.

Another complicated case is depositing an object that links to another, which in turn links to others, and so on. How many levels of links should be harvested and deposited in the repository?

The most problematic case is when a deposited object links to a digital object *outside* the repository. This could happen when the link is to an object that has copyright restrictions and therefore cannot be added to the repository.

Careful planning will help to achieve object integrity and should be addressed in the submission agreement negotiated between the producer and preservation repository administration. With the producer's cooperation, it should be possible for the administration to ensure integrity within and between objects in the repository. However, the preservation repository may not be able to ensure long-term integrity for relationships represented by links to materials outside the repository.

Data Migration Integrity Services

All policies and procedures for the preservation repository are designed to protect the integrity of deposited materials over the long-term. There will come times when it will become necessary to migrate digital objects to new, more efficient and cost effective technologies and data formats. The DPAC describes a migration as the ability to create an *exact copy* or *transformed version* of deposited materials. An *exact copy* is one in which the migrated metadata and content has not been changed (i.e., the bits that make up the metadata and digital content are the same). A *transformed version* of the original digital object is created when the bits that represent the object's metadata and content need to be changed.

Basic Migration Service

Two levels of migration service are proposed. The *basic migration service* provides for depositing, saving and returning exact copies of deposited materials, ***up to the point of a transformative data migration.*** The goal of a basic migration is to preserve the physical and linking integrity of the digital objects. Basic migrations, such as refreshing data or replicating it to new storage media, will be the responsibility of the preservation repository.

Transformative Migration Services

Due to advances in technology, changing the bits that make up the digital object will become increasingly necessary. It is only a matter of time before popular file formats, such as ASCII, TIFF and JPEG are replaced by new formats, which will necessitate migrating the currently used formats forward. A migration that changes the bits of the digital object is called a *transformative migration* and the output of this process is a *transformed version* of the original object. For example, a transformative migration to convert text in the 8-bit ASCII character set to the 16-bit UNICODE standard would be simple, but other transformative migrations could be considerably more challenging.

The goal of a transformative migration is to ensure the *intellectual integrity* of the digital object by retaining all the *essential information* it contains, as opposed to the actual bit sequences that comprise the original content.

If the producer provides all the necessary metadata required by the preservation repository to support long-term preservation, the administration will guarantee that they will work with the producer to plan a transformative data migration. Transformative data migrations are a shared responsibility of the preservation repository and the producer.

EDUCATION AND OUTREACH SERVICES

Education and outreach services promote the importance of digital preservation, explain policies and procedures that govern the responsibilities of the preservation

repository and the community using its services, and provide expert consultation and training on digital preservation issues.

The long-term preservation of digital materials requires input and action from many different stakeholders: faculty, library selectors, catalogers, technologists, preservation repository staff, etc. The preservation repository administration is in a unique position to take a proactive role in educating the University community to the risks inherent in the long-term preservation of digital materials, as well as actions to mitigate these risks. In particular, preservation repository staff could help to educate producers about their responsibilities in minimizing risks. For example, producer communities need to establish digital preservation policies, as well as workflows and best practices to capture and encode digital content and metadata, including preservation metadata. The staff could organize workshops and training sessions to help campuses identify digital preservation issues, develop responses, and then document these in submission agreements.

The DPAC recommends that the educational and outreach services be centralized in the preservation repository administration managed by the CDL. From a resources perspective, centralizing this highly specific expertise at the CDL will help to limit the need to develop these skills at every UC library.

DISCOVERY AND DISSEMINATION SERVICES

The preservation repository must provide a basic discovery and dissemination service so consumers can find and extract needed materials. ***The DPAC recommends that the request for an object be via a unique ID assigned to the object by the producer, and indexed in the preservation repository.*** This simple form of access will help minimize the complexity and costs of running the preservation repository.

In addition to consumers' discovery needs, the preservation repository administration may also index certain metadata to provide discovery services required to fulfill its mission (i.e., identifying what producer deposited an item, discovering objects that have a data type that needs to be migrated, etc.).

Advanced Discovery and Dissemination Services (optional)

Advanced search and display services are critical in developing any useful end-user system. The DPAC defines an *access system*, as one that is designed to fill a range of advanced discovery and dissemination needs for a particular end-user community. The DPAC views access systems as transient. That is, communities will create access systems when they have the need and the resources to do so, and will abandon these systems when the need or resources diminish.

As opposed to an access system, a preservation repository is optimized to preserve digital materials over time, returning exact or transformed copies of the deposited materials to the consumer. So, what is the relationship between access systems and the preservation repository?

First, the DPAC feels it's important to strictly limit end-user access to the preservation repository in order to contain costs. If a preservation repository begins adding advanced discovery and dissemination services for end-users, the complexity and the cost of the system will increase dramatically. No matter how many advanced services are added, there would be continued pressure to keep adding new features found in commercial access systems. The DPAC believes the preservation repository will collapse under its own complexity and cost, if it attempts to directly compete with end-user access systems

On the other hand, one problem in implementing a preservation repository is that depositing materials is an extra step for the producer. So, there needs to be sufficient motivation for producers to deposit their materials. The DPAC proposes a *real-time object export service* be explored as an advanced dissemination service of the preservation repository. Real-time object export would allow access systems to request objects from the preservation repository and therefore, not have to store these objects themselves. This service clearly delineates the relationship between access systems and the preservation repository. The real time object export service will continue to increase in importance, as very large digital audio and video objects start to populate our collections. Subsidizing the cost of the preservation repository storage also would add incentive to producers to deposit their materials.

DATA RESCUE SERVICE (optional)

The preservation repository could provide a data rescue service to help producers prepare ("rescue") digital materials that are not SIP-compliant. Examples include materials that were created before a submission agreement and SIP were established for that producer, or digital materials acquired by a producer in a non-SIP compliant format. ***If a data rescue service is established, the DPAC recommends it be a recharge operation.***

INTELLECTUAL PROPERTY AND COPYRIGHT ISSUES

The DPAC makes a distinction between intellectual property *access rights* and *preservation rights*. Access rights control what digital objects can be legally presented to each user and are most often granted through contract and licensing agreements. Therefore, access rights are managed by access systems, as opposed to the preservation repository, using rights management technology solutions (e.g., authentication/authorization systems, IP filtering).

Preservation rights address the legality of replicating digital objects for preservation purposes (i.e., exact copies or transformed versions). Without preservation rights, digital objects cannot be preserved over time. Preservation rights must be secured or granted by the producer and specified in the submission agreement. If these rights change for materials held in the preservation repository, the producer must alert the repository administration in a timely manner.

The DPAC recommends that UC libraries strive to only license digital materials for which preservation rights can be secured, and that the university administration and libraries work to ensure that intellectual property legislation does not impede the preservation of digital materials (e.g., restrict the right to make exact, transformed or derivative copies needed for preservation).

CENTRALIZED VS. DECENTRALIZED PRESERVATION REPOSITORY

The DPAC recommends that a centrally managed preservation repository be created for the UC libraries rather than attempting to coordinate several decentralized campus preservation repositories.

The DPAC further recommends that CDL act as the centralized preservation repository administration, as it is well placed to ensure that best practices are both negotiated in submission agreements, and followed in practice.

A centrally managed repository would reduce costs by implementing the repository and its services once for all UC libraries. In particular, storing very large digital objects once within the UC system would be cost efficient and would be particularly attractive to libraries if the *real-time object export* service were to be implemented.

While the DPAC recommends a centrally managed preservation repository, the repository contents should be replicated to limit the risk of loss to disaster. This replication could be as simple as having the preservation repository administration create tape copies of the digital objects, which would be stored in multiple locations. Or, if cost-efficient, a geographically replicated online version of the repository could be implemented.

If the preservation repository is serving content to access systems via the real-time object export feature, the DPAC recommends that the preservation repository be run as a "high availability" service (e.g. using clustered servers) to minimize unscheduled downtime. The costs and benefits of the high availability service would have to be determined to justify the service.

COSTS

The cost of an initial implementation of a preservation repository is separated into the one-time capital costs and recurring labor costs. Each cost element relates to an initial system scaled to store three currently existing collections: MARC records, EAD records, and the CDL digital objects. The initial scale of the system is:

- 30 million records
- 5 TB total capacity (last for 3-4 years)
- Access rate of 10% per year

- Approximately 11 submission agreements to start (one per library)

$377K is the hardware/software startup cost and $227K is the recurring salary expense needed to implement an off-line preservation repository. This implementation would not include the real-time object export service (i.e., providing digital objects to public access systems). An additional $87K startup expense would be required if the preservation repository were expanded to support real-time object export, for a total startup cost of $464K. The following explains the costs in more detail. All these costs assume the preservation repository runs in an established data center.

Governance – The DPAC recommends a governance/management structure modeled on the RLF boards (see Role of Management). The main operating expense would be travel costs to be covered by each participating campus.

Director of the Preservation Repository ($111K per year; includes 23% benefits) – A CRM II would be hired to manage the repository services, advocate use, and manage the interactions with the UC campus libraries, including negotiation of the submission agreements.

Production ($116 per year, includes 23% benefits) – A 0.5 PA IV ($55K) is needed to manage the database, schedule processing for the submission process, and manage the submission process. 0.25 FTE PA III ($23K) is needed as a system administrator to manage the accession, storage, and access platforms. 0.5 FTE PA II ($38K) is needed to support the submission and retrieval of digital objects.

The production workflow is structured around the submission agreements and submission of digital objects. A reasonable workload is to process a new submission agreement and associated digital objects every two weeks. The initial effort will take longer as the submission information packages are defined and the processing steps for each type of submission information package are developed. Once the primary submission types are established, the submission process can be automated.

Hardware and Software ($377K) – The basic hardware system is designed to support an off-line storage service. The components are:

- Accession platform ($48K) – includes a 2-processor server at $25K, an Oracle license at $18k, and 50 GB of disk space for the information catalog at $5k.
- Access platform ($48K) – the system also serves as the back-up system to the accessioning platform, and uses a duplicate hardware configuration.
- Storage system ($281K) – tape storage is assumed as the cheapest support mechanism for off-line access. The cost of tape media is currently $1,250 per TB. The cost of buying a tape robot is about $125K, and the cost of a hierarchical storage manager is about $100K. The cost of the CPUs for managing the system is about $50K. The capacity of the system is much greater than initially needed.

Expanded hardware and software system for a real-time object export service ($87K additional, for a total startup cost of $464K) – The basic hardware system would be augmented with a disk cache to support on-line access. Disk systems currently cost about $25,000 per TB. Given that the disk cache only needs to hold actively referenced data, the size of the disk cache should be about 35% of the archive size (set from the combined access rates). This implies a disk cache of 3.5 TBs at a cost of $87K.

APPENDIX A: METHODS TO MITIGATE THE RISK OF LOSING DIGITAL MATERIALS

All organizations that wish to mitigate the risk of losing digital materials should:

1) *Place the preservation repository in an institution that has preservation as a "core value" and has a stable long-term future.*
2) *Ensure the preservation repository service has a secure funding model.*
3) *Specify and collect the preservation metadata that will allow for future data migrations.* For example, information on how the original object was created, stored and rendered on access systems.
4) *Set standards, guidelines and best practices for the repository*
   For example, the CDL Digital Object Standard provides a single metadata/content encoding that can be used with any hierarchical object (book, journal, diary, correspondence, photograph, etc). Therefore, if someday TIFF is replaced, it will be easy to find all the TIFF files in each object.
5) *Limit "linking" outside the repository*
   Try to have a stored object fully contained inside the repository, as the repository service cannot control the preservation of materials on the outside. However, there may be times it is not appropriate, important or possible to do this (e.g., copyright issues over harvesting linked objects).
6) *Use standard file formats for digital content*
   There is a greater chance that popular file formats (TIFF, PDF, XML encoding, etc.) will be able to be migrated to new technologies.
7) *Store multiple content file formats where possible and economical*
   For example, it should be possible for a born-digital XML document (e.g. EAD) to be "printed" to disk to as a PDF document. Or, Word documents can be printed as PDF using Acrobat. Having more than one format increases the chance that the document can be migrated forward.
8) *Always preserve the original deposited material.*
   While it may not be technically possible or economical to migrate a document upon obsolescence of the format, it may at a future time.
9) *Describe the materials in enough detail so one can check authenticity.*
   One should be able to check a narrative metadata description against the content to help validate the material as being the correct item.
10) *Implement quality control measures.*
    For example, use digital signature techniques to ensure an object hasn't changed since its last back-up. Have people spot-check stored materials on a random basis.
11) *Replicate the preservation repository in multiple locations*
    Storing materials in different geographic locations minimizes the risk of losing materials in the primary repository to fire, flood, etc. It may also be desirable to

store the same materials in multiple repositories using different technologies. This would protect against a catastrophic software failure that damages materials in the primary repository and its replications.

12) *User Education*

Last but not least, it is critical that producers and consumers of preservation repository services be educated to its mission, policy and procedures, as well as their own responsibilities (e.g., developing and following standards, guidelines and best practices, negotiating submission and dissemination agreements with the repository, etc.).

APPENDIX B: OVERVIEW OF THE OAIS REFERENCE MODEL

Editor Note: This information was retrieved and reformatted from:
http://ssdoo.gsfc.nasa.gov/nost/isoas/dads/OAISOverview.html

IMPORTANT CONCEPTS FROM THE DRAFT ISO STANDARD "REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)"

The following figures and text provides a set of terms and concepts into which fundamental, long-term, archival activites may be mapped and thus compared. Long-term is long enough to be concerned about the impacts of changing technologies on the management, preservation, and distribution of archived information. This material has been adopted from the draft ISO standard entitled "Reference Model for an Open Archival Information System (OAIS)." For clarification of this material please see the full *reference model document (* http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html).

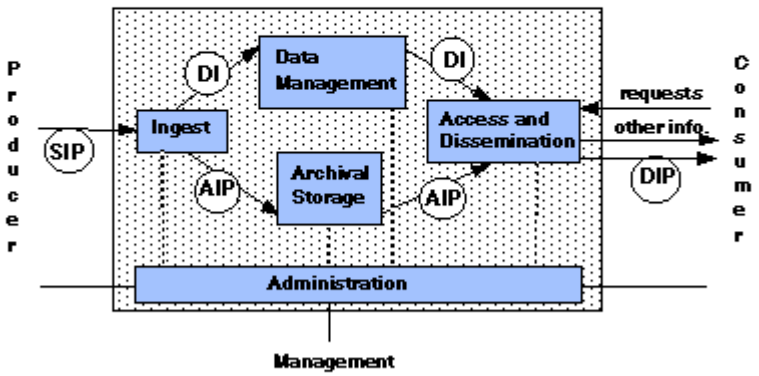


Figure 1: Functional component model for an OAIS

The functional areas of figure 1 include both systems and people needed to support the OAIS operation. The OAIS is an archive that meets a set of responsibilities as defined in the reference model document and this allows an OAIS archive to be distinguished from other uses of the term 'archive.'

Information objects (which may be any type of data), together with attributes needed for efficient ingest, archival preservation, and searching, are received from data Producers by the INGEST function using a Submission Information Package (SIP). The INGEST function does validation, adds supporting information as needed, ensures that the information is understandable to the designated Consumer communities, and peforms any transformations needed to put the information into archival storage forms. These transformations may include reorganizing and reformatting to meet archival storage and dissemination needs. The resulting information objects are sent to ARCHIVAL STORAGE using an Archival Information Package, and search information (e.g., Catalog data) used to support Consumer selection of archived data is sent to DATA MANAGEMENT as Descriptive Information.

ARCHIVAL STORAGE accepts Archival Information Packages, stores and manages them, and provides them to ACCESS and DISSEMINATION in response to requests. It also handles migrations of Archival Information Packages to new media when specialized domain oversight is not required to perform the migration.

DATA MANAGEMENT is the repository for all information used to support search aids, and for all information (outside of ARCHIVAL STORAGE) used to support the general operation of the archive. It stores all the Descriptive Information (catalog information) used to support searching and ordering. It stores all the request information generated by Consumers and by the archive in responding to requests. It stores all the information about Consumers.

ADMINISTRATION is responsible for coordinating daily operations of the archive, and for addressing the implementations of policy issues which impact multiple archival functions. In contrast, Management is a higher level function that oversees archival operations as only one of its responsibilities. ADMINISTRATION makes sure that necessary hardware and software are purchased and maintained, that security is maintained, and that the archive is using cost-effective technology and standards. It also oversees negotiations with data Producers on what is to be submitted to the archive, and it ensures that Consumers are generally satisfied with its services. It ensures the long-term preservation function is accomplished.

Consumers interact primarily with the ACCESS and DISSEMINATION function to find and receive information objects of interest. The finding aids used are supported by the catalog data (Descriptive Information) held by DATA MANAGEMENT. Requests to ARCHIVAL STORAGE yield Archival Information Packages (AIPs) which are processed as needed by ACCESS and DISSEMINATION to complete the order. Standing orders are processed automatically as the information becomes available and meets distribution requirements. Disseminations are provided as Dissemination Information Packages (DIPs) to the Consumer using some protocol (e. g., FTP, http, or tape).

Information Models

The other major dimension for the reference model is the modeling of information in the OAIS. A general model of archival information objects is shown in Figure 2. Much more extensive modeling is contained in the full document.

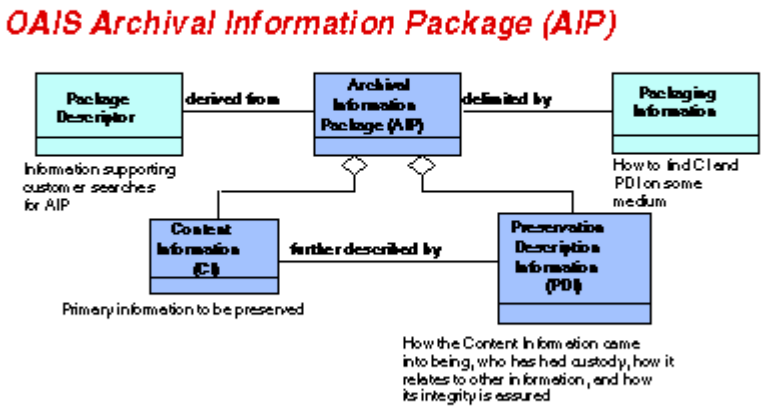


Figure 2: Archival Information Package Objects

The **Archival Information Package** (AIP) contains two primary information objects that are identified as **Content Information** (CI) and **Preservation Description Information** (PDI). The Content Information is that information which is the primary information submitted for preservation. The scope of what constitutes this information is agreed to between the archive and the Producer. To be complete, and preservable for the long-term, this information must include the associated Representation Information (or format information) that turns the Content Information bits into meaningful information.

Once the Content Information has been determined, it is possible to ask what constitutes the Preservation Description Information for that particular Content Information. The PDI includes several types of additional information that are needed to help preserve the Content Information. These are:

o **Reference**: How consumers can uniquely identify the Content Information from any other Content Information.

o **Provenance**: Who has had custody of the Content Information and what was its source. This would include the processing that generated it.

o **Context**: How the Content Information relates to other information objects, such as why it was created and how it may be used with other information objects.

o **Fixity**: Information and mechanisms used to protect the Content Information from accidental change.

The PDI information is needed for long-term preservation and its completeness is a key element in determining the quality of the archival function being performed.

Within the archive, the Content Information and Preservation Description Information need to be tracked and associated. This is done using the **Packaging Information**. For example, this may consist of some directory and file names, and their underlying implementations, on some medium. Or it may consist of a tar file together with some information relating the Content Information bits, its Representation Information, and Preservation Description Information.

Also associated with the Archival Information Package is the Descriptive Information. This is the information that is used to populate finding aids and is typically thought of as the catalogue information. It is this information that supports Consumer searches or that triggers the dissemination of information in response to a standing order.

APPENDIX C: AGREEMENT TO TRANSFER DIGITAL MATERIALS TO THE CDL PRESERVATION REPOSITORY

I: REPOSITORY INFORMATION

A: Transferring Agency

B: Institution / Repository Owning Collection/Item (if different than above):

C: Address

D: Contact Person

E: Title:

F: Email address:

G: Telephone Number:

H: Names and Contact Information for Persons / Agency appointed to succeed current custodian:

II: TERMS OF AGREEMENT

The materials covered by this Submission Agreement will be deposited in the CDL Preservation Repository. The materials are subject to the management policies of the CDL Preservation Repository and to technological modifications to the repository.

A: The Producer submitting the digital materials

1: agrees to submit digital objects according Preservation Repository object standards or to reimburse the Preservation Repository for upgrading the Producer's digital objects to object standards;

[Note: the upgrade provision assumes the Data Rescue Service is implemented and is a recharge service.]

2: will provide digital object metadata, including rights and license information, sufficient for use, preservation and efficient management of the digital objects;

3: will keep up to date all object versions stored in the Preservation Repository;

4: is responsible for all links, internal and external, contained in the deposited objects; and

5: will identify all persons having authority to submit and retrieve the Producer's digital objects stored in the Preservation Repository.

6: will ensure the preservation rights for all materials to be deposited have been secured and will inform the preservation repository administration, in a timely manner, if these rights change.

B: The Preservation Repository Administration:

1: will provide a unique identifier for each digital object submitted by the Producer;

2: will manage the Producer's digital objects in a manner that protects them from corruption and loss due to either vandalism, system failure, or advancement in technology, ensuring the content authenticity of the digital objects and their status as reasonable surrogates of their original sources;

3: will establish a disaster recovery backup of all digital objects;

4: will notify the Producer beforehand of any expected alterations to file formats, including upgrading to newer file formats, and to submission and dissemination information packages;

5. will ensure that the Producer maintains links contained in the objects, be they internal links to other objects in the Preservation Repository or external links to objects outside the Preservation Repository;

6: will supply to the Producer any metadata the Preservation Repository creates to manage the digital objects, using the

Dissemination Information Package (DIP) specified in this Submission Agreement;

7: ensures digital objects can be accessed and retrieved by the producer and any other authorized parties, using the specified Dissemination Information Package (DIP);

8: will report timely statistics to the Producer indicating all instances of authorized party uses of the Producer's digital objects; and

9: the Preservation Repository's management will assume the role of Producer for all digital materials "orphaned" by the dissolution of their Producer and / or assigned custodian.

Signature and dates of Submitters and Acceptors:

III: SUBMISSON AGREEMENT DESCRIPTION

A: Submission Agreement Title:

B: Brief Content Description:

C: Submission Agreement Date

D: SIP/DIP Type (Check all that apply):

___ CDL Digital Objects

___ MARC Records

___ EAD Finding Aids

E: Retention Period (check one):

___ Permanent

___ Specified number of years less than 15 years (or until a transformative migration is required)