



Overview of the UC Libraries Digital Preservation Repository

Long-term stewardship of digital collections requires a data management program that provides secure storage, monitoring of object integrity, physical security, access control, and format migration strategies.

The Digital Preservation Repository (DPR) serves the stewardship mission of the UC libraries by providing a single shared solution for the preservation, management, and controlled dissemination of digital collections that support research, teaching, and learning. The repository provides a set of self-service interfaces that the libraries use to deposit and manage digital objects, relieving individual libraries of the burden of creating and maintaining custom digital repositories. The services and storage are based at the California Digital Library (CDL).

This document provides a broad overview of the DPR. More detailed information is provided in the *Technical Overview of the Digital Preservation Repository* and *Preparing to Submit Objects to the Digital Preservation Repository*.

A Shared Service

The UC libraries share the DPR as a common and versatile service, each library adapting it to address local digital preservation needs.

The DPR was created to secure diverse types of digital information for the UC community. For some libraries the primary need is to preserve digitized versions of objects in their own collections, but the DPR serves a large variety of preservation needs as well, including:

- “Born digital” primary materials, including datasets produced by scholars in the course of their research
- Digital content created by UC libraries and their partners
- Exclusively web-based content, including web sites that provide supporting ephemera for historically significant collections

- The published record, including journal articles, monographs, technical reports, conference papers
- Online teaching and learning materials produced by UC scholars

Security

The fundamental security issues for digital storage are data integrity (will my data objects be the same when I retrieve them?) and data access (can I control who has access to my data objects?)

The California Digital Library is a UC institution established to assist the UC library community with the technology to manage collections of digital material. The DPR provides a secure service for libraries to store their digital collections.

Securing Data from Unauthorized Access

The library that owns the data collection determines who has access to the material stored in the repository. Access is password-controlled, and libraries can configure access profiles with different levels of access.

Secure Storage

The California Digital Library, at the University of California Office of the President, hosts the repository and controls physical access to it.

Currently, the repository uses daily incremental backup to tape, and preserves data objects in a storage area network (SAN) array. The repository is migrating to a replication model (configured as a RAID 1+0 array), which features full mirroring of the data and recoverability from hardware failure.

Plans for geographically remote replication are also currently underway. As a first step, the CDL has purchased disk storage space at UC Berkeley for this purpose and is implementing replication there, pending testing of replication software and strategies.

Monitoring Data Integrity

The repository validates the formats of digital objects at ingest. After checking the validity and well-formedness of an object, the repository generates a checksum value, which the data center periodically checks to make sure the object has not changed. The data center also produces a separate checksum value, which it also periodically validates.

Auditing

UC libraries can view administrative reports about their digital collections, ordered by access group, size, number of objects, disk space monitoring, and transaction.

Format Migration

A common issue in digital preservation is that as formats evolve, the formats of stored objects can become obsolete and eventually unusable. To support format migration for the collections it safeguards, the DPR provides:

- Object versioning, so that objects can be kept in old and new formats
- Tools for common format migrations (in development)
- Expert advice about format migration tools
- Format migration service (in development)

How the DPR Works

Authorized users at the UC libraries and affiliates use the DPR's services to submit or manage their digital collections, to request dissemination of material, or to request administrative reports.

A Self-Service Repository

The libraries retain control over the digital objects they store in the DPR. A library can replace or remove its objects, or maintain several versions of them.

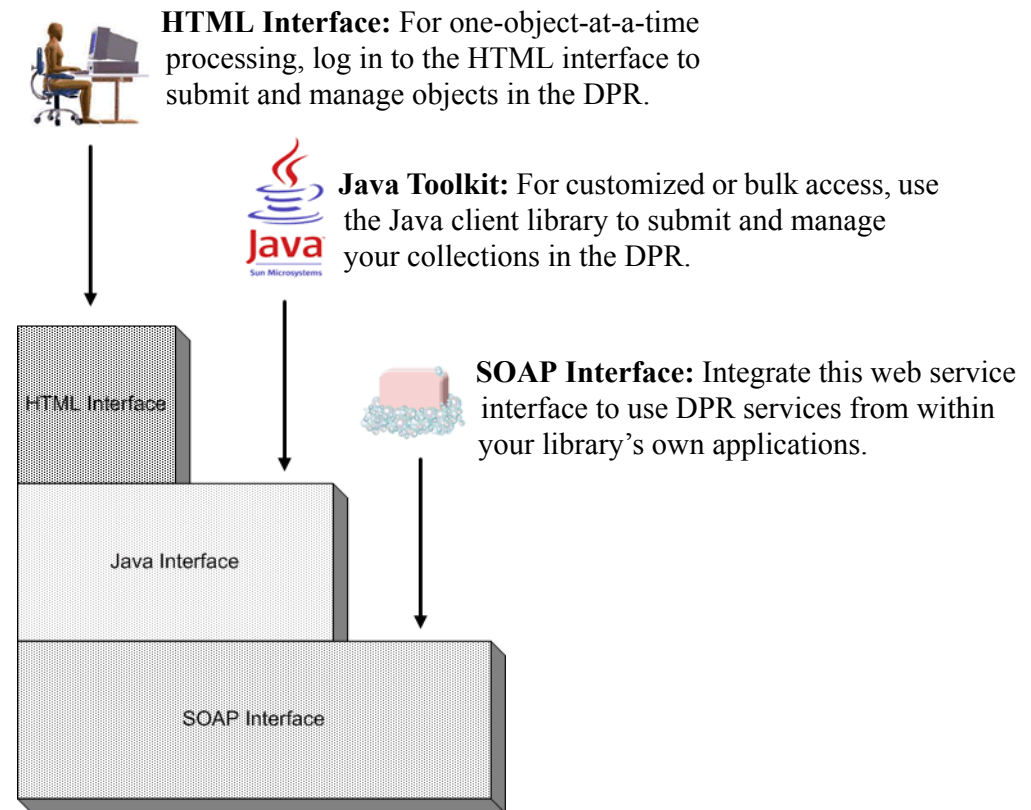
The UC libraries that use the DPR service are in charge of making intellectual decisions about the content. The role of the library is to select the objects or collections for storage in the repository, to secure the right to submit the objects, to prepare the objects for submission, and to manage the objects in the repository.

The role of the CDL is to ensure the technical integrity of the storage system and the data that it contains, to communicate and educate about the service, to develop and implement policies, procedures, and best practices to manage the service's architecture, to control user access, and to control and report on submission and dissemination.

How Libraries Use the Repository

The DPR offers three interfaces by which libraries can access its services. Libraries select the interface that best suits their collection needs and technical means. A simple HTML interface allows browsing and small-scale

submissions. A Java toolkit provides tools for large automated submissions. A library can plug the DPR SOAP interface into its own application to let librarians access the DPR from within a familiar local environment.



Interfaces to the Repository

For information about each of these interfaces, see <http://www.cdlib.org/inside/projects/preservation/dpr/>.

Submitting Content

To deposit an object or collection, the library starts by completing a submission agreement and an inventory of the collection. The agreement authorizes the DPR to archive the objects in the collection. The library specifies who will be allowed to deposit similar objects in the future and manage the deposited objects.

The library prepares the object for storage by creating a METS wrapper. (METS stands for Metadata Encoding and Transmission Standard, a standard for which the Library of Congress is the maintenance agency.) A METS

wrapper is an XML document format for encoding metadata for managing digital objects in a repository.

During the object's ingest process, the DPR also assigns the object a globally unique identifier called an Archival Resource Key (ARK). The ARK is used as an address for the object and its descriptive metadata. Once assigned, the persistent identifier is recorded so that it will never be reassigned, even if an object were removed entirely.

Maintaining a Collection

An authorized user logs in to the DPR's HTML interface and uses the ARK to locate and curate an item. This user makes decisions about re-versioning, replacing, preserving, and removing objects in the collection.

Administrative reports about the objects will be available to submitters.

Content and Formats

Types of content that may be submitted to the DPR include the following:

Articles and preprints	Images
Technical reports	Audio files
Working papers	Video files
Conference papers	Learning objects
E-theses	Reformatted digital library collections
Datasets	Web sites

Content Requirements

The requirements for submitted objects are these: the object must have persistent value to the UC community, the submitter must have the right to authorize the deposit for preservation and copying, and the object must conform to the CDL Digital Object Standard.

Value of the object: During the creation of the submission agreement, the teaching or learning value of the object to the University of California is validated.

Right to deposit the object: The library must have rights to store the collection. Rights should belong to one of the following four categories:

- The content is in the public domain
- The copyright is held by the submitter
- The library has obtained permission from the copyright holder to deposit the object

- The submission is an exercise of the depositor's rights of use under Sections 107 and 108 of the U.S. Copyright Act

CDL Digital Object Standard: The admission standards for digital objects are flexible, balancing the high value and risk level of some digital objects against the demands of meeting the standard. The more information that can be encoded into the descriptive metadata, the more reliably and flexibly the object can be used. Generally, the object is a package that includes:

- A file or some files comprising the content of the object
- A METS instance conforming to a declared or known METS profile, containing XML-encoded metadata about the object
- A unique identifier (ARK) supplied by the DPR

See the guidelines here: [CDL Guidelines for Digital Objects](#).

Formats

There are no format restrictions on digital objects deposited in the DPR.

To ensure the objects' long-term usability, however, the DPR does recommend that objects be in a format recognized by the JHOVE object format validator. JHOVE currently recognizes these digital formats:

- AIFF
- XML
- HTML
- JPEG 2000
- TIFF
- WAVE
- ASCII
- GIF
- JPEG
- PDF
- UTF-8

Given the broad range of file formats and standards used throughout the digital library world, the DPR is designed to accommodate variety. The system can be configured to accept new file formats, METS profiles, metadata schemas and standards, and object types. Unknown formats need to be negotiated during the pre-submission consultation, and may delay the submission of objects into the repository.

Some Holdings

The DPR currently safeguards or is in the process of ingesting several important collections on behalf of UC libraries.

- **The Legacy Tobacco Documents Library:** The Legacy Tobacco Documents Library (LTDL) contains more than 7 million documents related to the advertising, manufacturing, marketing, sales, and scientific research of tobacco products. The LTDL includes documents

posted on tobacco industry web sites as of July 1999 in accordance with the Master Settlement Agreement, additional documents added to those sites since that date, and the Mangini and Brown & Williamson document collections from the Tobacco Control Archives maintained by UC San Francisco.

- **The Hoover Collection: Images of UCLA:** Hoover (Thelner and Louise) Collection: Photographs of UCLA, 1927-1982. Photographs of UCLA dating from 1927-1982 by photographer, Thelner Hoover (UCLA Class of 1930); originals located in the UCLA University Archives.
- **Rats Spinal Cord Image Archive:** From the UCLA Digital Library, 640 images of neurons from spinal cords from a group of 36 rats. These images were created for use in Psychology 116, the psychobiology laboratory course, to train students in data analysis and interpretation. This image collection has been used to test the transfer of digital objects between the library's digital image collection and course management software systems.
- **Strachwitz Frontera Collection of Mexican and Mexican-American Recordings:** The Arhoolie Foundation's Strachwitz Frontera Collection of commercially produced Mexican and Mexican-American Recordings at the UCLA Library is possibly the largest repository of Mexican and Mexican-American vernacular recordings in existence.
- **The Local History Digital Resources Project (LHDRP):** Supported by a Library Services and Technology Act (LSTA) grant administered by the California State Library, the LHDRP leverages CDL infrastructure to provide long-term preservation and access to important and unique digitized historical materials housed in California's public libraries, museums, and other cultural heritage institutions. Content includes over 10,000 objects dealing with topics such as Chinese theater in California (San Francisco Performing Arts Library & Museum), the California border region (San Diego Historical Society), early California pioneer reminiscences (Society of California Pioneers), and local historical treasures from public libraries in communities as regionally varied as Santa Ana, Chula Vista, Oakland, and Marysville.

Standards

The DPR uses Java and XML, and relies on standards and models recognized in the digital-library community.

METS: the Metadata Encoding and Transmission Standard (see <http://www.loc.gov/standards/mets/>)

Dublin Core: a set of metadata elements for digital objects (see <http://dublincore.org/>)

JHOVE: an extensible format validator for digital objects (see <http://hul.harvard.edu/jhove/>)

ARK: the Archival Resource Key persistent object identifier scheme (see <http://www.cdlib.org/inside/diglib/ark/index.html>)

OAIS: an international standard reference model that describes the functions and organization of preservation repositories (described at <http://nssdc.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>)

Storage Resource Broker: a data grid management system from UCSD's San Diego Supercomputer (see http://www.sdsc.edu/srb/index.php/Main_Page)

RAID: data replication scheme described in Wikipedia (<http://en.wikipedia.org/wiki/RAID>)

CDL Guidelines for Digital Objects:
<http://www.cdlib.org/inside/diglib/guidelines/>

Future Plans

The California Digital Library continues to refine and improve the DPR. Ongoing activities include making it easier for libraries to use our services, adding to and enhancing our services, and developing and fully applying digital preservation standards and protocols.

Increasing Ease of Use

- Releasing shrink-wrapped DPR software, including documentation
- Streamlining methods for submitting digital collections
- Improving tools for managing content in the DPR
- Adding text indexing, global metadata searching, and metadata editing at the object and global levels

Developing and Applying Protocols

- Working with user groups to develop protocols that support digital preservation services
- Implementing a preservation monitoring and integrity service
- Using the RLG's [Audit Checklist for Certifying Digital Repositories](#) as a guide, working with a range of third-party experts in the preservation and digital library community to evaluate the DPR

- Developing policies

Adding and Enhancing Services

- Enhancing the DPR infrastructure to include a web archiving service that will enable libraries to capture, manage, and preserve web-published content
- Evaluating existing preservation strategies for e-journal preservation (mature and emerging), electronic theses and dissertations and implementing a strategy as necessary

Program Sustainability

- Working with UC libraries to develop and evaluate economic models to sustain DPR preservation services

Further Information

For further information, see:

- *DPR User Interface Guide*
- *DPR Java Developer's Toolkit Guide*
- CDL Digital Preservation Repository
<http://www.cdlib.org/inside/projects/preservation/dpr/>
- JSTOR/Harvard Object Validation Environment
<http://hul.harvard.edu/jhove/jhove.html>
- Nice Opaque Identifiers
<http://www.cdlib.org/inside/diglib/noid/>

For help with implementation:

- Contact dprsupport-l@ucop.edu