IR Digital Preservation Policies

UC LIBRARIES
Digital Preservation Repository

Overview of the UC Libraries Digital Preservation Repository (DPR)

The University of California Libraries Digital Preservation Repository (DPR) supports the long-term storage, and management of digital objects. Content submitted to the DPR must support research, teaching, or learning.

Preserving an object is the joint responsibility of the California Digital Library (CDL) and submitters. When negotiating a submission agreement and setting up an account, both the CDL staff and submitters work together to define the nature of the objects, establish rights, and determine user access roles.

The CDL is responsible for checking submission errors, controlling user access, and retaining deposited objects and versions in perpetuity. Submitters are responsible for choosing the objects they wish to deposit in the DPR and for making decisions about removing or replacing any of their objects or versions. If at some point submitters are unable to manage their objects, they can negotiate with the DPR to withdraw as active participants.

Submission and access to the DPR are possible through an HTML user interface, a CDL-supplied Java-client library, and a web-services interface that uses the SOAP protocol. A persistent identifier (ARK) is automatically assigned to any object that does not already have one upon submission. The services and associated storage are based at the California Digital Library (CDL).

Submission Considerations

When preparing objects for submission, consider the following:

Rights

The submitter must have the right to authorize the deposit of the digital objects (which includes the right to copy) for preservation purposes. There are three categories of rights:

- Content that is in the public domain.
- Copyright that is held by the submitter.

- Permission that is obtained by submitter from the copyright holder, allowing the deposit of the object.

METS Format

Each object, consisting of one or more files, is deposited with a "wrapper" data structure that includes some descriptive metadata and an inventory of the object's component files.

The XML-based "wrapper" format that is used is called the Metadata Encoding and Transmission Standard (METS), for which the Library of Congress is the maintenance agency.

The transmission of an object begins when the METS file is submitted to the DPR interface. This file references all the object's component files either on a web server or on physical media (CDs or hard drives) that the DPR can access.

Dublin Core Elements

In addition to the metadata listed on your submission agreement (for example, format and submitter's identifier), the DPR requires four elements from the submitter. These "kernel" elements, derived from a subset of the Dublin Core, are needed for minimal object description (who, what, when, and where). The DPR then records and may add information about the origin of the object ("provenance").

Unique Identifiers

An object saved in the DPR must have a globally unique persistent identifier in the form of Archival Resource Key (ARK) that provides a long-term, stable association between a string of characters and an object. It is used for addressing the object and for obtaining a basic description of it. Once assigned, the persistent identifier is flagged so that it is never reassigned, even if an object is removed entirely. A record will always be kept describing the object so that the association between the identifier string and the object will never be broken.

Object Versions

Using the same ARK, the object can exist in the DPR with as many versions as needed. Like the object itself, each new version must be self-contained: it must include new versions of all the object's component files. Versions can be removed and replaced at will. Access rights for a version are defined at the object level.

The submitter must determine if two submissions are versions of the same object or two different objects (with different ARKs). Therefore, the object can be entered into the DPR in one of two ways:

- A version of a previously deposited object
- A distinct object

Primary vs. Alternate Object Identifiers

Each object has a *primary object identifier*. Upon first deposit, if an object does not have an ARK as a primary identifier, one is assigned and returned to the submitter for future reference. The CDL can provide your institution with software for generating and maintaining your own ARKs if you wish, in which case you may submit your existing ARK with the object when you deposit it.

Each object can also be deposited with an *alternate object identifier*, which may be a local identifier that you want to continue using. You are responsible for the uniqueness and persistence of your alternate object identifier. You may use either identifier to reference your object.

Format

There are no format restrictions. However, in order to ensure the long-term usability of the object's content, structure, and functionality, it is recommended that each object's files be in a format recognized by JHOVE. This step examines the format and generates structural and technical metadata that could assist in future migration efforts.

DPR Standards and Practices

The DPR relies on community-wide standards and practices as much as possible in both the digital-library community (METS, Dublin Core, ARK) and the wider computing community (Java, XML). The design of the DPR has been influenced by two seminal efforts in the digital preservation community: the OAIS reference model and the PREMIS metadata activities.

Interfaces

Users can prepare objects for submission by interacting with the DPR in the following ways, depending on their needs.

Web-Based User Interface

The HTML browser interface that the DPR supplies is a simple option that is ideal for becoming familiar with DPR functionality.

Java API

The Java API employs a DPR-supplied Java client library that submitters can use with their own Java programs, allowing them to plug the DPR into systems

of their own, thereby tailoring it to meet their local needs. The Java-client library implements the upper layer of the DPR Application Program Interface (API). This is ideal for bulk processing.

Note: The following Java toolkit is used: Sun Microsystems' *Java Web Services Developer Pack (WSDP)*.

SOAP Interface

The DPR API relies on a lower layer CDL-supplied SOAP interface.

Note: SOAP is an XML-based protocol allowing interoperability between software programs. That is, programs written in one language can communicate with programs written in another language.

Therefore, applications written in any language can interact directly with the DPR using SOAP (with Attachments). For assistance with this implementation, contact dprsupport-l@ucop.edu.

Differences between File Backup Systems and DPR

File Backup Systems	DPR
Files are backed-up on recyclable tapes and are rarely retained for more than one year. Previous file versions, if saved at all, are recycled even sooner.	Objects and their versions are never deleted unless explicitly removed by the submitter.
Backups are created by recording updated computer files indiscriminately.	The DPR selectively records coherent, platform-independent objects. An initial examination during the submission agreement negotiation establishes that objects should be preserved if they benefit the UC community. Subsequently, the submitter has control over the objects it wishes preserved in the DPR.
Backups made on one platform may not work when retrieved onto a different platform. Even if you were able to recover a file containing digital content from your backup tapes, you might find it unusable without the supporting files and database records. There are no guarantees that older files available on specific computer platforms can be used today.	File and record dependencies are minimized by packaging related files in one object. Platform dependencies are minimized by requiring that deposited objects conform to the standards-based, non-proprietary specifications. Unlike ordinary computer files, each object is assigned a globally unique persistent identifier.

Future Plans

Many preservation issues are still being investigated by CDL and the rest of the digital library community.

- Migration Issues

  A primary reason that the DPR supports multiple object versions is to implement future format migration (that is, conversion to more viable contemporary formats).

  - DPR is generating (via JHOVE) standardized file format identifiers to anticipate this need. It is also looking into sustainable funding models to support potential migrations.

- Scheduled Check of DPR Objects

  DPR performs file checksums when objects are deposited. The goal is to create a schedule for periodically accessing each DPR object, recomputing the checksums, reporting discrepancies, and repairing any damage found.

- Replicated Storage of Objects

  Various storage systems will be tested that support redundancy and geographic replication.

  - The Storage Resource Broker (SRB) system, which is a product of the San Diego Supercomputer Center, is being explored as a way to replicate storage among the UC campuses and non-UC library partners.

    SRB currently functions as the DPR's storage layer, but its full functionality has not been utilized. It is important to gain some practical experience with SRB during the DPR pilot phase before exploring further possibilities.

- Refreshing (rewriting) objects periodically.
- Desiccated Data

  As a protection against technical or financial obstacles to future format migration, we are considering a supplemental strategy involving low-tech preservation derivatives for document-like objects. In particular, this calls for the generation of a small number of derivative versions in highly sustainable formats, including both plain text files and raster image files.

  These derivative versions are "dried up" representations of the original documents that undoubtedly lose many features in this process—animation and hotlinks in the case of a raster image, and fonts, colors, and inline-graphics in the case of a text—but they retain the essential document properties. This type of desiccated data has proven to be among the most long-lived digital formats. Ingested along with the original format, this strategy may provide a high degree of stability as we confront a future of uncertain funding for all but the most valuable of older materials.

- There are currently no plans to support emulation of objects.

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/
- Reference Model for an Open Archival Information System (OAIS)
  http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

For help with implementation:

- Contact dprsupport-l@ucop.edu

University of Illinois at Urbana-Champaign

IDEALS Digital Preservation Support Policy

Committed to building and maintaining collections for the use of students, faculty, scholars, and the public long into the future, the University of Illinois at Urbana-Champaign assumes an obligation to ensure long-term access to the materials deposited into IDEALS and their intellectual content, but also acknowledges the inherent challenges involved in preserving digital content.

To this end, the IDEALS Digital Preservation Support Policy defines the categories of preservation support available and provides specific information about where different file formats fit. Our ability to preserve digital objects deposited in IDEALS is dependent, among other things, on the file format used, whether it is proprietary or publicly available, the number of software platforms that support the file format, and whether there are embedded files or dynamic references within the digital resource. This policy is subject to change as new and emerging technologies impact our ability to preserve deposited content.

All digital objects deposited to IDEALS will receive a basic level of preservation. Basic preservation means that IDEALS will: strive to ensure that the bitstream (the 1's and 0's that make up the digital file) remains exactly the same over time, assign a persistent, permanent identifier, create preservation metadata, maintain onsite and offsite backup copies, and perform regular virus and file corruption checks and periodic refreshments by copying files to new storage media.

IDEALS categorizes digital objects into four categories of preservation support. These categories are defined below.

Category 1 – Full Support

Digital objects in Category 1 receive the highest level of preservation support. IDEALS will make its best effort to maintain full functionality of objects found to meet the Category 1 criteria. Full functionality is defined as preserving the *viability* of the original bitstream (the 1's and 0's that make up the digital file) so that information must be intact and readable, the *renderability* of the resource so that the resource is viewable by humans and processable by computers, and the *understandability* of the resource so that the resource is interpretable by humans. Formats fitting the Category 1 criteria will be monitored for changes that might warrant transformation or reassessment. An example of a Category 1 digital object is a tiff (.tiff) image file or a plain text file (.txt).

The criteria for Category 1 support are that the object:

- Is in a format this is publicly documented;
- Is in a format this is widely adopted;
- Is in a format that may be rendered by multiple software packages; and

- Contains no embedded files or dynamic references.

Category 2 – Intermediate Support with Assurance of Full Functionality

Digital objects in Category 2 receive intermediate preservation support and will be normalized or transformed to a format that will receive Category 1 support to ensure that the renderability and understandability of the object will be available via the normalized version. Intermediate preservation support means that the original digital object will receive the basic level of preservation, and the original digital object will be monitored for changes that might warrant migration or transformation to a newer or more stable format. An example of a Category 2 digital object is a Microsoft Word file without any embedded files.

The criteria for Category 2 support are that the object:

- Is in a format that is widely used;
- Is in a format that is of enough public and/or commercial interest that tools are likely to be available to migrate them to successor formats;
- Contains no embedded files, dynamic references, or other ; and
- Can be transformed to a Category 1 format without any loss of renderability or understandability.

Category 3 – Intermediate Support without Assurance of Full Functionality

Digital objects in Category 3 receive intermediate preservation support and may be normalized or transformed to a format that will receive Category 1 support with probable loss of functionality or renderability due to issues like embedded files or dynamic content. Intermediate preservation support means that the original digital object will receive a basic level of preservation, and the original digital object will be monitored for changes that might warrant migration or transformation to a newer or more stable format. An example of a Category 3 digital object is a Microsoft Excel file containing macros and references to external files that cannot be normalized to a comma separated text file without loss of functionality.

The criteria for Category 3 support are that the object:

- Is in a format that is widely used;
- Is in a format that is of enough public and/or commercial interest that tools are likely to be available to migrate them to successor formats; and
- **Cannot** be reliably transformed to a Category 1 format without any loss of renderability or understandability. In most cases, either tools are unavailable to ensure reliable and repeatable transformation to a Category 1 format, or the object contains embedded files or dynamic references which cannot be successfully transformed.

Category 4 – Basic Support Only

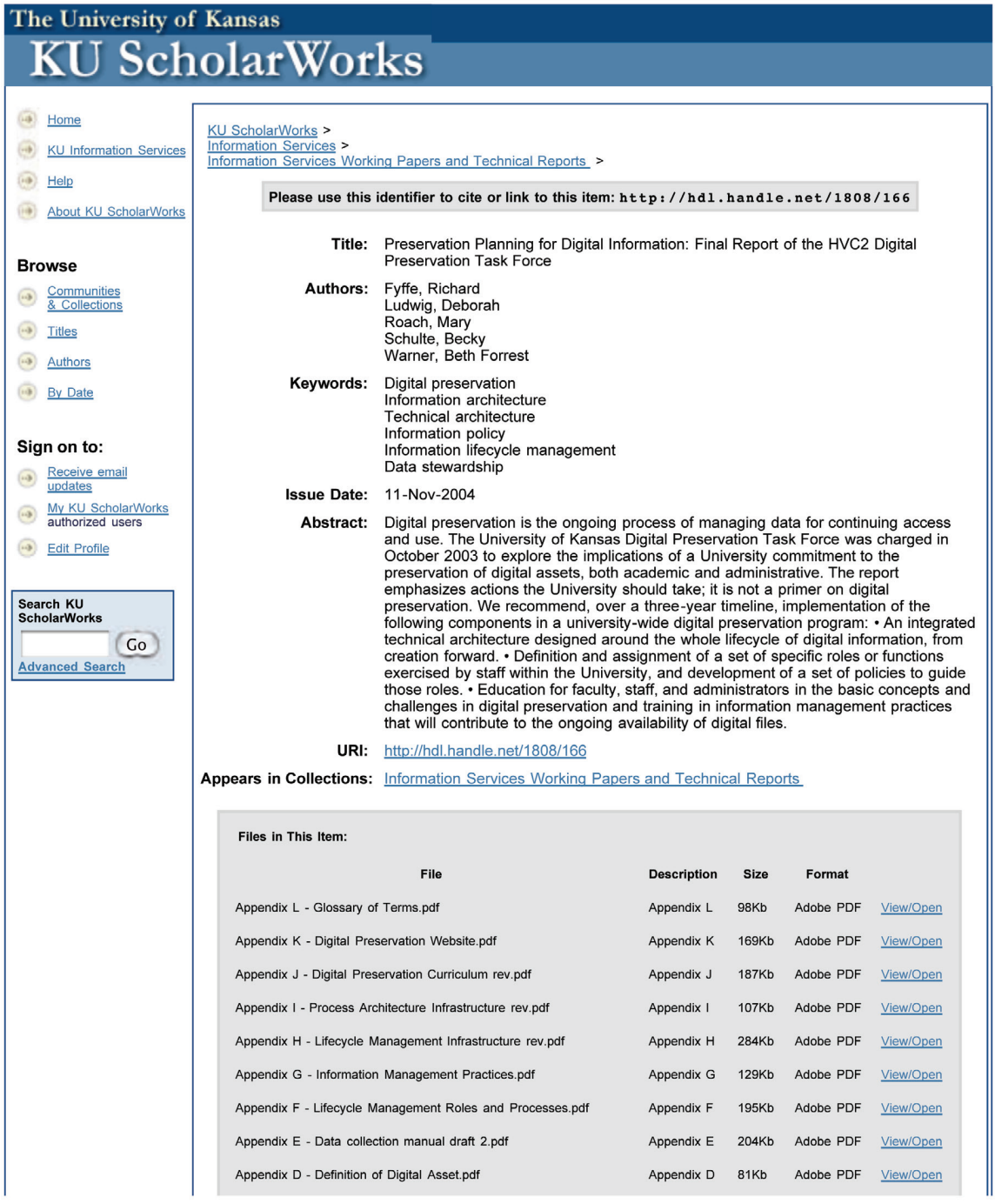
Digital objects in Category 4 receive only the basic preservation support available to all digital objects within IDEALS. Basic preservation means that IDEALS will: strive to ensure that the bitstream (the 1's and 0's that make up the digital file) remains exactly the same over time, assign a persistent, permanent identifier, create preservation metadata, maintain onsite and offsite backup copies, and perform regular virus and file corruption checks and periodic refreshments by copying files to new storage media. The original object will be preserved 'as-is', with no guarantee of transformation, migration to later formats, or renderability. This is often called 'bit-level' preservation, since only the original 'bits'- the 1's and 0's - are preserved. If the object depends on a particular version of software, there is no guarantee that the original object will still be usable when that software is no longer available. IDEALS makes no guarantees that objects with Category 4 support will be monitored for migration or transformation support. Category 4 support usually applies to digital objects written in highly specialized, proprietary formats (often usable only in a single software environment), formats no longer widely utilized, and/or formats about which little information is publicly available. An example of a Category 4 digital object is the Kodak Photo CD (.pcd) format that is proprietary.

The criteria for Category 4 support are that the object:

- Is in a highly specialized, proprietary format, often usable only in a single software environment;
- Is in a format about which little information is publicly available;
- Is in a format that is no longer widely utilized; or
- Does not meet the criteria for any of Categories 1-3.

Any format not yet reviewed and evaluated by IDEALS will receive Category 4 support on deposit. A different category may be assigned after format review takes place.

Preservation Action	Category 1	Category 2	Category 3	Category 4
Provision of persistent identifier for object and/or its metadata	X	X	X	X
Creation of preservation metadata	X	X	X	X
Secure storage and backup	X	X	X	X
Regular fixity checks	X	X	X	X
Regular virus checks	X	X	X	X
Periodic refreshment to new storage media	X	X	X	X
Transformation to a Level 1 format with full functionality	N/A	X		
Transformation to a Level 1 format without full functionality	N/A	N/A	X	
Storage of original digital object	X	X	X	X
Strategic monitoring of format for changes	X	X	X	
Migration to successive format upon obsolescence	X	X	X	



The University of Kansas
KU ScholarWorks

Home
KU Information Services
Help
About KU ScholarWorks

Browse

Communities & Collections
Titles
Authors
By Date

Sign on to:

Receive email updates
My KU ScholarWorks authorized users
Edit Profile

Search KU ScholarWorks
Go
Advanced Search

KU ScholarWorks >
Information Services >
Information Services Working Papers and Technical Reports >

Please use this identifier to cite or link to this item: `http://hdl.handle.net/1808/166`

Title: Preservation Planning for Digital Information: Final Report of the HVC2 Digital Preservation Task Force

Authors: Fyffe, Richard
Ludwig, Deborah
Roach, Mary
Schulte, Becky
Warner, Beth Forrest

Keywords: Digital preservation
Information architecture
Technical architecture
Information policy
Information lifecycle management
Data stewardship

Issue Date: 11-Nov-2004

Abstract: Digital preservation is the ongoing process of managing data for continuing access and use. The University of Kansas Digital Preservation Task Force was charged in October 2003 to explore the implications of a University commitment to the preservation of digital assets, both academic and administrative. The report emphasizes actions the University should take; it is not a primer on digital preservation. We recommend, over a three-year timeline, implementation of the following components in a university-wide digital preservation program: • An integrated technical architecture designed around the whole lifecycle of digital information, from creation forward. • Definition and assignment of a set of specific roles or functions exercised by staff within the University, and development of a set of policies to guide those roles. • Education for faculty, staff, and administrators in the basic concepts and challenges in digital preservation and training in information management practices that will contribute to the ongoing availability of digital files.

URI: http://hdl.handle.net/1808/166

Appears in Collections: Information Services Working Papers and Technical Reports

Files in This Item:

File	Description	Size	Format	
Appendix L - Glossary of Terms.pdf	Appendix L	98Kb	Adobe PDF	View/Open
Appendix K - Digital Preservation Website.pdf	Appendix K	169Kb	Adobe PDF	View/Open
Appendix J - Digital Preservation Curriculum rev.pdf	Appendix J	187Kb	Adobe PDF	View/Open
Appendix I - Process Architecture Infrastructure rev.pdf	Appendix I	107Kb	Adobe PDF	View/Open
Appendix H - Lifecycle Management Infrastructure rev.pdf	Appendix H	284Kb	Adobe PDF	View/Open
Appendix G - Information Management Practices.pdf	Appendix G	129Kb	Adobe PDF	View/Open
Appendix F - Lifecycle Management Roles and Processes.pdf	Appendix F	195Kb	Adobe PDF	View/Open
Appendix E - Data collection manual draft 2.pdf	Appendix E	204Kb	Adobe PDF	View/Open
Appendix D - Definition of Digital Asset.pdf	Appendix D	81Kb	Adobe PDF	View/Open

Appendix C- Working Summaries and Background Information.pdf	Appendix C	550Kb	Adobe PDF	View/Open
Appendix B - HVC2 Digital Preservation Task Force Membership.pdf	Appendix B	104Kb	Adobe PDF	View/Open
Appendix A - HVC2 Digital Preservation Charge 9-24-03.pdf	Appendix A	70Kb	Adobe PDF	View/Open
Preservation Planning for Digital Information.pdf	Report	674Kb	Adobe PDF	View/Open

Show full item record

VU e-Archive Supported Formats

The formats listed below as "supported" are those that Vanderbilt Libraries recognize and will support as part of the preservation and migration of the VU e-Archive. Formats listed as "known" are familiar, but their support and migration are not guaranteed. "Unknown" means that we do not recognize the format, and although we can store it, we will probably not be able to migrate it. If a format is not listed here, check with your VU e-Archive administrator.

FORMAT SUPPORT LEVELS

Name	Extensions	MIME Type	Support Level
Unknown		application/octet-stream	unknown
Adobe PDF	Pdf	application/pdf	supported
XML	Xml	text/xml	known
Text	txt, asc	text/plain	supported
HTML	htm, html	text/html	supported
Microsoft Word	Doc	application/msword	supported
Microsoft Powerpoint	Ppt	application/vnd.ms-powerpoint	known
Microsoft Excel	Xls	application/vnd.ms-excel	supported
MARC		application/marc	known
JPEG	jpeg, jpg	image/jpeg	supported
GIF	Gif	image/gif	known
image/png	Png	image/png	known
TIFF	tiff, tif	image/tiff	supported
AIFF	aiff, aif, aifc	audio/x-aiff	known
audio/basic	au, snd	audio/basic	known
WAV	Wav	audio/x-wav	known
MPEG	mpeg, mpg, mpe	video/mpeg	known
MP3	mp3	audio/video	known
RTF	Rtf	text/richtext	supported
Microsoft Visio	Vsd	application/vnd.visio	known
FMP3	Fm	application/x-filemaker	known
BMP	Bmp	image/x-ms-bmp	known
Photoshop	psd, pdd	application/x-photoshop	known
Postscript	ps, eps, ai	application/postscript	supported
Video Quicktime	mov, qt	video/quicktime	known
MPEG Audio	mpa, abs, mpega	audio/x-mpeg	known
Microsoft Project	mpp, mpx, mpd	application/vnd.ms-project	known
Mathematica	Ma	application/mathematica	known
LateX	Latex	application/x-latex	known

TeX	Tex	application/x-tex	known
TeX dvi	Dvi	application/x-dvi	known
SGML	sgm, sgml	application/sgml	known
WordPerfect	Wpd	application/wordperfect5.1	known
RealAudio	ra, ram	audio/x-pn-realaudio	known
Photo CD	Pcd	image/x-photo-cd	known

WHAT TO DO IF YOUR FORMAT ISN'T LISTED

Please contact your VU e-Archive Administrator if you have questions about a particular format.