OCLC Digital Archive Preservation Policy and Supporting Documentation

Last Revised: 8 August 2006

OCLC Online Computer Library Center, Inc.
Dublin, Ohio 43017-3395 USA

Table of Contents

OCLC Digital Archive Preservation Policy

Background

OCLC has been engaged in internationally based initiatives aimed at addressing the issues of digital preservation since 2001. OCLC participation in digital preservation committees, training initiatives, and research efforts during this time has informed and helped guide current best practices and related documentation identifying possible solutions for managing digital content. The initial development and ongoing support of the OCLC Digital Archive has been in direct response to the concerns for long-term trusted preservation of digital content expressed by our member institutions.

The OCLC Digital Archive Preservation Policy, (the Policy), is based upon two of the primary guiding documents in the field of digital preservation:

> Trusted Digital Repositories: Attributes and Responsibilities
> http://www.rlg.org/en/pdfs/repositories.pdf
>
> OAIS - Open Archival Information System Reference Model
> http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

The OCLC Digital Archive was launched in 2002 in response to the need expressed by libraries, archives, museums, and educational institutions for trustworthy long-term storage, management, and preservation of digital materials. The Digital Archive provides services to subscribing institutions interested in undertaking a practical approach to digital preservation. It allows institutions to ingest, manage, disseminate, and preserve their digital content using either a web interface or an offline process. Information about the Digital Archive services, tools, accepted file formats, and required metadata along with other resources can be found on the OCLC Digital Archive website. http://www.oclc.org/digitalarchive/default.htm A glossary of terms used in this policy can be found at the end of the document.

Purpose

The OCLC Digital Archive Preservation Policy outlines OCLC's approach to the preservation of digital content objects and related metadata deposited into the Digital Archive. This Policy is one component of OCLC's digital preservation program. The program is a fully integrated approach to reliably managing digital content for member institutions over time. The program utilizes the Digital Archive as an archival management environment in accordance with the requirements of the preservation service level assigned to the content by the subscribing institution and their Terms and Conditions (see Terms and Conditions documents in Supporting Documentation Section.)

The Policy includes the following information:

1. Service Levels and Preservation Activities available for digital materials deposited into the Digital Archive.

2. Preservation Strategy considerations for assessing risk, requirements for accessing deposited content, and action plans for preservation.
3. The Succession Plan which outlines the means for reliably transferring content from the Digital Archive to the original depositing institution should OCLC discontinue the Digital Archive for any reason.

The Policy is supported by documentation regarding

1. Data integrity and continuity.
2. Risk assessment.
3. A glossary
4. Policy and supporting documentation update procedures and notification process

Service Levels and Preservation Activities

Service levels for content objects in the Archive are: (1) bit preservation and (2) local. The depositor assigns service levels when content objects are ingested and can change service levels after digital materials are ingested into the Digital Archive. For details, refer to the Digital Archive Administration Module documentation. In the future, OCLC plans to offer full preservation for some content objects in the Digital Archive.

Bit Preservation. (currently available) Digital content objects and associated metadata are processed and ingested into the Digital Archive, and maintained in their original formats. Access to and management of the digital content objects is provided to the depositing institutions and their respective communities according to the Terms and Conditions.

Preservation activities for bit preservation include verifying the fixity of the objects and metadata and checking for viruses at the time of ingestion and annually on the anniversary of their ingest. The digital content objects are packaged with administrative and technical metadata into a standard dissemination package. "Bit preservation" is analogous to the "Store" service level in Digital Archive documentation.

The digital content objects and their associated metadata are backed up both onsite and offsite according to the processes outlined in Supporting Documentation, Data Integrity and Access Continuity Assurances, which documents data management, backup policies, storage facilities and security, and disaster prevention and recovery.

Local Preservation. (currently available) Digital content objects and associated metadata are processed and ingested into the Digital Archive so they may be disseminated to a local repository. Dissemination to the local repository must occur within 90 days after processing is complete or the service level automatically converts to bit preservation.

Preservation activities carried out by the OCLC Digital Archive prior to dissemination of objects for local storage include verifying the fixity of the objects and metadata as well as ensuring their freedom from viruses at the time of ingestion into the Digital Archive. The digital content objects are packaged with administrative and technical metadata into a METS encoded dissemination package.

Full Preservation. The goal of the OCLC Digital Archive is to ensure the long-term accessibility (i.e., full preservation) of content objects in the Archive. The development of preservation action plans for particular formats is key to providing full preservation. The preservation strategy below describes how preservation action plans will be created based on current knowledge and best practices in the digital preservation community.

Preservation Strategy

Assuring that the integrity of object bit streams is reliably maintained is an important component of digital preservation. Nonetheless, a preservation strategy must include more than just what can be achieved by good system back-up procedures. A strategy is needed also to ensure the long-term accessibility of digital content objects deemed to have enduring value.

A preservation strategy details the types of activities that will be undertaken to ensure reliable preservation of digital content objects. These activities include:

1. Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications.
2. Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied.
3. Determining the appropriate metadata needed for each object type and how it is associated with the objects.
4. Providing access to the content.

Data Format Risk Assessment

Data format risk assessment attempts to detect the timing and likelihood of changes in technology environments and file formats (e.g. PDF, HTML, TIFF) that will affect accessibility and long-term preservation of digital content objects. Risks may vary according to:

- The content of the object,
- The formats in which the content has been created,
- The ability to support hardware and software necessary to render the content usable,
- The needs of the users and depositors of the content,
- Institutional budgetary constraints, and
- The organization of management activities of the institution.

The current OCLC Digital Archive preservation strategy recommends risk assessment of file formats to determine what environments are needed for the most reliable access to the content and what actions need to be taken on a regular basis to ensure preservation of the digital content. This risk assessment will also take into account the content of the digital objects and the needs of the subscriber. The content access environments and the preservation action plans provide alternatives for future accessibility of the content objects.

Content Access Environments

A content access environment is a set of technology applications, operating systems, and hardware needed to render a content object. The OCLC preservation strategy defines two content access environments for each digital content object type held in the Digital Archive: the “current access environment” and the “open-source access environment.”

Current access environment. An environment created by a set of currently available technology applications, operating systems, and hardware used to render a digital content object. The current access environment is one commonly found in use today.

Over time, one or more of the current access environment components or file formats of digital content objects deposited into the Digital Archive may become outdated. This may cause one or more types of objects to become inaccessible via the current access environment, thereby requiring the open-source access environment to be used.

Open-source access environment. An environment defined by a set of primarily open-source technology applications, operating systems, and hardware needed to render a content object. The open-source access environment is defined to be used when the current access environment can no longer support the digital content object.

The open-source access environment is considered to be less volatile than the current access environment, due to open-source support in the community. However, both access environments (open-source and current) may change over time as their components become outdated or the file formats of digital content objects change. As these changes occur, a Preservation Action Plan may recommend changes to or emulation of the access environment and/or migration of the data to a new file format. The Preservation Action Plan for the file format will recommend specific preservation actions to take to provide ongoing access to content objects.

Preservation Action Plan

A Preservation Action Plan will be created for selected file formats. Content objects made up of files formats that have corresponding Preservation Action Plans are eligible for the Full Preservation service level in the Digital Archive. The Preservation Action Plan describes the preservation approach, including actions that are considered necessary for immediate, intermediate, and long-term preservation. For example, the Plan will state whether or not a format should be migrated to a different format upon ingestion into the Archive; if the format may be migrated at all in the future; or if the format can be migrated as needed to different formats to ensure access.

In considering what preservation actions are appropriate for which file formats, OCLC assumes the content in its original file format will remain in the Archive and able to be disseminated. Each Preservation Action Plan will be based, in part, on risk assessments. Once a Preservation Action Plan for a given file format is drafted, the Plan itself will also be risk assessed. All Plans will be reviewed periodically, in approximately 18-month cycles. The duration of the review cycle will depend largely on changes in the technological environment. Other factors that will be considered in determining the review cycle include available financial and human resources and the needs of Digital Archive subscribers.

OCLC will communicate the availability of new and updated Preservation Action Plans to depositors. Per the Terms and Conditions, OCLC will implement Preservation Action Plans only if the depositor has changed the Service Level of the digital content objects to Full Preservation. Preservation service levels are changed via the Administration Module. For details, refer to the Digital Archive Administration Module documentation.

Succession Plan

One of the attributes of a trusted digital repository is organizational viability. Preservation of content objects in the OCLC Digital Archive is a joint responsibility between OCLC and the depositing institution.

Should OCLC discontinue the Digital Archive for any reason, OCLC will disburse the content objects and preservation metadata to the depositors in a mutually agreed upon manner. At the time of transfer, OCLC will ensure the transfer media and the dissemination format will be relevant and compatible with current best practices and standards.

Should content objects be “abandoned” in the Digital Archive, the Terms and Conditions apply.

As the digital preservation community evolves in its understanding of digital repositories, OCLC will modify its succession plans to continue in its role as a trusted digital repository.

OCLC Digital Archive Preservation Policy Supporting Documentation

OCLC Online Computer Library Center, Inc.
Dublin, Ohio 43017-3395 USA

Data Integrity - Access Continuity Assurances

Last Updated: January 20, 2005

Purpose

This document provides general information about the methods by which the OCLC Digital Archive provides assurances for data integrity and the continuity of access to materials deposited into the Archive. These assurances cover the following activities:

- Data Management
- Backup Policies
- Storage Facilities and Security
- Disaster Prevention and Recovery
- Policy Updates

Data Management

All content objects receive a documented fixity check and virus check at the time of initial ingest into the Archive. A second documented fixity check and virus check is performed the day after ingest along with a documented object verification procedure, which ensures that all components of a given object are stored in the optimum manner needed to keep the object and metadata intact and that appropriate metadata exists for all components of the object. A fixity check, virus check, and object verification are also performed quarterly based on the anniversary of the object's ingest into the Digital Archive.

Content in the Digital Archive resides on RAID (Redundant Array of Independent Disks) storage. This system uses a combination of two or more drives to achieve superior fault tolerance and performance.

Backup Policies

Back-ups are created and maintained locally and off-site according to the following guidelines:

- New and changed content and metadata in the Digital Archive are immediately and redundantly logged to disk
- Recurring back-ups to non-disk media occur at systematic intervals throughout each day and are stored locally
- Daily back-ups are maintained for an extended period of time and are stored locally
- Weekly back-ups of system software and data are stored at off site facilities

In the event of a failure, the potential data loss depends on the nature of the failure. Two potential failures that could occur and their outcomes are listed below:

- A localized system failure could result in a loss of data that had not yet been backed up via recurring back-up to tape. This loss would be a few hours worth of data.

- A worst-case catastrophic local disaster could result in a potential data loss of several days worth of data as the data would have to be restored from weekly off site back-ups. The recovery time frame in any disaster involving retrieved backups depends on the nature of the failure, the quantity of data involved, and whether restoration is from on-site or off-site copies of data. Data recovery from backups requires twenty-four hours to six weeks, depending on the nature of the failure.

Storage Facilities and Security

OCLC stores content data, metadata, and system software at OCLC and off site facilities.

OCLC maintains staff solely dedicated to network and system security, including at least one Certified Information Systems Security Professional. At OCLC, backup tapes are stored in a building limited by a staff badge system. The tapes are stored in a computer room in a robotic tape silo. Access privileges to the computer room are limited and are reviewed every three months. Each access is logged, recording information such as the staff person entering, the door entered, and the time. Access to tapes in the storage silo is limited to staff trained to load and unload the tapes. All computer rooms are protected from fire by a halon gas fire suppression system. All computer rooms are climate-controlled with raised-floor environments.

Disaster tapes are sent to off-site storage facilities that meet the highest industry standards for safety and security. Tapes are always in the custody of OCLC staff or the off-site storage provider.

Disaster Prevention and Recovery Procedures

The goal of disaster prevention is to safeguard the data (content and metadata) in the Digital Archive and to safeguard the Digital Archive's software and systems. For disaster prevention and recovery, all data (content and metadata) is considered of equal value.

There are manual and automatic system restart and recovery processes in place for the Digital Archive in case of unintended power failures, and node, server, individual process, or other outages. Failover capabilities (redundancies) exist at three levels: network, host, and disk.

OCLC provides a business continuity solution via a remote site that provides hardware and software to run OCLC services. Regular tests are carried out with that facility.
If all major OCLC systems go down in a disaster situation, other systems may be restored before the Digital Archive. If other systems are restored first, the Digital Archive system functions (harvest, ingest, disseminate, etc.) will be restored within forty-eight hours thereafter. It may take up to several weeks to restore access to Digital Archive content stored before the disaster occurred.

Digital Archive Risk Assessment

Last Updated: January 20, 2005

Purpose

This document provides the criteria that will be used in the risk assessment process for:

- Organizations
- Preservation plans
- File format characteristics, software, and hardware

Risks for formats covered under a Full Preservation service level will be assessed in terms of their characteristics, required software, required hardware, and associated organizations such as software vendors, content depositors, etc. In addition, when a preservation action plan for a given format is drafted, it will be subjected to risk assessment.

Organizational Risks

Risk assessment is performed on all organizations that are related to any aspect of the preservation plans for a given digital format. Examples of related organizations are the software vendor or the content depositor. Organizational risks are always assessed in conjunction with the relevant format characteristic—software or hardware—that is under evaluation. For example, the maker of a particular software application is risk-assessed in conjunction with the risk assessment of that software application.

Not all risks are applied to all organizations. The following risks can vary according to the type of organization being risk-assessed.

- Organization has a small number of support staff.
- Organization is in danger of collapsing or going out of business.
- Organization is a target of acquisition.
- Support received from one organization cannot be easily replaced.
- No competitors exist (i.e., organization is unique in its class).
- Organizational budget is in danger or insufficient.
- Organization is subject to staff turnover and lack of continuity.

Preservation Action Plan Risks

- Loss of authenticity or ability to prove authenticity.
- Transformation could result in unauthorized changes to original content, either accidental or malicious.
- Transformation could result in unauthorized changes to transformed content, either accidental or malicious.
- Inadequate documentation of current transformation process creates problems for future transformations.
- Transformation requires a lengthy development cycle.

File Format, Software and Hardware Risks

Risk	File format	Software	Hardware
Royalties or license fees may be requested.	√	√	√
Specification or source code is not available for independent inspection.	√	√	√
Prior versions of the specification or source code are incompatible with new ones.	√	√	√
Specification or source code is very complex, large, ambiguous or poorly documented.	√	√	√
Specification or source code is not widely accepted, either *de jure* or *de facto.*	√	√	√
Specification or source code is unique in its class and cannot be mapped onto another; or, embedded metadata cannot be mapped onto other formats.	√	√	√
Specification or source code uses digital rights management schemes, signed envelopes, encrypted sections or watermarks.	√	√	√
Institution staff does not have sufficient expertise in the format, software, or hardware; staff is overwhelmed by rapidly changing specifications or source code.	√	√	√
Specification does not allow identical copies to be created, making traditional media refresh impossible.	√		
Specification allows extensions, such as executable sections (macros, javascript) or proprietary or narrowly supported features; specification contains objects in other formats, or links to objects in other formats.	√		
Specification defines structure (e.g., XSD, DTD), or uses external formatting styles such as fonts.	√		
Source code supports a limited number of file formats and as such offers little common capabilities.		√	
Source code depends on external libraries.		√	

Accepted Formats and Access Environment

Last Updated: January 20, 2005

Accepted Formats

Any content that is accessible via the http protocol can be harvested and ingested with the Digital Archive Web tools including Microsoft file formats, ZIP files, etc.

File formats that are served up via a protocol other than http, such as some audio and video files that require a streaming server, are not supported at this time via the harvesting tools and therefore cannot be ingested into the Digital Archive via the Web tools.

All file formats may be batch ingested into the Digital Archive, including audio and video file formats. File formats that require streaming support to view will not be viewable at this time in either the administration module or via OpenURL from the archive.

All file formats fall under the current disseminate limit of a 4 GB zip file. Zip files larger than 4 GB cannot be disseminated using the current tools.

Glossary

Last Updated: January 20, 2005

The definitions below are compiled from several sources listed at the end of the Glossary. Numbers in parentheses after the definitions indicate the source, if other than OCLC.

accepted format
One of the specified formats accepted into the Digital Archive for storage and preservation.

access
Continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential to the purposes the digital material was created and/or acquired for. (3)

access environment
A set of technology applications, operating systems, and hardware needed to render an object.

bit preservation
The processes used to ensure that the bits comprising a preserved object do not change over time. The processes include refreshing, media migration, and backups.

content
The intellectual substance of a digital object typically comprised of text, data, symbols, numerals, images, sound and moving images.

content object
One or more files that make up the content to be preserved.

current access environment
A set of current technology applications, operating systems, and hardware in common use that are needed to render an object.

deposit
Individual or batch ingest of digital objects into the OCLC Digital Archive that have been provided by a depositor.

deposit agreement
The terms and agreement for the Digital Archive services.

depositor
A representative of a subscribing organization who has designated authority to deposit individual and/or batch objects into the Digital Archive.

digital preservation
The series of managed activities necessary to ensure continued access to digital materials for as

long as necessary. (3)

disaster prevention and recovery plan
A set of responses based on sound principles and endorsed by senior management, which can be activated by trained staff with the goal of preventing or reducing the severity of the impact of disasters and incidents. (3)

fixity
The state or quality of being fixed or unchanged. Since digital objects are easily modified, a mechanism is necessary to maintain fixity over time, or to consciously document when a digital object has been altered. Technologies such as checksums and digital signatures are used to verify that a digital object retains its fixity, which helps maintain the object's authenticity and integrity. (1)

file format
A structure used for the interchange, storage, or display of data. Examples of formats include nonproprietary formats such as USMARC, HTML, and XML, along with proprietary formats such as Adobe Acrobat PDF or Microsoft Word. (2)

ingest
In the Open Archival Information System (OAIS) model, processes related to receiving information (content and metadata) from an external source and preparing it for storage and management within the archive. (2)

local service level
An object ingested with this service level is processed by the Digital Archive prior to dissemination for storage in a local archive within 90 days. The Digital Archiving processing includes fixity and virus checking, creation of administrative and technical metadata, and dissemination in a standard package.

media migration
The process of converting data from type of storage material to another to ensure continued access to the information as the material becomes obsolete or degrades over time. Examples of media migration include copying files from 5¼ floppy disks to 3½ floppy discs to CD to DVD then verifying that the copy was accurate. (2)

open-source access environment
A set of primarily (as far as it is possible and practicable) open-source applications, operating systems, and hardware needed to render an object.

preservation metadata
Information used for the long-term retention of an object. (2)

preservation action plan
The specific activities to be undertaken to ensure digital preservation for each file format in the Digital Archive.

preservation planning
The process of determining the preservation strategy for each file format in the Digital Archive and developing immediate, intermediate, and long-term preservation action plans for them.

preservation strategy
A technical approach to long-term digital preservation that includes technology preservation (hardware and software) , technology (software) emulation and data migration. (6)

refreshing
Copying digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bit stream (e.g. from a decaying 4mm DAT tape to a new 4mm DAT tape, or from an older CD-RW to a new CD-RW.) (1)

render
The process of displaying an image. The final and actual displayed image is said to have been rendered. (5)

risk assessment
The evaluation of the possibility of incurring a loss and a determination of the amount of risk that is acceptable for a given situation or event. (2—called “risk analysis”)

service level
How the content object will be managed in the Digital Archive. The service level is assigned by the depositor. Current service level designations are “Local” and “Bit Preservation.”

store service level
See bit preservation.

succession plan
Developed in consultation with community experts, depositors, and peer organizations that identifies all relevant content and designates trusted inheritors should the repository cease to exist. (6)

Glossary Sources

1. *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*, by Cornell University, 2003. Available at http://www.library.cornell.edu/iris/tutorial/dpm/.

2. *A Glossary of Archival and Records Terminology* by Richard Pearce-Moses, exposure draft, 2004. Available at http://www.archivists.org/glossary/.

3. *The Preservation Management of Digital Material Handbook* by Neil Beagrie and Maggie Jones, 2001. Available at http://www.dpconline.org.

4. *Trusted Digital Repositories: Attributes and Responsiblities by OCLC and RLG*, 2002. Available at http://www.rlg.org/longterm/repositories.pdf.

5. *Universal Format Preservation Glossary*. Available at http://info.wgbh.org/upf/glossary.html.

6. *Working Definitions of Commonly Used Terms (for the Purposes of the CEDARS Project),* updated 1999. Available at http://www.leeds.ac.uk/cedars/documents/PSW01.htm.

Policy and Supporting Documentation Updates

Last Updated: January 20, 2005

As new standards, best practices, and protocols for managing digital content emerge they will be reflected in the OCLC Digital Archive Preservation Policy. Changes to the Policy will be communicated through the OCLC website, newsletters, and appropriate email lists as they are implemented. The most current Policy will be made continually available through the OCLC website.

This document, and the policies explained in it, are considered dynamic in nature and will be updated periodically. Substantive revisions and the date on which they occur will be noted at the top of new versions of this document. Each time this document is changed, a notice will appear on the Digital Archive web site. Notification will be sent to all Archive subscribers via the Digital Archive listserv, to which subscribers are added when they initiate activity in the Digital Archive.

We encourage comments and questions about our preservation processes; please send all feedback to the Digital Archive, digitalarchive@oclc.org

OCLC Digital Archive
Terms and Conditions Documents