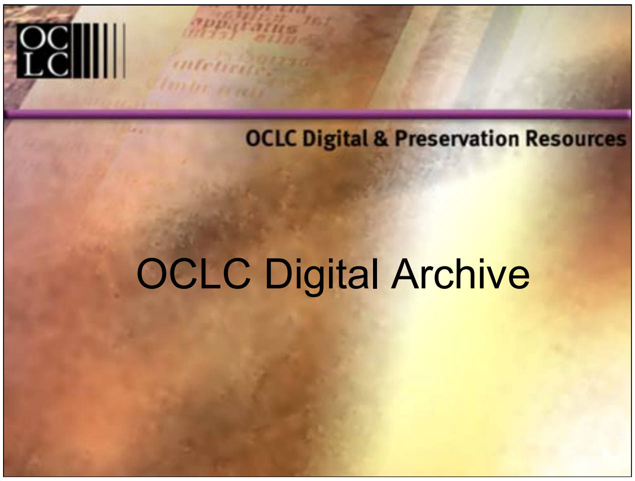
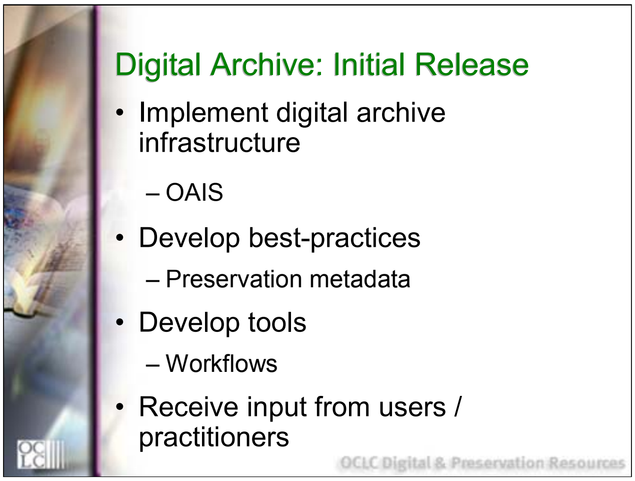


OCLC Digital & Preservation Resources

OCLC Digital Archive

he last element in the OCLC digitization and preservation suite of offerings is the Digital Archive.



Digital Archive: Initial Release

- Implement digital archive infrastructure
  - – OAIS
- Develop best-practices
  - – Preservation metadata
- Develop tools
  - – Workflows
- Receive input from users / practitioners

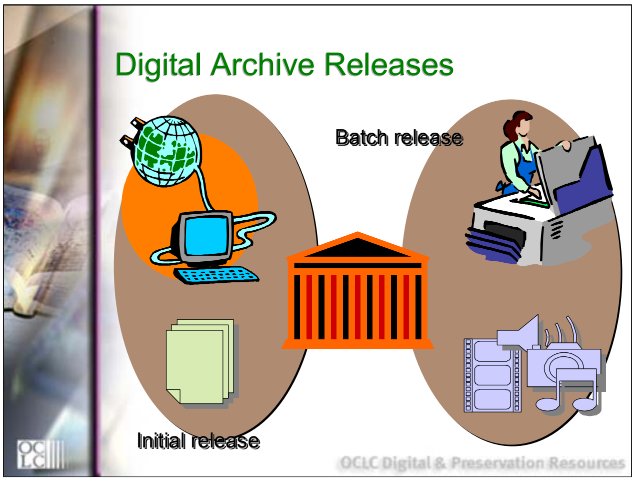
OCLC Digital & Preservation Resources

he initial release of the digital archive, which is the stage we're in now, facilitates capture of web-based ocuments, it enables creation of preservation metadata for digital objects, provides for ingest of hose objects into the archive, and allows for long-term retention of these digital objects. The verall goal of the archive is to take in digital objects and make them accessible for the long-term.

CLC has based the archive on the OAIS functional model. OAIS stands for Open Archival Information ystem, an ISO standard. The OAIS is a conceptual framework, a foundation on which to build best practices d standards relating to digital archiving. If you're building a digital archive, it outlines the elements of service u need to provide. It does not specify how to build the archive, only what entities compose the archive and w those entities interrelate. Tom will get into more detail on the OAIS model at the end if time permits.

CLC has used preservation metadata schemes under development around the world as sources while creating r own scheme. OCLC is committed to getting comments and input from practitioners in the field as the heme evolves.

e're also trying to develop the archive in a way that libraries can adapt it to various workflows.



Digital Archive Releases

Batch release

Initial release

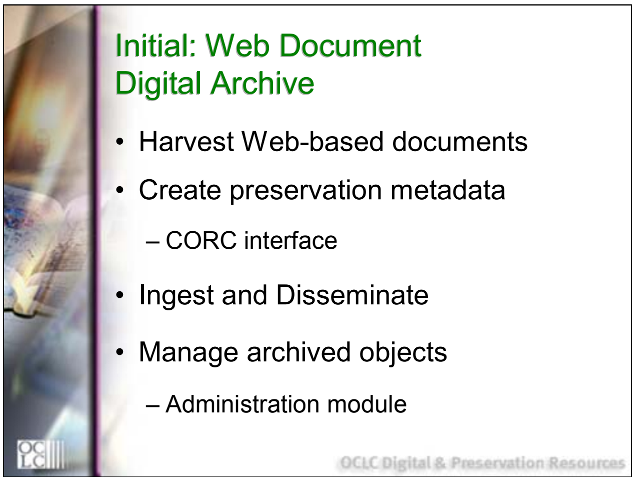
OCLC Digital & Preservation Resources

s I mentioned, we're in the initial release phase now, where pilot participants are ingesting documents from the eb and putting them into the archive, one logical object at a time at this stage.

ext we will be developing batch processing capabilities for various document types such as digital newspapers d contentdm collections.   I'll talk more about future plans later.

CLC is testing the initial release now, and it's a limited release to pilot participants, with general availability of e archive to anyone in September 2002. Newspaper archiving is tentatively set for this Fall, with special llections probably early in the next calendar year.

et me tell you a little about the pilot program.



Initial: Web Document Digital Archive

- Harvest Web-based documents
- Create preservation metadata
  - – CORC interface
- Ingest and Disseminate
- Manage archived objects
  - – Administration module

OCLC Digital & Preservation Resources

he pilot project is called the Web Document Digital Archive, or WDDA. The participants are GPO, the onnecticut State Library, the Library of Michigan, Arizona State Library and Archives, JERRI, made of the ate Library of Ohio, the Ohio Historical Society, the Ohio Supercomputing Center and Ohio General dministration, and the University of Edinburgh. Phase 1 of the pilot ran Sept. 2001 through March 2002. hase 2 is underway now and will run through summer 2002.

ne of the goals of the WDDA project was to work with government libraries who face the issue of fugitive ocuments. . .documents that agencies put on the web, that appear one day then are gone the next when the new sue is out. The Digital Archive provides these institutions with tools to put documents into the archive and to ccess them long-term in their original form.

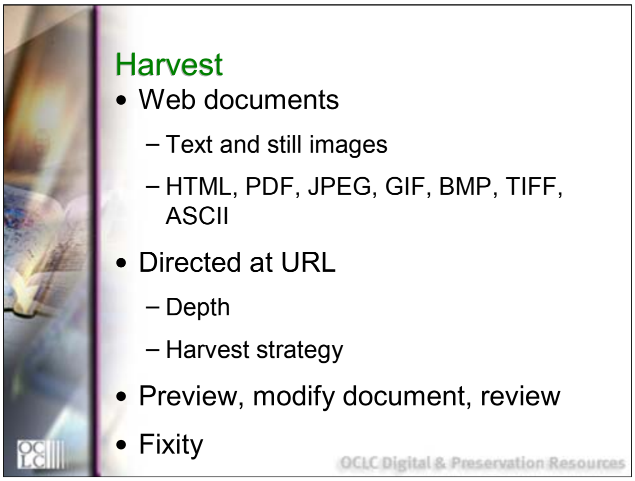
or the purposes of the WDDA project, a web document was defined as being analogous to a print document. It s defined edges, a beginning and an end. It isn't a website. So we're talking about text and still images at this oint.

sing the CORC interface (the Cooperative Resource Catalog), participants create digital archive records ntaining preservation metadata, they ingest the objects and associated records into the archive, and by eptember will be able to disseminate them. They manage the archived objects through an administration odule that provides ways to administrate who can get access, and we'll talk more about this whole process in a inute.

****

he archive will
eserve and provide long term retention to digital content in a changing technological environment.
pport a variety of content and data formats
pport the evolution and migration of formats and function
ndle large volumes of projects and data
ovide the ability to search for and provide access to this data
X 7 fully supported service.

rchive participants can choose service levels for the objects (service levels being storage or migrate, number of ews)



Harvest

- Web documents
  - Text and still images
  - HTML, PDF, JPEG, GIF, BMP, TIFF, ASCII
- Directed at URL
  - Depth
  - Harvest strategy
- Preview, modify document, review
- Fixity

OCLC Digital & Preservation Resources

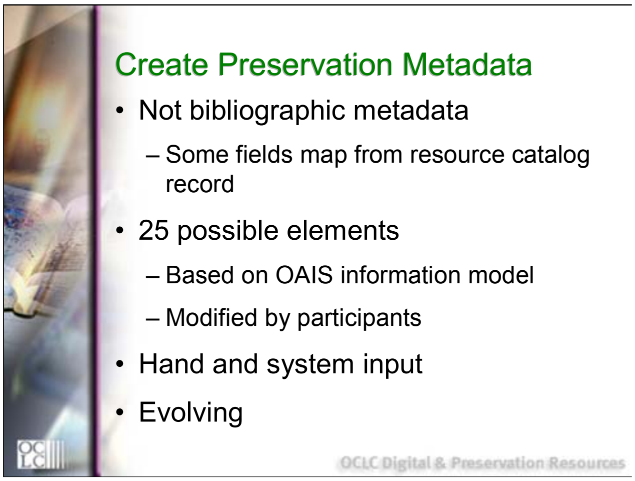
et′s start with harvesting digital objects.

eb documents that can be harvested include text and still images and a limited number of file types: HTML, DF, JPEG, GIF, BMP, TIFF, and ASCII text.

Once you′ve harvested a document , you bring it down to a holding pen at OCLC. This allows you to preview the object to see what you′re planning to put in the archive. Here you have a choice of harvest properties such as depth of harvest, or how deep into the web site do you want to harvest, how many levels? And you also select a harvest strategy of harvesting only the current path or harvesting all links.

You then preview the object. You can edit at that time if you like and change your harvest strategy before actually completing the harvest process.

Once you′ve harvested, you still have to ingest the object and the corresponding digital archive record. We′ll talk about that now.



Create Preservation Metadata

- Not bibliographic metadata
  - Some fields map from resource catalog record
- 25 possible elements
  - Based on OAIS information model
  - Modified by participants
- Hand and system input
- Evolving

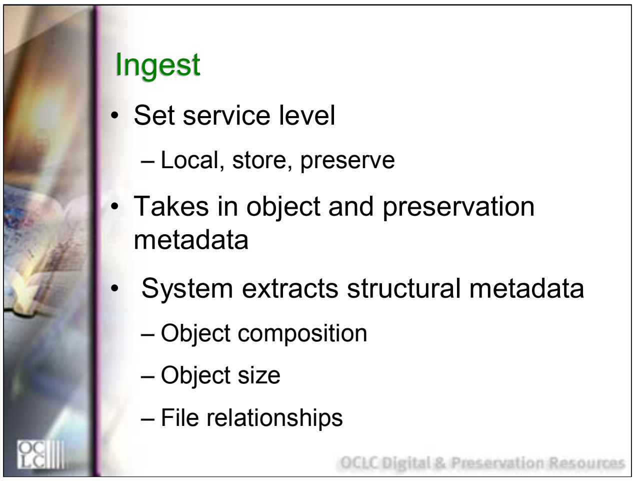
OCLC Digital & Preservation Resources

ve mentioned that you create a digital archive record containing preservation metadata. You do this in the ORC interface. There are 25 preservation metadata elements and they're published on the OCLC website. he preservation metadata set is being developed by an OCLC team and is based on the OAIS model, with input om other initiatives and the OCLC/RLG working group.

he digital archive record is a combination of information picked up from the CORC bibliographic record and me information that you input by hand. The digital archive record is used for collection management purposes d for grouping things together, and very importantly, to help preserve the item in the future.

****

e have tried to achieve a balance of what elements are essential in order to preserve and maintain access to an oject, what are the elements our users can practically create, and what elements the archive can extract or create.



Ingest

- Set service level
  - Local, store, preserve
- Takes in object and preservation metadata
- System extracts structural metadata
  - Object composition
  - Object size
  - File relationships

OCLC Digital & Preservation Resources

hen you're ready to actually ingest an object, you first need to set your service level. There are ree service levels, those being local, store and preserve.

ocal means that you just want to use the OCLC harvesting tools, but not store your digital objects at CLC, but instead you plan to store them in a local repository. For this service level, you'd estimate e number of ingests you think you'll do in a year, and you'd be charged an annual subscription fee.

he Store service level means that you want to use all the tools to bring your documents in and that ou want to store them in OCLC's digital archive. Fees for this level will be based on a monthly lculation of the gigabytes used for storage, plus an annual subscription fee.

bjects that are stored include benefits of OCLC's state-of-the-art storage management, such as:

Uninterruptable Power Supply (UPS)

Backups with Disaster Recovery Plans

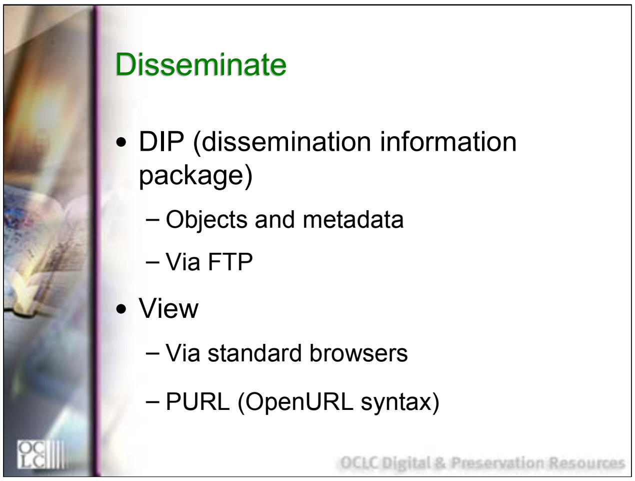
Raised-floor computer rooms

so a number of other services will be applied to documents in storage. These include checksum calculation for checking fixity. Also virus detection after an object has been archived.

he third service level is Preserve. If you select this option, which by the way is not available yet, CLC will provide the storage level I just mentioned, but also management of technological changes to nsure the continued access to objects as formats are abandoned or changed over time (for example ading a microsoft Word document 20 years from now). This might be done by migrating the bjects to another format, although probably the preferred way to handle antiquated formats will be onversion on the fly or derivative forms.

****

he third bullet simply says we know how to put the objects together, take them apart, put them back together gain and still have it work.



Disseminate

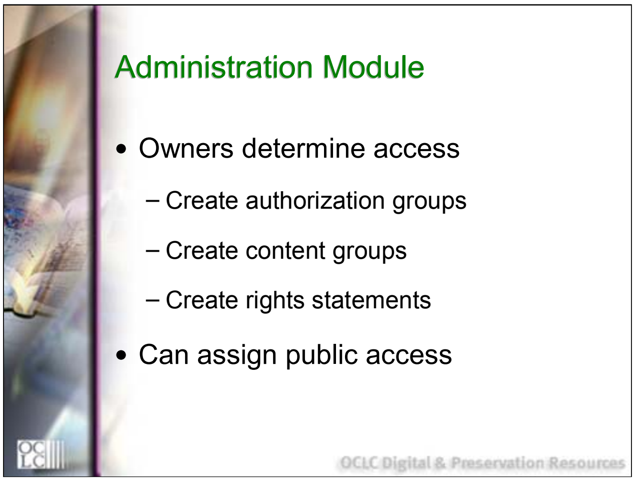
- DIP (dissemination information package)
  - Objects and metadata
  - Via FTP
- View
  - Via standard browsers
  - PURL (OpenURL syntax)

OCLC Digital & Preservation Resources

bjects and metadata are disseminated from the digital archive in two ways: (1) objects may be viewed via a andard browser, and (2) objects will be packaged into a DIP format and can be pulled from the archive via FTP.

**************************8

isseminate to a local archive" is important because we view this archive as interoperable and standards-based. e believe the OCLC archive will be one archive in a distributed network of archives.



Administration Module

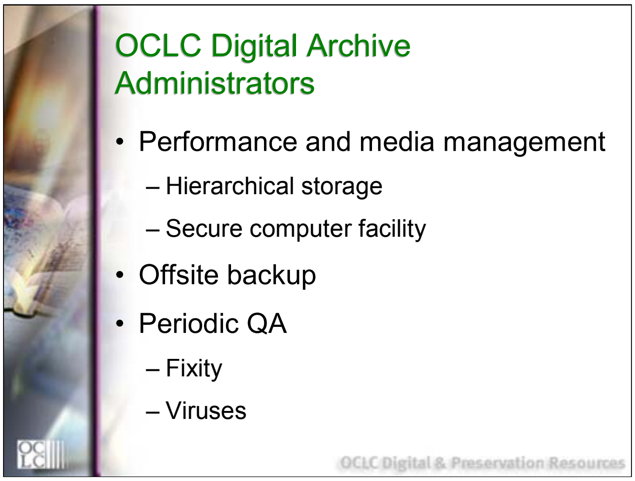
- Owners determine access
  - Create authorization groups
  - Create content groups
  - Create rights statements
- Can assign public access

OCLC Digital & Preservation Resources

he owners still own the object even though it’s in the archive. Through the flexible Administration Module, the wners control who has access. They can create authorization groups: these particular authorizations have cess, or the owner can assign public access which means anyone can see the digital object.

wners also can create content groups for collections, and these named groups have access control lists which efine who can access the account.

d can create rights management statements where any copyright statement you want to present can be input.



OCLC Digital Archive Administrators

- Performance and media management
  - Hierarchical storage
  - Secure computer facility
- Offsite backup
- Periodic QA
  - Fixity
  - Viruses

OCLC Digital & Preservation Resources

hink I covered all the points on this slide while talking about other slides, so let’s skip this one.

****

ll service levels include a basic set of functions for all objects to ensure fixity, multiple backups, virus checking and other activities revolving around keeping the object “safe”.

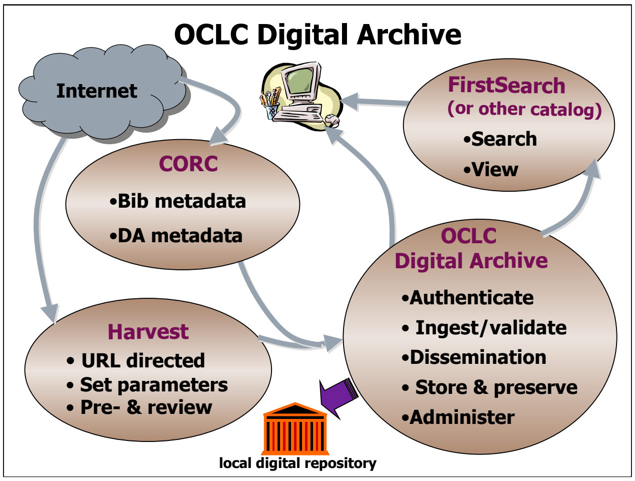
he archivist sets the service level for an object during ingest. The institution archive administrator may interactively change the service level for an object.

he institution administrator may group objects into named groups. These named groups have access control lists which define who may access the content.

hecksum and virus checking ensure integrity

ierarchical storage management (Tivoli storage management at OCLC)

telligently store objects on media depending on metadata (based on format, rights, service level, age, access



OCLC Digital Archive

Internet

CORC
•Bib metadata
•DA metadata

FirstSearch
(or other catalog)
•Search
•View

OCLC
Digital Archive
•Authenticate
• Ingest/validate
•Dissemination
• Store & preserve
•Administer

Harvest
• URL directed
• Set parameters
• Pre- & review

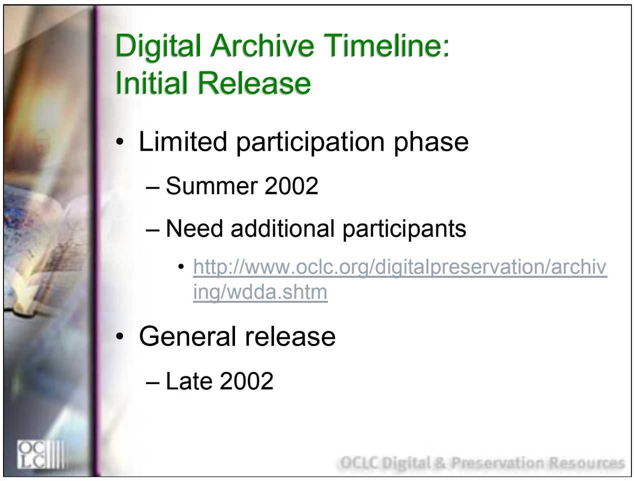
local digital repository

ve been giving you words about what we're doing, so let me show you how the digital archive works, generally.

et's say you're in firstsearch and you find a web document on the Internet that you want to archive. You start creating a bibliographic record in CORC, and a separate digital archive record with preservation metadata. he system is designed to be flexible. For instance, the preservation metadata could be created before the bliographic record or afterwards. In either case, the CORC interface maps some fields from one record to the her, eliminating rekeying.

ou then launch the harvester to bring the web document down to the holding pen at OCLC, and part of that ocess is setting parameters for depth of harvest – how deep into the website do you want to go. You preview e object, edit it if you like, and then ingest it into the digital archive.

uring the ingest process, the user is authenticated, and the object is validated. When the ingest is complete, the ject and record have been moved into the archive. The object and digital archive record can be disseminated to local repository, or stored and preserved at OCLC. Digital objects are not added to worldcat. Only the stitution sees their own digital archive record.

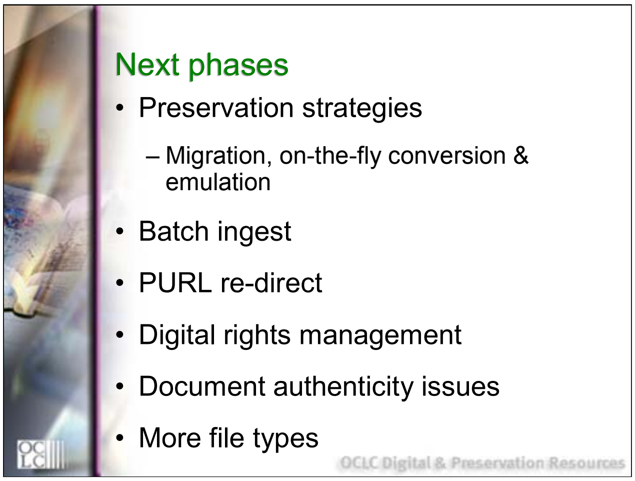


Digital Archive Timeline: Initial Release

- Limited participation phase
  - Summer 2002
  - Need additional participants
    - http://www.oclc.org/digitalpreservation/archiving/wdda.shtm
- General release
  - Late 2002

OCLC Digital & Preservation Resources

e’re in the limited participation phase now, with the general release planned for fall 2002.



Next phases

- Preservation strategies
  - – Migration, on-the-fly conversion & emulation
- Batch ingest
- PURL re-direct
- Digital rights management
- Document authenticity issues
- More file types

OCLC Digital & Preservation Resources

reserving the digital objects through migration, conversion and emulation are going to be part of the next phase. owever nothing is decided yet, other than OCLC is committed to doing this, and much investigation remains to done in this area.

ext phases will allow for batch ingest of various formats: serials, newspapers, special collections

he other bullets on the slide are issues that are community wide and OCLC will continue to work towards lping to develop best practices and standards in digital preservation.

et me check the time. . . Tom will talk next on the OAIS model.   OR, looks like we're running short of time, so t me finish here and we can take a break before the next event.
****

igration: Plans include research and collaboration PURL redirect: feature which allows the "document of cord" to remain on the official website, and then retrieve from the archive when unavailable at the website

apture

Detect changes to a site

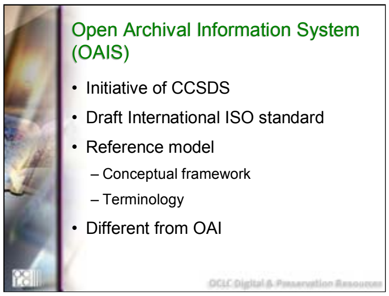
Detect changes to a document

Goal is unmediated harvest to archive

udio or video?

ghts management: containers for all or parts of docs, copyright fee management, watermarking for non-image, gital certificates

ocument authenticity is concerned with proving that the document is the "copy of record" at all points in the ocument lifecycle, including document harvest, display, and dissemination to other archives.



Open Archival Information System (OAIS)

- Initiative of CCSDS
- Draft International ISO standard
- Reference model
  - Conceptual framework
  - Terminology
- Different from OAI

OCLC Digital & Preservation Resources

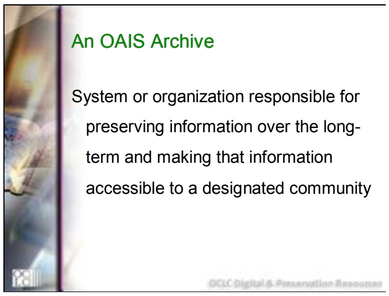
The CCSDS was formed in 1982 as an international forum to develop standards for handling data related to space research. In 1990, CCSDS entered into a cooperative agreement with ISO whereby the CCSDS recommendations would receive ISO review and evolve into an ISO standard.

The Open Archival Information System (OAIS) reference model, recommended by CCSDS, is a conceptual framework. It is a tool that increases awareness of concepts, defines terminology, entities and entity relationships. It is a foundation for creating standards related to digital archiving.

ISO = International Standardization Organization

CCSDS = Consultative Committee for Space Data Systems

OAI = Open Archive Initiative. It is an application protocol for the exchange of metadata. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. http://www.openarchives.org/



An OAIS Archive

System or organization responsible for preserving information over the long-term and making that information accessible to a designated community

OCLC Digital & Preservation Resources

When people speak of “an OAIS” archive, what is meant is a system or organization responsible for preserving information over the long term and making it accessible to a specified class of users (known as the Designated Community).

The OAIS reference model does not specify how to build the archive, only what entities compose the system and how those entities interrelate. It doesn’t specify what types of materials or how accessible those materials should be (“open” refers to the fact that the model was developed in open forums, as will be any future recommendations).

Therefore, as the reference model received attention from a wide, varied community and digital preservation projects have incorporated OAIS concepts or have been influence by them, the resulting archives are not all identical.

(Above info from Lavoie)

In the library community there are a variety of digital preservation projects based on OAIS, both in the U.S. and abroad. A brief, incomplete list is

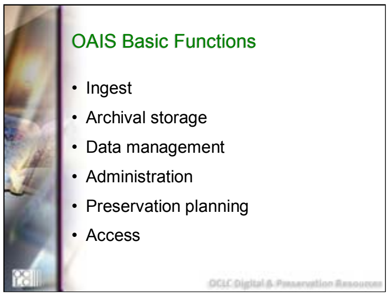
CEDARS (CURL Exemplars in Digital Archives)

NEDLIB (Networked European Deposit Library) --

NAL – National Agriculture Library

LC – Library of Congress

OCLC (Digital Archive)



OAIS Basic Functions

- Ingest
- Archival storage
- Data management
- Administration
- Preservation planning
- Access

OCLC Digital & Preservation Resources

As already mentioned, the OAIS reference model does not tell you how to build an archive. It describes what entities are necessary and how those entities should relate for the long-term preservation and accessibility of information.

Remember, not all these functions are machine or automated. Some are still people!

Ingest = Accepts the data object (I.e. information to be preserved) and any associated metadata into the archive

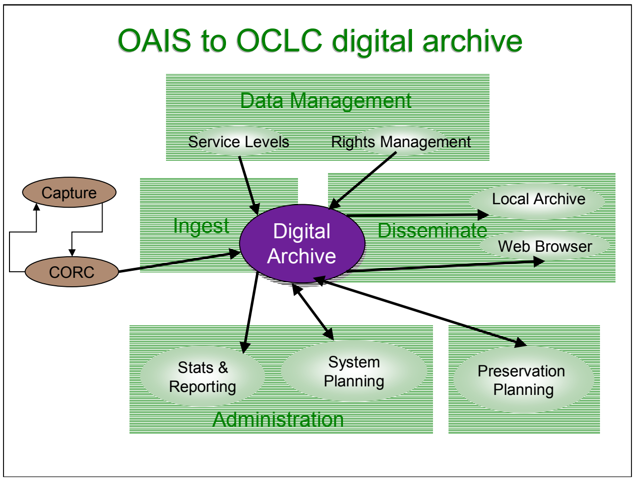
Archival Storage = storage, maintenance and retrieval of the archived data object

Data management = manages, maintains, and provides access to the metadata associated with the archived data object, and administrative data.

Administration = services & functions for the overall operation of the archive system

Preservation planning = monitoring environment, working with designated community, evaluating contents of archive, recommending methods to ensure long-term accessibility of the archive content

Access =



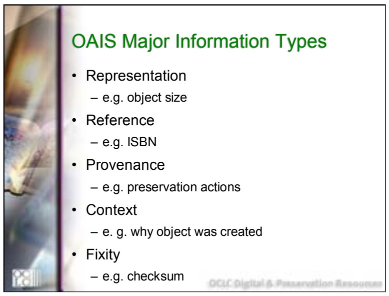
OAIS to OCLC digital archive

his document attempts to map the OCLC digital archive feature set we've been discussing into OAIS functions. is meant to be a loose interpretation for illustrative purposes, rather than a strict mapping. As you can see, the apture and CORC software exist outside of OAIS because the purpose of these components is to acquire and escribe the content.

issemination of OCLC digital archive objects is through either display to a web browser or creation of a DIP to e pulled by an institution via FTP.

he data management functions which are most pressing in getting the archive up and running are the Service evels and Rights Management features discussed previously.

kewise, in the area of administrative, we are focusing on the following areas: (1) statistics; (2) system onfiguration and capacity planning; and (3) Preservation Planning.



OAIS Major Information Types

- Representation
  - e.g. object size
- Reference
  - e.g. ISBN
- Provenance
  - e.g. preservation actions
- Context
  - e. g. why object was created
- Fixity
  - e.g. checksum

OCLC Digital & Preservation Resources

Besides describing the basic functions, the OAIS also provides an information model broadly describing the metadata requirements for long-term retention of digital objects these are the main information (or metadata) types identified by the OAIS. Each information type is followed by an example.

Representation: Representation Information is information necessary to render/display, understand, and interpret the Content Data Object. It includes technical & structural info about the content (data) object, such as structural type, file description, installation requirements, size, functionality, software environment & hardware environment.

Reference: assigned identifiers to unambiguously identify the content (data) object within and without the archive – locally and globally.

Provenance: history of the object, including origin, changes to object OR its content over time, chain of custody, preservation actions.

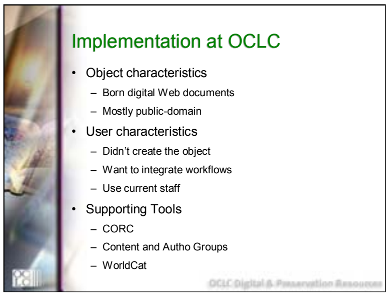
Context: relationship of object to its environment, including other objects (e.g. other manifestations of the same object; as part of a larger collection of objects that itself constitutes a whole), why this objects was created, etc.

Fixity: data integrity checks, validation/verification keys to ensure the object is what it purports to be.

These can be summarized as metadata about the object, its Identity, relationships, history, integrity.

(above from *Preservation Metadata for Digital Objects*, working group white paper).

Besides describing the basic functions, the OAIS also provides an information model broadly describing the metadata requirements for long-term retention of digital objects. The OAIS categorizes this metadata into four types of information objects.



Implementation at OCLC

- Object characteristics
  - – Born digital Web documents
  - – Mostly public-domain
- User characteristics
  - – Didn't create the object
  - – Want to integrate workflows
  - – Use current staff
- Supporting Tools
  - – CORC
  - – Content and Autho Groups
  - – WorldCat

OCLC Digital & Preservation Resources

As with other digital archive implementations, the characteristics of objects and user groups are major factors in metadata decisions and in the tools created to support the metadata creation process.

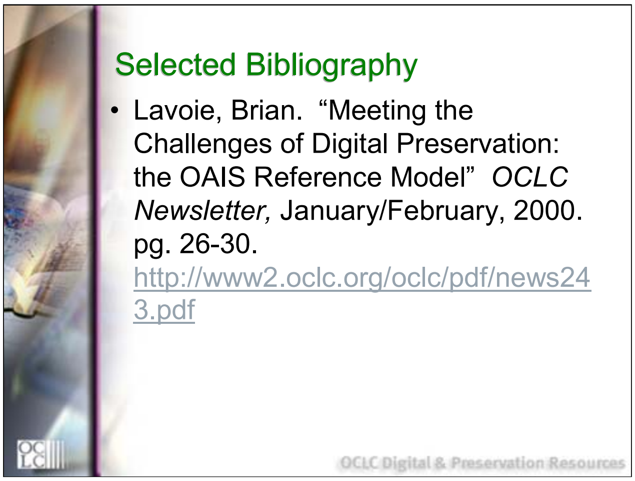
The first objects in the OCLC Digital Archive will be born-digital and mostly public-domain government documents published on the web consisting of text and still images. (HTML, PDF, JPEG, GIF, BMP, TIFF, and ASCII text)

For the most part, the users of our archive in phase 1 are capturing objects created by other people. The significance of this is that they may not know or be able to obtain some preservation metadata elements, for example the recommended hardware for rendering an object.

Also, our users want to integrate workflows: to select, capture, catalog, and archive in a streamlined fashion. They need this to as easy as possible so current staff can ingest objects into the archive with the necessary preservation metadata. They also want a choice of ingesting the object into the OCLC archive, a local archive, or both.

We have created some tools to make the metadata creation easier. We use CORC, OCLC's tool set for creating descriptive metadata for electronic objects, as our foundation. CORC now supports a preservation metadata record, which can be populated with data from a bibliographic record, updated with preservation data extracted from objects by the archive, as well as allowing the user to enter data by hand.

We've created a new harvester that is launched from CORC. We use tools within Oracle 9iFS to extract technical information about the object. We are building a management module to allow the user to assign objects to content groups and then specify the access to that group.

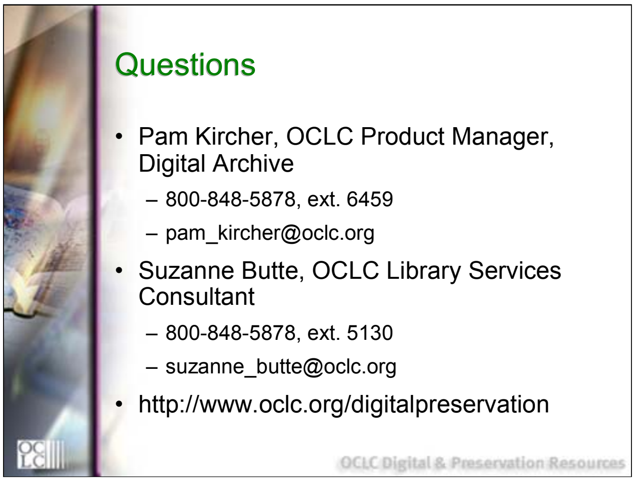


Selected Bibliography

- Lavoie, Brian. “Meeting the Challenges of Digital Preservation: the OAIS Reference Model” *OCLC Newsletter,* January/February, 2000. pg. 26-30. http://www2.oclc.org/oclc/pdf/news243.pdf

Selected Bibliography, continued

- Lavoie, Brian. “Metadata for Digital Preservation.” *OCLC Newsletter*, September/October 2001. p. 33-36. http://www2.oclc.org/oclc/pdf/news253.pdf
- OCLC/RLG Preservation Metadata Working Group. White Papers http://www.oclc.org/research/pmwg/



Questions

- Pam Kircher, OCLC Product Manager, Digital Archive
  - – 800-848-5878, ext. 6459
  - – pam_kircher@oclc.org
- Suzanne Butte, OCLC Library Services Consultant
  - – 800-848-5878, ext. 5130
  - – suzanne_butte@oclc.org
- http://www.oclc.org/digitalpreservation