

Digital Preservation Strategies



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Digital Preservation Strategies

Broadly speaking

- 1) Have a plan
- 2) Recognize the commitment
 - a) Commit the financial resources
 - b) Commit staff resources
- 3) Hire and retain qualified personnel
 - LIS + CS professionals
 - LIS Pros bring knowledge of standards and best practices
 - CS Pros bring knowledge of storage and systems
 - Programming and programmers



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Digital Preservation Strategies

The Digital Landscape

Expect that digital content can be anything, prepare for everything.

Digitized Physical Material

Manuscripts

Audio

Video

Born Digital Originals

Scholarship

Data sets

Email



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Digital Preservation Strategies

The Digital Landscape

Expect that digital content can be anything, prepare for everything.

Digitized Physical Material

- Manuscripts

- Audio

- Video

Born Digital Originals

- Scholarship

- Data sets

- Email

Cultural heritage and institution's societal contributions

- Need to provide long-term access

- Need to preserve indefinitely



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Digital Preservation Strategies

- Copying/Refreshing/Replication
- Migration
- Normalization
- Emulation
- Technology preservation
- Bitstream Identification & Integrity Checks
- Standards & Documentation
 - File Formats
 - Metadata



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Copying/Refreshing/Replication

Copying – the act of making an exact duplicate of the bitstream (i.e. any type of digital file).

Refreshing – the act of copying data from one long-term storage medium to another of the same type.

Replication – the act of copying data to another node.



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Migration

The act of moving data from one hardware and/or software configuration to another. It is more than copying, refreshing, or replicating the data.

Simple example: Creating a new version of an MS Word 2003 document by “migrating” it to the newer MS Word 2007 document format. MS Office software programmed to (presumably) faithfully render the old document into the newer format without information loss, including both data and style attributes.



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Normalization

The act of moving data from one hardware and/or software configuration to another configuration that adheres to a published standard. It is typically more than *migration*.

Simple example: All images converted to TIFFs upon ingest into a digital repository.

XENA – Digital Preservation software from Nat'l Lib of Australia



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Emulation

Replicating a hardware and software environment in which the data object is accessed. The hardware and software environment conforms in nearly all characteristics to an environment in which the data object may have been originally accessed.



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Technology Preservation

Preserving the original hardware (and software – operating systems and applications) needed to access an archived data object. This is a strategy for gaining *access* to old data, not for *preservation* of data.



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Bitstream Identification & Integrity Checks

File Identification - What type of file is it? Is the file well-formed? Is this the original? How old is it?

Information about the data object, such as file format, file size, data characteristics

JHOVE, Metadata Extraction Tool, DROID

File Integrity - Is the file, at bit level, unchanged?

Generally relies on capturing a *checksum* for a data object. A checksum is a unique string derived from a number of characteristics of a data object, such as filesize, created date, modified date, and the data itself. If the data object is altered, the checksum will change.

Built into software – Dspace, Fedora e.g.



Standards & Documentation

File Formats

Desirable		Less Desirable
Open (odt)	vs.	Proprietary (doc)
Widely Used (doc)	vs.	Smaller-adoption (odt)
Not Compressed (WAV)	vs.	Compressed (FLAC)
Lossy (mp3)	vs.	Lossless (WAV, FLAC)
DRM-free	vs.	DRM-protected
Nothing Embedded	vs.	Contains Embedded Content

These are guides, not prescriptions.



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Standards & Documentation

Metadata Standards

Metadata Types and Formats:

Descriptive – What is it?

DC

VRA4

CDWA

MODS

EAD

Darwin Core

MARCXML

Preservation – Who did what to the data object and when?

May also include provenance information

PREMIS



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Standards & Documentation

Metadata Standards

Metadata Types and Formats (cont'd):

Technical – What are the technical characteristics of the data object?

MIX

EXIF

Tools: JHOVE, Nat'l Lib NZ Metadata Extraction Tool, DROID

Use/Rights – What is the copyright status of the data object and how may it be used?

XrML

ODRL

Some metadata schemas often provide means to record Rights Metadata as part of that schema (see e.g. METS, MPEG-21)



Standards & Documentation

Metadata Standards

Metadata Types and Formats (cont'd):

Structural – Connecting the metadata to the data object and contextualizing the data object in its environment

METS
MPEG-21
FOXML



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Columbia College Chicago – Case Study

Organization

College Archives and Library Digital Collections

New library department from reorganization in late Spring 2009
Fair amount of overlap in objectives
Often an overlap in activities
Majority of special digital projects originates with archival material

Pros

Combine some resources (disk storage, file sharing)
Streamline procedures and activities, b/c now one department
Combine some labor (student workers)



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Columbia College Chicago – Case Study

Hardware & Software

Hardware

- Server hosting software
- Disk back-ups being made daily
- Tape back-up system making weekly back-ups (removed offsite)

Software

- Fedora Commons Software
 - Flexible repository architecture
 - Content agnostic
 - Digital Object Storage (all data, binary + metadata)
 - Relationships
 - Built-in ability to check bitstream integrity
- Custom front-end application for cataloging and access



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Columbia College Chicago – Case Study

Documentation

Archives

- Collection Policy
- Retention Policy and Schedule

Library Digital Collections

- Collection Policy
- Digitization & Best Practices Documentation
 - Images, Audio, Video (in progress), Email (in progress)

College Archives + Library Digital Collections

- Identify needs
- Develop digital preservation workflows
 - From triage of data to official ingest



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia College Chicago – Case Study

General Workflow – Audio Example

Type: Oral History

Received format: MP3

- 1) Copy mp3 to Library servers
- 2) Normalize the mp3 file, converting it to a WAV
(original mp3 is not discarded)
- 3) If necessary, modify WAV file to protect personal details of interviewee
- 4) Upload original file (mp3), unmodified WAV file, and modified WAV file (if required) to repository
 - a) Ingest date and time captured, including staff member responsible
 - b) Technical metadata captured. Metadata Extraction Tool used
 - c) Web accessible file created (mp3)



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Columbia College Chicago – Case Study

General Workflow – Audio Example (cont'd)

5) Metadata

- a) Structural metadata – FOXML used (automatic at time of upload)
- b) Descriptive metadata – Local scheme defined (added by cataloger) – mapped to Dublin Core
- c) Use/Rights metadata – Embedded in local scheme (added by cataloger)
- d) Technical metadata – Metadata Extraction Tool (captured at time of upload)
- e) Preservation metadata – Limited – Rely on Fedora Audit DS currently



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 

Columbia College Chicago – Case Study

Areas Requiring Attention

Preservation metadata
Separate and distinct Rights metadata

More formal documentation required



Digital Preservation Strategies

Ensuring Enduring Access: A Forum on Digital Preservation
21 July 2009 - iHotel - Champaign, Illinois

Kevin Ford
kford@colum.edu

Columbia COLLEGE CHICAGO 