Practical Digital Preservation 2015/16

• Welcome!

• PDP Online Workshops - with focus on records and email (Tuesdays 2-4pm Eastern)
  • Nov 10 2015: Part 1
  • Nov 17 2015: Part 2
  • Mar 08 2016: Part 1
  • Mar 15 2016: Part 2
  • May 10 2016: Part 1
  • May 17 2016: Part 2

• PDP “Hot Topic” Webinars (Tuesdays 2-3pm Eastern)
  • Achieving ISO Standards for your digital archive Oct 28 2015
  • Ingesting records from multiple sources and systems Dec 08 2015
  • Automating email archiving and preservation Feb 23 2016
  • New ways of providing public access to your archive Apr 26 2016
  • Real-world digital preservation and program/resources round-up Jun 21 2016

Workshop Objectives

Module 1 (last week)
Understand the **fundamentals of Digital Preservation** - moving beyond the main acronyms and theory by illustrating topics with examples and demonstrations of practical real-world digital preservation workflows and processes.

Module 2 (today)
Understand how Digital Preservation fits into the Information Governance **lifecycle** – including content ingested from other systems (e.g. long-term records and emails) - as well as how to provide greater “transparency” through controlled access to information for internal and public users.
Digital Content and Records are Fragile

- Need to protect from loss and degradation
- Need to proactively manage file formats & metadata
- Digital content already 10+ years old or needs to be retained for 10+ years is at risk
## Long-Term Records

<table>
<thead>
<tr>
<th>Example non-permanent records</th>
<th>Keep for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal health records</td>
<td>Must not be destroyed, or deleted, for the foreseeable future</td>
</tr>
<tr>
<td>Employee Health records</td>
<td>Review 9 and then 25 years</td>
</tr>
<tr>
<td>Public contract records</td>
<td>30 years</td>
</tr>
<tr>
<td>Pension Corporate records</td>
<td>25 years</td>
</tr>
<tr>
<td>Pension personal record</td>
<td>Until 100th Birthday</td>
</tr>
<tr>
<td>Student records</td>
<td>Life of student’ (which is taken to be 120 years from date of birth)</td>
</tr>
</tbody>
</table>
## Digital Records – Short Term

**Enterprise Content/Records Management**

| Content collaboration, governance, policy, retention and disposal |
| Bit-level storage |
| Short-term retention (< 10 years) |

- Structured management of operational content
- Short term retention
- Protect the file ‘bits’
### Digital Records – Long Term

**Digital Preservation**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-term usability and readability of digital content</td>
<td>Protect file accessibility</td>
</tr>
<tr>
<td>Protection against file format obsolescence</td>
<td>OAIS (ISO: 14721)</td>
</tr>
<tr>
<td>Long-term and permanent retention (&gt; 10 years)</td>
<td></td>
</tr>
</tbody>
</table>

- Long term & permanent retention of digital content
- Structured and unstructured content
- Protect file accessibility
- OAIS (ISO: 14721)
Making Digital Preservation part of the Information Governance Lifecycle

Enterprise Content Management System

Access Policy
Catalogue
Records Management

Content Management System
Information Manager

Digital Preservation System

Creation
Use / Maintain
Short Term Retention
Long Term Retention
Permanent Retention

Content Lifecycle (time)

Retention Policy
Disposal Appraisal

Digital Content Contributor

Digital Content Consumer
Making Digital Preservation Part of the Information Lifecycle

Digital Preservation System

Management | Preservation
--- | ---
Ingest | Access
Storage | Admin

Contributor

Website Harvesting

CALM

ArchivesSpace

AXIELL

SharePoint

Gmail

Outlook

EMC²

Documentum

FileNet

Lotus Notes

Alfresco

OPENTEXT

Oracle WebCenter

CoSA

Council of State Archivists

Preservica

Digital Preservation
Practical OAIS Digital Preservation

Jack O’Sullivan
Technical Consultant, Preservica
## Recap: Module 1
**Digital Preservation Fundamentals**

### Module 1

| Session 1 | Why do we need Digital Preservation?  
The fundamentals of preserving digital content |
|------------|---------------------------------------------------------------------------------|
| Session 2 | Understanding Metadata, Fixity and File Characterization  
- including example demonstration |

---

**CoSA**
Council of State Archivists

**Preservica**
Digital Preservation
## Agenda: Module 2
### DP in Information Governance Lifecycle

<table>
<thead>
<tr>
<th>Module 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Session 3</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Questions</strong></td>
</tr>
<tr>
<td><strong>Session 4</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Questions &amp; Close</strong></td>
</tr>
</tbody>
</table>
Digital Records: What’s Important?

- Authenticity
- Provenance

Module 1

- Preservation
- Retention & Disposition
- Access, Security & Privacy

Module 2
Workshop demonstrations

The live practical examples use the Preservica Preservation system

Other systems are available:
Session 3

How do we plan and act to mitigate digital preservation issues?
Planning to Have a Plan

Websites

CALM

Administration

Data Management

Ingest

Archival Storage

Access

Preservation Planning

Software Logos:
- DSpace
- CONTENTdm
- HP TRIM
- Lotus Notes
- Axiell
- SharePoint
- Outlook
- EMC Documentum
- FileNet
- OpenText
- CoSA
- Preservica
DEMO: MIGRATING CONTENT TO AN OAIS REPOSITORY
Passive Preservation

Passive Preservation is concerned with the secure storage of digital objects.

- Security and access control
- File integrity
- Storage management
- Media selection & refresh
- Disaster recovery
**Strategy**

**DEFINITION:** The organization charged with the preservation of permanent electronic government records must proactively mitigate the risks associated with technology obsolescence including plans related to periodic renewal of storage devices, storage media, and adoption of preferred preservation file formats.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The Archives/RM unit does not have a plan to address technology obsolescence.</td>
</tr>
<tr>
<td>1</td>
<td>The Archives/RM unit accepts electronic records in their native format on an ad hoc basis with the expectation that new software will become available to support these formats. Organizations must be aware that not all formats they accept in this manner might be accessible over time. <em>Preservation ready</em> policies can address which formats are acceptable and which are not.</td>
</tr>
<tr>
<td>2</td>
<td>The Archives/RM unit encourages records producers to retain records in formats that are compatible with the preservation-ready file formats. Preservation formats will depend on your organization’s preservation capabilities. Level 2a: The Archives/RM unit proactively and systematically monitors change that may impact the digital records collections and the archival repository. Level 2b: The Archives/RM unit develops and implements policies and procedures for selecting, acquiring, organizing, and preserving an electronic record collection.</td>
</tr>
<tr>
<td>3</td>
<td>The Archives/RM unit implements the transformation of selected native file formats to preferred formats when necessary.</td>
</tr>
<tr>
<td>4</td>
<td>The Archives/RM unit to produce and maintain records from records producers.</td>
</tr>
</tbody>
</table>

The Risk: Format Endangerment

Windows can't open this file:
File:  WordPerfect for MS-DOS 5.0.wpd

To open this file, Windows needs to know what program you want to use to open it. Windows can go online to look it up automatically, or you can manually select from a list of programs that are installed on your computer.

What do you want to do?

- Use the Web service to find the correct program
- Select a program from a list of installed programs

OK  Cancel
Active Preservation Planning

Active Preservation is concerned with the preservation of the underlying information.

- **Institutional risk register**
  - Is the information in a legacy system?
  - Can we identify it?
  - Can we validate?
  - Do we have rendering software?
  - Is software widespread in the designated community?
  - Is it a binary format?
  - Is it a proprietary format?
  - Is it well documented?

- **Identify alternative formats**
  - Do we have software that generate those formats?

- **Identify rendering tools**
Emulation
Normalisation

• Support a small number of file formats

• Migrate on ingest to these formats

• Enhance control of your repository

• Migration pathways not always available
  – You may end up rejecting submissions

• Always likely to incur some loss
DEMO: RUNNING A NORMALISATION INGEST
Migration

• Transform information from one file format into another

• Guard against format obsolescence

• Delays transformation until needed

• Always likely to incur some loss

• On-going process
Reasons for Migration

• Preservation
  – The file is unsupportable in it’s current format

• Presentation
  – We can support the format, but it’s not very useful for widespread dissemination
Mitigating Loss

• Changing format will normally lead to some loss of fidelity

• Validate – Can we accept the compromise?
  – Manually, small scale, run tests
  – Automatically, bulk migrations
Prisoners (Temporary Discharge for Ill-health) Act, 1913.

AN ACT

To provide for the Temporary Discharge of Prisoners whose further detention in prison is undesirable on account of the condition of their Health.

Chapter 4. 25th April 1913.
DEMO: FILE FORMAT MIGRATION
Questions?

& short break
Backup and Restore

Farm Backup and Restore
- Perform a backup
- Restore from a backup
- Configure backup settings
- View backup and restore history
- Check backup and restore job status

Granular Backup
- Perform a site collection backup
- Export a site or list
- Recover data from an unattached content database
- Check granular backup job status
Site Or List Export

Readiness
- No export is in progress.
- Timer service is running.

Site Collection
Select a site or list to export. First select the site collection that contains the site, then select a site to export. To export a list, select the site collection and site that contain it, and then select the list.

File location:
Specify the destination for the export package.

Export Full Security
Export full security of the site, including author, editors, created by times, and modified by times. This also includes all users in the sites.

Export Versions
Select the version history information to export.
Site Or List Export

SharePoint

Monitoring
Backup and Restore
Security
Upgrade and Migration
General Application Settings
Apps
Office 365
Configuration Wizards

Site:
List: Project Files 1995

File location:
Specify the destination for the export package.

Example: E:\Sharepoint\ProjectFiles.cmp

Export Full Security
Export full security of the site, including author, editors, created by times, and modified by times. This also includes all users in the sites.

Export Versions
Select the version history information to include for files and list items. You can include all versions, the last major version, the current version, or the last major and last minor versions.

Filename:
E:\Sharepoint\ProjectFiles.cmp

Overwrite existing files
Example: E:\backup\SharePoint\export.cmp

Export full security

Export versions
All Versions

Start Export  Cancel
Granular Backup Job Status

Readiness
- No site collection backup is in progress.
- An export is currently in progress.
- Timer service is running.

Refresh | Delete Export Job

Site Collection Backup

Current Job
Status: No operation in progress.

Previous Job
Status: No previous job.

Content Export

Current Job
Status: Operation initializing.
Requested By: PRESERVICA\esuj
Site Collection URL: http://sdb4:8080
Server Relative URL: /Project Files 1995
Filename: E:\SharepointDemo\ProjectFiles.cpp
Log File generated: E:\SharepointDemo\ProjectFiles.cpp.export.log
Overwrite: Yes
Recovery Step: To recover the data use the PowerShell import command Import-SPWeb. For more details, type Import-SPWeb -? at the PowerShell command prompt.

Previous Job
Status: Succeeded
Session 4:

Controlling access to digital content

Pulling it all together: Preservation of emails as a complex digital record
Accessing Packages

Repository

Request

Metadata

This file contains additional information, probably added from the digital camera or scanner used to create or digitize it. If the file has been modified from its original state, some details may not fully reflect the modified image.

<table>
<thead>
<tr>
<th>Camera manufacturer</th>
<th>Canon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera model</td>
<td>Canon EOS 50D</td>
</tr>
<tr>
<td>Author</td>
<td>unknown</td>
</tr>
<tr>
<td>Exposure time</td>
<td>1/150 sec (2.00000)</td>
</tr>
<tr>
<td>F-number</td>
<td>5.6</td>
</tr>
<tr>
<td>Date and time of date generation</td>
<td>2021-02-20</td>
</tr>
<tr>
<td>Lens focal length</td>
<td>600 mm</td>
</tr>
</tbody>
</table>

CoSA
Council of State Archivists

Preservica
Digital Preservation
DEMO: DIP EXPORT
Widespread Access
What are you protecting?

- Hardware/equipment
- Digital files
- Content of the digital files
  - Including sensitive information
- Your investment!
- Your reputation!
Protect Content from Ourselves

- Accidental
  - Deletion
  - Saving over original

- Intentional
  - Disgruntled employee causes harm
  - Abusing permissions to change/remove files
  - Misuse of sensitive information
Inappropriate Access

- Hackers
- Viruses
- Trojan horses
- Worms
**Security**

**DEFINITION:** Contemporary enterprise-wide information systems typically execute a number of shared or common services that may include inter-process communication, name services, temporary storage allocation, exception handling, role based access rights, security, backup and business continuity, and directory services, among others. An ISO 14721/ISO 16363 conforming archival repository is likely to be part of an information system that may routinely provide some or perhaps all of the core security, backup, and business continuity services including firewalls, role based access rights, data transfer integrity validations, logs for all preservation activities, including failures and anomalies to demonstrate an unbroken chain of custody.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Currently, the archival repository does not have formal disaster recovery, backups, or firewall procedures in place to protect the security of electronic records.</td>
</tr>
<tr>
<td>Level 1</td>
<td>The security of electronic records in the archival digital repository is protected through disaster recovery procedures.</td>
</tr>
<tr>
<td>Level 2</td>
<td>The security of electronic records in the archival digital repository is protected through a comprehensive firewall protection.</td>
</tr>
<tr>
<td>Level 3</td>
<td>The security of electronic records in the archival digital repository is protected through a comprehensive role based access rights management.</td>
</tr>
<tr>
<td>Level 4</td>
<td>The archival repository continuously monitors security protection processes and revises them in response to evolving technology capabilities and changing business requirements.</td>
</tr>
</tbody>
</table>

Security Models

- Classes of content
- Classes of user
- Classes of operation
- Policy:
  - User of class X has permission to perform operation of class Y on content of class Z
DEMO: CONTROLLING ACCESS
## Long-Term Non-Permanent Records

<table>
<thead>
<tr>
<th>Example non-permanent records</th>
<th>Keep for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal health records</td>
<td>Must not be destroyed, or deleted, for the foreseeable future</td>
</tr>
<tr>
<td>Employee Health records</td>
<td>Review 9 and then 25 years</td>
</tr>
<tr>
<td>Public contract records</td>
<td>30 years</td>
</tr>
<tr>
<td>Pension Corporate records</td>
<td>25 years</td>
</tr>
<tr>
<td>Pension personal record</td>
<td>Until 100th Birthday</td>
</tr>
<tr>
<td>Student records</td>
<td>Life of student’ (which is taken to be 120 years from date of birth)</td>
</tr>
</tbody>
</table>
## Permanent “Long Term” Records

<table>
<thead>
<tr>
<th>Example records</th>
<th>Long Term Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copyrighted Material</td>
<td>Restricted access for the lifetime of the copyright</td>
</tr>
<tr>
<td>Classified Material</td>
<td>Declassified after 10-75 years, reviewed at 25 yearly intervals</td>
</tr>
<tr>
<td>Donor restricted material</td>
<td>Possibly restricted for the lifetime of the donor or other person.</td>
</tr>
<tr>
<td>Unprocessed material</td>
<td>Held until processed</td>
</tr>
</tbody>
</table>
Classification & Policy

• Automatically classify and set access permissions during ingest

• Search by classification

• Automate future actions (e.g. appraise, delete) based on policy rules
EXAMPLE: EMAIL APPRAISAL
Automated Appraisal of Email - Classifying

What do we know?
Automated Appraisal of Email - Classifying

Personal Email

Object Metadata
From: john.doe@example.com

Ingest

Classification Step
‘From: john.doe@example.com’ = Personal

Preservica
Object Metadata
TR Classification: Personal
From: john.doe@example.com
Access restriction: closed
Automated Appraisal of Email - Policy

Rule 1: ‘Personal’ emails = DELETE

24hr

Flag for Appraisal

Delete
PRESERVATION USE CASE: COMPLEX RECORDS

WEB & EMAIL
Why do we need Web Preservation?

• The web is increasingly the record of our collective memory
What makes it so hard?

- 25 years of evolution
Why do we need Email Preservation?

- Email is increasingly the default medium of correspondence
What makes it so hard?
DEMO: EMAIL PRESERVATION
Overall Recap

Module 1
- Why do we need for digital preservation
- The fundamentals of preserving digital content
- The main standards for digital preservation
- Metadata, Fixity and Characterisation

Module 2
- Digital Preservation Planning and Action
- Controlling Access to digital content
- Part of the information lifecycle (SharePoint and eMail)
- Management of long-term non-permanent records
- Handling Complex formats (like eMail and Websites)
Digital Records: What’s Important?

<table>
<thead>
<tr>
<th>Module 1</th>
<th>Module 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Authenticity</td>
<td>• Preservation</td>
</tr>
<tr>
<td>• Provenance</td>
<td>• Retention &amp; Disposition</td>
</tr>
<tr>
<td></td>
<td>• Access, Security &amp; Privacy</td>
</tr>
</tbody>
</table>
Questions ?
Next Steps ……

- Next webinar: 2-3pm Eastern, Tuesday December 08 2015
  - Ingesting in multiple formats and from multiple systems

- Achieving a Step Change in Digital Preservation Capability

- Safeguarding your vital long-term electronic records
  https://preservica.com/resource/electronic-records-preservation/

- www.preservica.com/resources
Next Steps ……

Workshop Objectives

Understand **how Digital Preservation fits into the Information Governance lifecycle** – including content ingested from other systems (e.g. long-term records and emails) - as well as how to provide greater “transparency” through controlled access to information for internal and public users.

- We value your feedback ;-

- Please complete the short evaluation….
Thank you!

www.preservica.com
info@preservica.com
@preservica
@dPreservation

www.statearchivists.org/