CoSA & Preservica Practical Digital Preservation 2015/16

Practical OAIS Digital Preservation
Online Workshop
Module 1
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 (today)
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 (next Tuesday)
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
Preservica Company Milestones

- **2000**: Award winning project with TNA
- **2002-3**: Complete OAIS (ISO14721) compliance
- **2010**: National Archives & Libraries customers
- **2012**: Leadership in DP research & standards
- **2012**: Archives & Records Practical Digital Preservation Programs
- **2015**: 70+ Public sector & Commercial customers in 12 countries
- **2015**: Preservica New company formed
- **2015**: Public Access & ECM Integration
- **2015**: “Cool Vendor” in ECM

Preservica Digital Preservation

CoSA Council of State Archivists
## Preservica Users (partial list)

### National & Pan-National
- Dutch National Archives
- Malaysian Archives
- Swiss Federal Archives
- Latvian National Archives
- Estonian National Archives
- National Archives of Gibraltar
- UK National Archives
- Royal Commission on the Ancient and Historical Monuments of Scotland
- Austrian Archives
- Finnish National Archives
- HM Government of Gibraltar
- National Archives of Hungary

### Libraries, Museums & Education
- MoMA
- Museum of Modern Art, New York
- The Frick Collection
- Museum of Fine Arts, Houston
- Bates College
- University College London
- Emerson College

### State & Government
- UK Met Office
- UK Parliament
- State of Vermont Archives
- Wisconsin Historical Society
- Archives of Michigan

### Corporate Archives
- Unilever
- HSBC
- Lloyds Banking Group
- BT
- AP
- American Press

### Business & Corporate
- Boston City Archives
- Rotterdam City Archive
- Gemeente Rotterdam
- Boston Archives
- Dorset Council
- Oxfordshire County Council
- Massachusetts State Archives
- Alabama Department of Archives & History
- State Archives and Records Management
- Pennsylvania State Archives
- Kentucky Department for Libraries and Archives
- Texas State Library and Archives
- University of Dundee
- University of Glasgow
- University of Liverpool
- University of Northumbria
- University of Auckland

### National Archives
- National Archives of Hungary
- Finnish National Archives
- Latvian National Archives
- Estonian National Archives
- National Archives of Gibraltar

### European Commission
- European Commission

### Institution Logos
- Museum of Modern Art, New York
- The Frick Collection
- Museum of Fine Arts, Houston
- Bates College
- University College London
- Emerson College
- Unilever
- HSBC
- Lloyds Banking Group
- BT
- AP
- American Press
- Boston City Archives
- Rotterdam City Archive
- Gemeente Rotterdam
- Boston Archives
- Dorset Council
- Oxfordshire County Council
- Massachusetts State Archives
- Alabama Department of Archives & History
- State Archives and Records Management
- Pennsylvania State Archives
- Kentucky Department for Libraries and Archives
- Texas State Library and Archives
- University of Dundee
- University of Glasgow
- University of Liverpool
- University of Northumbria
- University of Auckland
Practical OAIS Digital Preservation

Jack O’Sullivan
Technical Consultant, Preservica
# Agenda: Module 1

## Digital Preservation Fundamentals

<table>
<thead>
<tr>
<th>Module 1</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Session 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Why do we need Digital Preservation?</td>
<td></td>
</tr>
<tr>
<td>• The fundamentals of preserving digital content</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Questions</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Session 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Understanding Metadata, Fixity and File Characterization</td>
<td></td>
</tr>
<tr>
<td>‒ including example demonstration</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Questions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Next Steps and Close</td>
<td></td>
</tr>
</tbody>
</table>
# Agenda: Module 2 (next week)
DP in Information Governance Lifecycle

<table>
<thead>
<tr>
<th>Module 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Recap</td>
<td>• Recap on Module 1 – and Intro to Module 2</td>
</tr>
<tr>
<td>Session 3</td>
<td>• Preservation Planning and Action</td>
</tr>
<tr>
<td></td>
<td>– Ingest and Preservation long-term records from SharePoint</td>
</tr>
<tr>
<td>Questions</td>
<td></td>
</tr>
<tr>
<td>Session 4</td>
<td>• Controlling access to digital content – including practical demonstration</td>
</tr>
<tr>
<td></td>
<td>• Ingesting and Preserving Complex formats</td>
</tr>
<tr>
<td></td>
<td>– Website harvesting (WARC files)</td>
</tr>
<tr>
<td></td>
<td>– Emails and Attachments – demonstration of record classification &amp; action</td>
</tr>
<tr>
<td>Questions &amp; Close</td>
<td>• Next Steps and Close</td>
</tr>
</tbody>
</table>
Workshop demonstrations

The live practical examples use the Preservica Preservation system

Other systems are available:
Session 1: Part 1

Why do we need digital preservation? (exploring the problem)
What is Digital Preservation?

*Digital preservation* is a process to ensure that digital information of continuing value remains accessible and usable.

It involves planning, resource allocation, and application of preservation methods and technologies.
What are the Risks and Consequences?

Institutional risks posed by poor preservation strategy:

- Failure to meet mandate
- Legal & financial liability
- Regulatory fine
- Reputational damage
- Cost of re-digitization
- Born-digital material lost forever
- Loss of knowledge
But also benefits!

Institutional benefits of a good preservation strategy:

• Make digital information more accessible (transparent)
• Enable future reuse of information
• Preserve heritage or corporate memory (brand value)
• Save time searching for information
• Respond quickly to legal and compliance challenges
• Retire existing legacy systems (saving cost)
• Support digital ways of working
Digital Records: What’s Important?

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

Today’s sessions

Next sessions
The Fragility of Digital Content

• Information does not exist in isolation

• Each part can be obsolete within information’s lifetime
Other Long Term Preservation Problems

Vital information on removable or unmanaged media

Lack of metadata to interpret the data

Without organisation, you cannot find your information

Column headings incl. units?
Does 1 = male or female?
Temperature experiment run at?
When was it carried out?
Why not just Print to Paper?

• Lose some key advantages:
  – Content searching
  – Easy to copy without loss
  – Variable content (e.g. database, tracked changes)
  – All behaviour
  – Multiple views

• Some digital records just can’t be printed!
  – CAD models, Video, Audio, GIS etc.
Complexity of Digital Content

Example: Web page

- Really two structures:
  - “Physical” (digital objects)
    - Understood by machines
    - Technology dependent
    - May need to be migrated
  - “Conceptual” (information objects)
    - Understood by humans
    - Technology independent
    - Needs to be preserved
Preservation Strategies

Emulation

Migration

Managed file formats

Usable application software

Compatible operating system

Compatible hardware
Case Study: Andy Warhol’s Amiga Images

• Created in 1985 by Andy Warhol

• Commissioned by Commodore to promote the Amiga 1000

• State of the art technology
  – Pre-release hardware & software

• Unreadable at discovery in 2011
Viking Mars Lander Missions (1)

- NASA’s 1975 Viking Lander probes collected data from the Martian surface
- Datasets were compiled and stored to magnetic tape for long term reuse
- Tapes were stored in climate controlled environments
Viking Mars Lander Missions (2)

- Tapes became brittle and cracked
- The formats could not be decoded in the 1990s
- Ultimately retyped using old printouts!
- Space community proposed what would become OAIS
Figure A-1: Composite of Functional Entities
Session 1: Part 2

The fundamentals of Digital Preservation (exploring the solution)
The Repository

- We need a preservation capable digital repository
Content Types

- **Digitised**
  - Historic Materials
  - Potentially expensive to reproduce (but often possible)

- **Born Digital**
  - Records Management systems
  - Email
  - Web
  - Impossible to reproduce
OAIS Terms

• Open Archival Information System (ISO 14721)

• Information = content + metadata

• Information Packages are the currency
OAIS Information Model

• **SIP**: (Submission Information Package)
  – The information supplied to the repository by a content producer

• **AIP**: (Archival Information Package)
  – The information stored within the repository

• **DIP**: (Dissemination Information Package)
  – The information supplied to a user by the repository to satisfy a request

• These *may* be mapped 1:1
OAIS Actors

• Content Producers
  – The people generating the information and submitting it for long term preservation

• Content Consumers
  – The “designated community”, people who want to use the information

• Content Managers
  – Archivists, collection managers, preservation manager
How do I know my system complies?

- OAIS ISO 14721

- Trusted Digital Repository (TDR)
  - ISO 16363
  - Requirements detailed as a checklist of metrics
    - *e.g.* The repository shall assign and maintain persistent identifier of the AIP and its components so as to be unique with the context of the repository

- Trustworthy Repositories Audit & Certification (TRAC)
  - ISO 16919
  - The process of auditing the repository for compliance to TDR
A Truly Trusted Repository

People

Policy & Process

Content

System Software
Questions?
Session 2

Understanding Metadata, Fixity and File Characterisation
Record Structure
DEMO: RECORD STRUCTURE AND HIERARCHY
Metadata

• Descriptive/Structural:
  – Context to make the information usable
  – Usually created by and for humans

• Access Rights:
  – What to make available and to whom

• Preservation Description Information:
  – Proving authenticity and provenance

• Technical:
  – Enable long term preservation
  – Risk assessment & Validation
The Information Package

- File size
- Image sizes
- Page count
- etc...

- Directories
- Parent/child relationship
- etc...

- Technical Metadata
- Preservation Descriptive Information

- Descriptive Metadata

- Structural Metadata

- Access Rights

- Content

- Author
- Title
- Date
- etc...

- Provenance
- Context
- Reference
- Fixity

- Permissions
- Embargoes
- Copyright
- etc...
Descriptive Metadata

The Archival packages will contain descriptive metadata, this needs to be created by humans.

There is no one standard way of including descriptive metadata

Lots of Standards to choose

- ISAD(G) (EAD)
- Dublin Core
- METS
- MODS
- MARC
- ...
OAIS Representation Information
Lives outside the package

- Tells us how to convert the binary representation of the content into something meaningful

representation Information

- Structural
- Semantic
- Other

- Data types
- File format
- Terminology
- Language
- Encryption
- Software
DEMO: FILE FORMAT REPRESENTATION INFORMATION - REGISTRY
Fixity

• Specific to Digital Material

• Measure to guarantee invariance of the bits
  – “Fingerprint”

• Sometimes to referred to as:
  – Hash (or Cryptographic Hash)
  – Digest (or Message Digest)
  – Checksum

• Popular algorithms MD5 & SHA
FIXITY EXAMPLE
Which is Authentic?

Standard Services Contract

This agreement was made between The General Services Company Inc (the “Supplier”) and the Commonwealth of West Australia (the “Buyer”) on the 31st day of October 2007.

1.2.1 Initial Payment
The Buyer agrees to pay an initial fee of $1,000 by 29th October 2007.

1.2.2 Payment
The Supplier agrees to collect an initial fee of $1,000 by 29th October 2007.

1.2.3 Recurring Collections
The Supplier agrees to collect further payments on the first Monday of each month, or the first business day thereafter.

1.2.4 Excision
If the monthly charge will be in effect for the first year of this contract, thereafter it is to be reviewed and agreed to by both parties on an annual basis, the first revision to be agreed by the 3rd day of October 2008.

1.3 Contract Period
This agreement shall be effective from the date it is made and shall continue in force for a period of five (5) years from the date it is made, and therefor for successive five (5) year terms unless and until terminated by one year prior notice in writing by either party.

Signed by: On behalf of the Provider On behalf of the Buyer

The date it is made and shall continue in force for a period of five (5) year terms unless and until terminated by one year prior notice in writing by either party.
Check the Fingerprints

MD5: 31d49ed2b739f0ae0e639ad1a7bba162   MD5: 078dd7ddf28b1bc4801c737d2ff1dee
SHA-1: f0eb43942966b7df6c3ec236a074652c6810d0b   SHA-1: be5d0f673fa884df3f07cdf4e7652fa333aec9a5
FILE CHARACTERISATION
What is our Information Object?

- What does the object claim to be?
- Do we believe it?
What else do we know about it?

• What makes our object different from other objects of the same type?
Above and below the file level
So what?

- Data is not information!

- Data is the bit-stream
  - The conveyor of information

- Information is what we glean from this data

- The intrinsic characteristics are a proxy for information
DEMO: INGEST & CHARACTERISATION
Questions ?
Next Steps ……

➢ **Module 2**: 2-4pm Eastern, Tuesday Nov 17 2015

➢ **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

➢ [www.preservica.com/resources](http://www.preservica.com/resources)
Workshop Objectives

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.

- We value your feedback ;-
- Please complete the short evaluation....
Thank you!

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

www.statearchivists.org/