Establishing a Digital Preservation Program

Dr. Luciana Duranti
Chair and Professor, Archival Studies, UBC
InterPARES Trust Project Director
Abu Dhabi, UAE National Archives, 28 January 2015
Permanent Digital Records

What are “permanent records” in a digital environment given the volatility of digital materials and the obsolescence of software and hardware?

Permanent digital records are records that have been technologically and formally stabilized and whose intellectual content is unchanged and unchangeable so that they remain accessible, understandable, and trustworthy during and across different generations of technology over time, irrespective where they reside.
Records Trustworthiness

**Reliability**
The trustworthiness of a record as a statement of fact, *based on:*
- the competence of its author
- the controls on its creation

**Accuracy**
The correctness and precision of a record’s content *based on:*
- the competence of its author
- the controls on content recording and transmission

**Authenticity**
The trustworthiness of a record that is what it purports to be, untampered with and uncorrupted *based on:*
- its identity
- its integrity
InterPARES Principles

• Technology must not determine the solution to the reliable and accurate creation of digital records or to their authentic preservation over the long term: **the creator’s needs** define the problem and **archival principles** establish the technical solution

• Solutions to digital challenges are inherently **dynamic** and **specific** to the cultural, disciplinary, administrative and legal situations

• Preservation of permanent digital records is a **continuous process that begins with records creation**

• We must be able to **presume records trustworthiness** till proof to the contrary is established
What is involved?

1. Establishing a digital preservation program
2. Appraising digital records along their life
3. Transferring permanent records to the archives custody as soon as possible
4. Processing permanent records
5. Creating archival descriptions
6. Storing permanent records in a trusted repository
7. Converting and migrating digital records on an established ongoing cycle to maintain accessibility
Establish a Digital Preservation Program

• Determine Scope and Objectives
• Acquire Resources
• Focus on Digital Records
• Offer Advice to Creators
• Set a Good Example
• Develop Procedures
• Implement Maintenance Strategies
Appraise Digital Records

1. As early as possible in the life of the records, **assess** their continuing **value** to the creator (after having identified multiple owners) and/or to society at large

2. Assess the **authenticity** of the records considered for continuing preservation, and document the process

3. Determine the **feasibility** of their preservation (after having identified all digital components) by the designated preserver (creator and/or archives); and,

4. Constantly **monitor** all the records of the creator and, if warranted by the changes that they have undergone through time, revise the appraisal decision

The greatest challenge of digital preservation is not preservation but appraisal
Transfer Permanent Records

- Develop shared plan for transfer
- Enforce standardised procedures
- Keep the oldest available logical format
- Eliminate duplicates
- Document all processing
Process Permanent Records

It involves the creation of authentic copies of the transferred records. Their authenticity is guaranteed by:

- a controlled process of migration of the records to the archives technological environment (always keeping the records in the format in which they were acquired)
- the accurate documentation of any change that the records undergo during such process and every time that the archives technological environment is upgraded
- the implementation and monitoring of privileges concerning the access, use and reproduction of the records within the archives
Process Permanent Records (cont.)

• and the establishment of procedures
  – to prevent, discover, and correct loss or corruption of records
  – to guarantee the continuing identity and integrity of the records against media deterioration and across technological changes (ongoing conversion and migration)
  – to assign responsibility for and means of authentication of individual records, when required; and
  – to ensure redundancy, internally and remotely.

The processing challenges related to formats and migration
Preservation Formats

Selection criteria
• widespread adoption
• non-proprietary origin
• published specifications
• interoperability (platform independence) and
• lack of compression or lossless compression

Standards:
• PDF 1.4 became PDF/A ISO 19005-1:2005
• PDF 1.7 became ISO 32000-1: 2008
• PDF/A-2 became ISO 19005-2:2011
• PDF/A-3 became ISO 19005-3:2012
Migration

- Migration always entails some degree of risk
- An essential element of any migration strategy is an analysis of
  - what the required functionalities of the old format are,
  - which functionalities need to be maintained through migration,
  - who has to use them, and
  - how to best address those needs in a new format
- Tools for migration are:
  - Conversions Software Registry (NCSA—National Center for Supercomputing Applications, “Conversions Software Registry - Query Conversions.”)
  - PRONOM’s DROID (UK The National Archives)
- There is no one-size-fits-all solution.
Archival Description

Archival description acquires a primary authentication function

- The authentication function of archival description is a collective attestation of the authenticity of the records in an archival *fonds* and of all their interrelationships
- Archival description provides an historical view of the records and of their transformations while maintaining the bond of their common provenance and destination
- Archival description of permanent digital records relies on evidence about a record’s identity and integrity from metadata
- Every record creator should have a metadata application profile for its records
Metadata Application Profile

The set of metadata elements, policies, and guidelines defined for a particular application or implementation. The metadata application profile should

• enable the functional requirements of the recordkeeping or records preservation system
• ensure interoperability
• support verification of authenticity
Authenticity: Identity

The whole of the attributes of a record that characterize it as unique, and that distinguish it from other records.

Identity metadata:

• names of the persons concurring in its creation
• date(s) and time(s) of issuing, creation and transmission
• the matter or action in which it participates
• the expression of its archival bond
• documentary form
• digital presentation
• the indication of any attachment(s)
• digital signature
• name of the person responsible for the business matter
A record has integrity if the message it is meant to communicate in order to achieve its purpose is unaltered.

**Integrity metadata:**
- name(s) of handling persons over time
- name of person responsible for keeping the record
- indication of annotations
- indication of technical changes
- indication of presence or removal of digital signature
- time of planned removal from the system
- time of transfer to a custodian
- time of planned deletion
- existence and location of duplicates outside the system
“A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.” RLG-OCLC report Trusted Digital Repositories: Attributes and Responsibilities

There two exemplary models that come both from research:

- Chain of Preservation—InterPARES generated. Integrally used in the NARA ERA and as a model by several others, such as the City of Vancouver.
TDR Requirements

- accept responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users;
- have an organizational system that supports not only long-term viability of the repository, but also the digital information for which it has responsibility;
- demonstrate fiscal responsibility and sustainability;
- design its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it;
- establish methodologies for system evaluation that meet community expectations of trustworthiness;
- be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly;
- have policies, practices, and performance that can be audited and measured;
Purposes of the OAIS Reference Model

In the words of the standard, this reference model provides:

• a framework for understanding archival concepts needed for Long Term digital information preservation and access;

• the concepts needed by non-archival organizations to be effective participants in the preservation process;

• a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives;

• a framework for describing and comparing different long-term preservation strategies and techniques;

• information that is Independently Understandable to the Designated Community;

• preservation against all reasonable contingencies, and dissemination of authenticated copies of the original.
Figure 4-1: OAIS Functional Entities
OAIS and Archives

• It is not by Archivists, or for Archivists

• It does not focus on “records,” “archival fonds,” or “metadata”

• It deals with “Content Information” and “Information Packages”

• A good start for a non-archival community
Based on the OAIS Model

ISO/TR 17068:2012 details the authorized custody services of a Trusted Third Party Repository (TTPR) in order to ensure provable integrity and authenticity of the clients' digital records and serve as a source of reliable evidence.

It describes the services and processes to be provided by a TTPR for the clients' digital records during the retention period, to ensure trust. It also details the criteria of "trustworthiness" and the particular requirements of TTPR services, hardware and software systems, and management.

ISO/TR 17068:2012 has the limitation that the authorized custody of the stored records is between only the third party and the client.
InterPARES 2
Chain of Preservation Model

1. Managing the framework for the chain of preservation
2. Managing records creation
3. Managing records in a recordkeeping system and
4. Preserving selected records.

Model

Explanation:
InterPARES 2: Chain of Preservation Model

<table>
<thead>
<tr>
<th>NODE</th>
<th>TITLE</th>
<th>NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>Manage Chain of Preservation</td>
<td></td>
</tr>
</tbody>
</table>
InterPARES 2: Chain of Preservation Model
Open Source

• Based on free source code, readily available
• Anyone can modify, change, radically alter it
  – Often with proviso that new code given back
• Often with original providers offering integration services as a cost
Pro-Open Source Arguments

• **Lower cost** – volunteers design the software, no centralized ‘company’ to collect revenue or pay employees
• **Customizable** – implementers can develop the system to do whatever is necessary
• **Nimble** – Adaptations implemented faster than typical commercial development cycle
• **Openness** – Easy to look under the hood
• **Faster update cycle** – bugs and security holes can be patched in-house
Anti-Open Source Arguments

- **Quality** – Without strong reporting structure, who ensures consistent, well-documented, bug free code?
- **Responsibility** – Lack of corporate structure means no finger pointing ability, no-one is responsible
- **Interoperability** – Lack of long term product vision may lead development down dead-ends, incompatibility with other products
- **Learning curve** – Lack of commonality with other organizational products may lead to steep learning curve
- **More time, money and effort spend in Research &Development**, testing and usability
Other options

• **Open specifications** – file formats; creator releases the documentation to allow for the creation of ‘home-grown’ tools

• **Shared Source** – Creator shares the source code with limited number of partners, community based/consortium/guild based

• **Escrow Accounts** – Creator maintains IP but places source code in third party for indemnification
When to use proprietary

- Staff skill set is in proprietary tools
- Compatibility with existing software requires it
- Mission critical operation needing immediate response and access (willing and able to pay $$$ for this service)
- Proprietary tools does not mean proprietary file formats!
  - Always make sure you know what the program is doing to your records and how they are being saved!
When to use open source

- When you are starting from scratch
- You have time to build your own
- Your staff skill set is LAMP (Linux, Apache, MySQL and PHP) based
- You have endorsement to provide your own support
- Understanding the ‘behind the scenes’ is vital (legal/evidential)
Looking Under the Hood

- As computer technology becomes more ubiquitous, pervasiveness of data is directly related to software.
- How to determine if software is ‘reliable’
  - Black box (functionalities) vs White box (structure—source code level) testing
  - Repeatability
  - Function by function analysis
- Very difficult with Commercial Off The Shelf
  - Court order to gain access to source code – EX: Breathalyzer
  - Some presumption based on wide distribution (millions of testers)
- Lack of accountability on Open Source Software
  - Requires more due diligence on adopters part
  - BUT, can perform a top down walk through
The Three Considerations

- Time
- Staff
- Money

- Each instance is unique, base judgment on current environment – not on rhetoric
- Understand Total Cost of Ownership
- Ensure that the TDR requirements are met
TDRs

• Microsoft SharePoint
• FEDORA
• Alfresco
• DSpace
• Archivematica
Microsoft SharePoint 2010

- Described as “Business Collaboration Platform for the Enterprise and the Web”
- One of the fastest growing server products in history

Wrox manual states:

“SharePoint is an enabling technology. Merely deploying SharePoint isn’t a silver bullet that will transform your organization. But consciously building a SharePoint environment and training end users how to use the tool appropriately can and will propel an organization to a higher level of productivity.”
FEDORA

- Originally developed at Cornell University (USA)
- Flexible Extensible Digital Object Repository Architecture
- Defines ‘abstractions’ for expressing digital objects, describing relationships, linking necessary ‘behaviors’ to objects
- OAI-PMH (Protocol Metadata Harvesting) and RDF (Resource Description Framework) support
- Fairly small install base, mostly universities and libraries
- Handles at the data-stream level, can ingest any type of content
Alfresco

- Open Source RMA
- Analyzed in InterPARES General Study 08: Open-Source Records Management Software
- Comes in two flavors: Community and Enterprise
- Community free, Enterprise requires support subscription
- Very little help with Community Edition, 32 page manual and little activity on listserv
DSpace

• Institutional repository
  “database with a set of services to capture, store, index, preserve and make accessible ‘university research’ in a digital format over the internet”

• Open Source, Java based
• Developed by MIT, HP
• Used extensively throughout archival field
  – 500 installations, 65+ countries
  – Predominately University based
Archivematica

- A **free and open-source digital preservation system** that is designed to maintain standards-based, long-term access to aggregations of digital objects.
- Uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the **ISO-OAIS** functional model.
- Users monitor and control the micro-services via a web-based dashboard.
- Archivematica uses METS, PREMIS, Dublin Core and other best practice metadata standards.
- It implements format policies based on an analysis of the significant characteristics of file formats.
- It just formed a formal partnership with DSpace. It is a partner of InterPARES Trust.
Archives Online

Preservation institutions are beginning to entrust their holdings to Internet Providers because:

- Many of the materials they are mandated to preserve already exist online
- Access is possible from any location to anyone who can use a browser
- A trusted digital repository satisfying ISO standards as well as basic preservation requirements is not affordable and is often inadequate to the challenge
- The knowledge to deal with the digital products of complex technologies is not commonly available among information professionals and is very expensive
- Strong protection measures are often confused with preservation measures
- but mostly because archives are confronted with…
A Creators’ Generational Change

Generation Y or Millennials -- Post-1981

- integration of private and public
- *produsing*, co-authoring, crowdsourcing
- co-owning, sharing
- distributed workforce vs. BYOD using multiple clouds vs. the Fog using multiple devices for storage
- media convergence
- constant connectivity
- visual language
- “liquid communication,” instantaneous impact,
- ephemeral output
A Real Life Scenario

- The Council of Prairie and Pacific University Libraries (COPPUL) is piloting a **cloud-based preservation service** using the Archivematica digital preservation system. The service is offered to COPPUL member institutions that wish to preserve digital holdings but are unable or unwilling to install and manage local Archivematica instances. This service is a joint effort of COPPUL, Artefactual Systems Inc. (Archivematica lead developers) and UBC Library (the cloud storage provider).

- **COPPUL is responsible for promoting the service**, signing up new institutions and seeding the one-time set-up costs; **Artefactual Systems provides account administration, installation, server administration and user technical support**; and **UBC Library provides fee-based server hosting and digital object storage service**.
Preservation As a Service for Trust (PAST)

InterPARES Trust has focused on modeling preservation online using UML (Unified Modeling Language), which can be easily translated into code and in the development of a prototype. In order to prepare a reference baseline, we are working with the Object Management Group (OMG) to develop an international model of Preservation as A Service for Trust (PAST) detailing all the functional requirements that providers must respect.

Several national archives as well as international organizations are participating to ensure that the concerns of every jurisdiction will be heard, and we will be able to have archival clouds ensuring the permanence of our digital materials.
An Archival Cloud

Transparency:
- An unbroken chain of legitimate custody would be demonstrable
- Records reliability could be inferred from known creation processes
- Records authenticity could be inferred from the documentary context and a known preservation process

Stability:
- An archival cloud would guarantee that each record’s context is defined and unchanged, with all its relationships intact.

Permanence:
- Retention and disposition plans would be integrated with migration plans
- There would be no risk of termination of contract or provider for the archival cloud

We do not have to renounce our preservation mission to protect tomorrow’s digital heritage. We have to frame it differently, with the help of interdisciplinary and international research.
Thank you!

luciana.duranti@ubc.ca
www.ciscra.org
www.interpares.org
www.interparestrust.org