Practical considerations for implementing preservation strategies

Michael Day,
Digital Curation Centre
UKOLN, University of Bath
m.day@ukoln.ac.uk

Driving the long-term preservation of electronic records, London, 27 September 2006

Session overview

- The fragility of digital content
- Digital preservation approaches
- The OAIS model
- Metadata and shared infrastructures
- Some final comments

The fragility of digital content

The main technical issues

General comments

- Digital information is dependent on its technical environment
- Physical objects are subject to:
  - Physical deterioration
  - Technology obsolescence
- Relatively short timescales

Media longevity

- Media has a short (or unknown) life
- Technical solutions:
  - Periodic copying of data bits on to new media or types of media (refreshing)
  - Longer lasting media
  - Migrating to good-quality microform or paper (!)
- In an organised preservation system, regular routines (quality checking, backup, replication, refreshing, etc.) will help solve the media longevity issue
- The key is having managed processes in place

Technology obsolescence (1)

- A set of much bigger problems
- Software dependence
  - Digital content is, at least in part, dependent on the configurations of hardware and software (applications and operating systems) that were originally used to interpret or display them
- Hardware and software obsolescence
  - Application software and operating systems are upgraded regularly
  - Hardware becomes obsolete or needs repair

Technology obsolescence (2)

- Technical solutions
  - Various preservation strategies have been developed to cope with the obsolescence problem
  - For the most part, these depend on the existence of a continual programme of active management (life cycle management)
  - Supported by systems that implement the various functional entities identified by the Reference Model for an Open Archival Information System (OAIS)
  - Preservation strategies can only be seen in this wider context

Multiple layers of meaning (1)

- Digital objects are logical entities not fixed to any one particular physical carrier
- Three layers (Thibodeau, 2002):
  - Physical objects: the actual bits stored on a particular medium
  - Logical objects: defines how these bits are used by application software, based on data types (e.g. ASCII); in order to understand (or preserve) the byte-streams, we need to know how to process them
  - Conceptual objects: what humans deal with in the real world, meaningful units of information

Multiple layers of meaning (2)

- On which of these layers should preservation activities focus?
  - We need to preserve the ability to reproduce the objects, not just the bits
  - In fact, we could change the bits and logical representation and still reproduce an authentic conceptual object

Authenticity and integrity

- Digital information can easily be changed (e.g., by design or accident)
- How can we trust that an object is what it claims to be?
- Mechanisms are available at the bit level (e.g. checksums), but will this be sufficient?

Problems of scale

- An increasing flood of 'born-digital' data
  - Data deluge in science and engineering
    - Petabytes generated by high throughput instruments, streamed from sensors and satellites, etc.
  - The World Wide Web
    - Comprises billions of pages + "deep Web"
    - Internet Archive = >1 petabyte, and growing @ 20 Tb. per month (http://www.archive.org/)
  - 5 exabytes of *new* information created in 2002:
    - http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Some general principles (1)

- Most of the technical problems associated with long-term digital preservation can be solved if a life-cycle management approach is adopted
  - i.e. a continual programme of active management
  - Ideally, combines both managerial and technical processes, e.g., as in the OAIS Model
  - Many current systems (e.g. repository software) are attempting to support this approach
  - Preservation strategies need to be seen in this wider context
  - Preservation needs to be considered at a very early stage in an object's life-cycle

Some general principles (2)

- Need to identify and understand the 'significant properties' (essence) of an object
  - Focuses on what is deemed essential (performance)
  - Helps with choosing an acceptable preservation strategy
  - The relative importance of:
    - Content
    - Appearance and behaviour (look and feel)
    - Context
    - Structure
  - This is the area where decision support tools might be most useful

Some general principles (3)

- Encapsulation may have some benefits
  - Surrounding the digital object - at least conceptually - with all of the information needed to decode and understand it (including software)
  - Produces autonomous 'self-describing' objects, reduces external dependencies; linked to the Information Package concept in the OAIS Reference Model
- We should keep the original byte-steam, just in case … (?)

Digital preservation strategies

Preservation strategies

- A continuum of strategies
  - Some approaches focus on preserving the essential characteristics of objects
    - Migration
    - Persistent archives
  - Some approaches focus on preserving aspects of the technology
    - Technology preservation
    - Technology emulation
  - Also includes:
    - Digital archaeology (emergency rescue)

Technology preservation

- The preservation of an information object together with all of the hardware and software needed to interpret it
  - Successfully preserves the look, feel and behaviour of the whole system (at least while the hardware and software still functions)
  - May have a role for historically important hardware
  - Problems with storage and ongoing maintenance, missing documentation
  - Would inevitably lead to 'museums' of "ageing and incompatible computer hardware" -- Mary Feeney
  - May have a short-term role for supporting the rescue of digital objects (digital archaeology)

Technology emulation (1)

- Preserving the original bit-streams and application software; running this on emulator programs that mimic the behaviour of obsolete hardware
- Emulators change over time
  - Chaining, rehosting
  - Emulation Virtual Machines
    - Running emulators on simplified 'virtual machines' that can be run on a range of different platforms
    - Virtual machines are migrated so the original bit-streams do not have to be

Technology emulation (2)

- Benefits:
  - Technique already widely used, e.g. for emulating different hardware, computer games
  - Retains use of the 'original' bytestream
  - Reduces the need for regular object transformations (but emulators and virtual machines may themselves need to be migrated)
  - Retains ‘look-and-feel’
  - May be the only approach possible where objects are complex or dependent on executable code
  - Less 'understanding' of formats is needed; little incremental cost in keeping additional formats

Technology emulation (3)

- Issues
  - Which organisations have the technical skills necessary to implement the strategy?
  - Preserving 'look and feel' may not be needed for all objects
  - It will be difficult to *know* definitively whether user experience has been accurately preserved
- Conclusions
  - Promising family of approaches
  - Needs further practical application and research

Information migration (1)

- Managed transformations
  - A set of organised tasks designed to achieve the periodic transfer of digital information from one hardware and software configuration to another, or from one generation of computer technology to a subsequent one - CPA/RLG report (1996)
  - Abandons attempts to keep old technology (or substitutes for it) working
  - A 'known' solution used by data archives and software vendors (e.g., a linear migration strategy is used by software vendors for some data types, e.g. Microsoft Office files)
  - Focuses on the *content* of objects

Information migration (2)

- Main types (from OAIS Model)
  - Refreshment
  - Replication
  - Repackaging
  - Transformation
- Issues
  - Labour intensive
  - There are severe problems with ensuring the 'integrity and authenticity' of objects
  - Transformations need to be documented (part of the preservation metadata)

Information migration (3)

- Uses
  - Seems to be suitable for dealing with large collections of similar objects
  - Migration can often be combined with some form of standardisation (normalisation) process, e.g., on ingest to:
    - ASCII (for text)
    - Bit-mapped-page images (for images)
    - Well-defined XML formats (for structured documents or datasets)
  - Migration on Request (CAMiLEON project)
    - Keep original bit-streams, migrate the rendering tools

Digital archaeology

- Not so much a preservation strategy, but the default situation if we fail to adopt one
- Using forensic techniques to recover digital content from obsolete or damaged physical objects (media, hardware, etc.)
  - A time consuming process, needs specialised equipment and (in most cases) adequate documentation
  - Considered to be expensive (and risky)
  - Remains an option for content deemed to be of value

Choosing a strategy (1)

- Preservation strategies are not in competition (different strategies will work together)
  - Suggestion that we should keep the original bits (with some documentation) in case better preservation technologies emerge in the future
- But the strategy chosen has implications for:
  - The technical infrastructure required (and metadata)
  - Collection management priorities
  - Rights management
    - e.g, Owning the rights to re-engineer software
  - Costs

Choosing a strategy (2)

- Decision support for preservation, e.g.
  - Preservation strategies
  - Target formats for transformations
- Examples:
  - Nationaal Archief (Netherlands) testbed project
  - Vienna University of Technology utility analysis-based metrics
  - Both developed further by the Digital Preservation cluster of the DELOS Network of Excellence
  - http://www.dpc.delos.info/

The OAIS Reference Model

OAIS background

- Reference Model for an Open Archival Information System (OAIS)
  - Development led by the Consultative Committee for Space Data Systems (CCSDS)
  - Issued as CCSDS Recommendation (Blue Book) 650.0-B-1 (January 2002)
  - Also adopted as: ISO 14721:2003
    - Currently under review
  - http://public.ccsds.org/publications/archive/650x0b1.pdf

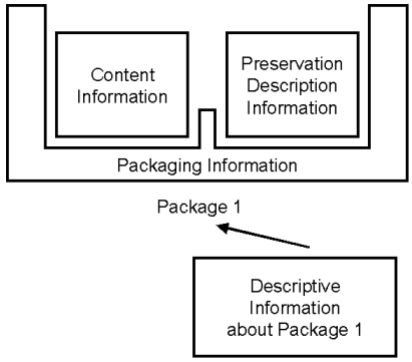
OAIS definitions

- Provides definitions of terms that need to have well-defined meanings, e.g.:
  - Archival Storage, Content Data Object, Designated Community (key term), Ingest, Metadata, Representation Information, etc.
  - OAIS = "An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community" (OAIS 1.7.2)

OAIS high level concepts (1)

- The *environment* of an OAIS (Producers, Consumers, Management)
- Definitions of *information*, Information Objects and their relationship with Data Objects
- Definitions of *Information Packages*, conceptual containers of Content Information and Preservation Description Information

OAIS high level concepts (2)



ationships (Figure 2-3)

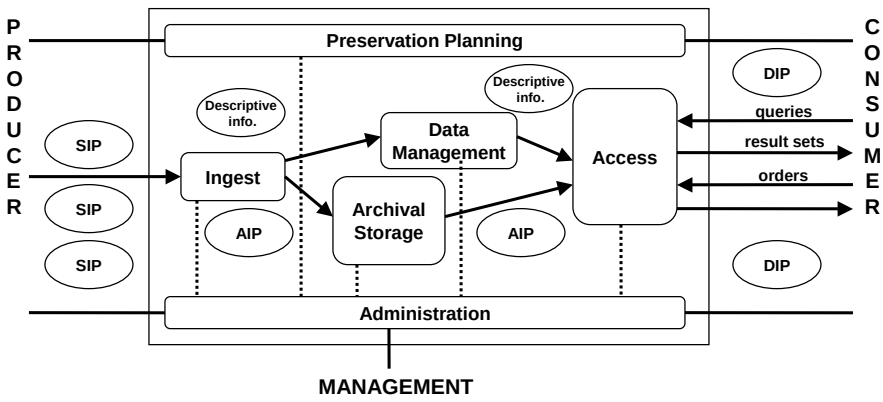
OAIS mandatory responsibilities

- Negotiating and accepting information
- Obtaining sufficient control of the information to ensure long-term preservation
- Determining the "designated community"
- Ensuring that information is **independently understandable**, i.e. without the assistance of those who produced it
- Following documented policies and procedures
- Making the preserved information available

OAIS Functional Model (1)

- Six entities
  - Ingest
  - Archival Storage
  - Data Management
  - Administration
  - Preservation Planning
  - Access
- Described using UML diagrams

OAIS Functional Model (2)



OAIS Functional Entities (Figure 4-1)

OAIS Information Model (1)

- Information Object (basic concept):
  - Data Object (bit-stream)
  - Representation Information (permits “the full interpretation of Data Object into meaningful information”)
- Information Object Classes:
  - Content Information
  - Preservation Description Information (PDI)
  - Packaging Information
  - Descriptive Information

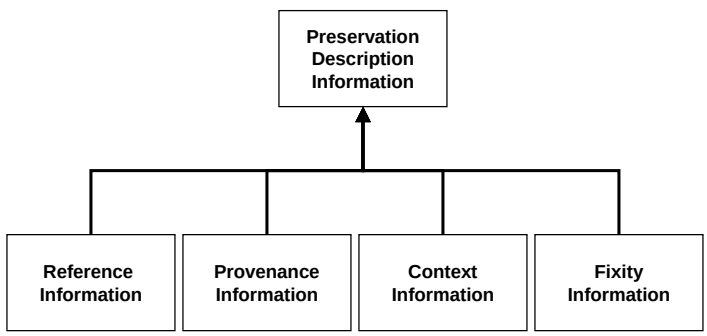
OAIS Information Model (2)

- Information package:
  - Container that encapsulates Content Information and PDI
  - Packages for submission (SIP), archival storage (AIP) and dissemination (DIP)
  - AIP = “... a concise way of referring to a set of information that has, in principle, all of the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object”

OAIS Information Model (3)

- Archival Information Package (AIP):
  - Content Information
    - Original target of preservation
    - Information Object (Data Object & Representation Information)
  - Preservation Description Information (PDI)
    - other information (metadata) “which will allow the understanding of the Content Information over an indefinite period of time”
    - A set of Information Objects
    - Based on categories discussed in CPA/RLG report: *Preserving Digital Information* (1996)

OAIS Information Model (4)



PDI Preservation Description Information (Figure 4-16)

OAIS Information Model (5)

- Also defines:
  - Archival Information Units and Archival Information Collections
  - Information Package transformations, e.g. for Ingest and Access
  - Preservation perspectives:
    - Migration, e.g refreshment, replication, repackaging, transformation
    - Preservation of look and feel (e.g., emulation, virtual machines)
  - Archive interoperability, e.g. federation

Implementing OAIS (1)

- Fundamentals:
  - OAIS is a reference model (conceptual framework), NOT a blueprint for system design
  - It informs the design of system architectures, the development of systems and components
  - It provides common definitions of terms … a common language, means of making comparison
  - But it does NOT ensure consistency or interoperability between implementations

Implementing OAIS (2)

- ISO 14721:2003, published in early 2003 - follows the text made available by the CCSDS
- However, the earlier versions of the model made available by the CCSDS informed implementations long before then
- Three broad areas of influence:
  - Preservation metadata schemas
  - Architecture and system design
  - Conformance criteria for repositories

Metadata and shared infrastructures

The importance of metadata (1)

- All digital preservation strategies depend - to a greater or lesser extent - on the creation, capture and maintenance of metadata
- Preservation metadata:
  - The "information a repository uses to support the digital preservation process," specifically "the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (PREMIS Data Dictionary, 2005)
  - Cuts across older categorisations of metadata (descriptive, administrative, structural)

The importance of metadata (2)

- Recordkeeping metadata:
  - "Structured or semi-structured information that enables the creation, registration, classificatoin, access, preservation and disposition of records through time and within and across domains" … [they] "can be used to identify, authenticate, and contextualize records; and the people, processes and systems that create, manage, maintain and use them and the policies that govern them"
  - The definition used in ISO 23081
  - Much stronger focus on organisational contexts

PREMIS metadata (1)

- PREMIS Working Group
  - Preservation Metadata: Implementation Strategies
  - Working Group sponsored by OCLC and RLG
  - Reviewed earlier Metadata Framework document and existing practice
  - Focused on implementation and definition of 'core' metadata
  - PREMIS Data Dictionary (May 2005)

PREMIS metadata (2)

- PREMIS Data Dictionary
  - Less explicitly based on OAIS Information Model structure than older OCLC/RLG Framework
  - Based on own data model
  - Defines some of the semantic units for: Objects, Events, Agents, Rights
  - Supports automatic capture, where possible
- PREMIS also provides:
  - An XML implementation, e.g. for use in a packaging format like METS (Metadata Encoding and Transmission Standard)
  - Maintenance activity (Library of Congress)

The need for shared infrastructures

- For example: registries for sharing information about, or for identifying or validating formats,
  - There is "… a pressing need to establish reliable, sustained repositories of file format specifications, documentation, and related software" (Lawrence, *et al*., Risk management of digital information (CLIR, 2000)
    - DSpace 'bitstream format registry'
    - Global Digital Format Registry (GDFR)
      - » Some components exist, e.g. Typed Object Model, JHOVE tool
    - DCC Representation Information registry

Some final comments

Some key points to remember (1)

- Most of the technical problems associated with long-term digital preservation can be solved if a life-cycle management approach is adopted
  - Preservation needs to be considered at a very early stage in an object's life-cycle
  - There also needs to be continuous re-evaluation of policies and practical approaches adopted
- There is a need to identify and understand the 'significant properties' of an object
  - There is a need to make (and articulate) difficult choices
  - Trade-offs with costs

Some key points to remember (2)

- Metadata is important
  - But is an area of much uncertainty
  - Evolving standards like the PREMIS Data Dictionary and ISO 23081 will help, but there is no substitute for detailed requirements analysis

Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union and other sources. UKOLN also receives support from the University of Bath, where it is based: http://www.ukoln.ac.uk/

The Digital Curation Centre is funded by the JISC and the UK e-Science Core Programme: http://www.dcc.ac.uk/