

**Susan Thomas,
Project Manager**

PARADIGM

**Practical Experiences with Personal Digital
Archives: The Paradigm project**

Data Standards Group, 16 November 2006



The University of Manchester



Summary of today's talk

- Comparing analogue with digital
- A rationale for early intervention in digital archives – two case studies of digital extraction
 - Digital rescue (modern)
 - Digital rescue (older)
- Paradigm (Personal ARchives Accessible in DIGital Media)
- Next steps

Comparing analogue with digital

<u>Intellectual manifestation</u>	Draft speech (Object A)	Recording of speech (Object B)	Draft speech (Object C)	Personal website (Object D)	<u>Risks</u>
<u>Physical manifestation</u>	Paper and ink	Audiotape	3" Amsoft disk	CD	<ul style="list-style-type: none"> ● Security ● Degradation ● Disaster
<u>Hardware stack</u>	X	✓	✓	✓	<ul style="list-style-type: none"> ● Obsolescence of 1+ components in the stack
<u>Software stack</u>	X	X	✓	✓	<ul style="list-style-type: none"> ● Obsolescence of 1+ components in the stack
<u>Representation</u>	X	X	1 locoscript file	1 css file 5 html files 6 jpg files 1 pdf file 1 javascript file	<ul style="list-style-type: none"> ● Relationships between component files broken

Case Study 1: Rescuing Modern Digital Stuff

Problem – accidental deletion and part-overwrite

- Family holiday photos deleted in rush to use camera for wedding
- Some wedding photos taken before camera owner realised error



Solution – Extraction & Analysis using Helix Incident Response and Computer Forensics Live CD

- Image of memory card taken using 'disk dump' (dd) command
- Image analysed for deleted files; files extracted from disk image

Outcome - success!

- 264 jpg/mpg files recovered; 25 files irrecoverable; 1 file corrupted
- Happy photographer

Case Study 1: The Corrupted Photo



<http://www.paradigm.ac.uk/>

Case Study 1: Rescuing Modern Digital Stuff

Lessons

- Lots of software and services for file recovery exist, especially for images and email
 - Must be a growing market for this
- It's easy enough to do this work in-house, once you know how
- It's only easy because the material is modern
 - No port/cable issues
 - No driver issues
 - No format obsolescence issues
 - Realistic learning curve
 - etc.
- Lesson reinforced – the delete key does not destroy

Case Study 2: Posthumous Digital Deposit

Problem

- Two older PCs
 - Apricot / Windows 95
 - Opus Technology / poss. Windows 3.?
- Several 3" disks



Decision

- explore potential of developing in-house expertise to work with older material
 - digital archaeology

Rationale

- Expect to get much more of this material in future
- Uncomfortable with sending third-party data to another third-party
- Wish to trust and understand processes involved
- Wish to document process for future scenarios

Case Study 2: Posthumous Digital Deposit

Approach

- Extract digital objects from storage media
- Create versions of digital objects in contemporary formats for cataloguing access
- By working with Jeremy John at the BL, and talking to computer enthusiasts

Outcome thus far

- Material on one hard disk extracted using 'forensic PC' running *Guidance Encase & AccessData Forensic Toolkit*
- Material on older hard disk to be extracted next month
- Sources of knowledge, hardware and software for data recovery from 3" disks, and migration pathway from *Locoscript* format identified

Case Study 2: Posthumous Digital Deposit

Digital Archaeology Lessons

- Data recovery for older material is more difficult
- Relative merits of commercial and open source forensic tools
- Drivers, connections & file systems are tricky when attempting to extract a disk image from an older system to a newer one
- Pooling expertise and resources across institutions is helpful, especially while services are immature
- Documentation for hardware and software is often difficult to locate or of poor quality, but there is much useful information on the web (that needs archiving – quickly!)
- The tacit knowledge, hardware and software to support a particular generation of computing is fragile
- Need to be able to do it, but best avoided if possible

Scenarios like Case Studies 1 & 2 suggest lifecycle management

- Archives traditionally reach a repository once an individual has retired or passed away – potentially a long time after creation
 - Physical survival of paper and parchment straightforward, but bit-level survival uncertain for digital objects of this age
 - If objects survive at bit level, digital archaeology may be required to liberate them
- Individuals have limited support from Information and Information Technology professionals
- Usage of third party storage solutions growing, so likelihood of capturing entire archive without active engagement reduces
- Reduce risk of loss and uncertainty of digital archaeology by bringing digital archives into a managed environment and/or providing advice while records still active

PARADIGM

- Funded for 2 years by the JISC, ends Feb. 2007
- Collaboration between Oxford University Library Services (lead) & John Rylands University Library, Manchester
- 1.5 fte archival, 1 fte developer plus input from Oxford Digital Library and Special Collections departments
- Explores digital preservation from 'personal' and 'collecting' perspectives in the context of a 'hybrid archive'
- To gain hands-on experience of:
 - an early-intervention approach to developing hybrid archive collections
 - soft issues - by working with politicians and their materials (selection and acquisition, creator attitudes, legal issues, etc.)
 - relevant technical issues and tools

Digital Preservation Alphabet Soup

PRONOM **JHOVE** **INTERPARES**

DPC **OAIS** **DD**

MODS **PREMIS** **TechMD**

DCC **METS** **NDIIPP**

METSRights **DROID**

AIP **VMware** **XENA** **GDFR** **PADI**

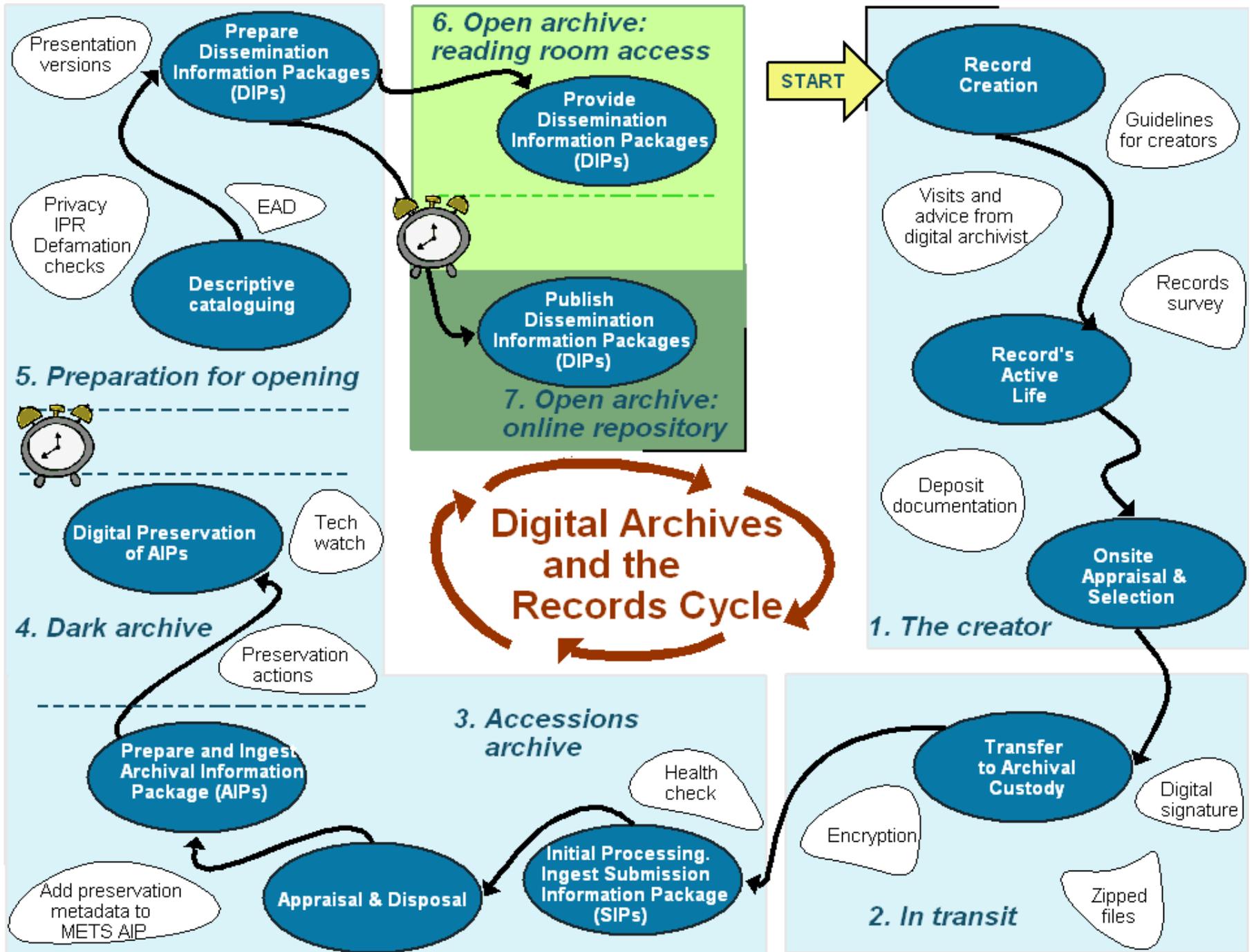
MIX **ERA** **SHA-1**



CC BY ♥ maggie, 2006
SOME RIGHTS RESERVED

Project Outcomes

- Have started to harmonise long-standing archival standards and workflows with digital repository standards and workflows
- Prototype preservation repository developed
 - Key standards: OAIS, Fedora, METS and PREMIS
 - Focus on acquisition, ingest and preservation rather than access
- Developed expertise in management of hybrid archives at partner institutions and provided a platform for future activity
- Have developed and shared strategies for personal digital archives
 - That are based on experiences with politicians and their digital archives
 - Through Paradigm's Online Workbook
<http://www.paradigm.ac.uk/workbook>
 - Through 'roadshow' events like this talk



Working with the Creator 1: Selection and Surveying

- Invited a range of politicians to participate in piloting early-intervention
 - Three political parties
 - MPs, MEP, Peers
 - Local, national and international portfolios
- Thoughts on selection
<http://www.paradigm.ac.uk/workbook/appraisal/index.html>
- Developed a records survey to identify:
 - Functions and roles
 - Technical environment
 - Working practices
 - Rights and responsibilities
 - Record series of historical interest and if and when they could be accessioned
- See <http://www.paradigm.ac.uk/workbook/record-creators/index.html> and <http://www.paradigm.ac.uk/workbook/introduction/structure.html>

Working with the Creator 2: Acquisition and Transit

- Needed to learn how to extract typical personal digital archives from popular desktop software and web services
 - Needed to develop transfer procedure and toolkit for secure and authentic transfer
 - Ideal process – use biometric protected USB-powered external hard-disk with forensic software
- 
- Captures material as structured by creator
 - Records checksums for each item acquired, which can be used to validate the continuing authenticity of items in the accession
- Ideal process not always possible. Depends on the hardware and software in place
- Digital archiving allows exact copies to be taken. The creator can therefore retain the material

Working with the Creator 3: Advice and Agreement

Provided guidance to creators as part of the process

- Reactive advice – responses to direct questions arising from the work
- Proactive advice – drafting basic advice leaflet for creators
- Advice sought on
 - safeguarding longevity and future accessibility of material, e.g. backup, filing and naming conventions, basic system administration
 - which series are historically significant

Developed deposit documentation (see Workbook)

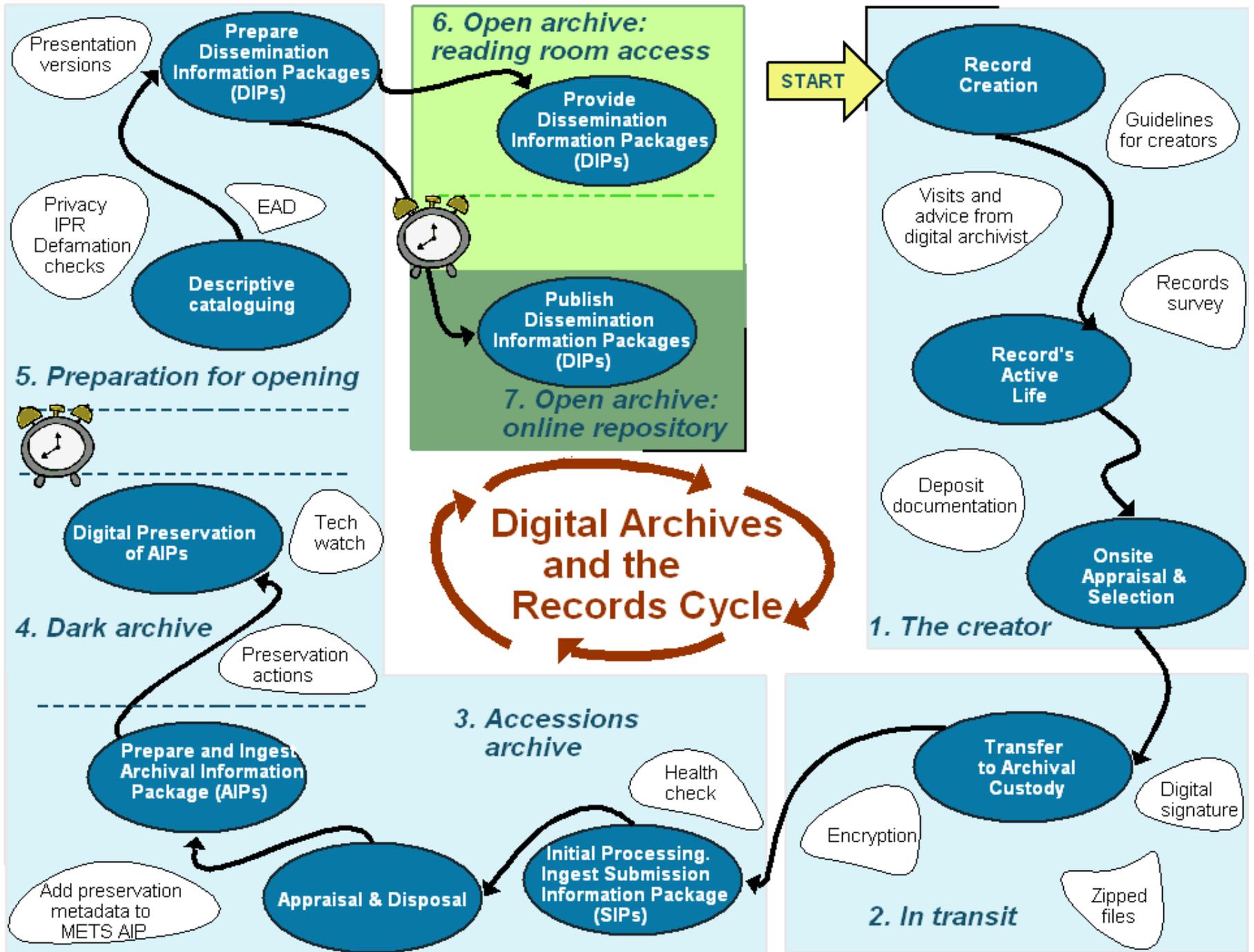
- Explicit permission to undertake preservation actions on digital material, from simple backup to migrating to new formats
- Explicitly document terms of agreement in relation to closure periods
- <http://www.paradigm.ac.uk/workbook/accessioning/documentation/index.html>
- Thoughts on legal issues
<http://www.paradigm.ac.uk/workbook/legal-issues/index.html>

Early-intervention pilot : Lessons

- Digital increasingly used as 'master', but poorly managed
- Poor understanding of archiving for historical purposes
- Privacy and security concerns – own and third party – increased by recent date of material. Reluctance to deposit some material now, or at all
- Repository must manage material with legal protections for longer
- Finding time for history in the present
- Authority to act
- Variety: individual concerns; technical set-up; organisational set-up; IT literacy or support
- Frequency and scope of accessions; dealing with duplication
- Can accession a copy of the archive
- What about the paper, audio, video, photographs, etc.?
- Opportunity to acquire valuable contextual information
- Contemporary formats are easier to access and normalise

Early-intervention: Conclusions

- A worthwhile approach
 - Individuals have lost material!
 - Can obtain excellent context
- But relies on
 - Headhunting individuals
 - Good will and trust of individuals
 - Sustaining relationships over long periods of time
 - May produce **different** collections
 - May not work so well in instances where archives are to be purchased
- Digital archaeology inescapable
- Need to repeat with other groups
- Not the only way. See 'Approaches to Collection Development' section in Paradigm Workbook
<http://www.paradigm.ac.uk/workbook/collection-development/index.html>



Processing New Accessions 1

- New accessions are copied to a stand-alone quarantined staging area
 - Authenticity of transfer can be validated using checksums generated at creator's premises
 - Material is virus checked
 - Material may be appraised to identify archival files and dispose of others
 - The file formats in the accession are identified and validated using various tools - DROID/PRONOM, misc registries, and JHOVE
- Must ensure that incidental copies of archives are securely deleted
- Delete duplicate files, system and software files
- Add information on new formats encountered to PRONOM, etc.
- Assemble preservation metadata to submit with digital objects to the digital archive repository

Processing New Accessions 2: Preservation metadata (PREMIS)

- Digital archives need lots of metadata!
- PREMIS – a preservation metadata standard devised to cover all the things a preservation repository needs to know to support and document the digital preservation process:
 - Provenance: *Who has had custody/ownership of the digital object?*
 - Authenticity: *Is the digital object what it purports to be?*
 - Preservation Activity: *What has been done to preserve the digital object?*
 - Technical Environment: *What is needed to render and use the digital object?*
 - Rights Management: *What intellectual property rights must be observed?*
- Aim – to make digital object self-documenting over time
- See <http://www.loc.gov/standards/premis/>

Processing New Accessions 3: Preservation metadata (specific)

- Repositories also need to record metadata specific to different object types
- Different object types have different characteristics
- Example metadata standards
 - MIX for images <http://www.loc.gov/standards/mix/>
 - TextMD for text <http://dlib.nyu.edu/METS/textmd.xsd>
 - VideoMD for moving images
http://www.loc.gov/rr/mopic/avprot/DD_VMD.html
- Tools to extract some metadata required by PREMIS and these standards exist, but there are some problems:
 - Duplication between tools
 - Tools use their own metadata schemas
 - No mapping between tool output schemas and standard schemas
 - Requires co-ordinated use of multiple tools and assembly of their output
 - Some tools not very user-friendly
 - Sustainability of tools and the schema of their output uncertain

METS – wrapping it all up in an AIP

- METS & OAIS Information Packages
- Unites metadata in one XML file
- Not the only way of creating an AIP



 By J. McPherson, 2006

Advantages

- Flexible - can accommodate all the metadata required by a digital archive in one file
- Increasing user-community
- Several institutions are now developing METS templates for preservation
- Maintained by LoC

Disadvantages

- Flexible – requires strong implementation guidelines
- Existing profiles and tools geared towards dissemination rather than preservation
- Need to learn how to use it!

<http://www.loc.gov/standards/mets/>

<http://public.ccsds.org/publications/archive/650x0b1.pdf>

<http://www.paradigm.ac.uk/>

Storing Digital Objects in a Suitable Managed Environment

- Paradigm uses the open-source *Fedora* digital repository software. Developed at Cornell and Virginia.
 - <http://www.fedora.info/>
- Fedora can import and export METS, but stores digital object metadata in FOXML (its equivalent to METS)
- It can store digital objects and metadata, or just metadata about digital objects which refers to content held externally
- Fedora supports relationships between objects
- Fedora maintains an audit trail of actions performed on an object
- Fedora is very flexible - requires business rules and development work to act as a trusted repository for preserving digital archives

Preservation Strategy 1: Possibilities

- Most digital archives will undertake to preserve digital objects at bit-level; i.e. to preserve the digital object in the form it was deposited
- Digital preservation should also seek to preserve access to the digital objects

Possible preservation strategies

- *Migrate* – recreate the object
 - To preferred formats on ingest
 - To single format on ingest (XML)
 - To preferred formats on obsolescence
 - To preferred formats on request
- *Emulate* – recreate the environment
 - Recreate the environment not the object
- *Preserve the Technology*
 - Maintain all of the software and hardware stack needed to access objects

Preservation Strategy 2: Recommendations

Recommend that preservation strategies be developed

- In-line with community practice
 - Need for shared knowledge base
 - Dependence on community for some tools
- Metadata should support multiple strategies (PREMIS)
 - Don't know what tools will be available in future
 - Strategies may change
- Technology Watch should be:
 - Local (knowledge of collection profile)
 - Distributed (sum of parts greater than the whole)
- Timing of preservation interventions dependent on format risk assessment
 - Normalisation on ingest for high risk (older, obscure, opaque) formats
 - Delay intervention for low risk (open, well-supported) formats until 'at risk'

Next
Steps



Some of the Challenges Ahead

- Simplify ingest for archivists
- Develop formal content models for our objects



<http://cairo.paradigm.ac.uk>

- Bring preservation monitoring/actions to the repository
- Work with other kinds of creator and their archives
- Integrate digital archives into existing policies for archives
- Provide controlled reading room access
- Create and enhance directories of conversion tools, etc.

Questions?

- Ask me now

- Or later:

Susan Thomas (Project Manager, Paradigm & Cairo)

Oxford University Library Services

Osney One Building, Osney Mead

OXFORD, OX2 0EW

Web : <http://www.paradigm.ac.uk>

<http://cairo.paradigm.ac.uk>

Email : susan.thomas@ouls.ox.ac.uk

- Tel: 01865 283821