

Experiences with Personal Digital Archives

Susan Thomas

Group for Literary Archives and Manuscripts,

16 March 2007



The University of Manchester



PARADIGM



Summary of today's talk

- Comparing analogue with digital
- Case study – traditional archival deposit including digital archives
- Case study – working with contemporary creators and digital materials
- Processing hybrid personal archives
- Next steps

PARADIGM

- Funded for 2 years by the JISC, ended 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
- Explored 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, metadata, tools and digital repository software
- Harmonise archival principles and workflows with digital curation standards
- Develop prototype digital archive repository
- Share lessons in an online Workbook <http://www.paradigm.ac.uk/workbook>

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: Posthumous Digital Deposit

Problem

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



Decision

- Explore potential of developing in-house expertise to work with older material
- 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

Progress

- Material on hard disk extracted using 'forensic PC' running *Guidance Encase & AccessData Forensic Toolkit* at BL
- Sources of knowledge, hardware and software for data recovery from 3" disks, and migration pathway from *Locoscript* format identified
- Useful acquisitions made via eBay!

Useful acquisition 1

Amstrad PCW 8512



...and it works!

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|----------------------------|--|--|--|------------|--|-----------------|--|-----------------|--|--|--|----------------------------------|--|--|--|------------|--|-------------------|--|--|--|--|--|-----------------------------|--|--|--|--|--|-------------|--|--|--|--|--|----------------|--|--|--|--|--|------------|--|--|--|--|--|
| Disc management. | | | | | | | | | | | | Printer idle. Using M: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| C=Create new document | | | | | | E=Edit document | | | | | | P=Print document | | | | | | D=Direct printing | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| f1=Actions | | | | | | f2=Disc | | | | | | f3=File | | | | | | f4=Group | | | | | | f5=Document | | | | | | f6=Settings | | | | | | f7=Disc change | | | | | | f8=Options | | | | | |
| Drive A: | | | | | | | | | | | | Drive B: | | | | | | | | | | | | Drive M: | | | | | | | | | | | | | | | | | | | | | | | |
| 167k used 6k free 15 files | | | | | | | | | | | | empty 0k used 0k free 0 files | | | | | | | | | | | | 54k used 238k free 10 files | | | | | | | | | | | | | | | | | | | | | | | |
| group 0 167k | | | | group 4 0k | | | | | | | | group 0 54k | | | | group 4 0k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| group 1 0k | | | | group 5 0k | | | | | | | | group 1 0k | | | | group 5 0k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| group 2 0k | | | | group 6 0k | | | | | | | | group 2 0k | | | | group 6 0k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| group 3 0k | | | | group 7 0k | | | | | | | | group 3 0k | | | | group 7 0k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| A: group 0 15 files | | | | | | | | | | | | M: group 0 10 files | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 linbo files | | | | | | | | | | | | 0 linbo files | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| DXR10 B.EXR 4k | | | | | | | | DXR10 B.EXR 4k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| DXR12 B.EXR 5k | | | | | | | | DXR12 B.EXR 6k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MATRIX .#SS 11k | | | | | | | | MATRIX .#SS 12k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MATRIX .#ST 12k | | | | | | | | MATRIX .#ST 12k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MATRIX .#XR 1k | | | | | | | | MATRIX .#XR 2k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MATRIX .#XS 1k | | | | | | | | MATRIX .#XS 2k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PHRASES .STD 1k | | | | | | | | 4 hidden 16k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| SETTINGS.STD 4k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 hidden 128k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Useful acquisition 2

Locolink – PCW to PC



Digital Archaeology Lessons

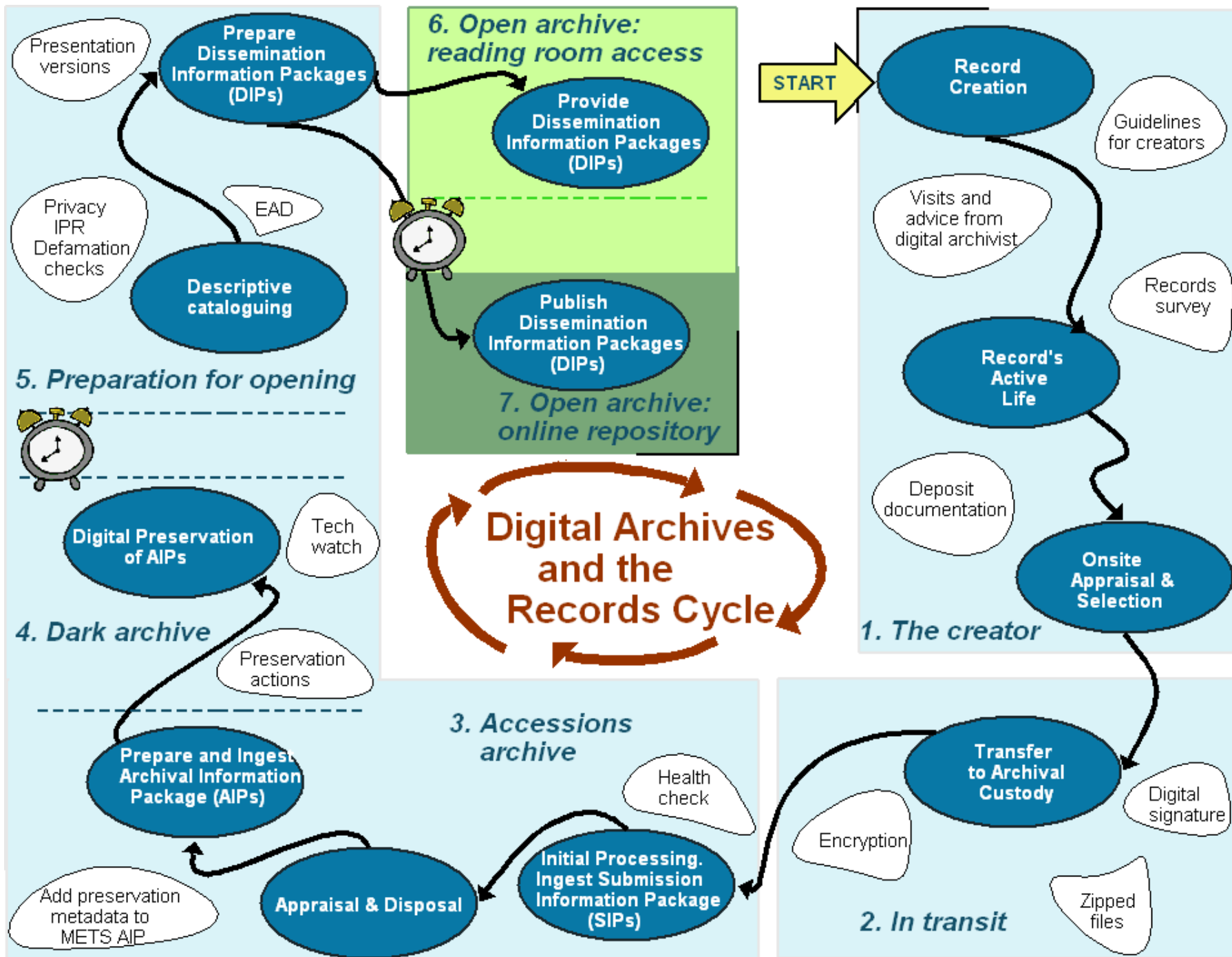
- Data recovery for older material is more difficult
- Little chance of knowing whether the effort is worth it beforehand
- Need to be able to do it, but best avoided if possible
- Relative merits of commercial and open source forensic tools
- Pooling expertise and resources across institutions is helpful, 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
- Drivers, connections & file systems are tricky when attempting to extract a disk image from an older system to a newer one

Evolving connectors



Lifecycle management for personal archives?

- 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
 - Hardware and software obsolescence may render archives inaccessible
 - What does an archive created with a lifetime of technologies look like?
- Individuals have limited support from Information and Information Technology professionals, but must 'curate' their own digital archives
- 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



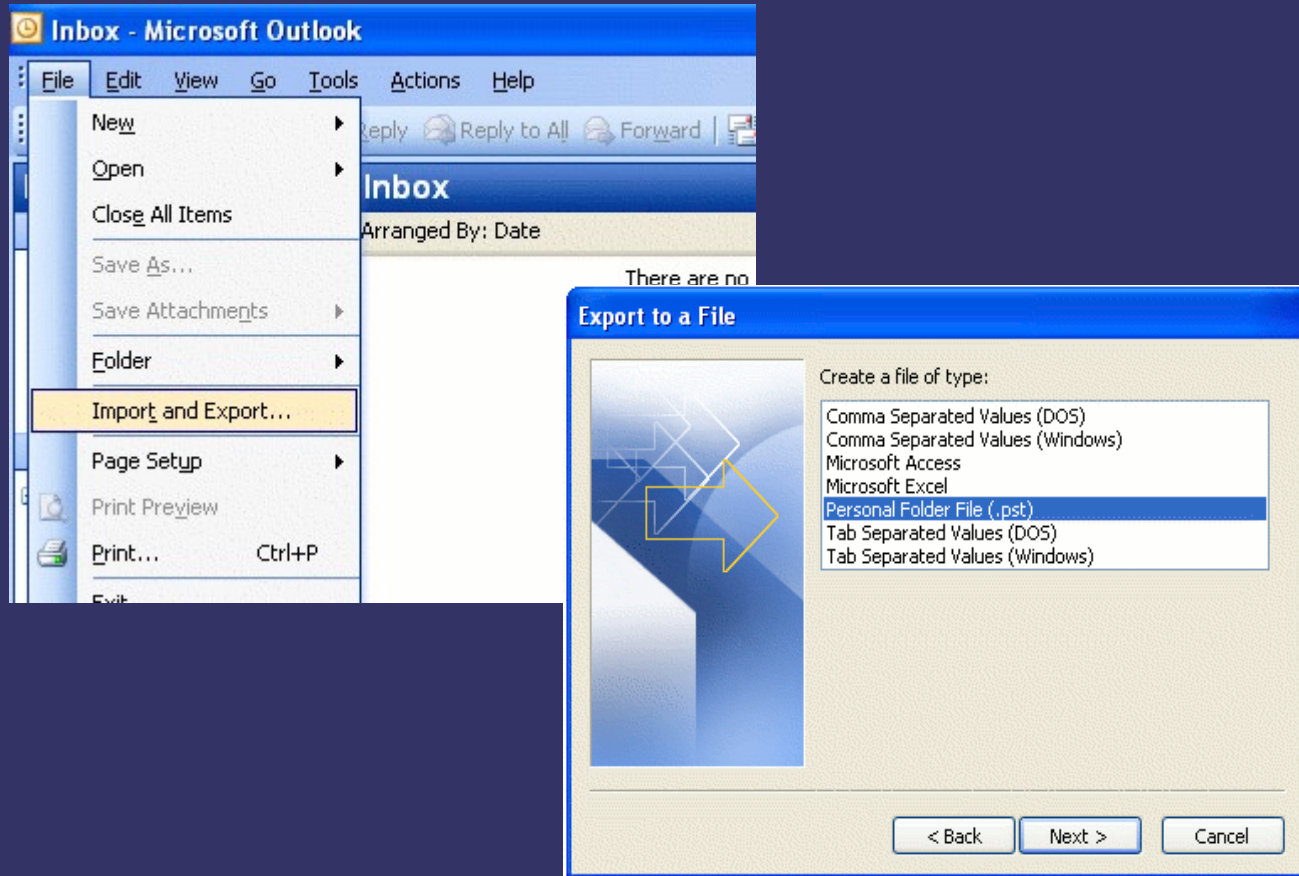
Selection and Surveying

- Invited a range of politicians to participate in piloting early-intervention approach to collection development
 - 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>

Archive extraction

- Digital files are simple to extract
 - selective copy and paste
 - use digital forensics tools to capture logical grouping of files
- Other digital archives more complex
 - email
 - diaries
 - websites
 - content stored by online services
 - personal digital assistants
- Archivists need to learn how to extract typical personal digital archives from popular desktop software and web services
- For material stored on a the computer, digital forensics tools can be used to create a bit-for-bit image of the computer's hard disk


Extracting email from clients



Extracting email from webmail services

YAHOO! MAIL Sign In
New User? Sign Up

Mail Home - Help



Want more?
Get it with
Yahoo! Mail Plus.

**Yahoo! Mail Plus
free trial* offer**

Get Started


Try it free for 30 days (limited time offer)
\$19.99/year thereafter!
Offer details


Yahoo! Mail Plus gives you more control and freedom

- No graphical ads**
- Offline access (with POP) and mail forwarding
- Even fewer unwanted mail with SpamGuard™ Plus
- More capacity with larger attachments and more storage

| Features | YAHOO! MAIL PLUS | YAHOO! MAIL |
|-----------------------------------|---------------------|----------------|
| Message size | 20MB | 10MB |
| Storage | 2GB | 1GB |
| Filters | 50 | 15 |
| Spam Protection | SpamGuard Plus | SpamGuard |
| AddressGuard/Disposable addresses | ✓ | |
| No graphical ads** | ✓ | |
| Offline access with POP | ✓ | |
| Mail forwarding | ✓ | |

Yahoo! Mail Plus goes with you

 Check email from any mobile phone

 With POP, download your emails into Outlook™ or other email desktop applications

Extracting personal movie collections

The screenshot shows a YouTube page with the 'Quicklists' banner. The video 'The Dead - Billy Collins Animated Poetry' is the main focus. It has a simple line drawing of a person lying down. The video player shows a progress bar at 00:25 / 00:51. Below the video, there are options to login to rate, save to favorites, share, and flag. The video has 2379 ratings, 460,545 views, 1071 comments, and 2421 favorites. To the right of the video, there is a 'Related' section with three videos: 'Forgetfulness - Billy Collins Animated Poetry' (01:51, 76866 views), 'Forgetfulness - Billy Collins Animated Poetry' (01:49, 54273 views), and 'The Dead -- Billy Collins Animated Poetry' (00:54, 19715 views). On the far right, there is a 'Director Videos' section with three videos: 'Mookie and Sam (EP 2)' (04:36), 'Regular Hamburger' (01:17), and 'Intro to MacBook Pro (and iLife '06)' (14:22).

You Tube
Broadcast Yourself™

[Sign Up](#) | [My Account](#) | [History](#) | [QuickList \(0\)](#) | [Help](#) | [Log In](#)

[Videos](#) [Categories](#) [Channels](#) [Community](#) [Upload Videos](#)

Quicklists

... make it easy to collect the videos you want to see, then sit back and enjoy!
[About Quicklists >](#)

The Dead - Billy Collins Animated Poetry

Added **February 13, 2007**
From [JWNTNY](#) to JWNTNY
Billy Collins, former US Poet Laureat... [\(more\)](#)
Category [Film & Animation](#)
Tags [poet](#) [laureat](#) [poetry](#) [Billy](#) [\(more\)](#)
URL <http://www.youtube.com/watch?v=iuTNdHdwI>
Embed `<object width="425" height="350"><param name="`

[SUBSCRIBE](#)

Director Videos

[Mookie and Sam \(EP 2\)](#)
04:36
From: [FullPic](#)

[Regular Hamburger](#)
01:17
From: [TheWoodcreekFaction](#)

[Intro to MacBook Pro \(and iLife '06\)](#)
14:22
From: [lininju](#)

[Login to rate](#)
★★★★☆
2379 ratings

[Save to Favorites](#) [Share Video](#) [Flag as Inappropriate](#)
[Add to Groups](#) [Post Video](#)

Views: **460,545** | Comments: **1071** | Favorited: **2421** times

Related [More from this user](#) [Playlists](#)

Showing 1-20 of 31 [See All Videos](#)

[Forgetfulness - Billy Collins Animated Poetry](#)
01:51
From: [smiwt](#)
Views: 76866

[Forgetfulness - Billy Collins Animated Poetry](#)
01:49
From: [JWNTNY](#)
Views: 54273

[The Dead -- Billy Collins Animated Poetry](#)
00:54
From: [smiwt](#)
Views: 19715


Extracting personal photo collections


flickr GAMING

You aren't signed in [Sign In](#) [Help](#)


[Home](#) [Learn More](#) [Sign Up!](#) [Explore](#) ▾


Search putt1ck's photos [Search](#) ▾

 **putt1ck's photos** [Sets](#) [Tags](#) [Archives](#) [Favorites](#) [Profile](#)


[View as slideshow](#)
(New window )


IMG00807




 © All rights reserved.
Uploaded on [May 22, 2005](#)
[0 comments](#)


IMG00804




 © All rights reserved.
Uploaded on [May 22, 2005](#)
[0 comments](#)


IMG00802



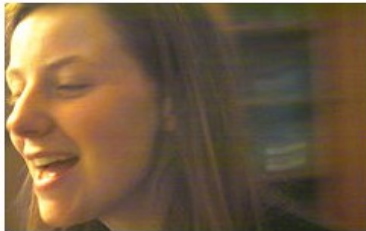
 © All rights reserved.
Uploaded on [May 22, 2005](#)
[0 comments](#)


IMG00800




 © All rights reserved.
Uploaded on [May 22, 2005](#)
[0 comments](#)


IMG00775




 © All rights reserved.
Uploaded on [May 22, 2005](#)
[0 comments](#)

IMG00753



 © All rights reserved.
Uploaded on [May 22, 2005](#)
[0 comments](#)

Acquisition

- Need 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

Developing Deposit Agreements for personal archives with digital bits

- Explicit permission to undertake preservation actions on digital material, from simple backup to migrating to new formats
 - only covers materials in which donor holds rights
 - recommendation of the Gowers review may render this unnecessary in future
- Seek permission for third parties to process/store archives if needed
- Seek rights to hold the sole research copy of the archive
- Seek permission to appraise and securely return/dispose of material
- 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>

Guidance for Creators 1

- 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
 - Identifying historically significant materials
- Flexible and general rather than prescriptive - users to pick and choose the advice they follow
- Hardware/software neutral
- Designed to facilitate not burden
- Enable creators to make informed decisions about using services, hardware, software and formats
- Enable creators to make informed decisions about archiving
 - Hidden material
 - Data mining

Guidance for Creators 2

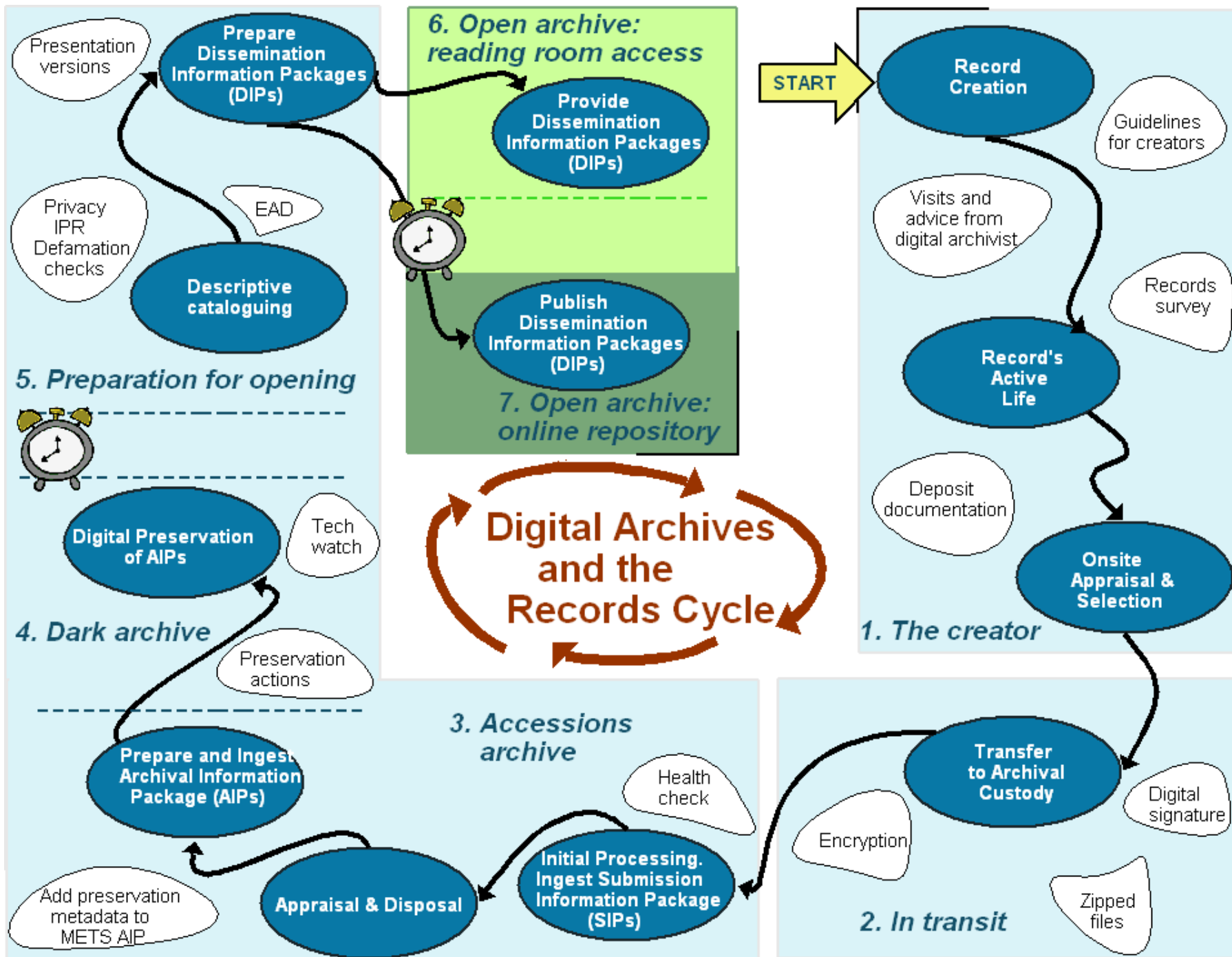
- Paradigm has produced generic guidance for creators which covers:
 - Backup
 - Caring for hardware and media
 - Administering your system
 - Selecting file formats
 - Filing and naming
 - Passwords and encryption
 - Keeping up-to-date
 - Managing emails
 - Handling legacy material
 - Ask digital curators for advice
 - Rights
- Would be useful to supplement this with domain-specific guides
- Best-supplemented by guidance tailored to an individual's needs
- Archivists supporting creators must be familiar with these issues and find way to communicate such guidance

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 Reception

- Usual processes occur alongside traditional archives
 - Record in accessions register
 - Create file for correspondence/agreement, etc.
- Digital material transferred to Digital Archive
- Compile basic inventory of carrier media/hardware
- Extract data from carriers to backed up environment ASAP
- Use fireproof data safe for unprocessed materials

Processing New Accessions

What do we have?

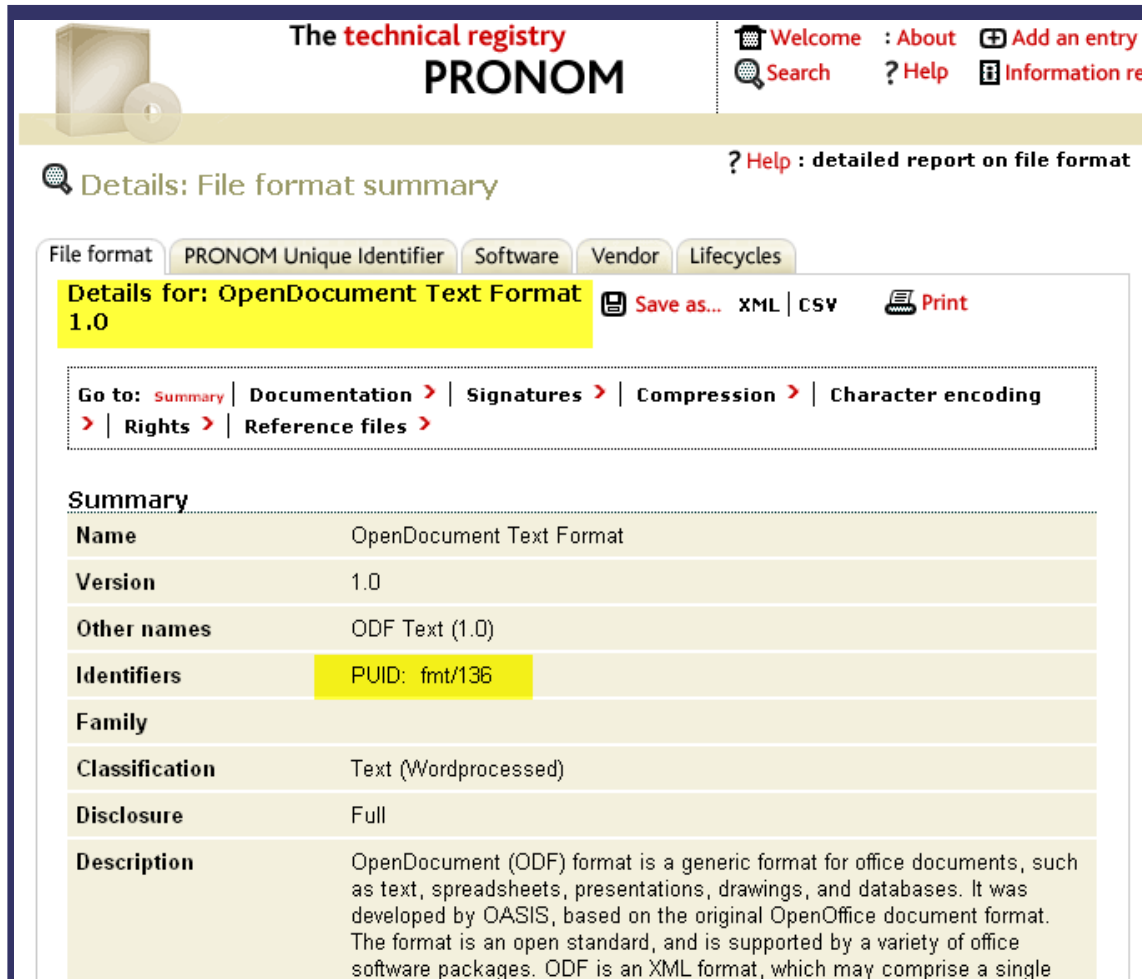
- Assessment
 - what *media* and *formats* do we have?
 - what *hardware* and *software* do we need?
 - What knowledge do we need?
 - Is material historically valuable?
 - Is material in good condition?
- Processes will improve through documentation of past scenarios for future reference – how-to guides

Processing New Accessions

Post extraction

- 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

Format registry – PRONOM



The screenshot displays the PRONOM technical registry website. The header includes the site logo, the title 'The technical registry PRONOM', and navigation links: 'Welcome', 'About', 'Add an entry', 'Search', 'Help', and 'Information re'. Below the header, a breadcrumb trail reads 'Details: File format summary'. A tabbed interface shows 'File format' as the active tab, with other tabs for 'PRONOM Unique Identifier', 'Software', 'Vendor', and 'Lifecycles'. The main content area is titled 'Details for: OpenDocument Text Format 1.0' and includes links for 'Save as...', 'XML', 'CSV', and 'Print'. A navigation bar below the title lists links: 'Go to: Summary', 'Documentation >', 'Signatures >', 'Compression >', 'Character encoding >', 'Rights >', and 'Reference files >'. The 'Summary' section is expanded, showing a table with the following details:

| | |
|-----------------------|---|
| Name | OpenDocument Text Format |
| Version | 1.0 |
| Other names | ODF Text (1.0) |
| Identifiers | PUID: fmt/136 |
| Family | |
| Classification | Text (Wordprocessed) |
| Disclosure | Full |
| Description | OpenDocument (ODF) format is a generic format for office documents, such as text, spreadsheets, presentations, drawings, and databases. It was developed by OASIS, based on the original OpenOffice document format. The format is an open standard, and is supported by a variety of office software packages. ODF is an XML format, which may comprise a single |

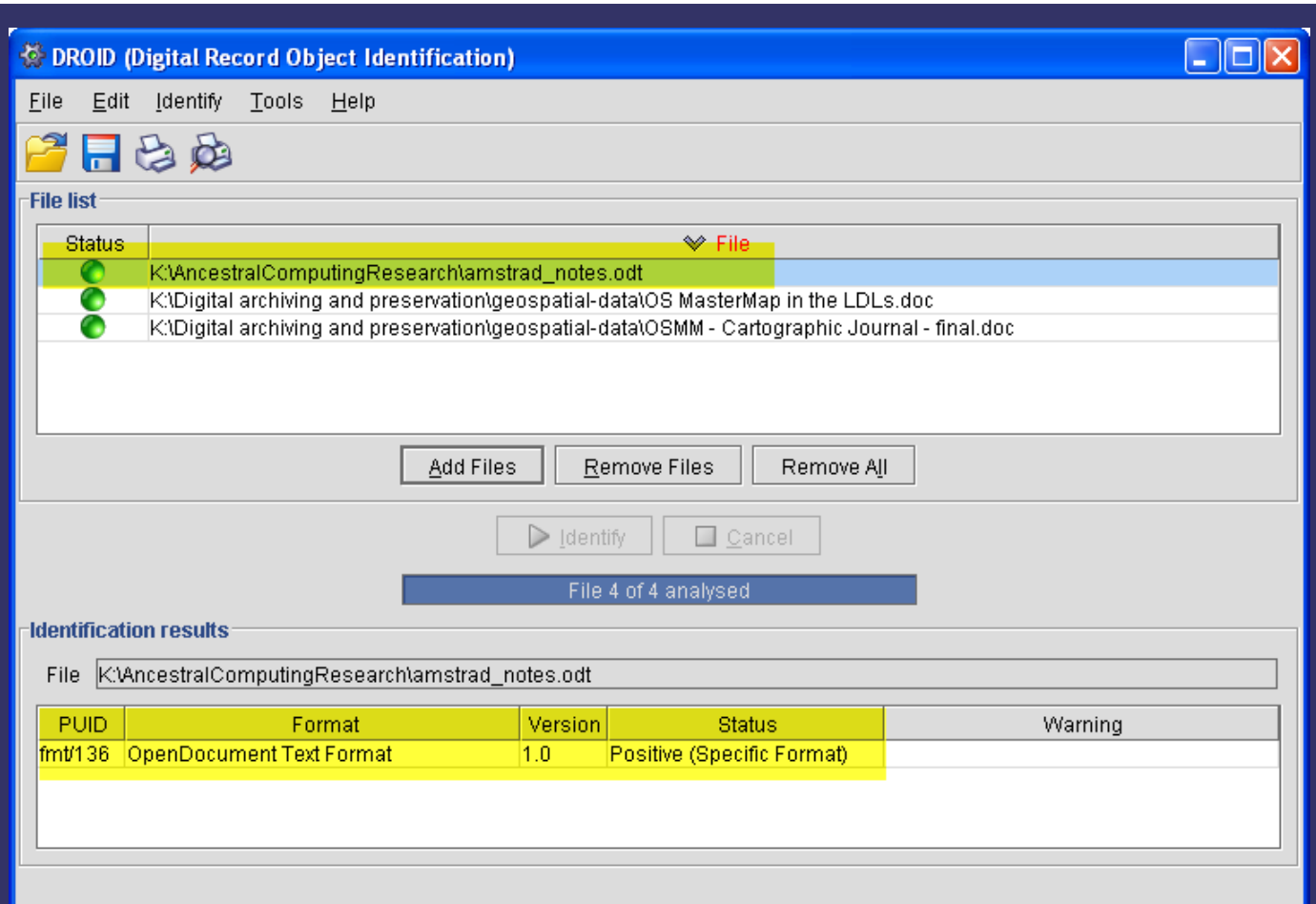
- Searchable database
- 587 formats to-date
- Provides framework for useful info
- More entries needed
- Fuller entries needed

Some formats listed in PRONOM

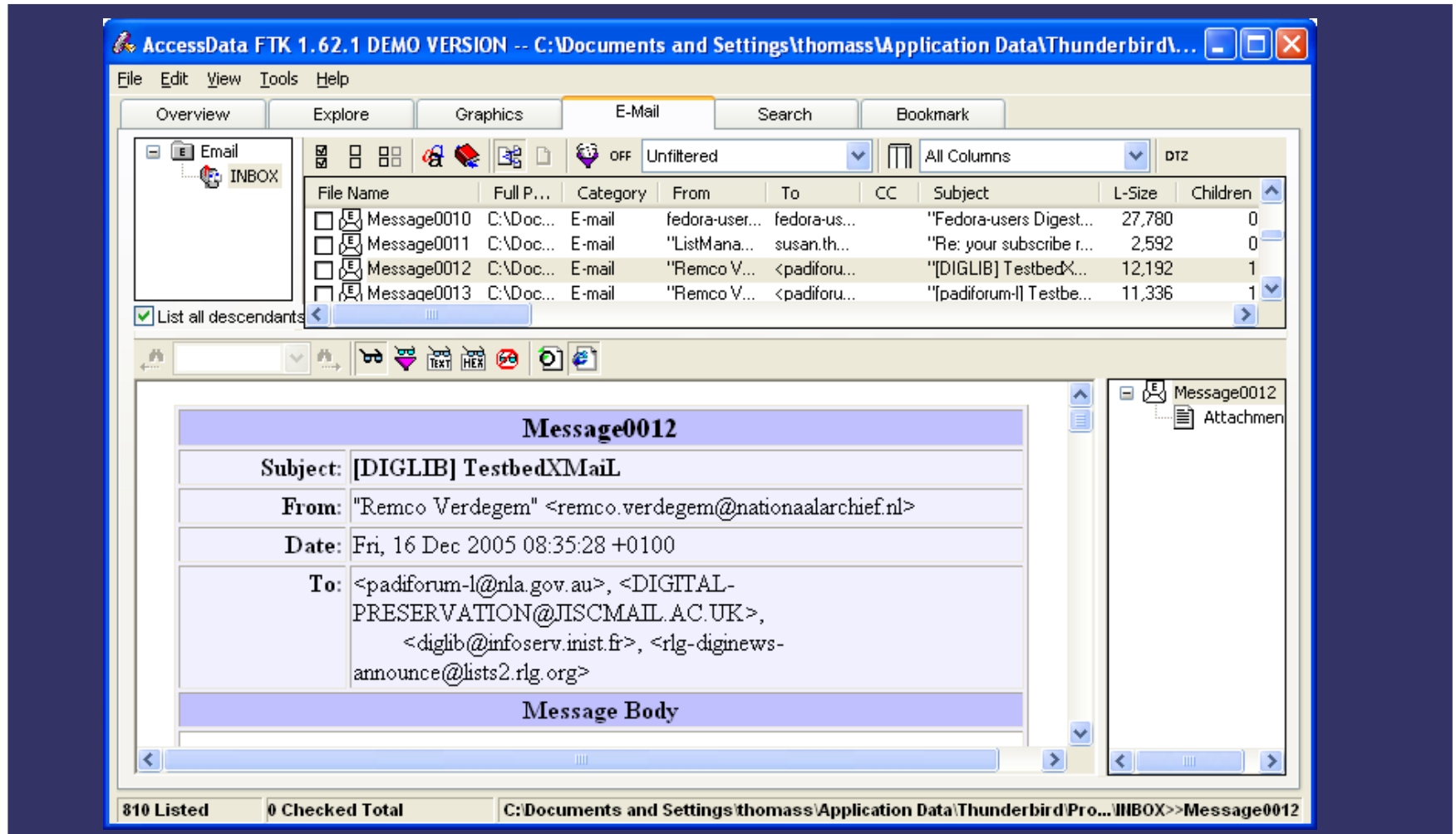
| PUID | Format Name | Format Version | Extension |
|---------|---|----------------|---------------|
| fmt/136 | OpenDocument Text Format | 1 | odt,ott, |
| fmt/137 | OpenDocument Spreadsheet Format | 1 | ods,ots, |
| fmt/138 | OpenDocument Presentation Format | 1 | odp,otp, |
| fmt/139 | OpenDocument Drawing Format | 1 | odg,otg, |
| fmt/14 | Portable Document Format | 1 | pdf, |
| fmt/140 | OpenDocument Database Format | 1 | odb, |
| fmt/141 | Waveform Audio (PCMWAVEFORMAT) | | wav,wave, |
| fmt/142 | Waveform Audio (WAVEFORMATEX) | | wav,wave, |
| fmt/143 | Waveform Audio (WAVEFORMATEXTENSIBLE) | | wav,wave, |
| fmt/144 | PDF/X-1:1999 | | pdf, |
| fmt/145 | PDF/X-1:2001 | | pdf, |
| fmt/146 | PDF/X-1a: 2003 | | pdf, |
| fmt/147 | PDF/X-2: 2003 | | pdf, |
| fmt/148 | PDF/X-3: 2003 | | pdf, |
| fmt/15 | Portable Document Format | 1.1 | pdf, |
| fmt/150 | JPEG-LS | | .jls, |
| fmt/151 | JPX (JPEG 2000 Extended) | | .jpx,.jpf, |
| fmt/153 | Tagged Image File Format for Image Technology (TIFF/IT) | | tif,tiff, |
| fmt/154 | Tagged Image File Format for Electronic Still Picture Imaging (TIFF/EP) | | tif,tiff, |
| fmt/156 | Tagged Image File Format for Internet Fax (TIFF-FX) | | tif,tiff, |
| fmt/17 | Portable Document Format | 1.3 | pdf, |
| fmt/18 | Portable Document Format | 1.4 | pdf, |
| fmt/19 | Portable Document Format | 1.5 | pdf, |
| fmt/2 | Broadcast WAVE | 1 | wav, |
| fmt/20 | Portable Document Format | 1.6 | pdf, |
| fmt/3 | Graphics Interchange Format | 1987a | gif, |
| fmt/4 | Graphics Interchange Format | 1989a | gif, |
| fmt/40 | Microsoft Word for Windows Document | 97-2003 | doc, |
| fmt/41 | Raw JPEG Stream | | jpe,jpg,jpeg, |
| fmt/42 | JPEG File Interchange Format | 1 | jpeg,jpe,jpg, |
| fmt/43 | JPEG File Interchange Format | 1.01 | jpg,jpe,jpeg, |
| fmt/44 | JPEG File Interchange Format | 1.02 | jpg,jpe,jpeg, |
| fmt/45 | Rich Text Format | 1 | rtf, |
| fmt/46 | Rich Text Format | 1.1 | rtf, |

Format identifier – DROID v. 1.1

- Identifies the format of files
- Uses PRONOM signature files
- Can output to CVS file



Access Data Forensic ToolKit



Processing New Accessions

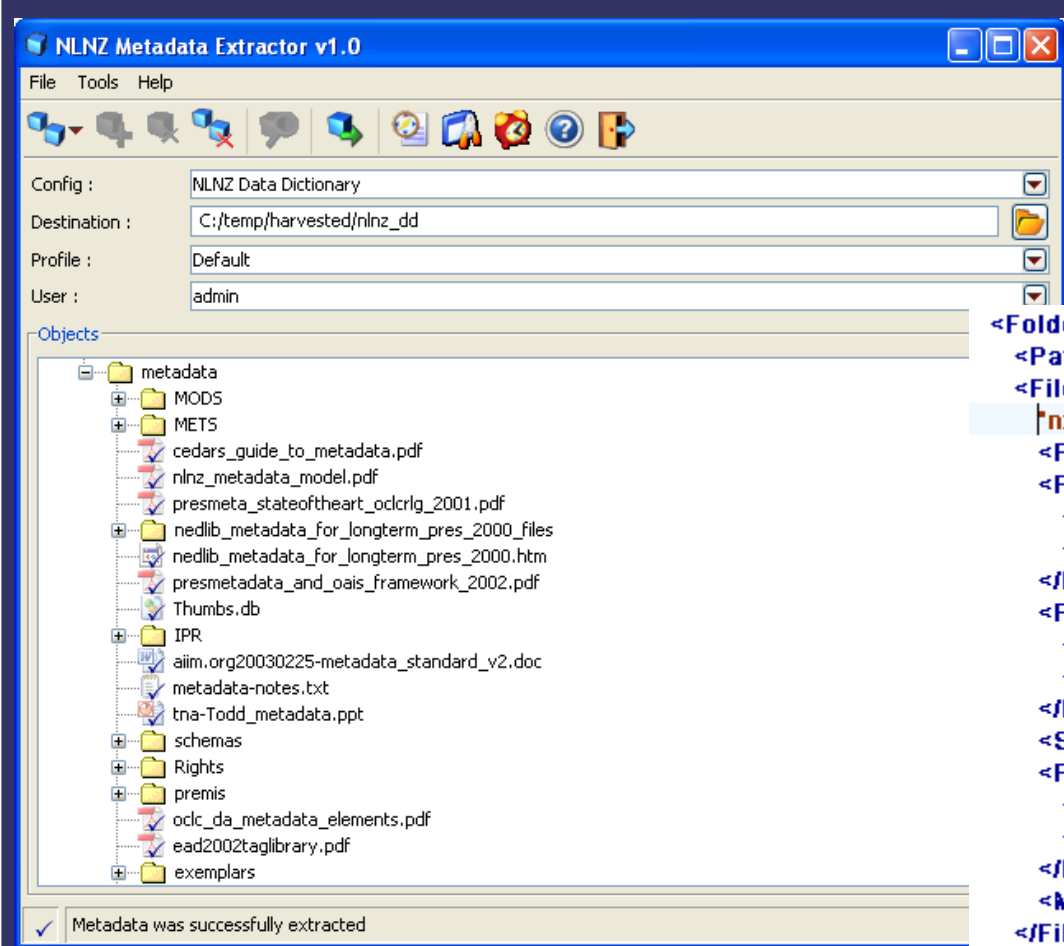
Preservation Metadata - Generic

- 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/>

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

Metadata extract – NLNZ tool v. 1



The screenshot shows the NLNZ Metadata Extractor v1.0 application window. The interface includes a menu bar (File, Tools, Help), a toolbar with various icons, and a configuration section with the following settings:

- Config: NLNZ Data Dictionary
- Destination: C:/temp/harvested/nlnz_dd
- Profile: Default
- User: admin

The 'Objects' section displays a tree view of the metadata structure:

- metadata
 - MODS
 - METS
 - cedars_guide_to_metadata.pdf
 - nlnz_metadata_model.pdf
 - presmeta_stateofheart_oclcrlg_2001.pdf
 - nedlib_metadata_for_longterm_pres_2000_files
 - nedlib_metadata_for_longterm_pres_2000.htm
 - presmetadata_and_oais_framework_2002.pdf
 - Thumbs.db
 - IPR
 - aiim.org20030225-metadata_standard_v2.doc
 - metadata-notes.txt
 - tna-Todd_metadata.ppt
 - schemas
 - Rights
 - premis
 - oclc_da_metadata_elements.pdf
 - ead2002taglibrary.pdf
 - exemplars

A status bar at the bottom indicates: ☒ Metadata was successfully extracted

On the right side of the interface, an XML snippet is displayed, showing the structure of the extracted metadata for the 'intro_to_mods.pdf' file:

```
<Folder>
  <Path>MODS</Path>
  <File xmlns:nz_govt_natlib_xsl_XSLTFunctions=
    |nz_govt_natlib_xsl.XSLTFunctions">
    <FileIdentifier>0-0</FileIdentifier>
    <Filename>
      <Name>intro_to_mods.pdf</Name>
      <Extension>pdf</Extension>
    </Filename>
    <FormerFilename>
      <Name/>
      <Extension>pdf</Extension>
    </FormerFilename>
    <Size>36793</Size>
    <FileDateTime>
      <Date format="yyyyMMdd">20050119</Date>
      <Time format="HHmmssSSS">112809000</Time>
    </FileDateTime>
    <Mimetype>application/pdf</Mimetype>
  </File>
</Folder>
```

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>



A Managed Environment for Storing Digital Files & Metadata

- Paradigm uses the open-source *Fedora* digital repository software. Developed at Cornell and Virginia. See <http://www.fedora.info/>
- Fedora associates a digital object with any kind of valid XML metadata the user wants to add. Wraps this in its METS-like FOXML, but can import and export METS files
- 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

Sample object in Fedora v. 2.2

The screenshot displays the 'Object - paradigm:399' window in the Fedora v. 2.2 interface. The 'Properties' tab is active, showing a tree view on the left with 'objectFile1' selected. The main area contains the following details:

- ID:** objectFile1
- Control Group:** Managed Content
- State:** Active
- Versionable:** Updates will create new version
- Created:** 2007-02-07T17:46:06.150Z
- Label:** Logo for Cairo project
- MIME Type:** image/gif
- Format URI:**
- Alternate IDs:**
- Fedora URL:** http://shuttle.paradigm.ac.uk:8080/fedora/get/paradigm:399/objectFile1
- Checksum:** SHA-1 94cee34930a7afae3b50b15c20fb4e715aa6e5da

Below the metadata is a preview of the logo, which features the word 'cairo' in a stylized font with a row of small figures above it. At the bottom of the window are buttons for 'View', 'Import...', 'Export...', 'Purge...', 'Save Changes...', and 'Undo Changes'.

Sample object in Fedora v. 2.2

Object - paradigm:399

Properties Datastreams Disseminators

DC

PREMISObject

objectFile1

objectFile2

objectFile3

New...

ID **objectFile2**

Control Group Managed Content

State **Active**

Versionable Updates will create new version

Created 2007-03-13T10:29:08.943Z

Label Logo for Cairo project


MIME Type image/gif

Format URI

Alternate IDs

Fedora URL <http://shuttle.paradigm.ac.uk:8080/fedora/get/paradigm:399/objectFile2>

Checksum **SHA-1** 94cee34930a7afae3b50b15c20fb4e715aa6e5da



View Import... Export... Purge...

Save Changes... Undo Changes

Sample object in Fedora v. 2.2

Object - paradigm:399

Properties Datastreams Disseminators

DC

PREMISObject

objectFile1

objectFile2

objectFile3

New...

ID **objectFile3**

Control Group Managed Content

State **Active**

Versionable Updates will create new version

Created 2007-03-13T10:36:47.985Z

Label Logo for Cairo project


MIME Type image/gif

Format URI

Alternate IDs

Fedora URL http://shuttle.paradigm.ac.uk:8080/fedora/get/paradigm:399/objectFile3

Checksum **SHA-1** 829f9254f78deb9c789b28e61056246edd1e4825



View Import... Export... Purge...

Save Changes... Undo Changes

Sample object in Fedora v. 2.2

Object - paradigm:399

Properties | Datastreams | Disseminators

DC

- PREMISObject
- objectFile1
- objectFile2
- objectFile3
- New...

ID: PREMISObject

Control Group: Internal XML Metadata

State: Active

Versionable: Updates will create new version

Created: 2007-03-13T10:51:02.973Z

Label: PREMIS Object metadata for objectFile1

MIME Type: text/xml

Format URI:

Alternate IDs:

Fedora URL: http://shuttle.paradigm.ac.uk:8080/fedora/get/paradigm:399/PREMISObje

Checksum: DISABLED none

```
<objectIdentifier>
  <objectIdentifierType>Local</objectIdentifierType>
  <objectIdentifierValue>paradigm:399/objectFile1</objectIdentifierValue>
</objectIdentifier>
<preservationLevel>Full</preservationLevel>
<objectCategory>File</objectCategory>
<objectCharacteristics>
  <compositionLevel>0</compositionLevel>
  <fixity>
    <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
    <messageDigest>0dad8190e6505fd3791abae600c3d235</messageDigest>
    <messageDigestOriginator>Paradigm preservation repository</messageDigestOriginator>
  </fixity>
  <fixity>
    <messageDigestAlgorithm>SHA-1</messageDigestAlgorithm>
    <messageDigest>94cee34930a7afae3b50b15c20fb4e715aa6e5da</messageDigest>
  </fixity>
</objectCharacteristics>
</objectIdentifier>
```

Edit Import... Export... Purge...

Save Changes... Undo Changes

Storage questions

Should storage be networked, stand-alone or offline?

- Security – preventing unauthorised access and misuse
- Automated integrity checking – guard against corruption
- Backup routines
- Fit with existing system administration?
- Searchability – dealing with enquiries and FOI requests
- Preservation monitoring – can be done via metadata
- Scalability

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'

Arrangement and Description

- Digital Archivist supplies Cataloguer with
 - Files in accessible formats
 - Digital provenance information
 - Background information about computing environment(s)
 - An inventory of files in spreadsheet form
- Cataloguer uses this to:
 - Appraise – marks items for disposal
 - Arrange material in series with traditional materials
 - Mark items for closure, with review dates
 - Allocate reference numbers for researcher access (shelfmarks)
- Hybrid cataloguing
 - Should represent the balance of materials in the archive
 - Simplest when paper and electronic materials clearly belong in separate series (provenance)
 - split monitors useful for cataloguing digital materials

Sample Arrangement Worksheet

samplearrangementworksheet.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

Quantum CY nonprop.001\Partition 1 [2014MB]\WINDOWS 95 [FAT32]\[root]\My Documents\Prog wc 20Nov.doc

| | A | B | C | D | E | F | G | H |
|---|-------------------|--------------|---|---------------|---------|--|------------------|-----------------------------------|
| 1 | MD5 | SHA1 | Filename | Notes | Series | Shelfmark/folio (ie number in sequence of shelfmark) | Restricted (Y/N) | If restricted, record review date |
| 2 | c66f861954301fda | 1898de67bb3 | Quantum CY nonprop.001\Partition 1 [2014MB]\WINDOWS 95 [FAT32]\[root]\My Documents\Conference Wednesday.doc | 27 Sept. 2000 | Diaries | MS. Castle digital 2/5 | No | |
| 3 | 13fa78fb43d49add | 21aeb5ddeca | Quantum CY nonprop.001\Partition 1 [2014MB]\WINDOWS 95 [FAT32]\[root]\My Documents\Prog wc 27Nov.doc | 27 Nov. 2000 | Diaries | MS. Castle digital 2/11 | No | |
| 4 | 08fd6b455fafcfebe | 1dab433ff706 | Quantum CY nonprop.001\Partition 1 [2014MB]\WINDOWS 95 [FAT32]\[root]\My Documents\Prog wc 13 Nov.doc | 13 Nov. 2000 | Diaries | MS. Castle digital 2/9 | No | |
| 5 | 670ce1084639ea0 | 6004dadceec | Quantum CY nonprop.001\Partition 1 [2014MB]\WINDOWS 95 [FAT32]\[root]\My Documents\Wendy\Monthly Plan.doc | 16 Apr. 2001 | Diaries | MS. Castle digital 2/18 | No | |

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