

# **Digital Preservation: From Theory to Practice**

Instructor: Evelyn McLellan

AABC pre-conference workshop  
April 28, 2011



Peter Van Garderen  
President / Systems Archivist



Evelyn McLellan  
Systems Archivist



Jessica Bushey  
Systems Archivist



MJ Suhonos  
Systems Librarian /  
Software Engineer



David Juhasz  
Software Engineer



Austin Trask  
Systems Engineer



Jesús García Crespo  
Software Engineer



Joseph Perry  
Software Engineer

# Agenda

1. Welcome & Introductions
2. What is digital preservation?
3. The Open Archival Information System (OAIS)
4. Metadata
5. About free and open-source software
6. The Archivemata project
7. Preservation planning in Archivemata
8. Using Archivemata (hands-on training)

# Introductions

- Name
- Institution
- Job title / responsibilities
- Nature of digital preservation experience / interest

What is digital preservation?

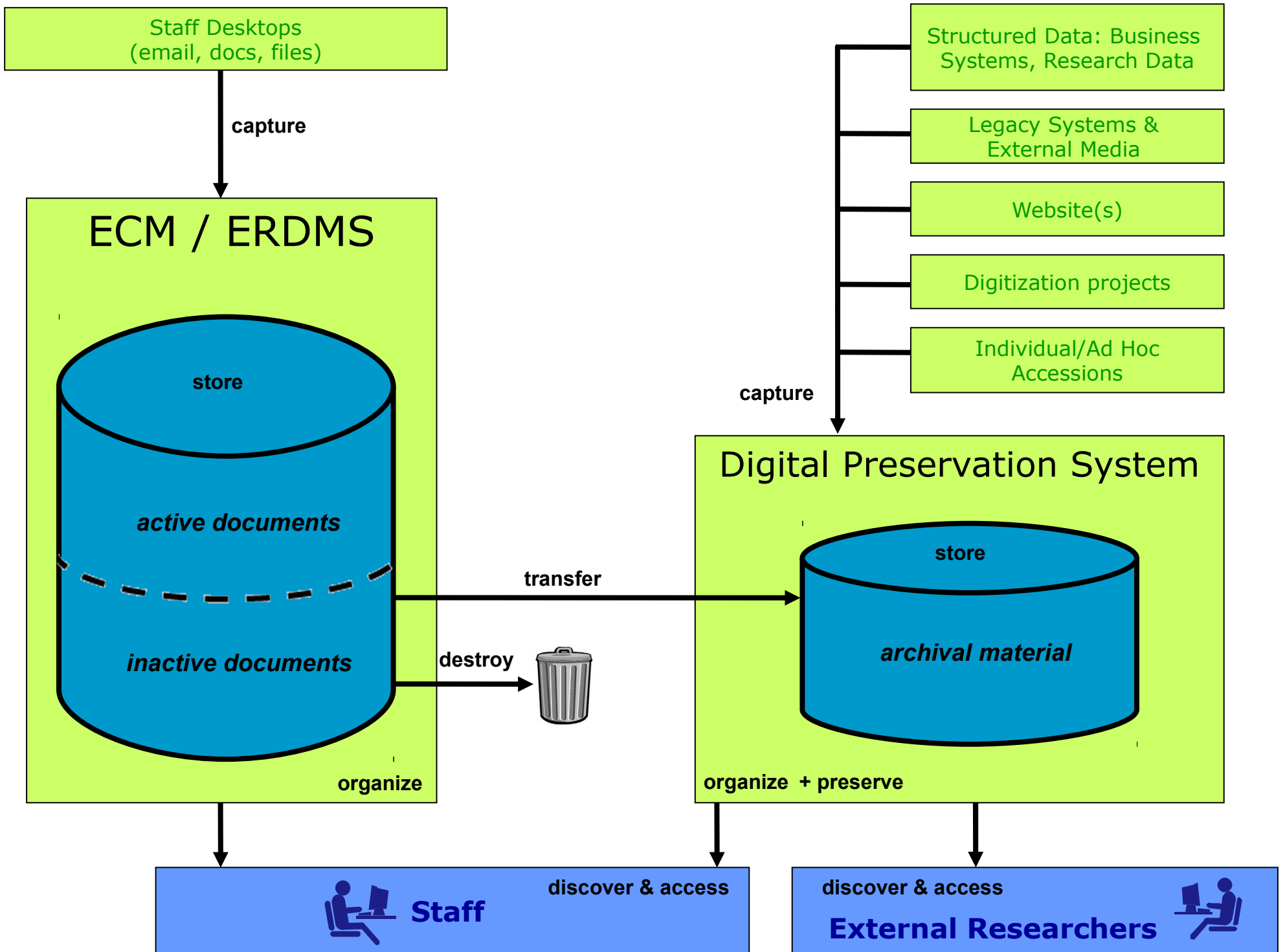
**Digital Preservation:**  
planning for the long-term  
accessibility and usability of  
authentic digital information

# The digital preservation problem

- The complexity of digital information
- Rapid technological change
- Lack or loss of adequate metadata
- Incompatible, obsolete, obscure or proprietary file formats
- Fragility of digital storage media
- The volume of digital information
- Lack of responsibility and resources

# Making the case for digital preservation

- “Don't we already have backup and a business continuity plan?”
- “Don't we just upgrade the software?”
- “Storage is cheap”
- “We'll just index everything”
- “Why can't we use the EDRMS/ECM system for this?”



# Business Case Opportunities

- ERDMS, ECM, DAM implementation
- Enterprise search implementation
- Business process/records scheduling analysis
- Archiving and storage pressure
- Audits
- FOI and disclosure/transparency initiatives
- Open Data / Open Government initiatives
- High-profile e-records transfer

## Authenticity (InterPARES):

The trustworthiness of a record as a record; i.e., the quality of a record that it is what it purports to be and that is free from tampering or corruption

# The authentic record

- What is a record and what makes it authentic?
- Knowing this tells us what to preserve

# Assessing authenticity

- *Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records*
  - Allows the preserver to determine authenticity based on how the records were created and maintained
  - Requires the preserver to establish the identity and integrity of the records

# Assessing authenticity: benchmark requirements

- Identity:
  - Who created the record?
  - Why was it created?
  - Who received it?
  - When was it created and received?
  - What other records does it relate to?

# Assessing authenticity: benchmark requirements

- Integrity:
  - What was the office of primary responsibility?
  - Who has had access to the record?
  - How has it been protected?
  - Have there been any technical modifications to the record?
  - How has it been transferred to the preserver?

# Assessing authenticity: baseline requirements

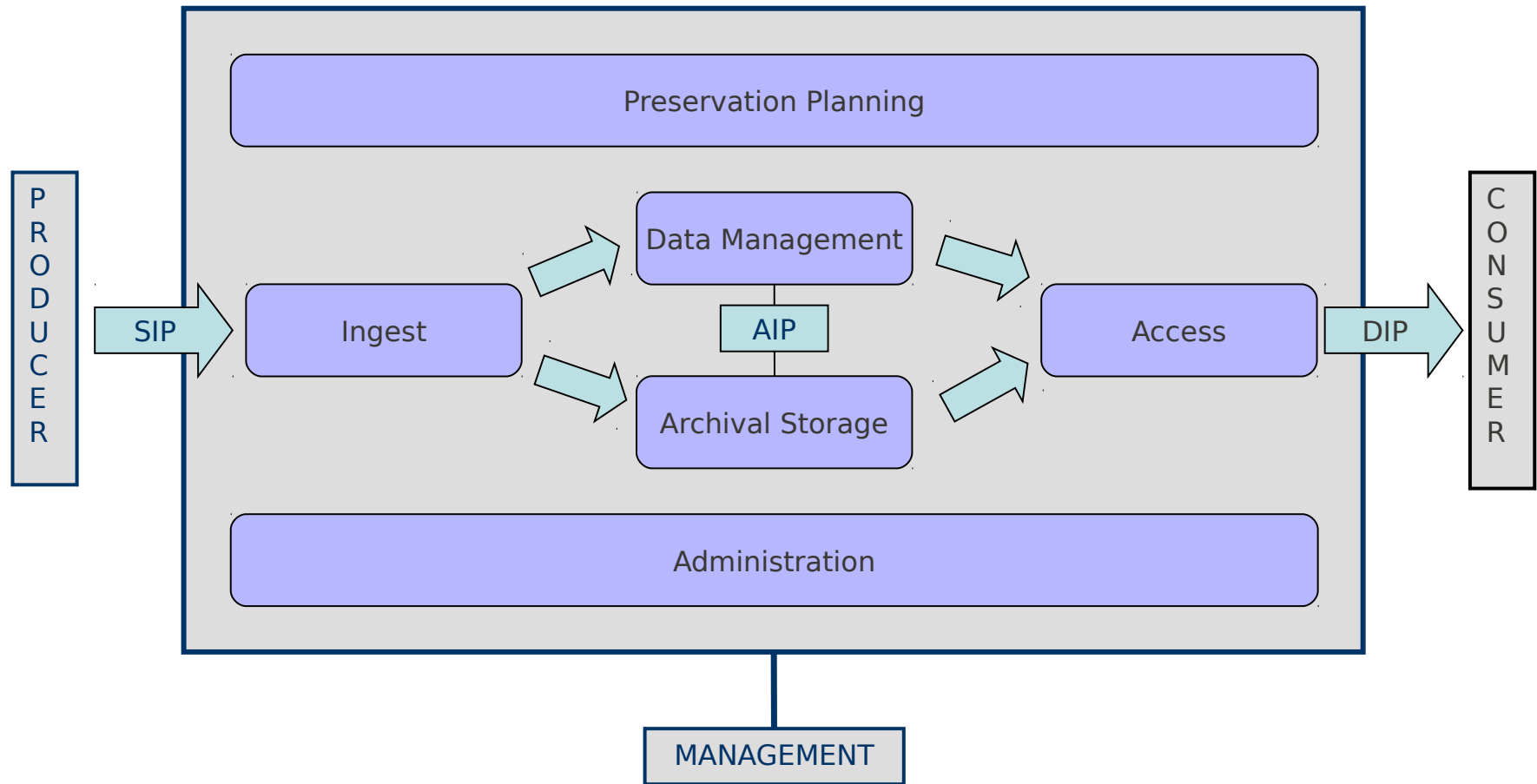
- *Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records*
- Allows the preserver to determine and evaluate minimum long-term preservation requirements

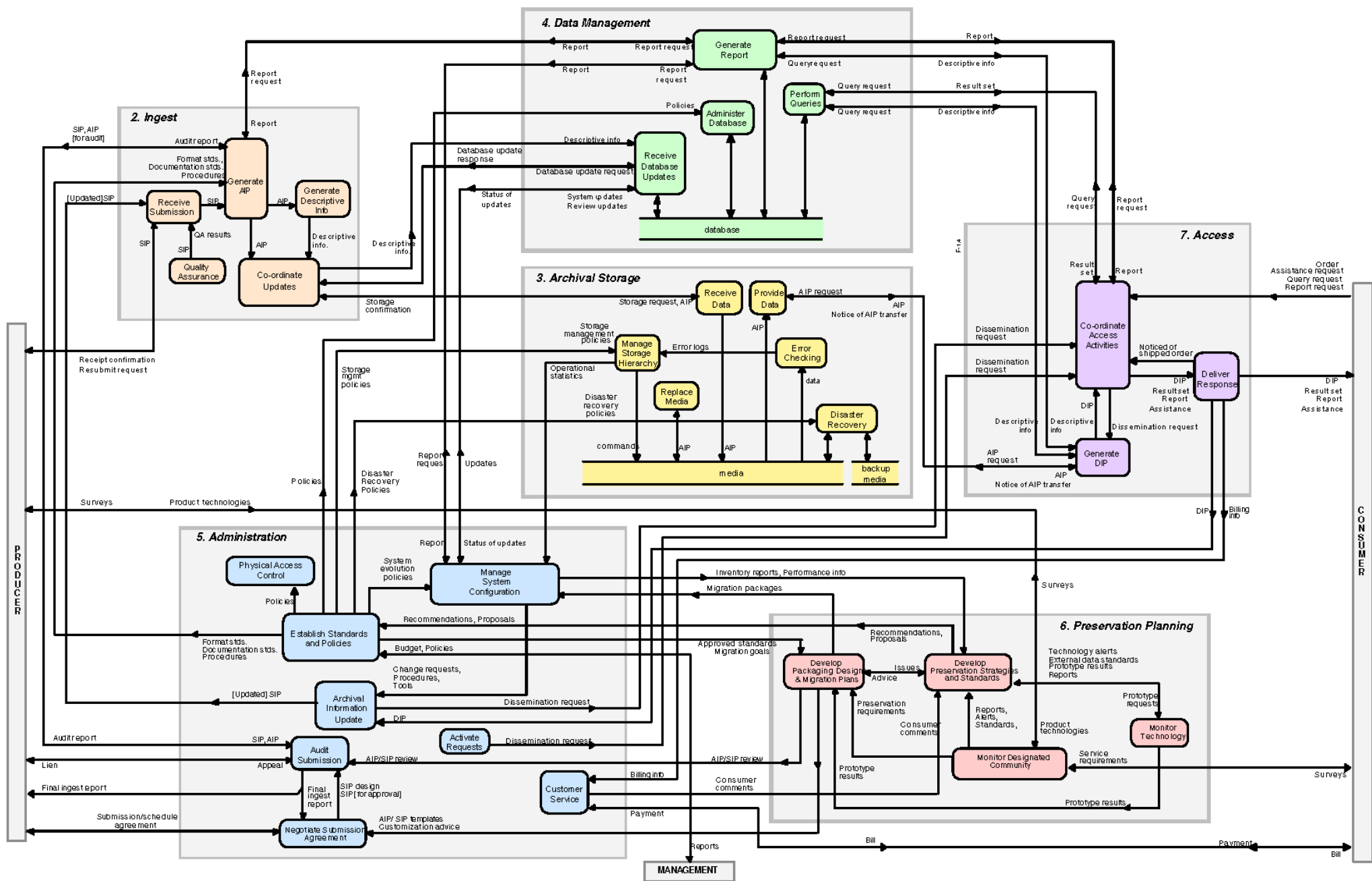
# Assessing authenticity: baseline requirements

- Maintain the chain of custody
- Keep the records secure
- Document all activities
- Describe the records

# The Open Archival Information System (OAIS)

# ISO 14721: Open Archival Information System





# OAIS definitions

- Submission Information Package (SIP)
  - A body of digital objects and associated metadata transferred from the Producer
- Archival Information Package (AIP)
  - A package derived from the SIP containing the digital objects and associated metadata that is preserved in the digital preservation system
- Dissemination Information Package (DIP)
  - A package derived from the AIP containing digital objects and associated metadata delivered to the consumer

# What does OAIS tell us to do?

- “Ingest” a SIP
  - Accept a SIP from a Producer
  - Prepare an AIP
- “Preserve” the AIP
  - Ensure that the objects are securely stored
  - Ensure the ongoing ability to access and use the objects
- Manage information about the objects
- Disseminate the objects

# Breaking it down further

- Ingest a SIP
  - Accept a SIP from a Producer
  - Verify that the transfer was successful
  - Verify that the SIP conforms to a Submission Agreement
  - Check the objects for viruses/malware
  - Identify file formats
  - Validate files against format specifications
  - Extract descriptive and technical metadata
  - Implement preservation plans
  - Create an AIP: the ingested objects, normalized versions of the objects, metadata about the objects, fixity information (checksums or hash values)

# Breaking it down further con't

- Preserve the AIP
  - Place the AIP in storage
  - Make backup copies
  - Periodically check integrity
  - Refresh storage media
  - Implement preservation plans

# Breaking it down further con't

- Manage information about the objects
  - Maintain databases
  - Run queries
  - Generate reports
  - Update metadata

# Breaking it down further con't

- Disseminate the objects
  - Manage access requests
  - Generate access copies
  - Deliver access copies

# Metadata

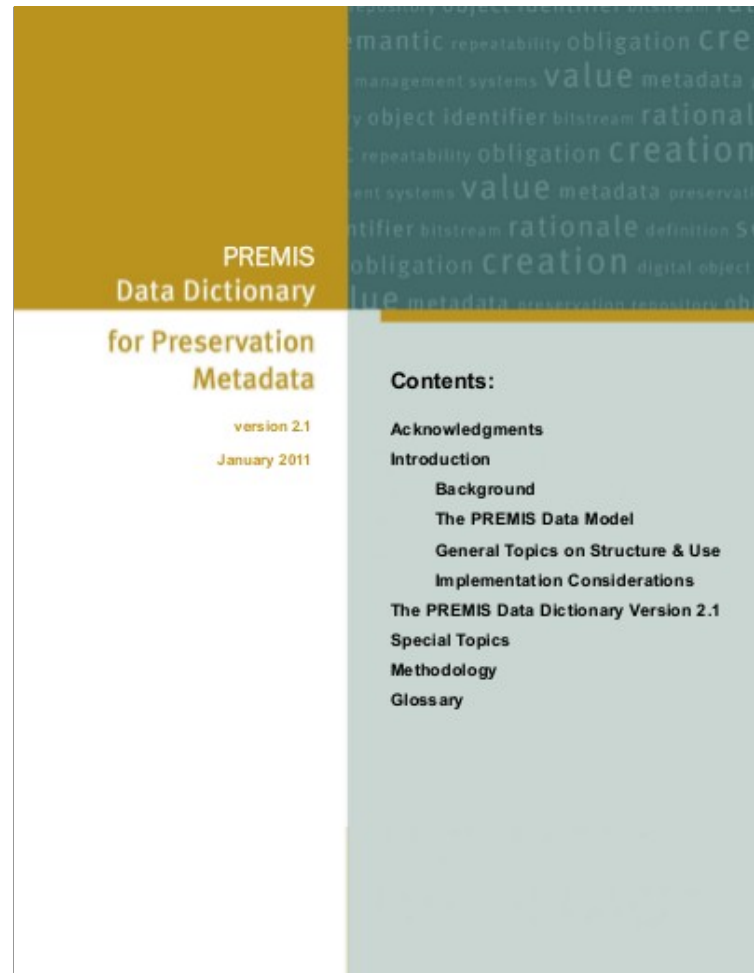
# What is metadata?

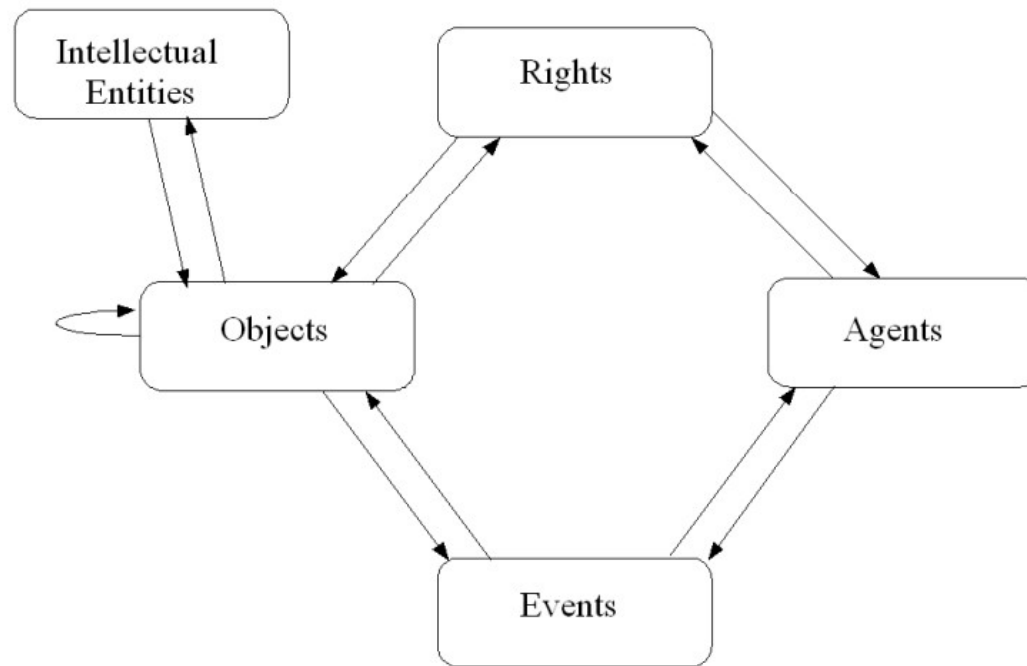
- Information about information
- In this case, information about digital objects
- Types of metadata:
  - Descriptive metadata
  - Preservation metadata
  - Structural metadata

# Descriptive metadata

- Dublin Core
- MODS
- RAD
- EAD

# Preservation metadata





**Figure 1: The PREMIS Data Model**

# Objects

- Identifier
- Category
- Composition level
- Size
- Fixity
- Format
- Characteristics
- Relationships

```
-<mdWrap MDTYPE="PREMIS">
  -<xmlData>
    -<premis:premis xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standards/premis/premis.xsd" version="2.0">
      -<object>
        -<objectIdentifier>
          <objectIdentifierType>UUID</objectIdentifierType>
          <objectIdentifierValue>37c9bd24-e929-474a-b643-a9a7f0d6b3e9</objectIdentifierValue>
        </objectIdentifier>
        <objectCategory>file</objectCategory>
      -<objectCharacteristics>
        <compositionLevel>0</compositionLevel>
        -<fixity>
          <messageDigestAlgorithm>sha256</messageDigestAlgorithm>
          -<messageDigest>
            b6ffa7c9c86e601957534b737bd177efede0102feb19a6d5a9b1e8409691a49f
          </messageDigest>
        </fixity>
        <size>787510</size>
      -<format>
        -<formatDesignation>
          <formatName>Windows Bitmap</formatName>
          <formatVersion>3.0</formatVersion>
        </formatDesignation>
        -<formatRegistry>
          <formatRegistryName>PRONOM</formatRegistryName>
          <formatRegistryKey>fmt/116</formatRegistryKey>
        </formatRegistry>
      </format>
      -<objectCharacteristicsExtension>
        -<fits xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd"
          version="0.5.0" timestamp="4/27/11 11:06 AM">
          -<identification>
```

```

    </mdWrap>
    </amdSec>
    <fileSec>
      <fileGrp ID="Land_image-4e6ad548-df36-4cf9-9cf5-31ca017cfeec" USE="Objects package">
        <fileGrp USE="directory" ID="">
          <xlink:fits ID="file-LAND2-68518e65-25e5-47b3-90ff-31519e0e3afc.tif-e4ff692a-db2d-433f-8e42-a053f5b98f89" ADMID="digiprov-LAND2-68518e65-25e5-47b3-90ff-31519e0e3afc.tif-e4ff692a-db2d-433f-8e42-a053f5b98f89">
            <pre>
</pre>
            </xlink:fits>
          </fileGrp>
        </fileGrp>
      </fileSec>
    </amdSec>
  </digiprovMD>
</mdWrap>
</xmlData>
</premis:premis>
+<agents></agents>
+<events></events>
</object>
</relationship>
</relatedEventIdentification>
</relatedEventIdentifierValue>68518e65-25e5-47b3-90ff-31519e0e3afc</relatedEventIdentifierValue>
</relatedEventIdentifierType>UUID</relatedEventIdentifierType>
</relatedObjectIdentification>
</relatedObjectIdentifierValue>e4ff692a-db2d-433f-8e42-a053f5b98f89</relatedObjectIdentifierValue>
</relatedObjectIdentifierType>UUID</relatedObjectIdentifierType>
</relationshipSubType>is source of</relationshipSubType>
</relationshipType>derivation</relationshipType>
</relationship>
</originalName>objects/LAND2.BMP</originalName>
</objectCharacteristics>
</objectCharacteristicsExtension>
</mdWrap>
</amdSec>
</fileSec>

```

# Events

- Ingestion
- Message digest calculation (fixity)
- Quarantine
- Unpacking
- Virus check
- Format identification
- Format validation
- Normalization

```
</object>
- <events>
  + <event></event>
  + <event></event>
  + <event></event>
  + <event></event>
  + <event></event>
  + <event></event>
  + <event></event>
  + <event></event>
  </events>
+ <agents></agents>
</premis:premis>
```

```
----- ,-----
+<event></event>
+<event></event>
-<event>
  -<eventIdentifier>
    <eventIdentifierType>UUID</eventIdentifierType>
    <eventIdentifierValue>a7baf91f-c12a-45e0-a0a4-214eb246430d</eventIdentifierValue>
  </eventIdentifier>
  <eventType>virus check</eventType>
  <eventDateTime>2011-04-27T18:06:08.583333</eventDateTime>
  -<eventDetail>
    program="Clam AV"; version="ClamAV 0.96.5"; virusDefinitions="ClamAV 0.96.5"
  </eventDetail>
  -<eventOutcomeInformation>
    <eventOutcome>Pass</eventOutcome>
    -<eventOutcomeDetail>
      <eventOutcomeDetailNote/>
    </eventOutcomeDetail>
  </eventOutcomeInformation>
  -<linkingAgentIdentifier>
    <linkingAgentIdentifierType>preservation system</linkingAgentIdentifierType>
    <linkingAgentIdentifierValue>Archivematica-0.7</linkingAgentIdentifierValue>
  </linkingAgentIdentifier>
```

# Agents

- Who or what is doing all these things to the digital objects?
  - Organizations
  - Individuals
  - Software

```
+<event></event>
+<event></event>
+<event></event>
+<event></event>
</events>
-<agents>
  -<agent>
    -<agentIdentifier>
      <agentIdentifierType>preservation system</agentIdentifierType>
      <agentIdentifierValue>Archivematica-0.7</agentIdentifierValue>
    </agentIdentifier>
    <agentName>Archivematica</agentName>
    <agentType>software</agentType>
  </agent>
  -<agent>
    -<agentIdentifier>
      <agentIdentifierType>repository code</agentIdentifierType>
      <agentIdentifierValue>ORG</agentIdentifierValue>
    </agentIdentifier>
    <agentName>Your Organization Name Here</agentName>
    <agentType>organization</agentType>
  </agent>
</agents>
</premis:premis>
```

# Rights

- Copyright
- Licenses
- Statutes
- Rights granted

# Structural metadata

- METS – Metadata Encoding and Transmission Standard
  - Can be used to link multiple objects together, to lay out structural relationships between objects, to describe the relationships between all the elements of the AIP

# About free and open-source software

# Definition

- Open-source software (or “free and open-source software”) is software which can be freely used, modified and redistributed through access to its source code.
- A number of different types of licenses govern the use of the software. But the core is that the code can be freely modified and redistributed.



**Free Beer!**

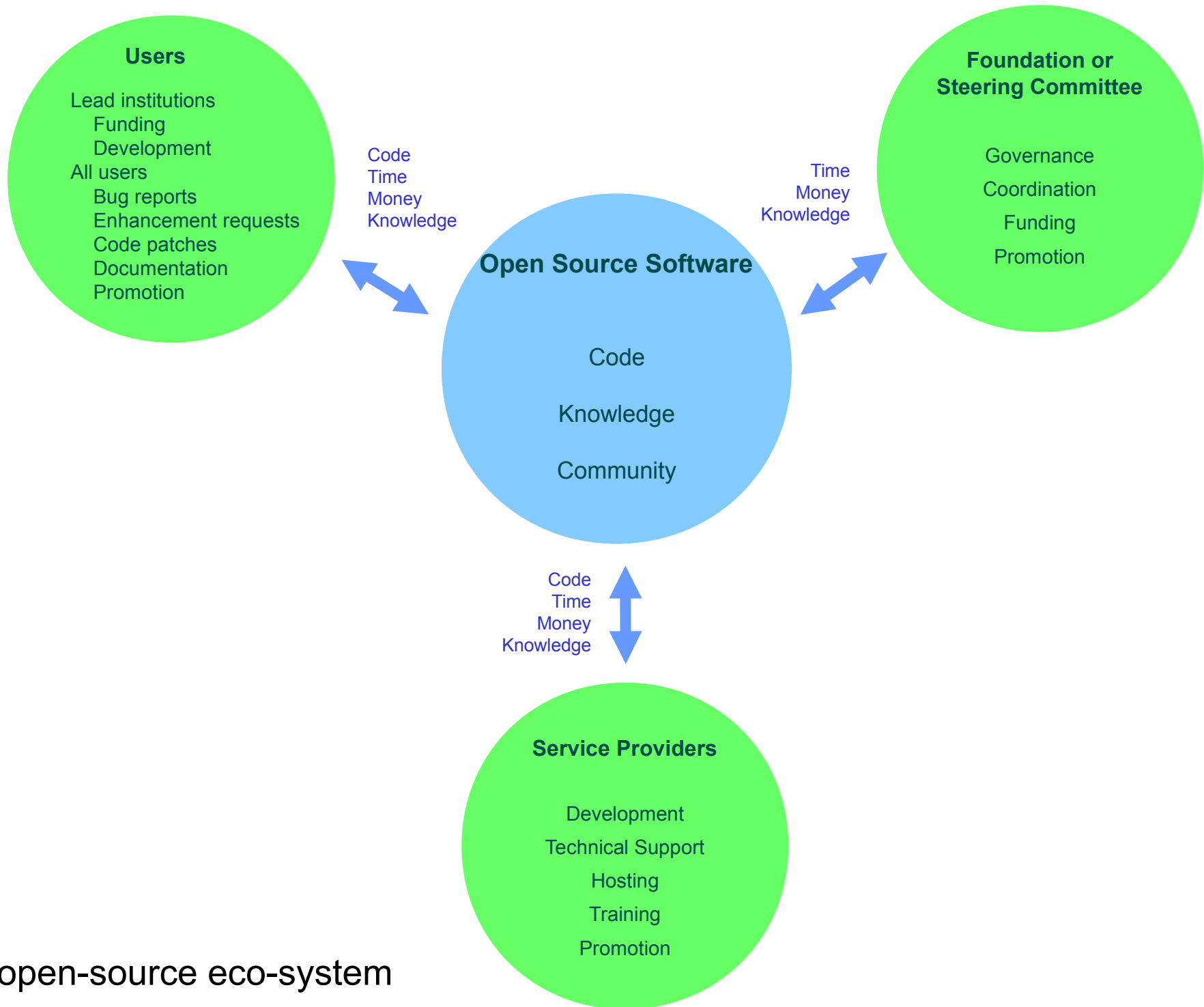
# “They’ll never take our freedom”







**Free Puppy!**



The open-source eco-system

# Some open-source software tools for digital preservation

- Repository software:
  - Fedora, DAITSS, LOCKSS, DSPace, RODA
- Plato (preservation planning tool)
- DROID (file format identification tool)
- JHOVE (file format validation tool)
- FITS (identification, validation, metadata extraction)
- Xena (normalization tool)
- Dioscuri (emulation tool)
- Many others

# The Archivematica project

# What is Archivematica?

- Archivematica is a comprehensive digital preservation system.
- Archivematica uses a micro-services design pattern to provide an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model.
- Archivematica implements media type preservation plans based on an analysis of the significant characteristics of file formats.

# Where did Archivemática come from?

- Artefactual Systems
- City of Vancouver Archives
- UNESCO Memory of the World
- International Monetary Fund Archives
- Rockefeller Archives Center
- University of British Columbia Library
- ?

**MEMORY OF THE WORLD**

**Towards an Open Source Repository  
and Preservation System**

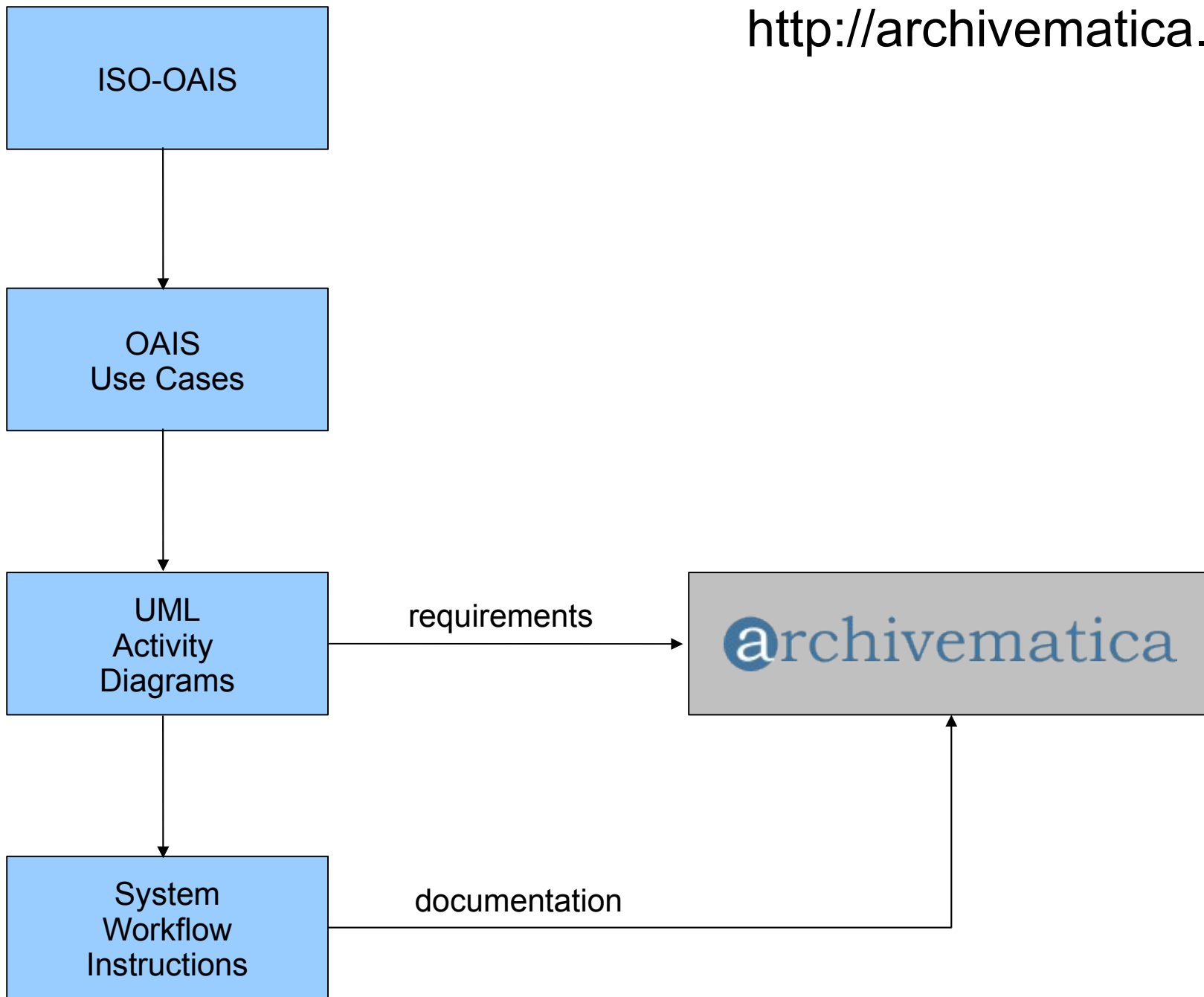
Recommendations on the Implementation of an Open Source  
Digital Archival and Preservation System and on Related  
Software Development



By

**Kevin Bradley**  
National Library of Australia  
UNESCO Memory of the World Sub-Committee on Technology

**Junran Lei,**  
Australian Partnership for Sustainable Repositories,  
**Chris Blackall,**  
Australian Partnership for Sustainable Repositories



# Agile development method

- System releases
  - Feb 2009: Release 0.1-alpha
  - May 2010: Release 0.6-alpha
  - December 2010: Release 0.6.2-alpha
  - February 2011: Release 0.7-alpha
- Each iteration leads to updated and improved:
  - Requirements
  - Software
  - Documentation
  - Development resources



Search go

CDL Home > Services and Projects > UC3 > Curation

<b>University of California Curation Center</b>
Merritt
EZID
Web Archiving Service
Digital Preservation Repository
Consultation Services
<b>Curation Micro-services</b>
Identity Service
Storage Service
Fixity Service
Replication Service
Inventory Service

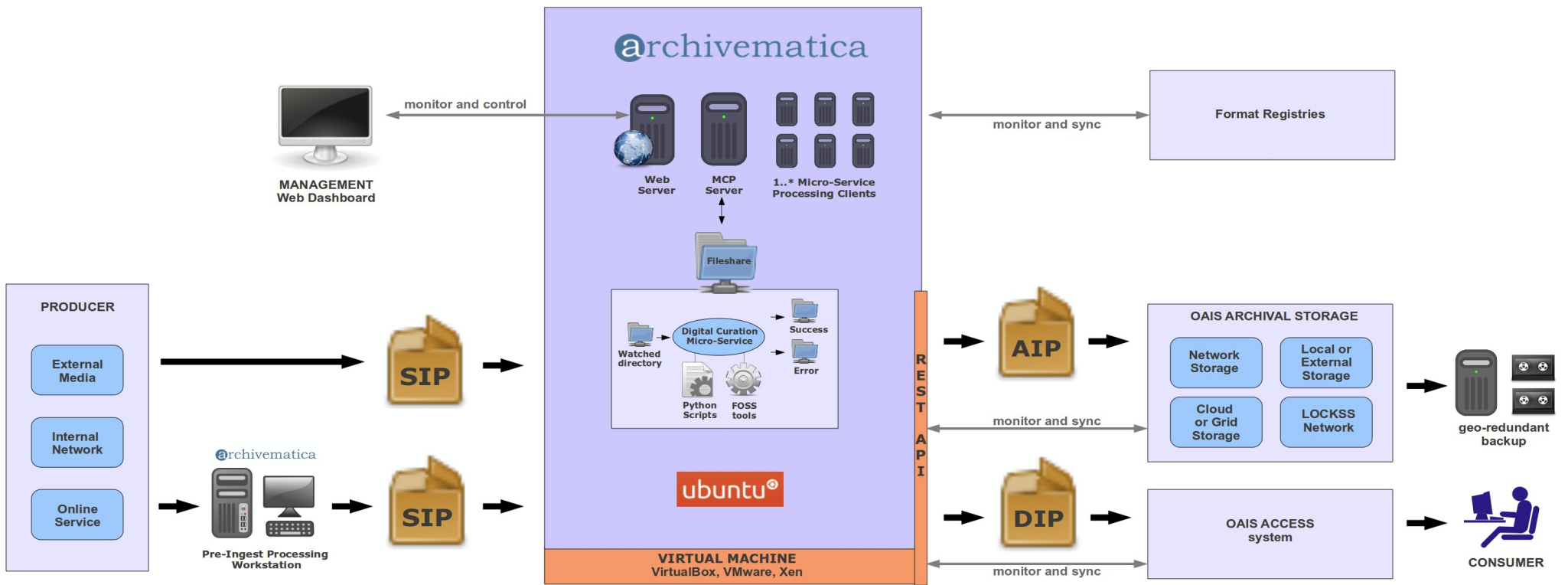
## Curation Micro-Services

Micro-services are an approach to digital curation based on devolving curation function into a set of independent, but interoperable, services that embody curation values and strategies. **Since each of the services is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance. Equally as important, they are more easily replaced when they have outlived their usefulness.** Although the individual services are narrowly scoped, the complex function needed for effective curation emerges from the strategic combination of individual services.

Micro-services provide a curation environment that is comprehensive in scope, yet flexible with regard to local policies and practices and the inevitability of disruptive technological change. Micro-services can be deployed in environments in which it makes most sense, both technically and administratively. UC3 will use micro-services as the basis for its centrally-managed curation activities (for example, the [Digital Preservation Repository](#)); micro-services can also be operated in local campus environments either individually or in strategic combinations.

The initial set of micro-services can be grouped into four categories that provide incrementally increasing levels of preservation assurance and curation value. For more information and documentation, see the [UC3 Curation wiki](#).

- Latest News**
- Deposit, save, share, find that content and data: new UC3 services launch
  - UC3 to host Curate Camp 2010!
  - Webinar of "Merritt" repository development
- [More ...](#)



# Preservation planning in Archivematica

# Digital Preservation Strategies

bitstream preservation

technology preservation

emulation

migration

normalization

# Defining normalization

- What is it?
  - Normalization means converting ingested objects into a small number of pre-selected formats
- Why do it?
  - Some formats are easier to preserve than others
  - A smaller number of formats means fewer preservation actions required

# Normalization vs. migration

- Migration is similar to normalization in that it involves converting ingested objects into preservation-friendly formats
- Unlike normalization, migration is typically done only when the format is at risk of obsolescence
- Migration as a strategy means adopting a wait and see approach

# Normalization vs. emulation

- Emulation means using virtualization to render the object in its original format
- Emulation does not require conversion of the ingested objects
- Emulation is appealing but not yet practical for many types of objects

# Disadvantages of normalization

- It requires more planning up front to implement
- Re-normalization may be required as better target formats or conversion tools become available

# Advantages of normalization

- Taking preservation action on ingest helps define and manage risk
  - Adopting a wait and see approach means putting off an *undefined* amount of work for an *indefinite* period of time at an *unknown* cost
- Normalization does not preclude the future use of migration or emulation

# Criteria for choosing formats

## 1. The format must be non-proprietary

- There must be no associated licenses or patents or the possibility of there being such licenses or patents in the future

# Criteria for choosing formats

## 2. There must be freely available specifications

- A specification is a document that explains exactly how the format is structured and rendered
- This specification must be freely available to all and not subject to copyright or other restrictions

# Criteria for choosing formats

## 3. The format should be widely endorsed and/or adopted

- Other established repositories should be using or have endorsed the format
- Formats that have been approved as international standards are particularly desirable

# Criteria for choosing formats

4. For images and audio files there should be no compression
5. For video files any compression should be completely lossless

# Criteria for choosing formats

6. There should be writing and rendering tools available for the format

- Idealized standards must be matched by practical tools
- The tools must reliably meet the requirements of the format specifications and must produce normalized objects that are faithful representations of the original objects

# Choosing formats

- The Archivemática approach: develop media type preservation plans
  - That is, break the various formats into groups and develop normalization plans for each group

# Media types

- Audio files
- Video files
- Raster images
- Vector images
- Databases
- Text files
- Websites
- Office documents:
  - Word processing files
  - Spreadsheets
  - Presentation files
  - PDF documents

# Case study: raster images

- Preferred format is uncompressed tiff 6.0
  - The format is non-proprietary
  - The specification is freely available
  - The format is used and endorsed by the digital preservation community
  - There are numerous tools capable of writing and rendering the format

# Raster images con't

- Great! So how do we convert from a source format to uncompressed tiff?
  - We convert using ImageMagick, a file conversion tool which is open-source and able to run from a Linux command line

# Raster images con't

- How can we tell whether the conversion was successful?
  - To test the quality of the conversion, we determine what the significant characteristics of the original file are, and measure them pre- and post-conversion
  - We also check that the normalized version renders properly!

# Significant characteristics of raster images

- From Florida Digital Archive preservation action plan for TIFF 6.0:
  - Image height, image width, sequence of images, X sampling frequency, y sampling frequency, samples per pixel, bits per sample, extra samples

## JPEG to TIF using ImageMagick

[Main Page](#) > [Documentation](#) > [Media type preservation plans](#) > [Joint Photographic Experts Group](#) > [JPG to TIF using ImageMagick](#)

- File used was <http://www.archivemata.org/wiki/index.php?title=File:003.JPG> [↗](#)
- Command used was `convert 003.JPG +compress 003.tif`

Property	Original	Normalized
FileSize	1361321 bytes	29965001 bytes
ImageWidth	3648	3648
ImageHeight	2736	2736
BitsPerSample	8	8
Compression	Huffman coding, Baseline DCT	Uncompressed
XSamplingFrequency	72	72
YSamplingFrequency	72	72
SamplesPerPixel	3	3

# Case study: office documents

- Office documents include word processing documents, spreadsheets, presentation files and pdf files

# Preservation formats

- For word processing documents, spreadsheets and presentation files, the Open Document Format (ODF) is an accepted international standard
- For all office documents, PDF/Archival (PDF/A) is also an accepted international standard

# Conversion tools

- Open-source tools exist to convert office documents to these formats
  - The most well-known of these is OpenOffice
  - In Archivematica we have added a tool called Unoconv, which batch converts files using OpenOffice

# Testing the quality of the conversion

- Remember that for raster images we determined the significant characteristics of the files and measured them pre- and post-conversion
- Significant characteristics for these files include image dimensions, resolution, samples per pixel etc.
- These are all easy to measure!

# Significant characteristics of a word processing file

- What are the significant characteristics and how do we measure them?
  - Page count, word count, character count, line count, presence of tables, presence of graphics, font types
- These are hard to measure accurately
- Even if they are measured accurately the elements may be in the wrong place or poorly rendered

# Converting an MS Word file to Open Document Format

- The good:
  - Can convert easily using OpenOffice
  - Can batch convert using Unoconv with OpenOffice

# Converting an MS Word file to Open Document Format

- The bad:
  - The metadata extracted from the files during ingest don't include the significant characteristics
  - There are differences in the way they look on the screen – why is that?

# Converting an MS Word file to Open Document Format

- The ugly:
  - The conversion is problematic because OpenOffice is reverse-engineering from closed specifications
  - The best ODF conversions would come from directly within the native application, but:
    - We can't add Microsoft Office to Archivematica, and
    - Microsoft's support for ODF is relatively weak anyway

# What about Office Open XML?

- Could we use Office Open XML (OOXML) instead of ODF as our preservation format?
  - OOXML is Microsoft's answer to ODF; it was approved as an international standard in 2008

# The problems with OOXML

- It is an extremely lengthy, complex standard
- It is very new and largely untested
- There are no reliable open-source tools to write to it
- It would presumably work well for Microsoft files but what about WordPerfect, OpenOffice and other formats?

# What about PDF/A?

- PDF/Archival is an approved international standard based on PDF 1.4 (PDF/A-1)
- PDF/A is well accepted in the digital preservation community
- Unoconv/OpenOffice can batch convert to PDF/A

# The problems with PDF/A

- It is very difficult to determine and measure significant properties for comparing pre- and post-conversion results
- OpenOffice is not converting from within the native application, so the same problems that are in ODF appear in PDF/A
  - As with ODF, the best conversions would come from directly within the native application – eg via an Adobe Distiller plugin within Microsoft Office

# More problems with PDF/A

- PDF/A-1 does not accept transparencies, and a lot of office files have graphics with transparencies
- PDF/A-1 does not preserve functionality such as animation and slide transitions in presentation files or calculations and macros in spreadsheets

# Solutions?

- The default normalization path for office documents is ODF; some errors in representation are expected
- Documents ingested in OOXML are left in that format
- We are still considering PDF/A-1 but may have to wait for better conversion tools, and it will never be the sole preservation format
- We will consider PDF/A-2 when it is approved

# Creating access formats

- In addition to preservation masters, we need access formats for our Dissemination Information Packages
- We don't want to use copies of the preservation masters for access because they're usually large files

# Creating access formats

- Criteria for selecting access formats
  - They should be small
  - They should be widely used
  - There should be open-source tools available to create them
  - They should look good / sound good / work well
  - They can be proprietary
- Unlike preservation masters, access formats are ephemeral and disposable

# Creating access formats

Original format	Access format
raster images	jpeg
audio files	mp3
video files	mpg-1/mp2
word processing files	pdf
presentation files	pdf
spreadsheets	original format
vector images	pdf
text	original format
databases	no access format
websites	no access format

## Attribution

Title: Digital Preservation: From Theory to Practice (course slides)  
Creator: Peter Van Garderen and Evelyn McLellan, Artefactual Systems Inc.  
Date: April 28, 2011



The original content in this presentation is Copyright Artefactual Systems Inc. 2011. Workshop participants and the general public may freely re-use this content under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike 3.0 license