Preservation Metadata
CARLI Metadata Matters series
December 7, 2010

Claire Stewart
Head, Digital Collections
Northwestern University Library

http://tinyurl.com/carli-claire-digpres
THE HOTTEST YEAR

The release of climate-science e-mails last November ripped apart Phil Jones’s life. He’s now trying to patch it back together.

Published online 15 November 2010 | Nature 468, 362-364 (2010) |
doi:10.1038/468362a
Science forgotten in climate emails fuss
No one identifies any scientific flaws in Phil Jones's work, yet the 'fallen idol' narrative is too alluring for the media to resist.

Myles Allen
guardian.co.uk, Friday 11 December 2009 12.30 GMT
Artica history

Take, for example, the "trick" of combining instrumental data and tree-ring evidence in a single graph to "hide the decline" in temperatures over recent decades that would be suggested by a naive interpretation of the tree-ring record. The journalists repeating this phrase as an example of "scientists accused of manipulating their data" know perfectly well that the decline in question is a spurious artefact of the tree-ring data that has been documented in the literature for years, and that "trick" does not mean "deceit". They also know their readers, listeners and viewers won't know this: so why do they keep doing it?

“Climategate”
November 2009
What would “Climategate” for a library look like?

Hypothetical situation:

• We hold the papers of a former member of the faculty; decades after her death, interest in her researchskyrockets
• Most of the material still exists in print form
• Some of the originals have somehow disappeared, but digital surrogates were created when the collection was processed
• Given the high stakes, how can we state, with confidence, that the digital surrogates are reliable, authentic, have not been tampered with, etc.?

PRESERVATION METADATA TO THE RESCUE

Preservation Metadata, CARLI Metadata Matters, December 7, 2010
Quick poll

Are you currently creating/capturing/storing preservation metadata?

Yes = green check, No = red “x”
Quick poll (2)

Are you creating/storing PREMIS?

Yes = green check, No = red “x”

Also:
if you have specific things you’re interested in talking about today, this would be a good time to shout (type) them out
Outline

• Conceptual and introductory stuff
• Strategizing an approach to preservation metadata
• PREMIS overview
• PREMIS examples
  – Northwestern Books
  – Portico
  – HathiTrust
• Open discussion
Curation. The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.

Archiving. A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.

Preservation. An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

The DCC Curation Lifecycle Model

CONCEPTUALISE

CREATE OR RECEIVE

TRANSFORM

CURATE

APPRaise & SELECT

PRESERVATION PLANNING

DESCRIPTION

Data
(Digital Objects
or Databases)

REPRESENTATION INFORMATION

COMMUNITY WATCH & PARTICIPATION

PRESERVE

INGEST

PRESERVATION ACTION

STORE

ACCESS, USE & REUSE

MIGRATE

DISPOSE

and

and
Figure 4-1: OAIS Functional Entities
A few terms

- SIP: Submission Information Package
- AIP: Archival Information Package
- DIP: Dissemination Information Package (from the OAIS reference model)

- METS = Metadata Encoding and Transmission Standard
Quick poll (3)

Are you creating/storing METS?

Yes = green check, No = red “x”
Preservation metadata supports activities intended to ensure the long-term usability of a digital resource.

-Priscilla Caplan, Understanding PREMIS

Before any preservation metadata is created: WHY?

• Ability to share between archives is critical to long-term health

• Sharing depends on clear understanding of what the objects are and how they were created (who did what and when?)

• Everything is potentially worth documenting: how and why was object created, what metadata decisions were made and why, etc.
PREMIS
PREMIS data model

PREMIS Data Dictionary for Preservation Metadata version 2.0.
Example: a (super-simplified) PREMIS XML document

<premis version = "2.0">
<object xsi:type="file">
...
</object>
<event>
...
</event>
<agent>
...
</agent>
</premis>
Slightly expanded view

<premis version = “2.0”>
<object xsi:type=“file”>
   <objectIdentifier> ... </objectIdentifier>
   <objectCharacteristics>
      <fixity>
         <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
         <messageDigest>7b7c655e2e25867e1ed1062f7102e5ef</messageDigest>
         <messageDigestOriginator>Archive</messageDigestOriginator>
      </fixity>
   </objectCharacteristics>
   ... 
</object>
<event>
   <eventIdentifier> ... </eventIdentifier>
   <eventType>Creation</eventType>
   <eventDateTime>2010-12-07T00:22:46-05:00</eventDateTime>
   <eventOutcomeInformation> ... </eventOutcomeInformation>
   ... 
</event>
<agent> ... </agent>
</premis>

Loosely based on PREMIS generated by FCLA’s DAITSS2 Format Description Service: http://description.fcla.edu/

Preservation Metadata, CARLI Metadata Matters, December 7, 2010
What’s NOT defined in PREMIS?

- Format-specific metadata
- Implementation-specific
- Descriptive metadata
- Details about media, hardware
- Detailed agent information
- Details about rights, permissions not affecting preservation functions

from *Understanding PREMIS*
Preservation Metadata

Detailed info on Agents, Rights, Media or Hardware

Descriptive Metadata

Repository Business Rules

PREMIS

Format-specific Technical Metadata

from Understanding PREMIS
Case study: PREMIS for Northwestern Books

Preservation Metadata, CARLI Metadata Matters, December 7, 2010
Scan → Identify edges → Batch-process to crop, remove background → Clean up and quality control
Pre-BWI: each step with human intervention

- Pages missing from original: request from ILL
- Send for rescans
- Problem with one or few pages
  - Crash during insert or page rearrangement, book file corrupted, recovery not possible, start over

- Scan
- Identify edges
- Batch-process to crop, remove background
- Clean up, insert foldouts and rescans, quality control

- Foldouts removed
- Foldouts scanned separately

Tracking database and shared file server
Book Workflow Interface (BWI)
Fedora Digital Object
List Datastreams

Object Identifier (PID): inu:inu-mntb-0005628062-pg-e52b9e2a-f29d-48b2-a4e6-194a57e23340

<table>
<thead>
<tr>
<th>Datastream ID</th>
<th>Datastream Label</th>
<th>MIME Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>Default Dublin Core Record</td>
<td>text/xml</td>
</tr>
<tr>
<td>MODS</td>
<td>MODS XML</td>
<td>application/xml</td>
</tr>
<tr>
<td>RELS-EXT</td>
<td>Fedora Object-to-Object Relationship Metadata</td>
<td>text/xml</td>
</tr>
<tr>
<td>PDF</td>
<td>OCR PDF File</td>
<td>application/pdf</td>
</tr>
<tr>
<td>OCR_XML</td>
<td>OCR XML File</td>
<td>text/xml</td>
</tr>
<tr>
<td>OCR_TEXT</td>
<td>OCR Plaintext File</td>
<td>text/plain</td>
</tr>
<tr>
<td>ARCHV-IMG</td>
<td>Source Camera Image File</td>
<td>image/jpeg</td>
</tr>
<tr>
<td>ARCHV-TECHMD</td>
<td>MIX Technical Metadata</td>
<td>text/xml</td>
</tr>
<tr>
<td>ARCHV-EXIF</td>
<td>Camera EXIF Technical Metadata</td>
<td>text/xml</td>
</tr>
<tr>
<td>PROC-IMG</td>
<td>Corrected Image File</td>
<td>image/jpeg</td>
</tr>
<tr>
<td>PROC-TECHMD</td>
<td>Corrected Image MIX Technical Metadata</td>
<td>text/xml</td>
</tr>
<tr>
<td>DELIV-IMG</td>
<td>Jpeg2000 Image File</td>
<td>image/jp2</td>
</tr>
<tr>
<td>DELIV-TECHMD</td>
<td>Jpeg2000 Image MIX Technical Metadata</td>
<td>text/xml</td>
</tr>
<tr>
<td>DELIV-OPS</td>
<td>SVG for image delivery mechanism</td>
<td>image/svg+xml</td>
</tr>
<tr>
<td>HTML</td>
<td>Viewer html</td>
<td>text/html</td>
</tr>
</tbody>
</table>
Fedora Digital Object
List Datastreams

Object Identifier (PID): inu:inu-mntb-0005628062-bk
Version Date: current

<table>
<thead>
<tr>
<th>Datastream ID</th>
<th>Datastream Label</th>
<th>MIME Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>Default Dublin Core Record</td>
<td>text/xml</td>
</tr>
<tr>
<td>MODS</td>
<td>MODS XML</td>
<td>application/xml</td>
</tr>
<tr>
<td>MARC</td>
<td>MARC XML</td>
<td>application/xml</td>
</tr>
<tr>
<td>BOOKSTRUCT</td>
<td>Book Struct METS XML</td>
<td>text/xml;charset=utf-8</td>
</tr>
<tr>
<td>RELS-EXT</td>
<td>Fedora Object-to-Object Relationship Metadata</td>
<td>text/xml;charset=utf-8</td>
</tr>
<tr>
<td>HTML</td>
<td>Book viewer html</td>
<td>text/html</td>
</tr>
<tr>
<td>KIRTA_00000972-001</td>
<td>Tar file of Kirtas files</td>
<td>application/x-tar</td>
</tr>
<tr>
<td>PDF</td>
<td>Entire Book PDF</td>
<td>application/pdf</td>
</tr>
<tr>
<td>PROVENANCE</td>
<td>PREMIS XML</td>
<td>text/xml;charset=utf-8</td>
</tr>
</tbody>
</table>
Guiding questions for Books PREMIS

What is actually possible to capture?
• Every event easily tied to a step in the Crabcake system
• Basic information about every user of the Crabcake system

What do we absolutely need to capture?
• Any event that results in a new digital file

What would be nice to have but is not possible in this grant timeframe?

What do we NOT want to capture?
• Any information that is already captured elsewhere:
  – Version information stored in Fedora
  – Metadata creation timestamps
  – Specifics of process and editing stored in jobs.xml or batchprocess.xml
  – Relationships between files as reflected in the RELS-EXT
Design evolved gradually

• Version 1: November 2008
  – Intellectual entity: representation of book
  – Objects: Fedora book object, maybe page objs?
  – Events: Fedora creation, scanning, each step in the image post-processing, rescans, foldout insertions, creation of METS book structure, OCR, permanent URL (Handle), JP2 file creation, ingestion
  – Agents: each piece of software used, each person operating
  – Rights: copyright status of book as a whole, detailed notes about likely expiration, investigative steps, etc.
Design evolved gradually

• Versions 2,3,4: December 2008
  – ADDED events specific to new BWI system
    • book files copied between processing workstations
    • “associate rescans” (page replacement)
  – Included specific object references to each page object in any event where they were implicated
Design evolved gradually

• Version 5: September 2009
  – All events eliminated except “Approve”
  – No references to objects other than the book object
  – Simplified rights statement
Design evolved gradually

• Version 6: February 2010
  – Same extremely lightweight structure: one object, one event, one agent
  – Further simplified rights statement to bring it in line with University of Michigan’s codes developed in Google scan evaluation
  – This is the version actually implemented
FINAL (1 of 3): object

  <object xsi:type="representation">
    <objectIdentifier>
      <objectIdentifierType>hdl:2166</objectIdentifierType>
      <objectIdentifierValue>inu:inu-mntb-0005834078-bk</objectIdentifierValue>
    </objectIdentifier>
    <preservationLevel>
      <preservationLevelValue>Full</preservationLevelValue>
    </preservationLevel>
  </object>
</premis>
<event>
  <eventIden'fier>
    <eventIden'fierType>hdl:2166</eventIden'fierType>
    <eventIden'fierValue>inu-event-0005834078-6946be40-ace8-4fd9-bff8-1b6edd319bfa</eventIden'fierValue>
  </eventIden'fier>
  <eventType>Approve Book</eventType>
  <eventDateTime>2010-12-06T13:00:41</eventDateTime>
  <eventDetail>Book Approved</eventDetail>
  <eventOutcomeInforma'on>
    <eventOutcome>0</eventOutcome>
  </eventOutcomeInforma'on>
  <linkingAgentIden'fier>
    <linkingAgentIden'fierType>hdl:2166</linkingAgentIden'fierType>
    <linkingAgentIden'fierValue>inu-agent-software-mountingbooks</linkingAgentIden'fierValue>
    <linkingAgentRole>Executing program</linkingAgentRole>
  </linkingAgentIden'fier>
  <linkingObjectIden'fier>
    <linkingObjectIden'fierType>hdl:2166</linkingObjectIden'fierType>
    <linkingObjectIden'fierValue>inu:inu-mntb-0005834078-bk</linkingObjectIden'fierValue>
    <linkingObjectRole>outcome</linkingObjectRole>
  </linkingObjectIden'fier>
</event>
<rights>
  <rightsStatement>
    <rightsStatementIdentifier>
      <rightsStatementIdentifierType>hdl:2166</rightsStatementIdentifierType>
      <linkingObjectIden'tierValue>inu-rights-0005834078-9ea15dd9-5ece-48b1-aec9-f0dd6161552</linkingObjectIden'tierValue>
    </rightsStatementIdentifier>
    <rightsBasis>Copyright</rightsBasis>
    <copyrightInforma'tion>
      <copyrightStatus(pd,cdpp</copyrightStatus>
      <copyrightJurisdiction(us</copyrightJurisdiction>
      <copyrightStatusDeter'minationDate>2010-12-06T12:54:27</copyrightStatusDeter'minationDate>
      <copyrightNote/>
    </copyrightInforma'tion>
    <licenseInforma'tion>
      <licenseTerms>null,null</licenseTerms>
      <licenseNote/>
    </licenseInforma'tion>
    <rightsGranted>
      <act>world</act>
      <termOfGrant>
        <startDate>2010-12-06T12:54:27</startDate>
        </termOfGrant>
      </rightsGrantedNote/>
    </rightsGranted>
  </rightsStatement>
</rights>
Lessons learned

• “Simple” is relative
• If you don’t have tools, you will probably have to be flexible to accommodate the capabilities of those who will implement them for you
• “Revisit it later” can be dangerous
• Missing vocabularies: kind of a problem!
Other examples: event logging

• PORTICO event logging  (PPT from PREMIS Imp Fair 2009)
• HathiTrust  (PPT from PREMIS Imp Fair 2010)

(for CARLI webinar purposes, selected slides from these two presentations are copied into this deck; the following slides are created by others, follow above links for the originals )
PREMIS fair 2009

Begin slides from Evan Owens, Portico
Bring together information common to all the events from a given processing pass; e.g., initial ingest, future migration, etc.
Not a real event!

Example XML serialization showing all possible child elements to illustrate the information model
Event Types

• Check: Virus, Fixity, ...
• Characterize: File, ...
• Generate: Desc. MD, Tech. MD, Fixity, ...
• Edit: Desc. MD, ...
• Set: Status, Format, Preservation Level, ...
• Ingest: into Archive
• Add, Create, Remove File
Mapping PMD 2.0 to PREMIS

<table>
<thead>
<tr>
<th>Portico 2.0 Event Model</th>
<th>PREMIS Event Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique ID</td>
<td>eventIdentifier</td>
</tr>
<tr>
<td>Timestamp</td>
<td>eventDateTime</td>
</tr>
<tr>
<td>Type of Event</td>
<td>eventType</td>
</tr>
<tr>
<td>Rationale for the Event</td>
<td>eventDetail</td>
</tr>
<tr>
<td>Agent</td>
<td>—</td>
</tr>
<tr>
<td>User Info</td>
<td>linkingAgentIdentifier; linkingAgentRole</td>
</tr>
<tr>
<td>Processing Record</td>
<td>(not sure where to put this yet...)</td>
</tr>
<tr>
<td>Process</td>
<td>—</td>
</tr>
<tr>
<td>Arguments</td>
<td>(not sure where to put this yet...)</td>
</tr>
<tr>
<td>Input objects</td>
<td>linkingObjectIdentifier, linkingObjectRole</td>
</tr>
<tr>
<td>Output objects</td>
<td>linkingObjectIdentifier, linkingObjectRole</td>
</tr>
<tr>
<td>Tool info</td>
<td>(not sure where to put this yet...)</td>
</tr>
<tr>
<td>Outcome</td>
<td>—</td>
</tr>
<tr>
<td>Result</td>
<td>eventOutcome</td>
</tr>
<tr>
<td>Details</td>
<td>eventOutcomeDetailNote</td>
</tr>
</tbody>
</table>
Observations

• Large-scale automated events feel very different from human events
• ITHAKA archive will quadruple in 2010
  – Likely 3-5 billion events . . .
• Every bit of metadata has to be need justified
• Events have proved their value
  – An entire talk on that subject alone
• Nothing is easy in quantities of billions
• We still have to work on full lifecycle events
• THIS IS STILL A WORK IN PROGRESS!
PREMIS fair 2009

End slides from Evan Owens, Portico
PREMIS fair 2010

Begin slides from Shane Beers, University of Michigan
Use of PREMIS for Internet Archive AIPs

September 22, 2010
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>Internet Archive</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>capture1</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>capture</PREMIS:eventType>
  <PREMIS:eventDateTime>2008-08-04T19:50:13</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Initial capture of item</PREMIS:eventDetail>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>AgentID</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>Internet Archive</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>scribe7.la.archive.org</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>image capture</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UM</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>fixity check1</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>fixity check</PREMIS:eventType>
  <PREMIS:eventDateTime>2010-04-27T16:34:02</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Calculation of md5 hash values for downloaded IA files, comparison with pre-download md5 values</PREMIS:eventDetail>
  <PREMIS:eventOutcomeInformation>
    <PREMIS:eventOutcome>warning</PREMIS:eventOutcome>
    <PREMIS:eventOutcomeDetail>
      <PREMIS:eventOutcomeDetailNote>files failed checksum validation</PREMIS:eventOutcomeDetailNote>
    </PREMIS:eventOutcomeDetail>
  </PREMIS:eventOutcomeInformation>
  <HT:fileList status="failed">
    <HT:file>arcanacelestiah03swed_files.xml</HT:file>
    <HT:file>arcanacelestiah03swed_meta.xml</HT:file>
  </HT:fileList>
</PREMIS:event>
PREMIS fair 2010

End slides from Shane Beers,
University of Michigan
Hathi ingesting IA files: events

- IA capture of item (book scanning)
- Hathi fixity check on IA files
- Inspecting IA METS for missing files
- Rewriting IA image headers
- Rename files to Hathi conventions
- Split a single OCR file into plain text and XML
- Create IA METS
- Calculate page MD5 checksums
- Validate METS
PREMIS in METS

<METS:mets>
<METS:metsHdr>...
<METS:dmdSec>...
<METS:amdSec>
  <METS:techMD>...
  <METS:digiprovMD>...
    <PREMIS:premis>...
  </METS:digiprovMD>
</METS:amdSec>
<METS:fileSec>...
<METS:structMap>...
</METS:mets>
Practical considerations

• What is your environment capable of supporting?
• Are you duplicating information stored elsewhere? (and if so, do you care?)
• What are your tools?
• Walk before you run
Science forgotten in climate emails fuss

No one identifies any scientific flaws in Phil Jones's work, yet the 'fallen idol' narrative is too alluring for the media to resist.

Myles Allen

guardian.co.uk, Friday 11 December 2009 12.30 GMT

Article history

About this article

Science forgotten in climate emails fuss | Myles Allen

This article was published on guardian.co.uk at 12.30 GMT on Friday 11 December 2009. It was last modified at 15.06 GMT on Friday 11 December 2009.
Discussion