Preservation Metadata
CARLI Metadata Matters series
December 7, 2010

Claire Stewart

Head, Digital Collections

Northwestern University Library

http://tinyurl.com/carli-claire-digpres

nature news

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.

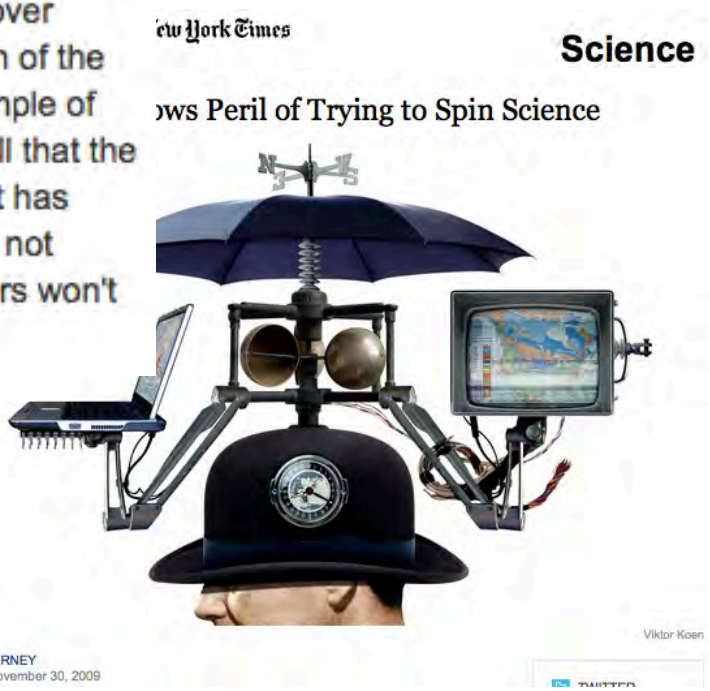
Published online 15 November 2010 | *Nature* **468**, 362-364 (2010) | doi:10.1038/468362a

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
Article 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 research skyrockets
- 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

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

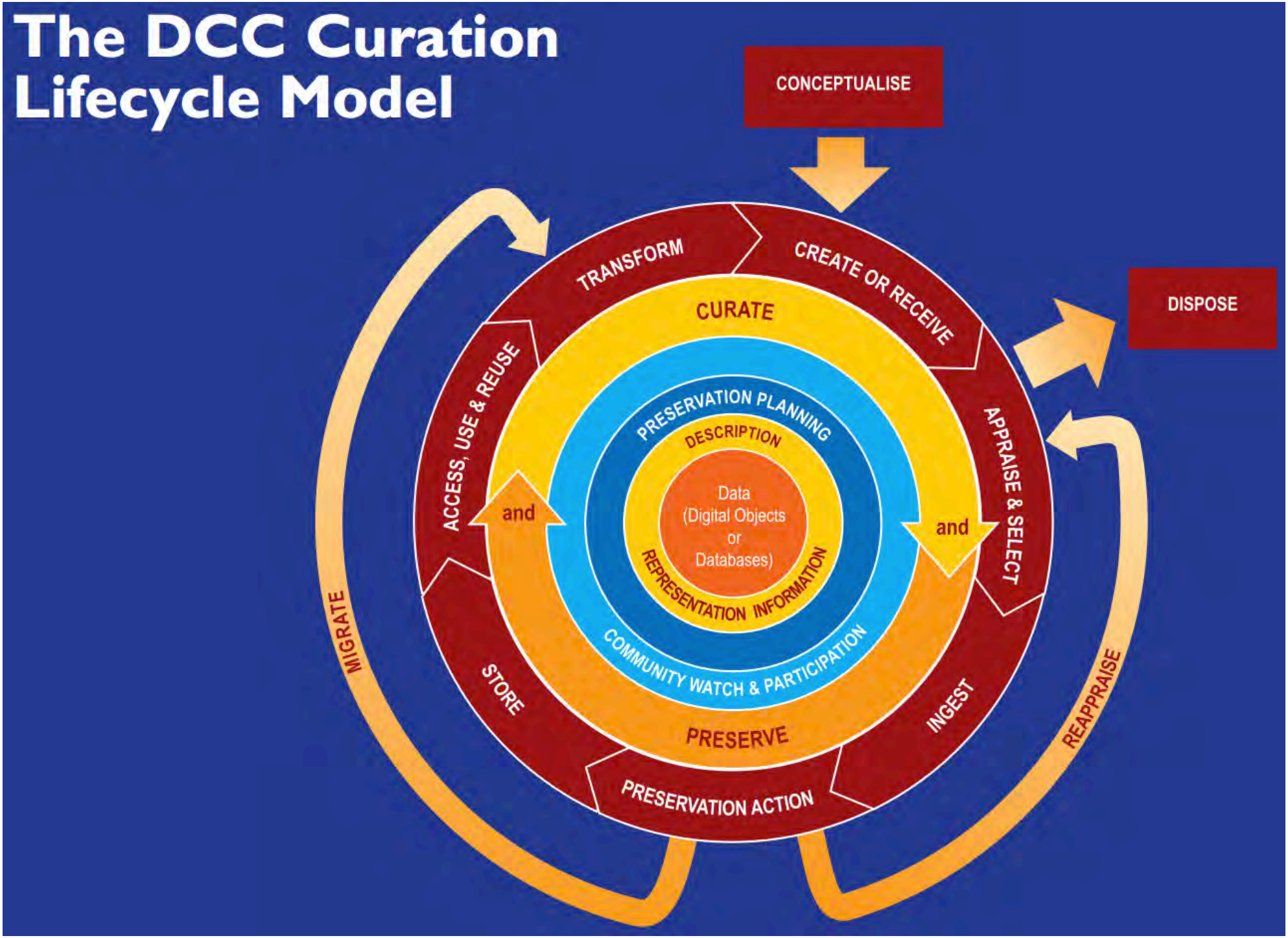
Curation. The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation 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.

Lord, Philip, and Alison Macdonald. 2003. *Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, prepared for The JISC Committee for the Support of Research (JCSR)*. http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf.

The DCC Curation Lifecycle Model



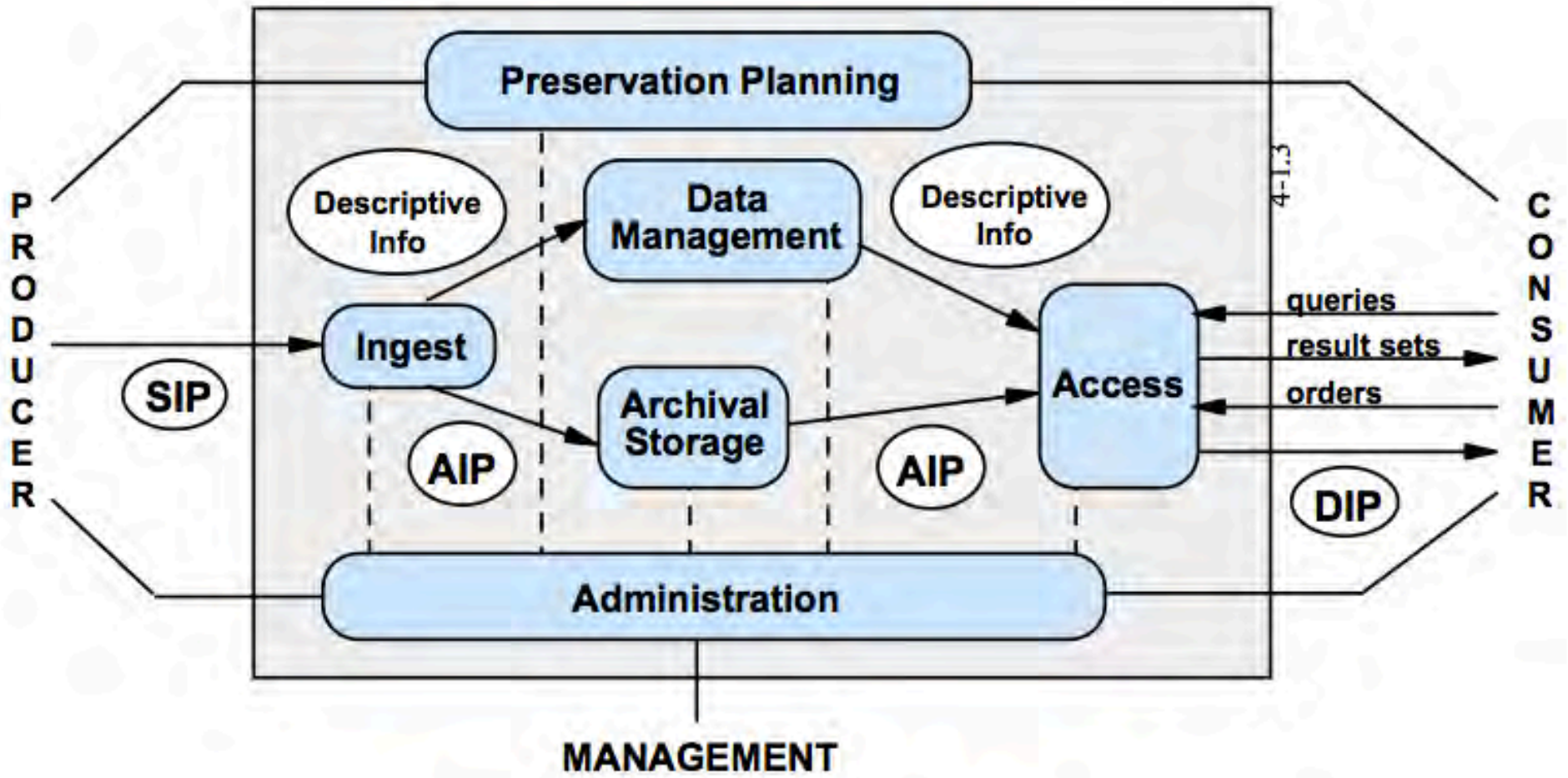


Figure 4-1: OAIS Functional Entities

A few terms

- SIP: **S**ubmission **I**nformation **P**ackage
- AIP: **A**rchival **I**nformation **P**ackage
- DIP: **D**issemination **I**nformation **P**ackage (from the OAIS reference model)

- METS = **M**etadata **E**ncoding and **T**ransmission **S**tandard

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*

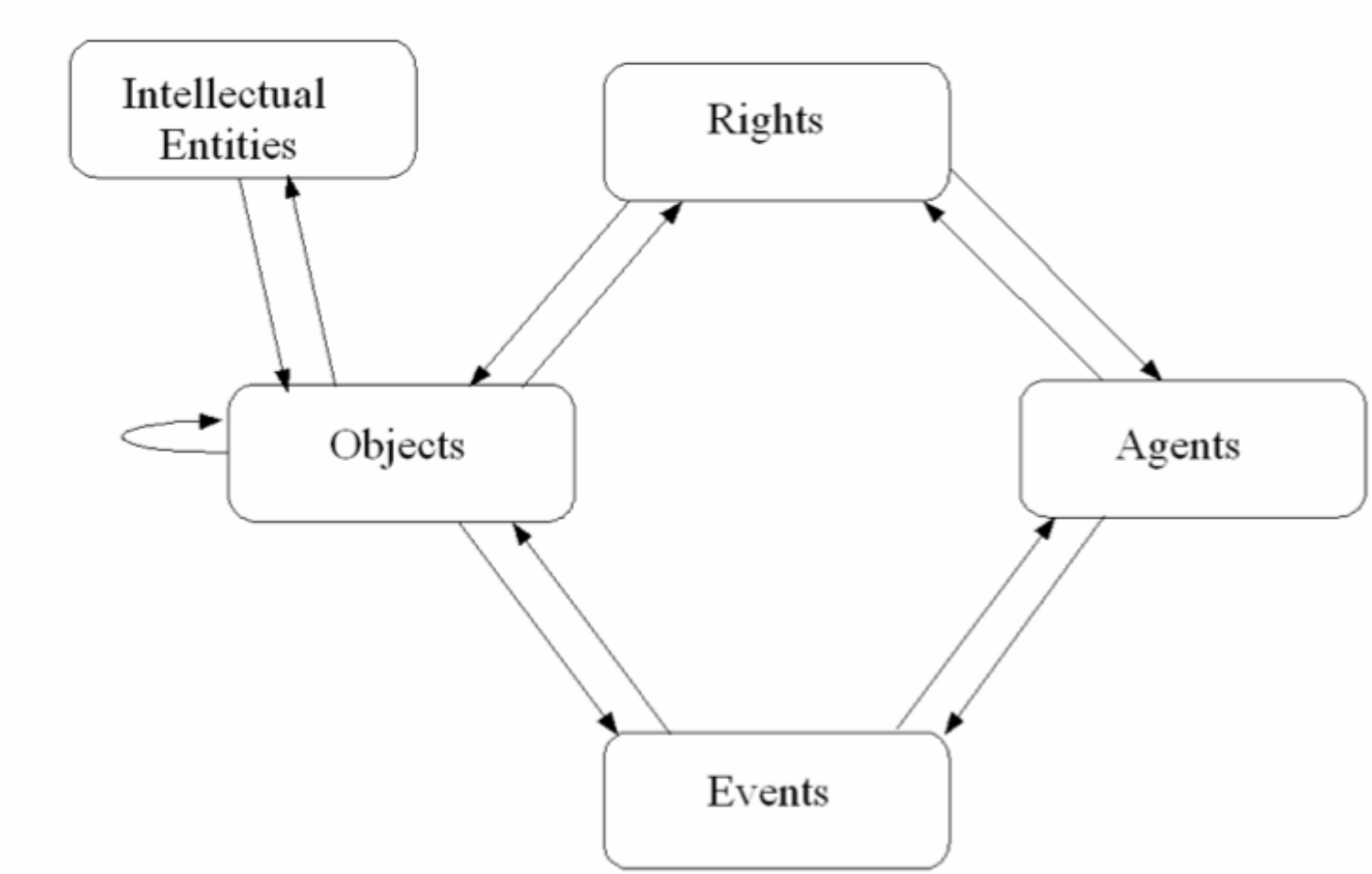
2009. Library of Congress Network Development and MARC Standards Office.
http://www.loc.gov/standards/premis/understanding-premis.pdf.

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.
http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

Example: a (super-simplified) PREMIS XML document

```
<premis version = “2.0”>
<object xsi:type=“file”>
    …
</object>
<event>
    …
</event>
<agent>
…
</agent>
</premis>
```

Slightly expanded view

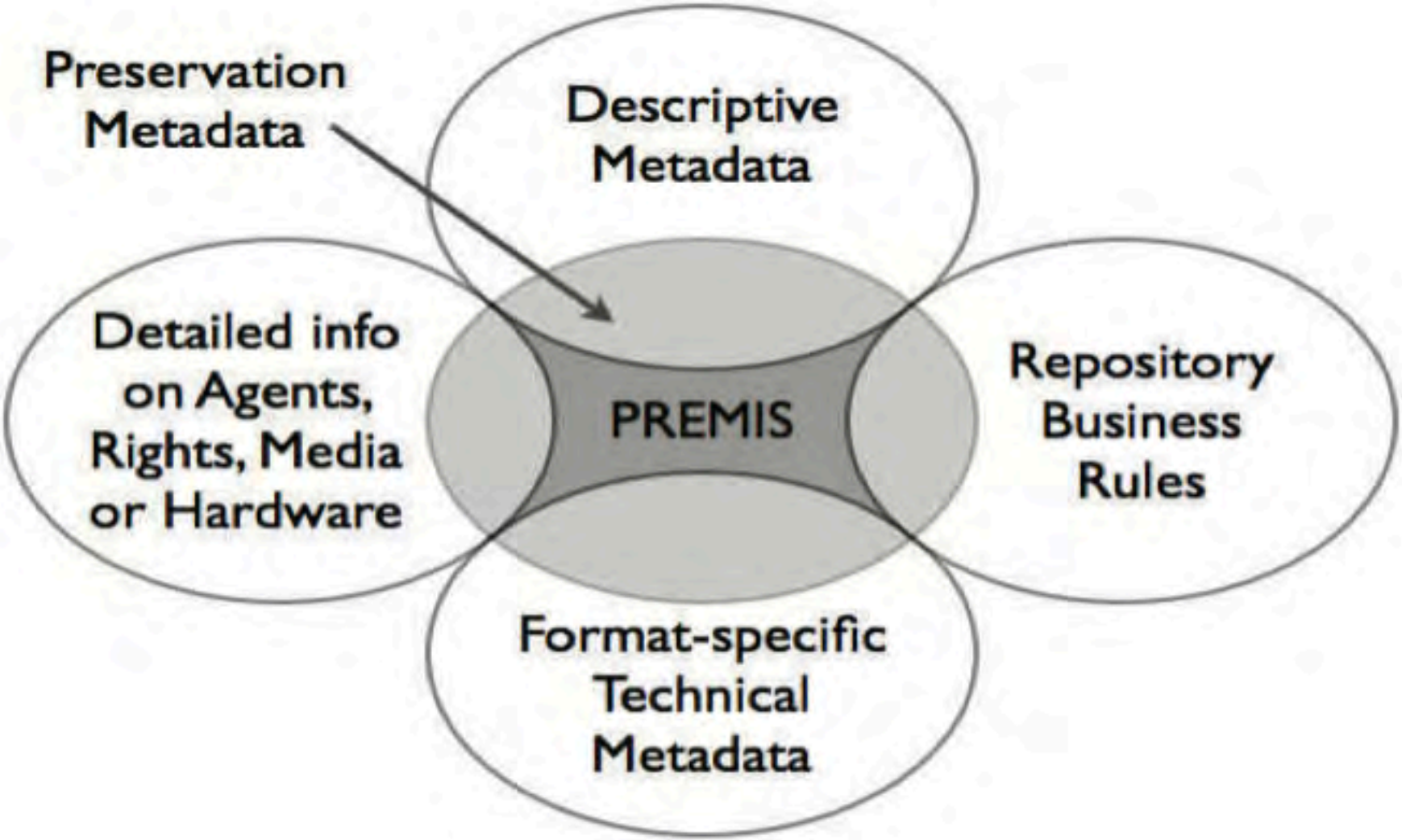
```
<premis version = “2.0”>
<object xsi:type=“file”>
        <objectIdentifier> … </objectIdentifier>
        <objectCharacteristics>
                <fixity>
                <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
                <messageDigest>7b7c655e2e25867e1ed1062f7102e5ef</messageDigest>
                <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/

What's NOT defined in PREMIS?

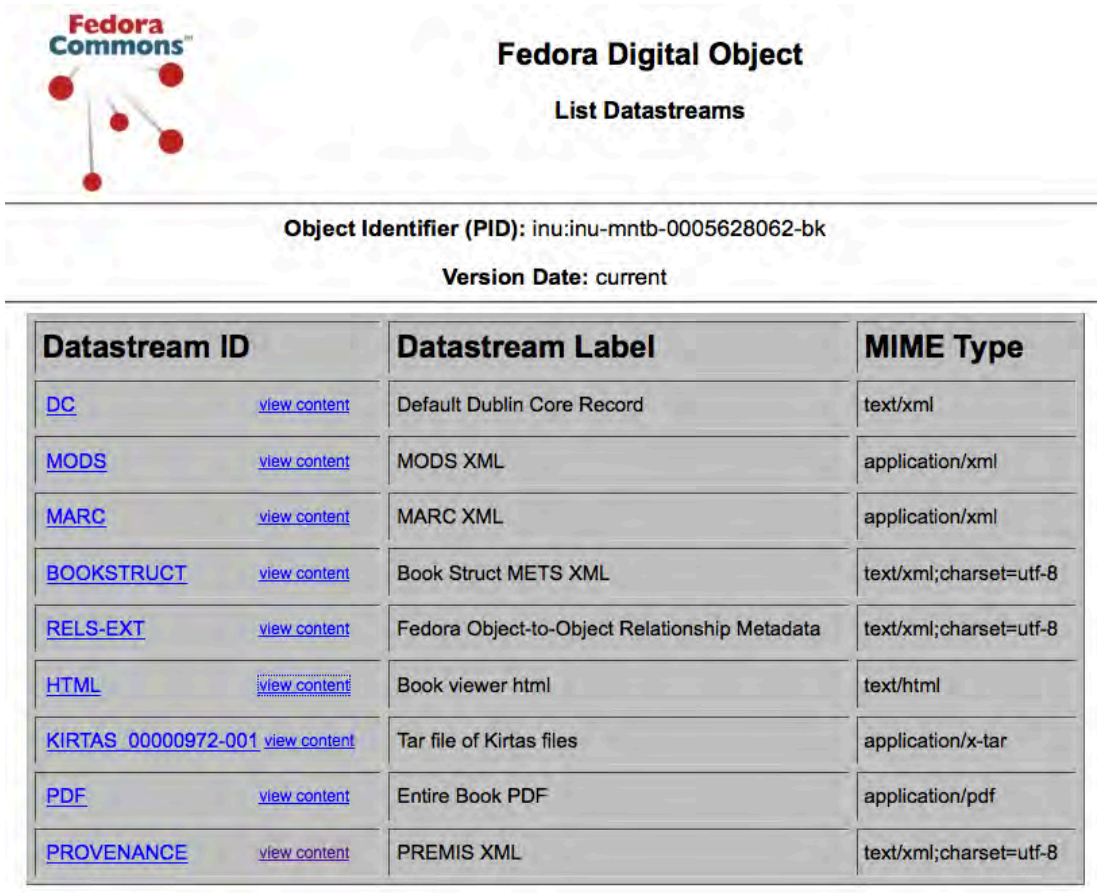
- Format-specific metadata
- Implementation-specific
- Descriptive metadata
- Details about media, hardware
- Detailed agent information
- Details about rights, permissions not affecting preservation functions

from *Understanding PREMIS*



from *Understanding PREMIS*

Case study: PREMIS for Northwestern Books



Fedora Commons

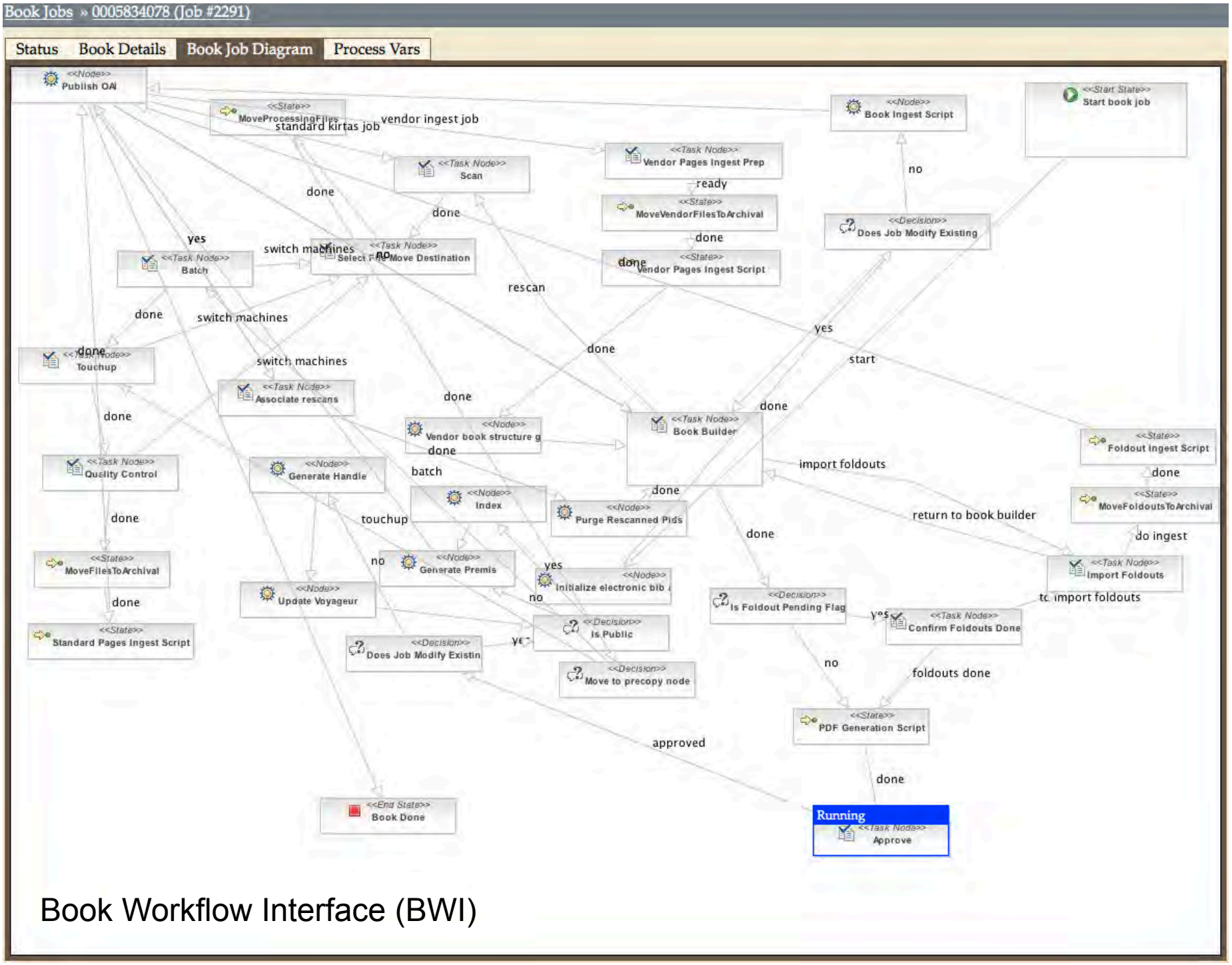
Fedora Digital Object

List Datastreams

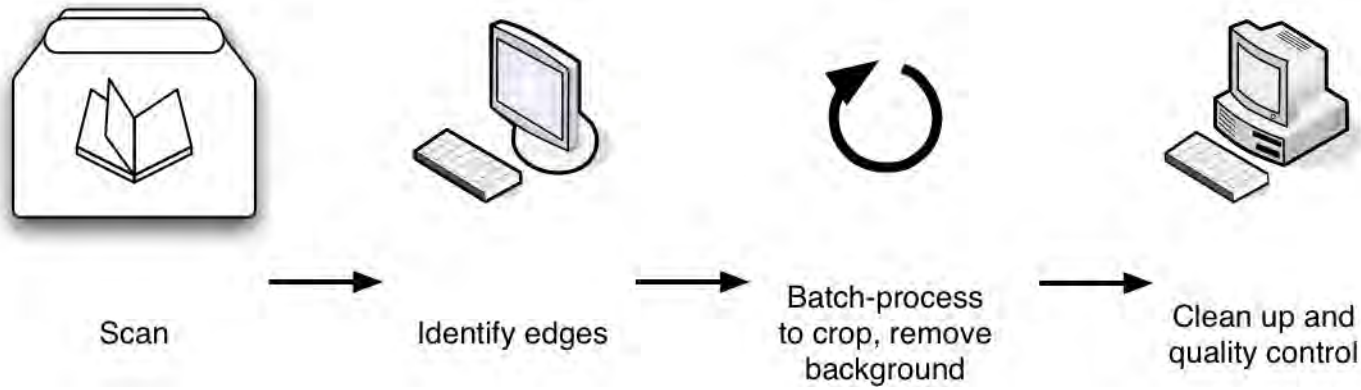
Object Identifier (PID): inu:inu-mntb-0005628062-bk

Version Date: current

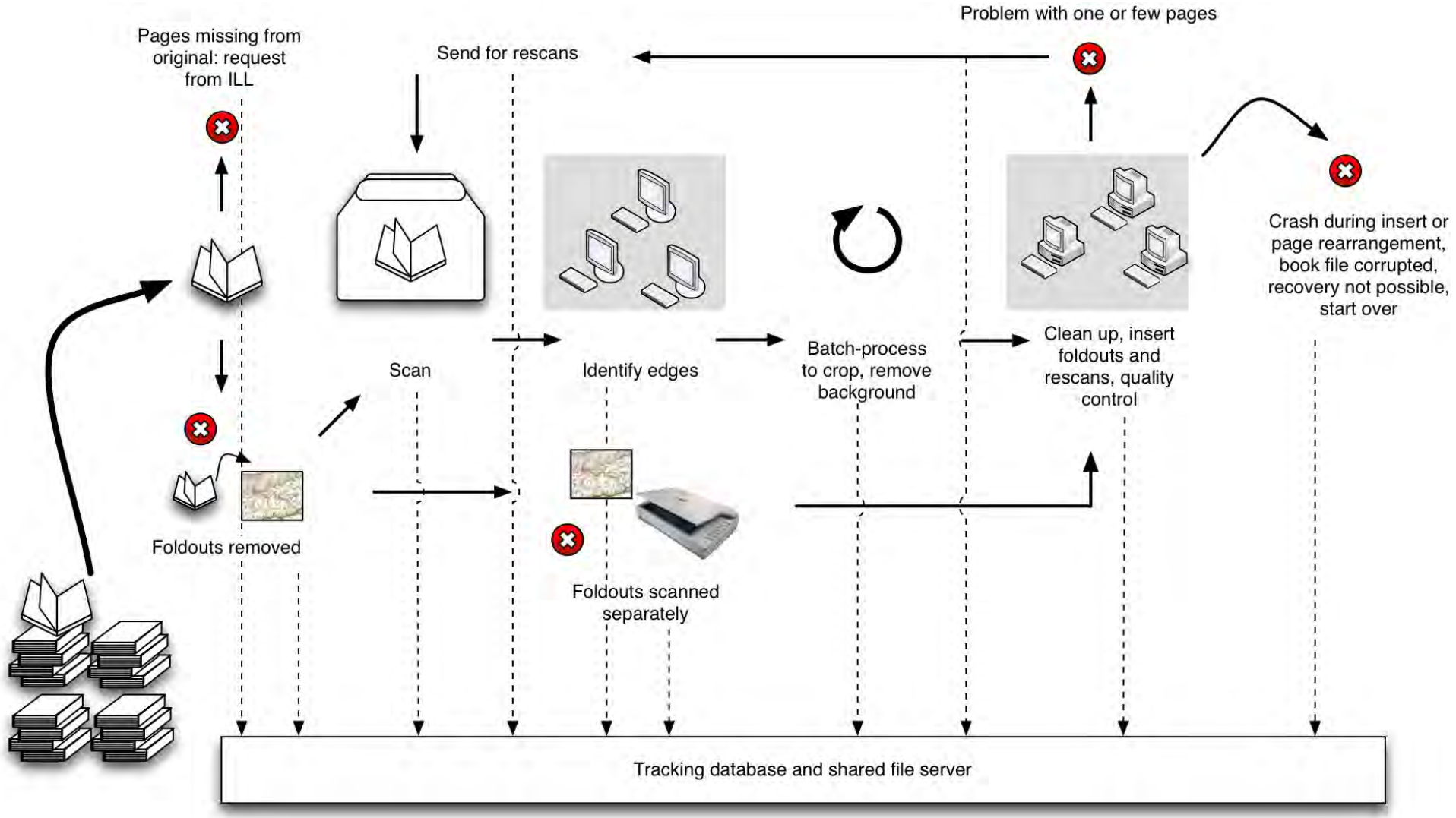
Datastream ID		Datastream Label	MIME Type
DC	view content	Default Dublin Core Record	text/xml
MODS	view content	MODS XML	application/xml
MARC	view content	MARC XML	application/xml
BOOKSTRUCT	view content	Book Struct METS XML	text/xml;charset=utf-8
RELS-EXT	view content	Fedora Object-to-Object Relationship Metadata	text/xml;charset=utf-8
HTML	view content	Book viewer html	text/html
KIRTAS_00000972-001	view content	Tar file of Kirtas files	application/x-tar
PDF	view content	Entire Book PDF	application/pdf
PROVENANCE	view content	PREMIS XML	text/xml;charset=utf-8

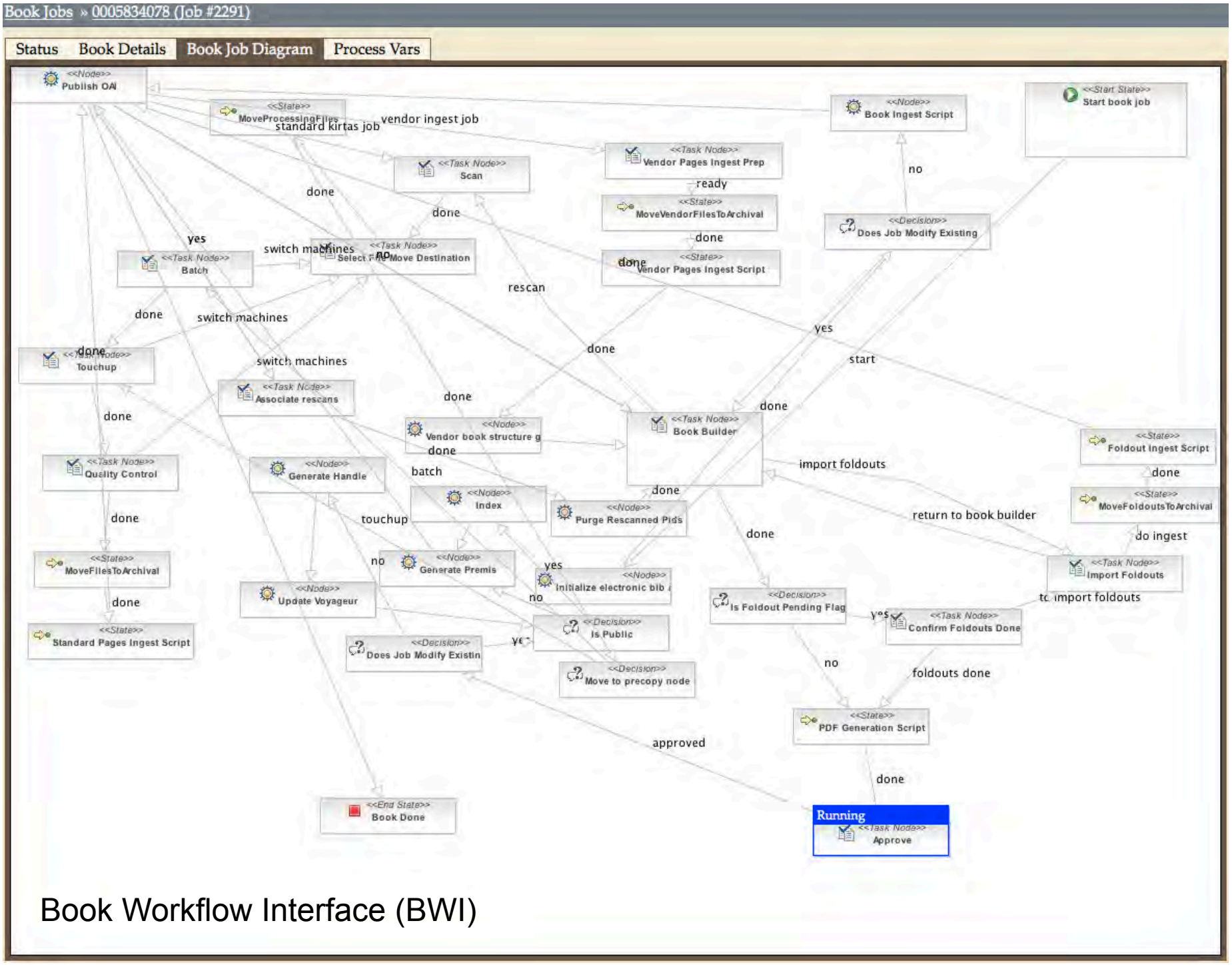


Book Workflow Interface (BWI)


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

Pre-BWI: each step with human intervention





Book Workflow Interface (BWI)

Fedora Digital Object

List Datastreams

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

Version Date: current

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

Fedora Digital Object

List Datastreams

Object Identifier (PID): inu:inu-mntb-0005628062-bk

Version Date: current

Datastream ID		Datastream Label	MIME Type
DC	view content	Default Dublin Core Record	text/xml
MODS	view content	MODS XML	application/xml
MARC	view content	MARC XML	application/xml
BOOKSTRUCT	view content	Book Struct METS XML	text/xml;charset=utf-8
RELS-EXT	view content	Fedora Object-to-Object Relationship Metadata	text/xml;charset=utf-8
HTML	view content	Book viewer html	text/html
KIRTAS_00000972-001	view content	Tar file of Kirtas files	application/x-tar
PDF	view content	Entire Book PDF	application/pdf
PROVENANCE	view content	PREMIS XML	text/xml;charset=utf-8

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

```
<premis xmlns="http://info:lc/xmlns/premis-v2" xmlns:xsi="http://
www.w3.org/2001/XMLSchema-instance" version="2.0"
xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/
standards/premis/premis.xsd">
  <object xsi:type="representation">
    <objectIdentifier>
      <objectIdentifierType>hdl:2166</objectIdentifierType>
      <objectIdentifierValue>inu:inu-mntb-0005834078-bk</
objectIdentifierValue>
    </objectIdentifier>
    <preservationLevel>
      <preservationLevelValue>Full</preservationLevelValue>
    </preservationLevel>
  </object>
```

FINAL (2 of 3): event

```
<event>
  <eventIdentifier>
    <eventIdentifierType>hdl:2166</eventIdentifierType>
    <eventIdentifierValue>inu-event-0005834078-6946be40-ace8-4fd9-bffb-1b6edd319bfa</eventIdentifierValue>
  </eventIdentifier>
  <eventType>Approve Book</eventType>
  <eventDateTime>2010-12-06T13:00:41</eventDateTime>
  <eventDetail>Book Approved</eventDetail>
  <eventOutcomeInformation>
    <eventOutcome>0</eventOutcome>
  </eventOutcomeInformation>
  <linkingAgentIdentifier>
    <linkingAgentIdentifierType>hdl:2166</linkingAgentIdentifierType>
    <linkingAgentIdentifierValue>inu-agent-software-mountingbooks</linkingAgentIdentifierValue>
    <linkingAgentRole>Executing program</linkingAgentRole>
  </linkingAgentIdentifier>
  <linkingObjectIdentifier>
    <linkingObjectIdentifierType>hdl:2166</linkingObjectIdentifierType>
    <linkingObjectIdentifierValue>inu:inu-mntb-0005834078-bk</linkingObjectIdentifierValue>
    <linkingObjectRole>outcome</linkingObjectRole>
  </linkingObjectIdentifier>
 </event>
```

FINAL (3 of 3): rights

```
<rights>
  <rightsStatement>
   <rightsStatementIdentifier>
    <rightsStatementIdentifierType>hdl:2166</rightsStatementIdentifierType>
    <linkingObjectIdentifierValue>inu-rights-0005834078-9ea15dd9-5ece-48b1-aec9-f0ddb6161552</linkingObjectIdentifierValue>
   </rightsStatementIdentifier>
   <rightsBasis>Copyright</rightsBasis>
   <copyrightInformation>
    <copyrightStatus>pd,cdpp</copyrightStatus>
    <copyrightJurisdiction>us</copyrightJurisdiction>
    <copyrightStatusDeterminationDate>2010-12-06T12:54:27</copyrightStatusDeterminationDate>
    <copyrightNote/>
   </copyrightInformation>
   <licenseInformation>
    <licenseTerms>null,null</licenseTerms>
    <licenseNote/>
   </licenseInformation>
   <rightsGranted>
    <act>world</act>
    <termOfGrant>
     <startDate>2010-12-06T12:54:27</startDate>
    </termOfGrant>
    <rightsGrantedNote/>
   </rightsGranted>
  </rightsStatement>
 </rights>
</premis>
```

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

Processing Record
"master" for each processing pass

```
<ProcessingRecord objID="ark:/27927/pfgz7c2vg" created="2008-10-16T13:46:45.785-04:00">
  <Timestamp>2008-07-07T17:05:06.707-04:00</Timestamp>
  <Rationale>Processing of New Content</Rationale>
  <System name="ConPrep" version="1.1.4"/>
  <Workflow name="E-Journal_CreateMD_CopyMD_OSMD Workflow" version="1.0"/>
  <Batch objID="ark:/27927/pfgz7c2vj"/>
  <Profile name="APS Digitized Profile" version="N/A"/>
  <RuntimeEnv OS="SunOS:sparc:5.10" VM="Java:Sun Microsystems Inc.:1.6.0_06"/>
</ProcessingRecord>
```

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

```
<EventSet objID="ark:/27927/tcdgf45g" created="2008-06-18T10:13:57.384-04:00"
      processingRecordID="ark:/27927/tcdf34g">
     <Event objID="ark:/27927/tcdgf45c" eventType="Sample Event Type">
       <Timestamp>2008-07-07T17:11:40.581-04:00</Timestamp>
       <Rationale>[why goes here]</Rationale>
       <RationaleDetail>[additional details of why goes here]</RationaleDetail>
       <Agent>[user id here]</Agent>
       <Input objID="ark:/27927/tcdgf45c"/>
       <Output objID="ark:/27927/tcdgf45c"/>
       <ArgList>
         <Arg name="Format" value="Elsevier DTD 4.1.x"/>
         <Arg name="Mime Type" value="application/xml"/>
       </ArgList>
       <ToolRegistry version="1.1.1" date="2008-01-01"/>
       <RuntimeEnv OS="SunOS:sparc:5.10" VM="Java:Sun Microsystems Inc:1.6.0_06"
         other="Other runtime info here"/>
       <ToolWrapper name="Sgml_TransformTool:1.0:2007-03-28"/>
       <ToolComponentList>
         <ToolComponent name="nrccrp2portico.xsl"/>
         <ToolComponent name="insert-portico-doctype.xsl"/>
       </ToolComponentList>
       <Outcome>Success</Outcome>
       <OutcomeDetailList>
         <OutcomeDetail name="Format" value="Elsevier DTD 4.1.x"/>
         <OutcomeDetail name="Mime Type" value="application/xml"/>
         <OutcomeDetail name="Format Assessment" value="Well Formed and Valid"/>
       </OutcomeDetailList>
     </Event>
</EventSet>
```

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

Portico 2.0 Event Model	PREMIS Event Entity
Unique ID	eventIdentifier
Timestamp	eventDateTime
Type of Event	eventType
Rationale for the Event	eventDetail
Agent	—
UserInfo	linkingAgentIdentifier; linkingAgentRole
Processing Record	*(not sure where to put this yet…)*
Process	—
Arguments	*(not sure where to put this yet…)*
Input objects	linkingObjectIdentifier; linkingObjectRole
Output objects	linkingObjectIdentifier; linkingObjectRole
Tool info	*(not sure where to put this yet…)*
Outcome	—
Result	eventOutcome
Details	eventOutcomeDetailNote

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

HATHI TRUST

A Shared Digital Repository

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:linkingAgentIdentifier
      Value>
      <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:eventOutcomeDetailExtension>
       <HT:fileList status="failed">
        <HT:file>arcanacaelestiah03swed_files.xml</HT:file>
        <HT:file>arcanacaelestiah03swed_meta.xml</HT:file>
       </HT:fileList>
    ....
```

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:metsHdr>
<METS:dmdSec>…</METS:dmdSec>
<METS:amdSec>
    <METS:techMD>…</METS:techMD>
    <METS:digiprovMD>…
        <PREMIS:premis>…</PREMIS:premis>
    </METS:digiprovMD>
</METS:amdSec>
<METS:fileSec>…</METS:fileSec>
<METS:structMap>…</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 Close

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.

the

Discussion