HathiTrust PREMIS Implementation

Aaron Elkiss
Technical Lead, Repository Services, HathiTrust
iPRES 2018

HATHI TRUST Digital Library

HathiTrust Background

- Founded Oct 2008 with CIC (now BTAA) institutions & Univ. of California
- Now around 150 institutions
- Originally a preservation repository for Google-digitized books from the libraries
- Now 16.7 million volumes digitized by over 50 different organizations
- Focus is still on book and book-like material

HATHI TRUST Digital Library

HathiTrust Services

- Access as well as preservation
  - Full-text access for public domain creative-commons licensed material
  - Search across entire corpus
  - Search within user-created collections
- Collaborative programs around repository
  - Computational analysis (HathiTrust Research Center)
  - Copyright Review
  - Federal Documents
  - Shared Print

HATHI TRUST Digital Library

HathiTrust PREMIS

- Evolving needs and priorities over 10+ years
- Focus on scalability, uniformity, controllability
- Hard to know what to record / anticipate future needs
- PREMIS v1 -> PREMIS v2 (2012)
- Not many changes since ~2014
- Many decisions are probably not the ones we'd make in 2018...

HATHI TRUST Digital Library

Use cases

- Can inspect individual object METS files
- No way to query/collect information across repository except by crawling it.
- Very little is exposed in the interface
- Most common use: determine ingest history
- Identify scopes of problems
  - bug in particular versions of software
  - date range particular actions happened

HATHI TRUST Digital Library

PREMIS Storage

- PREMIS in METS (XML)
- Temporary storage for events (during ingest) in relational database
- Descriptive metadata (MARC) stored & managed separately
- Updated when item is (re-)ingested
- Or during mass migration
- Implications
  - Events are stored per-object in the repository
  - Only successful (or at least non-fatal) events are preserved
  - Audit events aren't preserved

HATHI TRUST Digital Library

PREMIS object

- Bare-bones: most of the object-type information is in the METS

```
<PREMIS:object xsi:type="PREMIS:representation">
  <PREMIS:objectIdentifier>
    <PREMIS:objectIdentifierType>HathiTrust</PREMIS:objectIdentifierType>
    <PREMIS:objectIdentifierValue>mdp.39015000375116</PREMIS:objectIdentifierValue>
  </PREMIS:objectIdentifier>
  <PREMIS:significantProperties>
    <PREMIS:significantPropertiesType>file count</PREMIS:significantPropertiesType>
    <PREMIS:significantPropertiesValue>360</PREMIS:significantPropertiesValue>
  </PREMIS:significantProperties>
  <PREMIS:significantProperties>
    <PREMIS:significantPropertiesType>page count</PREMIS:significantPropertiesType>
    <PREMIS:significantPropertiesValue>120</PREMIS:significantPropertiesValue>
  </PREMIS:significantProperties>
</PREMIS:object>
```

HATHI TRUST Digital Library

PREMIS Event

```
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>b9a0995c-9d2d-3758-a0cc-dbb3a51343ca</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>validation</PREMIS:eventType>
  <PREMIS:eventDateTime>2016-10-10T04:15:46Z</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Validation of technical characteristics of image and OCR files</PREMIS:eventDetail>
  <PREMIS:eventOutcomeInformation>
    <PREMIS:eventOutcome>pass</PREMIS:eventOutcome>
  </PREMIS:eventOutcomeInformation>
  ...
</PREMIS:event>
```

HATHI TRUST Digital Library

PREMIS Event: Linking agents

```
<PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>HathiTrust Institution ID</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>umich</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>feedd.pl v1.3.65</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>JHOVE 1.11</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>Xerces-C 3.1</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
```

PREMIS event types

- Some come from early controlled vocabularies
- Lots of discussion every time we needed to create a new one
- Some combine action and object - better approach would be to include linkingObjectIdentifier.
- Structured PREMIS 3 eventDetails? (normalization w/ details)

- capture
- creation
- decryption
- deletion
- file rename
- fixity check
- image compression
- image header modification
- ingestion
- manual inspection
- message digest calculation
- mets migration
- ocr split
- package inspection
- page feature mapping
- premis migration
- source mets creation
- validation
- zip archive creation
- zip file message digest calculation

HATHI TRUST Digital Library

Deletion (tombstone) event

```
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>2F3CF520-EBC1-4519-80FF-D2B0AB5B1744</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>deletion</PREMIS:eventType>
  <PREMIS:eventDateTime>2014-01-17T21:02:19Z</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Deletion of content data object from repository</PREMIS:eventDetail>
  <PREMIS:eventOutcomeInformation>
    <PREMIS:eventOutcomeDetail>
      <PREMIS:eventOutcomeDetailNote>Missing pages; another scan at mdp.39015035431017</PREMIS:eventOutcomeDetailNote>
      <PREMIS:eventOutcomeDetailExtension>
        <HT:deleteReason xmlns:HT="http://www.hathitrust.org/premis_extension">quality</HT:deleteReason>
      </PREMIS:eventOutcomeDetailExtension>
    </PREMIS:eventOutcomeDetail>
  </PREMIS:eventOutcomeInformation>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>MARC21 Code</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>MiU</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
```

Event with eventOutcomeDetail

```
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>2d22b4b4-333e-3c5d-9f38-4d2d5e29b442</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>image header modification</PREMIS:eventType>
  <PREMIS:eventDateTime>2014-02-07T17:44:27Z</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Modification of image headers to meet HathiTrust conventions</PREMIS:eventDetail>
  <PREMIS:eventOutcomeInformation>
    <PREMIS:eventOutcome>warning</PREMIS:eventOutcome>
    <PREMIS:eventOutcomeDetail>
      <PREMIS:eventOutcomeDetailNote>Image creation date metadata may not be accurate: estimated from file
timestamp</PREMIS:eventOutcomeDetailNote>
      <PREMIS:eventOutcomeDetailExtension>
        <HT:fileList status="estimated capture date">
          <HT:file>00000652.jp2</HT:file>
          <HT:file>00001058.jp2</HT:file>
        </HT:fileList>
      </PREMIS:eventOutcomeDetailExtension>
    </PREMIS:eventOutcomeDetail>
  </PREMIS:eventOutcomeInformation>
  ...
<PREMIS:event>
```

Event with eventOutcomeDetail

```
<PREMIS:event>
  ...
  <PREMIS:eventOutcomeInformation>
    <PREMIS:eventOutcome>warning</PREMIS:eventOutcome>
    <PREMIS:eventOutcomeDetail>
      <PREMIS:eventOutcomeDetailNote>Original scanning artist unknown; digitization was performed under the direction
of the recorded artist.</PREMIS:eventOutcomeDetailNote>
      <PREMIS:eventOutcomeDetailExtension>
        <HT:fileList status="default artist">
          <HT:file>00000652.jp2</HT:file>
          <HT:file>00001058.jp2</HT:file>
        </HT:fileList>
      </PREMIS:eventOutcomeDetailExtension>
    </PREMIS:eventOutcomeDetail>
  </PREMIS:eventOutcomeInformation>
<PREMIS:event>
```

HATHI TRUST Digital Library

Event driving ingest decisions

```
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>f348f01d-01fa-3bd6-9720-d1cd2c18c4fc</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>manual inspection</PREMIS:eventType>
  <PREMIS:eventDateTime>2014-10-21T15:34:35Z</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Manually inspect item for completeness and legibility</PREMIS:eventDetail>
  <PREMIS:eventOutcomeInformation>
    <PREMIS:eventOutcome>validation exception granted</PREMIS:eventOutcome>
    <PREMIS:eventOutcomeDetail>
      <PREMIS:eventOutcomeDetailExtension>
        <HT:exceptionsAllowed category="jpeg2000_size"/>
      </PREMIS:eventOutcomeDetailExtension>
    </PREMIS:eventOutcomeDetail>
  </PREMIS:eventOutcomeInformation>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>Person</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>aelkiss@umich.edu</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Inspector</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
```

Migrations

- Make event and agent identifiers & types more consistent
- Duplicate event detection
  - UUID generated based on object ID, event type, date/time
- Migration uses the same METS / PREMIS generation as re-ingest
  - Unpack item from repository
  - Generate new METS+PREMIS based on unpacked item & existing METS
  - Compare to old METS+PREMIS

HATHI TRUST Digital Library

Opportunities

- Drive activity based on identified user & partner requirements
- Broader needs around item-level metadata
- Use linkingObjectIdentifier and/or structured eventDetails to reduce the number of distinct eventTypes
- agentVersion
- PREMIS Rights?

HATHI TRUST Digital Library

Additional reading

https://www.hathitrust.org/digital_object_specifications

HATHI TRUST Digital Library