PREMIS in Archivematica

PETER VAN GARDEREN
Artefactual Systems Inc.

American Library Association
New Orleans - June 24, 2011
open-source software for archives and libraries
digital preservation consulting services
http://artefactual.com
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 allow users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model.

Users monitor and control the microservices via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core and other best practice metadata standards. Archivematica implements media type preservation plans based on an analysis of the significant characteristics of the formats.

The overview section provides a detailed description of Archivematica’s functionality and technical architecture. This screencast gives a demo of the Archivematica 0.7-alpha release.

Free and open source

Archivematica is free and open source software. The software applications integrated into Archivematica are each released under their own open source license. These are checked for license compatibility before they are integrated into the project. A full list of applications with their respective license is available on the external software tools page.
Who is using Archivematica?

- Artefactual Systems clients
  - City of Vancouver Archives
  - International Monetary Fund Archives
  - Rockefeller Archives Center
  - University of British Columbia Library
  - Simon Fraser University Archives
Who is using Archivematica?

- Archivematica community
  - 20-30 pilot testers
  - Documentation contributions
  - Media preservation plan testing
  - Hydra fork: ‘Rubymatica’
  - Education: University of Toronto, University College London, SAA
- 10+ workshops in past 12 months
monitored and controlled in the archivematica dashboard.

The dashboard displays detailed information about the Archivematica project, including metadata, file information, and status updates. Each task is represented as a row in the table, with columns for the task description, status, and start time.

The dashboard integrates with a variety of clients and services, including web servers, microservices, and file shares. The integration enables seamless monitoring and control of the Archivematica project. The diagram also illustrates the workflow, showing how tasks are initiated, processed, and executed, with options for success and error handling.

Success or Error directory is a watched directory for each micro-service. Chain directories into multiple, custom workflows. Defined in simple XML configuration file for each Watched directory.
<table>
<thead>
<tr>
<th>Submission Information Package</th>
<th>UUID</th>
<th>Ingest start time</th>
</tr>
</thead>
<tbody>
<tr>
<td>🟢 Special projects</td>
<td>db720a77-657f-4272-8121-f3311bc7a48d</td>
<td>2011-06-16 14:23</td>
</tr>
<tr>
<td>Micro-Service: Appraise SIP for submission</td>
<td></td>
<td>Requires approval</td>
</tr>
<tr>
<td>Micro-Service: Set file permissions</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>Micro-Service: Create Dublin Core template</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>Micro-Service: Remove thumbs.db files</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>Micro-Service: Verify metadata directory checksums</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>Micro-Service: Assign file UUIDs and checksums</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>Micro-Service: Verify SIP compliance</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>Micro-Service: Create SIP backup</td>
<td></td>
<td>Completed successfully</td>
</tr>
<tr>
<td>✗ Planning Projects-Jan-Jun-94</td>
<td>11c3ebd8-b2d3-49b7-a393-9c7bd3009cc7</td>
<td>2011-06-16 13:50</td>
</tr>
<tr>
<td>🟢 Zoning hearings</td>
<td>cf7aa7a2-b0be-413f-8906-66c7259abb85</td>
<td>2011-06-16 13:49</td>
</tr>
<tr>
<td>Micro-Services</td>
<td>FOSS Tools</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------</td>
<td></td>
</tr>
<tr>
<td>receiveSIP</td>
<td>UUID</td>
<td></td>
</tr>
<tr>
<td>verifyChecksums</td>
<td>EasyExtract</td>
<td></td>
</tr>
<tr>
<td>generateChecksums</td>
<td>Detox</td>
<td></td>
</tr>
<tr>
<td>reviewSIPcompliance</td>
<td>MD5</td>
<td></td>
</tr>
<tr>
<td>quarantineFiles</td>
<td>ClamAV</td>
<td></td>
</tr>
<tr>
<td>extractPackagedFiles</td>
<td>FITS</td>
<td></td>
</tr>
<tr>
<td>assignUniqueIdentifiers</td>
<td>Jhove</td>
<td></td>
</tr>
<tr>
<td>removeProhibitedCharacters</td>
<td>Droid</td>
<td></td>
</tr>
<tr>
<td>virusScan</td>
<td>FFIdent</td>
<td></td>
</tr>
<tr>
<td>identifyFormat</td>
<td>File</td>
<td></td>
</tr>
<tr>
<td>validateFormat</td>
<td>Unocnv</td>
<td></td>
</tr>
<tr>
<td>identifyPreservationPlan</td>
<td>OpenOffice</td>
<td></td>
</tr>
<tr>
<td>transcodeFormat</td>
<td>Ghostscript</td>
<td></td>
</tr>
<tr>
<td>parseMetadata</td>
<td>FFmpeg</td>
<td></td>
</tr>
<tr>
<td>generatePREMIS</td>
<td>ImageMagick</td>
<td></td>
</tr>
<tr>
<td>generateMETS</td>
<td>DigiKam</td>
<td></td>
</tr>
<tr>
<td>createAIP</td>
<td>Inkscape</td>
<td></td>
</tr>
<tr>
<td>createDIP</td>
<td>FITS</td>
<td></td>
</tr>
<tr>
<td>storeAIP</td>
<td>Jhove</td>
<td></td>
</tr>
<tr>
<td>uploadDIP</td>
<td>Droid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bagit</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ICA-AtoM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incron</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Django</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Twisted</td>
<td></td>
</tr>
</tbody>
</table>

VIRTUAL MACHINE
VirtualBox, VMware, Xen
Open Archival Information System
2011 Development Priorities:
0.8 beta → production

- Interoperable AIP structure
- AIP Indexing (Lucene/ElasticSearch)
- SWORD REST APIs
- Persistent URIs (dns/uuid)
- Format registry integration (Open Planets Foundation)
- ContentDM, Dspace, XTF, TRIM & ICA-AtoM integration (Fedora?)
- Email + attachments preservation plan
- Transfer: SIP preparation (data visualization, keyword filtering)
Free Beer!
“They’ll never take our freedom”
The open-source eco-system
What is PREMIS good for?

The PREMIS data dictionary defines what a preservation repository needs to know.

The primary uses of PREMIS are for repository design, repository evaluation, and exchange of archived information packages among preservation repositories.

--Caplan, Understanding PREMIS (2009)
What is PREMIS good for?

Authenticity: establish identity and integrity

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

- Semantic unit values managed as SQL data while going through Archivematica ingest

- Output as XML into AIP upon ingest completion
PREMIS in Archivematica: AIP

- Bagit package (.zip, .tar optional):
  /data
  /logs
  /metadata
  /objects
  mets.xml

- Whole PREMIS record in METS digiprovMD:
  <mets><amdSec><digiprovMD><mdWrap>
  <XMLdata><premis><object><event><agent>
  <rights>
  </objectCharacteristicsExtension><FITS>
Objects

- Identifier
- Category
- Composition level
- Size
- Fixity
- Format
- Characteristics
- Relationships
<fileGrp ID="Land_image-4e6ad548-df36-4cf9-9cf5-31ca017cfec" 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"/>
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>
  + <event></event>
  </events>
</object>

+ <agents></agents>
</premis:premis>
<event>
  <eventIdentifier>
    <eventIdentifierType>UUID</eventIdentifierType>
    <eventIdentifierValue>a7ba91f-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>
</event>
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>
PREMIS in Archivematica: Next Steps

- Rights metadata: sync with security classification, FOIPA, accessioning, licensing vocabularies.
- Flexible structure for defining rights:
  <rightsExtension>
  <rightsGranted><act><restriction><termOfGrant>

- EAC for Agents?

- Indexing AIP metadata: use PREMIS entities as domain/document model
Attribution
Title: PREMIS in Archivematica: ALA 2011 New Orleans
Creator: Peter Van Garderen, Artefactual Systems Inc.
Date: June 25, 2011

The original content in this presentation is copyright Artefactual Systems Inc. 2011. You may freely re-use this content under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike 3.0 license.