The Archivematica project: Meeting digital continuity's technical challenges

Peter Van Garderen, Paul Jordan, Evelyn McLellan and Courtney C. Mumma

UNESCO Memory of the World in the Digital Age
Wednesday, September 26, 2012
Overview

- Archivematica history
- Requirements analysis
- System architecture
- Dashboard, PREMIS and METS
- IMF Archives pilot project
- Future development
- Developing nation context
http://artefactual.com

digital preservation consulting services

open-source software for archives and libraries

Evelyn McLellan, MAS
Director, Consulting Services

David Juhasz
Director, Technical Services

Courtney Mumma, MAS
Product Manager, Archivematica

Jessica Bushey, MAS
Product Manager, ICA-AtoM

Austin Trask
Systems Engineer

Jesús García Crespo
Software Engineer

Joseph Perry
Software Engineer

Mike Cantelon
Software Engineer

Peter Van Garderen, MAS
President / Systems Analyst
@pjvangarderen
Archivematica at the City of Vancouver Archives

Submitted by Sue Bigelow  Sue Bigelow's blog  Add new comment

The City of Vancouver Archives has been contributing to the development of the Archivematica digital preservation system for the past several years and we have just started using the 0.8 alpha release for production. This is an overview of why we got involved and where we are now.

Who we are

We’re a municipal archives, part of the City of Vancouver government. We follow the Canadian total archives concept of acquiring the records of both government and the private sector, and a wide variety of media. Our holdings include textual records, still and moving images, audio, maps and plans, and documentary art. We also have born-digital materials, the most significant being the records of the Vancouver Organizing Committee for the 2010 Olympic and Paralympic Winter Games (VANOC). We have the first Games records that are mostly digital, with more than 25TB acquired. We had been looking for a digital preservation system for many years, and in 2008, thanks to funding from the Olympic Legacy Reserve, we were able to take action.

What we were looking for
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

United Nations Educational,
Scientific and Cultural Organization

Paris, June 2007
Open Archival Information System

- ISO 14721
- High level reference model
- Default language of digital preservation world
- Key concepts:
  - Mandatory Responsibilities
  - Functional Entities
  - Information Packages
  - Actors
## UC-1.1 Receipt of SIP

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
</tr>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td><strong>Context</strong></td>
</tr>
<tr>
<td><strong>Sub-Use Cases</strong></td>
</tr>
</tbody>
</table>

### Overview
- **Producer** or **Administration > Archival Information Update** transfers **Submission Information Package (SIP)** to **Receive Submission**. **Receive Submission** provides the appropriate storage capability or devices to receive a SIP.

### Actors
- **Producer**
- **Receive Submission**
- **Administration > Archival Information Update**

### Description

#### Preconditions
- Establish a **Submission Agreement** with the **Producer**. Create a virtual Submission Agreement if necessary. (see OAIS, page 2-9, section 2.3.2)

#### Trigger
- A **Data Submission Session** is started by the **Producer**. OAIS or Management. (see OAIS, page 2-9, section 2.3.2)

#### Successful outcome
- **Receive Submission** sends SIP to **Quality Assurance**

#### Steps
1. **Producer** (or **Administration > Archival Information Update**) transfers SIP to **Receive Submission** (OAIS Page 4-5, Section 4.1.1.2)
2. **Receive Submission** provides the appropriate storage capability or devices to receive a SIP from **Producer** (OAIS Page 4-5, Section 4.1.1.2)
3. **Receive Submission** transmits SIP to **Quality Assurance** (OAIS Figure 4-2: Functions of Ingest Page 4-5)
<table>
<thead>
<tr>
<th>Submission Information Package</th>
<th>UUID</th>
<th>Ingest start time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pictures</td>
<td>4eac3ade-93ae-4608-9931-bcfe0df42d61</td>
<td>2012-08-15 13:57</td>
</tr>
</tbody>
</table>

- **Micro-service: Normalize**
  - Job: Normalize [?]
    - Awaiting decision
  - Job: Find options to normalize as
    - Completed successfully
  - Job: Move to workFlowDecisions-createDip directory
    - Completed successfully
  - Job: Grant normalization options for no pre-existing DIP
    - Completed successfully
  - Job: Check for Access directory
    - Completed successfully
  - Job: Check for Service directory
    - Completed successfully
- **Micro-service: Clean up names**
- **Micro-service: Remove cache files**
- **Micro-service: Include default SIP processingMCP.xml**
- **Micro-service: Rename SIP directory with SIP UUID**
- **Micro-service: Verify transfer compliance**
- **Micro-service: Verify SIP compliance**
digital files & metadata

Success or Error directory is Watched directory for next Micro-Service. Chain directories into multiple, custom workflows. Defined in simple XML.

or

SIP

Web Server

MCP Server

1..* Micro-Service Processing Clients

Fileshare

Digital Curation
Micro-Service

Watched directory

Python Scripts

FOSS tools

Success

Error

monitor and control

MANAGEMENT
Web Dashboard
<table>
<thead>
<tr>
<th>Tool</th>
<th>Version</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>BagIt®</td>
<td>4.1</td>
<td>Standard and script to package digital objects and metadata for archival storage</td>
<td>BSD License</td>
</tr>
<tr>
<td>Clam AV (anti-virus)®</td>
<td>0.97.5</td>
<td>Anti-virus toolkit for UNIX</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>digikam DNG Converter®</td>
<td>0.8.3</td>
<td>Converts RAW images to Digital Negative (DNG)</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>Digital Collection Builder®</td>
<td>1.0.9</td>
<td>Web-based digital collection description and access tool</td>
<td>GNU General Public License v2 (GPL2)</td>
</tr>
<tr>
<td>EXT4 filesystem®</td>
<td>12.04</td>
<td>Local storage, organization and batch processing of digital objects within the virtual appliance</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>FFmpeg®</td>
<td>2.0.3</td>
<td>Converts a wide variety of audio and video formats</td>
<td>GNU Lesser General Public License (LGPL)</td>
</tr>
<tr>
<td>File Information Tool Set (FITS)®</td>
<td>0.6.1</td>
<td>File format identification and validation software integration tool</td>
<td>GNU Lesser General Public License (LGPL)</td>
</tr>
<tr>
<td>freetype®</td>
<td>2.2.0</td>
<td>File lock utility used to queue objects through the workflow and prevent processing overload</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>Hypermail®</td>
<td>1.3</td>
<td>Web-based archival description and access tool</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>ImageMagick®</td>
<td>0.6.6</td>
<td>Converts a wide variety of bitmap images</td>
<td>GPL compatible Imagemagick licensed®</td>
</tr>
<tr>
<td>Inkscape®</td>
<td>0.48</td>
<td>Converts vector images to Scalable Vector Graphic (SVG) format</td>
<td>GNU General Public License (GPL) version 2.</td>
</tr>
<tr>
<td>LibreOffice®</td>
<td>3.5</td>
<td>Provides support for Normalization of documents.</td>
<td>LGPL3</td>
</tr>
<tr>
<td>MD5®</td>
<td>7.4</td>
<td>MD5 Checksum generation and verification scripts</td>
<td>“This program is a work of the US Government. In accordance with 17 USC 105, copyright protection is not available for any work of the US government. As such this code is considered public domain.”</td>
</tr>
<tr>
<td>NFS-common®</td>
<td>1.1.4</td>
<td>Network File System Access - allows access to files on network storage devices.</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>OpenOffice.org®</td>
<td>3.0.1</td>
<td>Office software suite; invoked by py-uno scripts to normalize office documents to ODF</td>
<td>GNU Lesser General Public License (LGPL)</td>
</tr>
<tr>
<td>PyODBC Converter Daemon®</td>
<td></td>
<td>Provides support for Normalization via OpenOffice.</td>
<td>Posted by Andrew Z. at Wednesday, February 27, 2008 LGPL®</td>
</tr>
<tr>
<td>Poster®</td>
<td>0.5</td>
<td>Provides support for both streaming POST requests as well as multipart/form-data encoding of string or file parameters</td>
<td>MIT License</td>
</tr>
<tr>
<td>Python®</td>
<td></td>
<td></td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>ReadPST/Outlook®</td>
<td>0.6.02</td>
<td>convert PST (MS Outlook Personal Folders) files to mbox and other formats</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>Samba/CIFS®</td>
<td>3.3.2</td>
<td>Linux to Windows interface; allows applications in the Ubuntu virtual appliance environment to use network storage made available via the Windows CIFS protocol</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>sha256®</td>
<td></td>
<td>Sha256 Checksum generation and verification scripts</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>TheX file manager®</td>
<td>1.2.3</td>
<td>Graphical file manager interface which end-user can use to move digital objects or trigger processing scripts.</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>UUID®</td>
<td>1.0.2</td>
<td>command line interface (CLI) for the generation of DCE 1.1, ISO/IEC 11578:1998 and IETF RFC 4122 compliant Universally Unique Identifier (UUID).</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>XUbuntu Linux®</td>
<td>12.04</td>
<td>Interface with computing hardware. Ubuntu Linux using Xfce desktop.</td>
<td>GNU General Public License (GPL)</td>
</tr>
<tr>
<td>Zip®</td>
<td>3.0</td>
<td>Utility called by BagIt to create AIP packages</td>
<td>Info-ZIP license: “Permission is granted to anyone to use this software for any purpose, including commercial applications, and to alter it and redistribute it freely”</td>
</tr>
<tr>
<td>Django®</td>
<td>1.3.1</td>
<td>Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design.</td>
<td>BSD License</td>
</tr>
<tr>
<td>Gearman®</td>
<td>0.27</td>
<td>Gearman provides a generic application framework to farm out work to other machines or processes that are better suited to do the work.</td>
<td>BSD License</td>
</tr>
<tr>
<td>p7zip®</td>
<td>9.2</td>
<td>7-Zip is a file archiver with a high compression ratio. (LZMA)</td>
<td>GNU General Public License (GPL)</td>
</tr>
</tbody>
</table>
### Transfer

<table>
<thead>
<tr>
<th>Micro-service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approve Transfer</td>
<td>This is the approval step that moves the transfer into the Archivematica processing pipeline.</td>
</tr>
<tr>
<td>Verify transfer compliance</td>
<td>Moves the transfer to a processing directory based on selected transfer type (standard, zipped bag, unzipped bag, DSEP, peer export or multi). Verifies that the transfer conforms to the folder structure required for processing in Archivematica and repositories if required. The structure is as follows: ( \backslash \text{metadata} \backslash \text{submissionDocumentation} \backslash \text{objects} ).</td>
</tr>
<tr>
<td>Rename with transfer UUID</td>
<td>Directly associates the transfer with its metadata by appending the transfer UUID to the transfer directory name.</td>
</tr>
<tr>
<td>Include default Transfer processing/MCP.xml</td>
<td>Adds a file named processing/MCP.xml to the root of the transfer. This is a configurable file to pre-configure processing decisions. It can configure workflow options such as creating transfer backups, quarantining the transfer and selecting a SIP creation option.</td>
</tr>
<tr>
<td>Assign file UUIDs and checksums</td>
<td>Assigns a unique universal identifier and sha-256 checksum to each file in the objects directory.</td>
</tr>
<tr>
<td>Verify transfer checksums</td>
<td>Checks any checksum files that were placed in the destination folder of the transfer prior to moving the transfer into Archivematica.</td>
</tr>
<tr>
<td>Generate METS.xml document</td>
<td>Generates a basic METS file with a basic structure to record the presence of all objects in the objects directory and their locations in any subdirectories. Designed to capture the original order of the transfer. The user chooses subsequently to delete, rename or move files or break the transfer into multiple SIPs. A copy of the METS file is automatically added to any SIP generated from the transfer.</td>
</tr>
<tr>
<td>Extract packages</td>
<td>Extracts objects from any zipped files or other packages. Extracts attachments from archive transfers.</td>
</tr>
<tr>
<td>Scan for viruses</td>
<td>Uses ClamAV to scan for viruses and other malware. If a virus is found, the transfer is automatically placed in ( \backslash \text{virusDirectory} \backslash \text{failed} ) and all processing on the transfer is stopped.</td>
</tr>
<tr>
<td>Clean up names</td>
<td>Some file systems do not support unicode or other special characters in filenames. This micro-service removes prohibited characters and replaces them with dashes. Original filenames are preserved in the PREMS metadata.</td>
</tr>
<tr>
<td>Characterize and extract metadata</td>
<td>Identifies and validates formats and extracts object metadata using the File Information Tool Set (FITS). Also identifies file extensions, which are used for selecting normalization paths.</td>
</tr>
<tr>
<td>Complete transfer</td>
<td>Indexes transfer contents, then marks the transfer as complete.</td>
</tr>
<tr>
<td>Create SIP from Transfer</td>
<td>This is the approval step that moves the transfer to the SIP packaging micro-services (Ingest).</td>
</tr>
</tbody>
</table>

### Ingest

<table>
<thead>
<tr>
<th>Micro-service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verify SIP compliance</td>
<td>Verifies that the SIP conforms to the folder structure required for processing in Archivematica. The structure is as follows: ( \backslash \text{metadata} \backslash \text{submissionDocumentation} \backslash \text{objects} ).</td>
</tr>
<tr>
<td>Rename SIP directory with SIP UUID</td>
<td>Directly associates the SIP with its metadata by appending the SIP UUID to the SIP directory name.</td>
</tr>
<tr>
<td>Include default SIP processing/MCP.xml</td>
<td>Copies the processing setup file added to the transfer in Include default Transfer processing/MCP.xml, above, to the SIP.</td>
</tr>
<tr>
<td>Remove cache files</td>
<td>Removes any Thumbs.db files.</td>
</tr>
<tr>
<td>Normalize</td>
<td>Determines which normalization options are available for the SIP and presents them to the user as choices. Normalizations (i.e., generates preservation and access copies) based on selection. Thumbmed files are also generated during this micro-service.</td>
</tr>
<tr>
<td>Process submission documentation</td>
<td>Processes any submission documentation included in the SIP and adds it to the objects directory.</td>
</tr>
<tr>
<td>Prepare DIP</td>
<td>Creates a DIP containing access copies of the objects, thumbnails and a copy of the METS file.</td>
</tr>
<tr>
<td>Upload DIP</td>
<td>Allows the user to choose to upload the DIP to either ICA-AtoM or CONTENTdm.</td>
</tr>
<tr>
<td>Upload DIP to ICA-AtoM</td>
<td>The user uploads the DIP to a selected destination in ICA-AtoM.</td>
</tr>
<tr>
<td>Upload DIP to CONTENTdm</td>
<td>The user uploads the DIP to a selected destination in CONTENTdm.</td>
</tr>
<tr>
<td>Prepare AIP</td>
<td>Creates an AIP in BagIt format, indexes the AIP, then losslessly compresses it.</td>
</tr>
<tr>
<td>Store AIP</td>
<td>Moves the AIP to ( \backslash \text{objectsDirectory} \backslash \text{metadata} \backslash \text{AIPs} ) or another specified directory. Once the AIP has been stored, a copy of it is extracted from storage to a local temp directory, where it is subjected to standard BagIt checks: verify.xml, check_bagit_contents, verify_complete, verify_bagit_transfers, verify_bagit_points.</td>
</tr>
</tbody>
</table>
Free software

1. The freedom to run the program for any purpose
2. The freedom to study how the program works, and adapt it to your own needs, meaning that easy access to the source code must be provided
3. The freedom to redistribute copies to help friends, family, colleagues or society in general
4. The freedom to improve the program, and release your own improvements to the public, so that the whole community benefits. Again, easy access to the source code is a precondition for this.
Free Beer!
“They’ll never take our freedom”
Community Support
We will try to answer fairly straight-forward questions from the open source community about installing and configuring our software. We are client contract service provider and we do what we can between client deadlines.

Due to our limited resources, community forum questions cannot become too specific, in-depth, or time-consuming for our support team. When we think a particular query is beyond these free support parameters we will inform the user that it may be necessary to address it as paid, commercial support.

Commercial Support
Our software is always free and open source, but with our hosting and support services, the Artefactual development team will assist a client with more in-depth questions to get the software installed and operating as required, whether on one of our servers or their own.
Open Source Software

- Code
- Knowledge
- Community

Users
- Lead institutions
- Funding
- Development
- All users
- Bug reports
- Enhancement requests
- Code patches
- Documentation
- Promotion

Foundation or Steering Committee
- Governance
- Coordination
- Funding
- Promotion

Service Providers
- Development
- Technical Support
- Hosting
- Training
- Promotion

Resources:
- Time
- Money
- Knowledge
<table>
<thead>
<tr>
<th>Transfer</th>
<th>UUID</th>
<th>Transfer start time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pictures</td>
<td>ba113e3a-f809-462d-8e4d-34f7fbb4f7649</td>
<td>2012-08-15 10:15</td>
</tr>
<tr>
<td>▶ Micro-service: Approve transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job: Approve transfer</td>
<td></td>
<td>Awaiting decision</td>
</tr>
<tr>
<td>- Good names</td>
<td>4bb41b9e-1c6b-4cb1-9e41-56fe0ba3dd64</td>
<td>2012-08-14 15:44</td>
</tr>
<tr>
<td>▶ Micro-service: Create SIP from Transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Complete transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Characterize and extract metadata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Clean up names</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Scan for viruses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Extract packages</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Generate METS.xml document</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Verify transfer checksums</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Assign file UUIDs and checksums</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Include default Transfer processingMCP.xml</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Rename with transfer UUID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Approve transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Bad names</td>
<td>747da9a9-46f6-4d77-bf98-caa7d5e29f8</td>
<td>2012-08-14 14:32</td>
</tr>
<tr>
<td>▶ Micro-service: Create SIP from Transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Micro-service: Complete transfer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Metadata
Tutorial Office Docs

Applies to
Tutorial Office Docs
In Archivematica 0.9 metadata can be added at the SIP level only

Title
Tutorial for 0.9 Office Docs Sample

Creator

Subject

Description

Publisher
## Browse archival storage

**Total size:** 107.77 MB  **Total files:** 27 indexed

<table>
<thead>
<tr>
<th>AIP</th>
<th>Size</th>
<th>UUID</th>
<th>Date stored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td>24.86 MB</td>
<td>0e3b54c7-5a5f-487d-9475-487a9fb3e5aa</td>
<td>2012-08-21 17:34</td>
</tr>
<tr>
<td>Multimedia</td>
<td>82.13 MB</td>
<td>1fa4989f-9bcd-42b0-a7af-1d2da852934e</td>
<td>2012-08-21 17:35</td>
</tr>
<tr>
<td>Tutorial Office Docs</td>
<td>0.78 MB</td>
<td>da3edfa3-762b-48f0-84aa-c61190c95d02</td>
<td>2012-08-21 17:26</td>
</tr>
</tbody>
</table>
The Archivematica Archival Information Package (AIP)

- The ingested digital objects
- Preservation versions of the objects
- Submission documentation
- Logs
- The METS file
The METS file

<dmdSec> (descriptive metadata)
Dublin Core XML

<amdSec> (administrative metadata)
<techMD>
  PREMIS: object
<digiProvMD>
  PREMIS: events
  PREMIS: agents
<rightsMD>
  PREMIS: rights

<fileSec> (a list of the files and their roles and relationships)

<structMap> (a representation of the physical structure of the AIP)
PREMIS events

- Ingestion
- Fixity check
- Message digest calculation
- Identifier assignment
- Unpacking

- Virus check
- Name cleanup
- Format identification
- Validation
- Normalization
Preservation planning

• A two-pronged approach:
  • Normalization on ingest
  • Preservation of the original file to support future strategies such as emulation
Format policies

• Normalization is based on format policies for various media types

• Criteria for selecting formats:
  • Non-proprietary
  • Freely available specifications
  • Widely used/endorsed by major repositories
  • No compression/lossless video compression
  • Tools available to write and render the format
<table>
<thead>
<tr>
<th>Media type</th>
<th>Extension</th>
<th>Normalization description</th>
<th>Command</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>ac3</td>
<td>Transcoding to mp3 with ffmpeg</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Audio</td>
<td>ac3</td>
<td>Transcoding to wav with ffmpeg</td>
<td>Show</td>
<td>preservation</td>
</tr>
<tr>
<td>Audio</td>
<td>aif</td>
<td>Transcoding to mp3 with ffmpeg</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Audio</td>
<td>aif</td>
<td>Transcoding to wav with ffmpeg</td>
<td>Show</td>
<td>preservation</td>
</tr>
<tr>
<td>Audio</td>
<td>aiff</td>
<td>Transcoding to mp3 with ffmpeg</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Audio</td>
<td>aiff</td>
<td>Transcoding to wav with ffmpeg</td>
<td>Show</td>
<td>preservation</td>
</tr>
<tr>
<td>Audio</td>
<td>mp3</td>
<td>Transcoding to mp3 with ffmpeg</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Audio</td>
<td>mp3</td>
<td>Transcoding to wav with ffmpeg</td>
<td>Show</td>
<td>preservation</td>
</tr>
<tr>
<td>Audio</td>
<td>wav</td>
<td>Transcoding to mp3 with ffmpeg</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Audio</td>
<td>wma</td>
<td>Transcoding to mp3 with ffmpeg</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Audio</td>
<td>wma</td>
<td>Transcoding to wav with ffmpeg</td>
<td>Show</td>
<td>preservation</td>
</tr>
<tr>
<td>Office Open XML</td>
<td>pptx</td>
<td>Transcoding to pdf with openOffice</td>
<td>Show</td>
<td>access</td>
</tr>
<tr>
<td>Portable Document Format</td>
<td>pdf</td>
<td>Transcoding to pdfa with Ghostscript</td>
<td>Show</td>
<td>preservation</td>
</tr>
<tr>
<td>Presentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A user's perspective
Paul Jordan: paul@larryjordan.biz

• International Monetary Fund pilot project
• Investigated and processed actual IMF collections
• Analyzed issues surrounding classification
Client partners
How we decide what's next

- In-house research
  - evolving best practices
  - new tools
- Client priorities
- User community input
  - Email list
  - Issues list
User email list
## Issues list

<table>
<thead>
<tr>
<th>ID</th>
<th>Milestone</th>
<th>Component</th>
<th>Type</th>
<th>Status</th>
<th>Priority</th>
<th>Owner</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1221</td>
<td>Release-1.0</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:joseph@artefactual.com">joseph@artefactual.com</a></td>
<td>Transfer names are not sanitized.</td>
</tr>
<tr>
<td>1220</td>
<td>Release-1.0</td>
<td>---</td>
<td>Enhancement</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:nicole@example.com">nicole@example.com</a></td>
<td>Add assign transfer backlog location instructions to Administrator manual on wiki</td>
</tr>
<tr>
<td>1219</td>
<td>Release-1.0</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>Medium</td>
<td><a href="mailto:nicole@example.com">nicole@example.com</a></td>
<td>Task reports with large number of files are slow to load.</td>
</tr>
<tr>
<td>1217</td>
<td>Release-1.0</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>Medium</td>
<td><a href="mailto:joseph@artefactual.com">joseph@artefactual.com</a></td>
<td>Normalization report - resource usage and time with large zip files.</td>
</tr>
<tr>
<td>1216</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>Critical</td>
<td><a href="mailto:austin@artefactual.com">austin@artefactual.com</a>@gtmpeascount.com</td>
<td>Add undecode as package</td>
</tr>
<tr>
<td>1215</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>New-Feature</td>
<td>QA-Reviewer</td>
<td>Medium</td>
<td><a href="mailto:nicole@example.com">nicole@example.com</a></td>
<td>Ability to rebuild ES index from AIP store.</td>
</tr>
<tr>
<td>1214</td>
<td>Release-1.0</td>
<td>---</td>
<td>New-Feature</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:courtney@example.com">courtney@example.com</a></td>
<td>Evaluate SKCore toolkit.</td>
</tr>
<tr>
<td>1213</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:joseph@artefactual.com">joseph@artefactual.com</a></td>
<td>Non ascii chars in transfer name cause error.</td>
</tr>
<tr>
<td>1212</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:david@example.com">david@example.com</a></td>
<td>DIP metadata not indexed in AtOM.</td>
</tr>
<tr>
<td>1211</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:nicole@example.com">nicole@example.com</a></td>
<td>DIP Upload shows failed in Dashboard when no intermediate level selected for AtOM</td>
</tr>
<tr>
<td>1210</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>Medium</td>
<td><a href="mailto:ophulien@example.com">ophulien@example.com</a></td>
<td>Failing email extraction - failing to extract all attachments - problems with attached email messages.</td>
</tr>
<tr>
<td>1209</td>
<td>Release-1.0</td>
<td>---</td>
<td>Enhancement</td>
<td>New</td>
<td>High</td>
<td><a href="mailto:nicole@example.com">nicole@example.com</a></td>
<td>Add setting in admin tab to set PREMIS agent Information.</td>
</tr>
<tr>
<td>1208</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Defect</td>
<td>Accepted</td>
<td>Medium</td>
<td><a href="mailto:nicole@example.com">nicole@example.com</a></td>
<td>Dashboard using 100% cpu after long period of inactivity.</td>
</tr>
<tr>
<td>1207</td>
<td>Release-0.9.1</td>
<td>---</td>
<td>Enhancement</td>
<td>Accepted</td>
<td>High</td>
<td><a href="mailto:austin@artefactual.com">austin@artefactual.com</a>@gtmpeascount.com</td>
<td>Evaluate performance of indexing Files. CurrentLocation.</td>
</tr>
</tbody>
</table>
Agile development model

• Release early, release often
  • Feb 2009: Release 0.1-alpha
  • June 2011: Release 0.7.1-alpha
  • January 2012: Release 0.8-alpha
  • August 2012: Release 0.9-beta
  • February 2013: Release 1.0

• Each iteration leads to updated and improved:
  • Requirements
  • Software
  • Documentation
  • Development resources
On the horizon

• Manage a backlog of transfers
• Create SIPs using visualization and analysis tools
• Basic arrangement and description
• Ingest email
• Format Policy Registry (FPR)
Development roadmaps

- On the wiki
- Usually available for the next two releases
- Link to requirements and other resources
- Guides agile development model
0.9 Transfer workflow

- Allow MCP access to media or storage where transfer is located
- Select preconfigured transfer type (generic, de-cap, frag, etc.)
- Enter transfer name
- Enter accession number
- Browse to transfer and select
- Browse to all submission documentation and select
- Start transfer
- Create structured transfer folder
- Assign transfer UUID
- Assign file UUIDs and checksums
- Log directory structure
- Quarantine
- Log directory structure
- Extract packages
- Assign file UUIDs and checksums to extracted files
- Verify checksums included in transfer
- Scan for viruses (generic tsp)
- Reject transfer or remove infected files and continue processing

Go to next page
0.9-beta release features

- Ubuntu 12.04 LTS as the base operating system
- Web browser dashboard interface
- DIPs can be uploaded to CONTENTdm
- All AIP metadata indexed and searched
- Rights module is updated to PREMIS 2.2
- Prototype ingest of maildir for email
0.9-beta cont'd

- User accounts can be created.
- Automatic restructuring of transfers for compliance
- Jobs are grouped into micro-services
- Ingest of Library of Congress Bagit format
- Nightly backup of MCP MySQL database
- Scalability enhancements
1.0 development roadmap
1.0 new features

- Format Policy Registry (FPR) & updates
- File identification triggers format policy actions
- Manual normalization workflow
- Improved email handling
- Format policy editing via dashboard
- Advanced AIP search
1.0 cont'd

- Generate DIPs from dashboard
- Visualization of transfers
- File-level DC and PREMIS MD entry
- BitCurator tools for SIP creation
- Custom reports
- Improved AIP delivery
1.0 cont'd

- Enterprise ID to connect AIPs and DIPs
- Enhanced CONTENTdm DIP upload
- DIP upload to Archivist's Toolkit
- DSpace ingest using OAI-PMH
- Edit processing workflows in the dashboard
- Improve multi-node processing and scalability
Digital preservation in developing world context

Technical infrastructure
Human capacity
Funding
Demand
Thank you!

Questions and discussion time.