Repository Interoperability and Preservation: The Hub and Spoke Framework

Robert Manaster, Tom Habing, William Ingram, Myung-Ja Han, and Patricia Hswe

University of Illinois at Urbana-Champaign
What Is Hub and Spoke?

- Repository Interoperability Architecture
- Process
- Preservation of Digital Objects
This Road to Preservation...
OAIS archive: An organization of people and systems that will preserve information and make it available for designated communities.
What we started with

- Sample data & digital repositories: DSpace, FEDORA, EPrints and Greenstone

What we continued with

- In context of preservation, developed checklist
- Moving sample data between digital repositories

What we ended with

- Beginnings of Hub and Spoke architecture
- No long-term preservation archive out there....
Limitations of Digital Repository

- Propriety storage
- No guaranteed viability or interoperability of software
- Little or no Intellectual Property Rights management
- Little or no Provenance
- Unsupported Digital Objects
- Not OAIS compliant
Repository interoperability is essential for the preservation of digital objects
What We Have...

- Plethora of repositories
- Overabundance of data sources
- Current integration solutions are local and ad hoc
- Not much preservation
• Preservation is taken into account
• Builds on existing infrastructure

• A common METS–based profile
• A standard programming API
• A series of scripts that use the API and METS profile for creating AIPS as well as SIPs & DIPs which can be used across different repositories
For N different repositories that need to interoperate, this model reduces the complexity from $N^N(N-1)$ to $2^N N$.

This simple idea is the rationale for many different standards that aim to promote interoperability.
Preserving the Representations of Intellectual Entities

- Representations are encapsulated in:
  - structMap
  - structLink
  - dmdSec
  - fileSec
  - behaviorSec

- Metadata about representations:
  - metsHdr
  - amdSec
  - Various attributes of the structMap, structLink, dmdSec, fileSec, and behaviorSec
• Non-prescriptive in regards to structure or file formats
• Intended to overlay other profiles which specify case-specific needs (i.e. web captures)
• PREMIS
• MODS
  – Must conform to the DLF Aquifer profile
• File-format specific technical metadata
  – MIX, VIDEOMD, AUDIOMD, others as appropriate
Master METS + Snapshots

Master METS
(includes provenance and technical metadata about each snapshot)

mptr

METS Snapshot 1

Mets Snapshot 2

Mets Snapshot 3

Files and Bitstreams:

File 1

File 2

File 3

File 4
Technical Metadata

Generation/Augmentation

• JHOVE Output + Custom XSLT
• Java “Applicators” for specific technical metadata schemas
  – MIX
  – TEXTMD
  – AUDIOMD
  – PREMIS
  – Class hierarchy to support new Applicators
METS Programming API

- Open Source
- Java
- XMLBeans
METS Profiles

- **Generic**
  - [http://www.loc.gov/standards/mets/profiles/00000015.xml](http://www.loc.gov/standards/mets/profiles/00000015.xml)

- **Web Capture**
  - [http://www.loc.gov/standards/mets/profiles/00000016.xml](http://www.loc.gov/standards/mets/profiles/00000016.xml)

- **Master METS**  – (not yet registered)
  - [http://dli.grainger.uiuc.edu/echodep/METS/DRAFTS/MasterMETSProfile.xml](http://dli.grainger.uiuc.edu/echodep/METS/DRAFTS/MasterMETSProfile.xml)
Technical Overview
Workflow

PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y.
LRCRUD
Service
Lightweight Repository
Create Retrieve Update Delete
REST
Representational State Transfer
<table>
<thead>
<tr>
<th>CRUD Action</th>
<th>HTTP Method</th>
<th>URL Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
<td>POST</td>
<td>/collection_id</td>
</tr>
<tr>
<td>Retrieve</td>
<td>GET</td>
<td>/item_id</td>
</tr>
<tr>
<td>Update</td>
<td>PUT</td>
<td>/item_id</td>
</tr>
<tr>
<td>Delete</td>
<td>DELETE</td>
<td>/item_id</td>
</tr>
</tbody>
</table>
LRCRUD Client communicates with the LRCRUD Service via HTTP methods, status codes, and headers.
GET /dspace-lrcrud/2135.89342 HTTP/1.1

1. LRCRUD Client
2. LRCRUD Service
3. DSpace native request for handle 2135.89342
4. Repository Server
Hub and Spoke Tool Suite

Workflow

GET: Repository package retrieved and sent to Hub Packager

LRCRUD Service for Repository X

Internet

LRCRUD Client

Retrieve Repository Package from Repository X via LRCRUD Service

To-Hub Packager

Create Hub Package METS Document from Repository Package

TechMD Augmenter

Generate technical metadata; augment Hub Package METS Document

Profile Validator

Validate Hub Package METS Document against METS Profile

From-Hub Packager

Create Repository Package from Hub package

LRCRUD Service

Send Repository Package to Repository Y via LRCRUD Service

PUT: Repository Package sent to Repository Y

PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
To-Hub Packager
Hub Package

Master METS
(includes provenance and technical metadata about each snapshot)

mptr

METS Snapshot 1

Files and Bitstreams

METS Snapshot 2

File 1

File 2

METS Snapshot 3

File 3

File 4
To-Hub Packagers

DSpace-to-Hub
EPrints-to-Hub
Fedora-to-Hub
Directory-to-Hub
Hub and Spoke Tool Suite

Workflow

GET: Repository package retrieved and sent to Hub Packager

LRCRUD Service for Repository X

Internet

LRCRUD Client

Retrieve Repository Package from Repository X via LRCRUD Service

To-Hub Packager

Create Hub Package METS Document from Repository Package

TechMD Augmenter

Generate technical metadata; augment Hub Package METS Document

Profile Validator

Validate Hub Package METS Document against METS Profile

From-Hub Packager

Create Repository Package from Hub package

LRCRUD Service for Repository Y

Internet

PUT: Repository Package sent to Repository Y

LRCRUD Client

Send Repository Package to Repository Y via LRCRUD Service

PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
TechMD Augmenter
JSTOR/Harvard Object Validation Environment (JHOVE)

Saved in METS as PREMIS object technical metadata
Hub and Spoke Tool Suite

Workflow

GET: Repository Package retrieved and sent to Hub Packager

LRCRUD Service for Repository X

Internet

LRCRUD Client

Retrieve Repository Package from Repository X via LRCRUD Service

To-Hub Packager

Create Hub Package METS Document from Repository Package

TechMD Augmenter

Generate technical metadata; augment Hub Package METS Document

Profile Validator

Validate Hub Package METS Document against METS Profile

From-Hub Packager

Create Repository Package from Hub package

LRCRUD Client

Send Repository Package to Repository Y via LRCRUD Service

PUT: Repository Package sent to Repository Y

PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
Profile Validator
Valid Hub & Spoke METS files:

- Aquifer MODS as primary descriptive metadata
- PREMIS object technical metadata for each file
- Valid PREMIS event elements for provenance metadata
- Every referenced file is present and has correct checksum, file-size, and mime-type values
Hub and Spoke Tool Suite

Workflow

GET: Repository Package retrieved and sent to Hub Packager

1. LCRUD Service for Repository X
2. Internet
3. LCRUD Client
4. Retrieve Repository Package from Repository X via LCRUD Service

5. To-Hub Packager
6. Create Hub Package METS Document from Repository Package

7. TechMD Augmenter
8. Generate technical metadata; augment Hub Package METS Document

9. Profile Validator
10. Validate Hub Package METS Document against METS Profile

11. From-Hub Packager
12. Create Repository Package from Hub package

13. LCRUD Client
14. Send Repository Package to Repository Y via LCRUD Service

PUT: Repository Package sent to Repository Y

PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
From–Hub Packager
From–Hub Packagers

Hub-to-DSpace

Hub-to-EPrints

Hub-to-Fedora

Hub-to-Bagit
Hub and Spoke Tool Suite

Workflow

GET: Repository package retrieved and sent to Hub Packager

Internet

LRCRUD Service for Repository X

LRCRUD Client

To-Hub Packager

Create Hub Package METS Document from Repository Package

TechMD Augmenter

Generate technical metadata; augment Hub Package METS Document

Profile Validator

Validate Hub Package METS Document against METS Profile

From-Hub Packager

Create Repository Package from Hub package

LRCRUD Client

LRCRUD Service for Repository Y

Internet

PUT: Repository Package sent to Repository Y

PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
LRCRUD
(again)
DEMO
SWORD
Hub-to-SWORD Packager

- application/zip
- METS
- SWAP metadata
- Content files

SWAP: Scholarly Works Application Profile
Got SWORD? Enter SWAP!
Scholarly Works Application Profile
SWAP: Scholarly Works Application Profile

• What it describes

• What it enables

• Context for purpose
  – Intute repository search service (http://www.intute.ac.uk/)
Negotiating a Crosswalk

MODS → SWAP
SWAP Visualized

Description Set

Scholarly Work

Expression

Manifestation

Copy

Identifier

Title

Creator

IsExpressedAs

Genre

Language

IsManifestedAs

Format

IsAvailableAs

AccessRights
## MODS to SWAP

<table>
<thead>
<tr>
<th>MODS</th>
<th>DC/QDC</th>
<th>FRBR Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;titleinfo&gt; &lt;title&gt;</td>
<td>&lt;title&gt;</td>
<td>Work</td>
</tr>
<tr>
<td>&lt;subject&gt; &lt;topic&gt;</td>
<td>&lt;subject&gt;</td>
<td>Work</td>
</tr>
<tr>
<td>&lt;temporal&gt; &lt;geographical&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;genre&gt;</td>
<td>&lt;type&gt;</td>
<td>Expression</td>
</tr>
<tr>
<td>&lt;mimeType&gt;</td>
<td>&lt;format&gt;</td>
<td>Manifestation</td>
</tr>
<tr>
<td>&lt;accessCondition&gt;</td>
<td>&lt;accessRights&gt;</td>
<td>Expression (but maybe Item)</td>
</tr>
</tbody>
</table>
• Need for more use cases

• SWAP and FRBR
  • Works in progress
  • FRBR – geared toward monographs

• Hub and Spoke preservation packages – variety of content, not just scholarly works.
Repository systems provide low out-of-the-box support for interoperability and emerging preservation standards.

Being able to move digital packages between repositories facilitates the long-term preservation of those objects.
Hub and Spoke tool suite facilitates content management across multiple repository systems while preserving valuable preservation metadata.

It uses a common packaging format in which METS files containing PREMIS metadata are treated as first class objects that are preserved along with the content.
Developers

Tom Habing
thabing@illinois.edu

Bill Ingram
wingram2@illinois.edu

Robert Manaster
manaster@illinois.edu

Project Manager

Patricia Hswe
phswe@illinois.edu

Metadata Librarian

Myung-Ja Han
mhan3@illinois.edu

University of Illinois Library at Urbana Champaign