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…

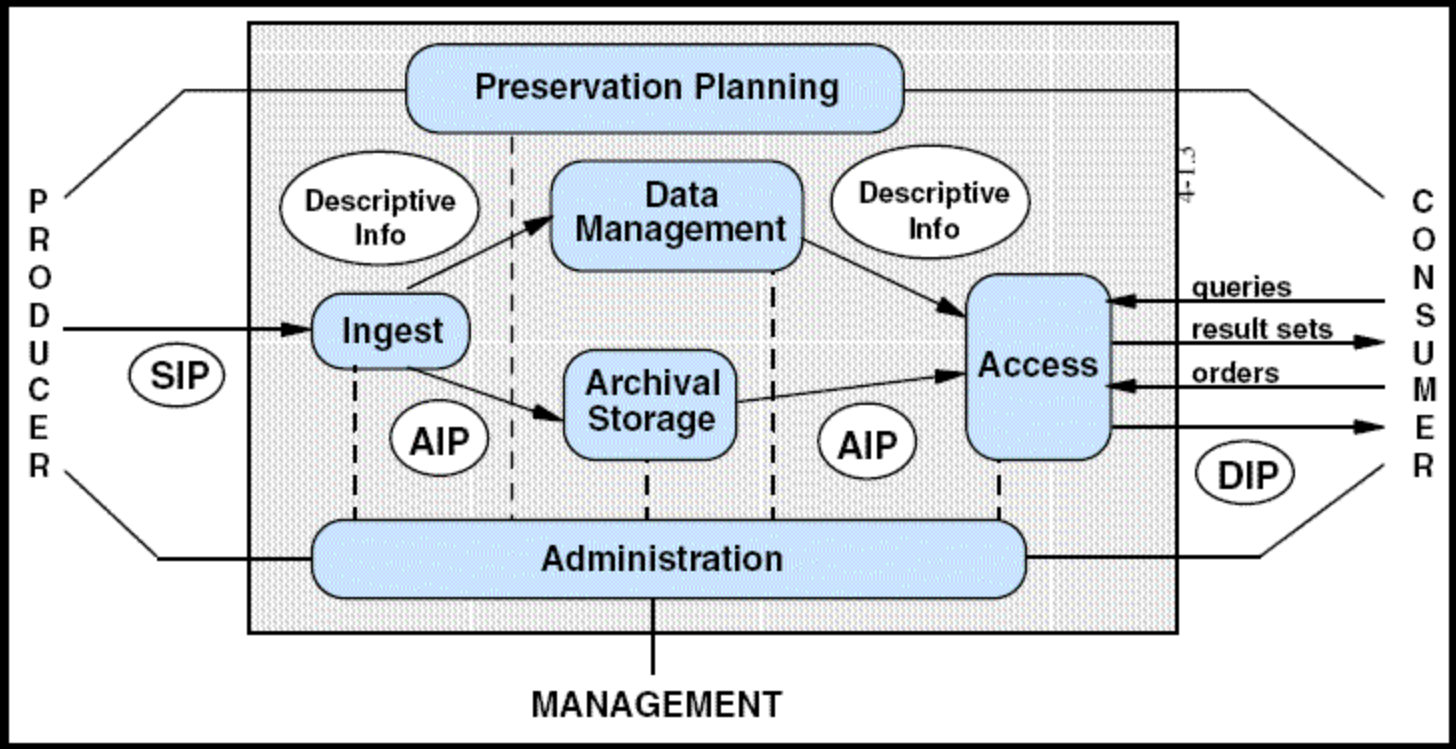
Digital PRESERVATION

National Digital Information Infrastructure & Preservation Program:
A Collaborative Initiative of the Library of Congress

Foundation of Preservation

Open Archival Information System (OAIS)

OAIS archive: An organization of people and systems that will preserve information and make it available for designated communities.



Evaluation of Repositories

- 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

Conclusion…

Repository interoperability is essential for the preservation of digital objects

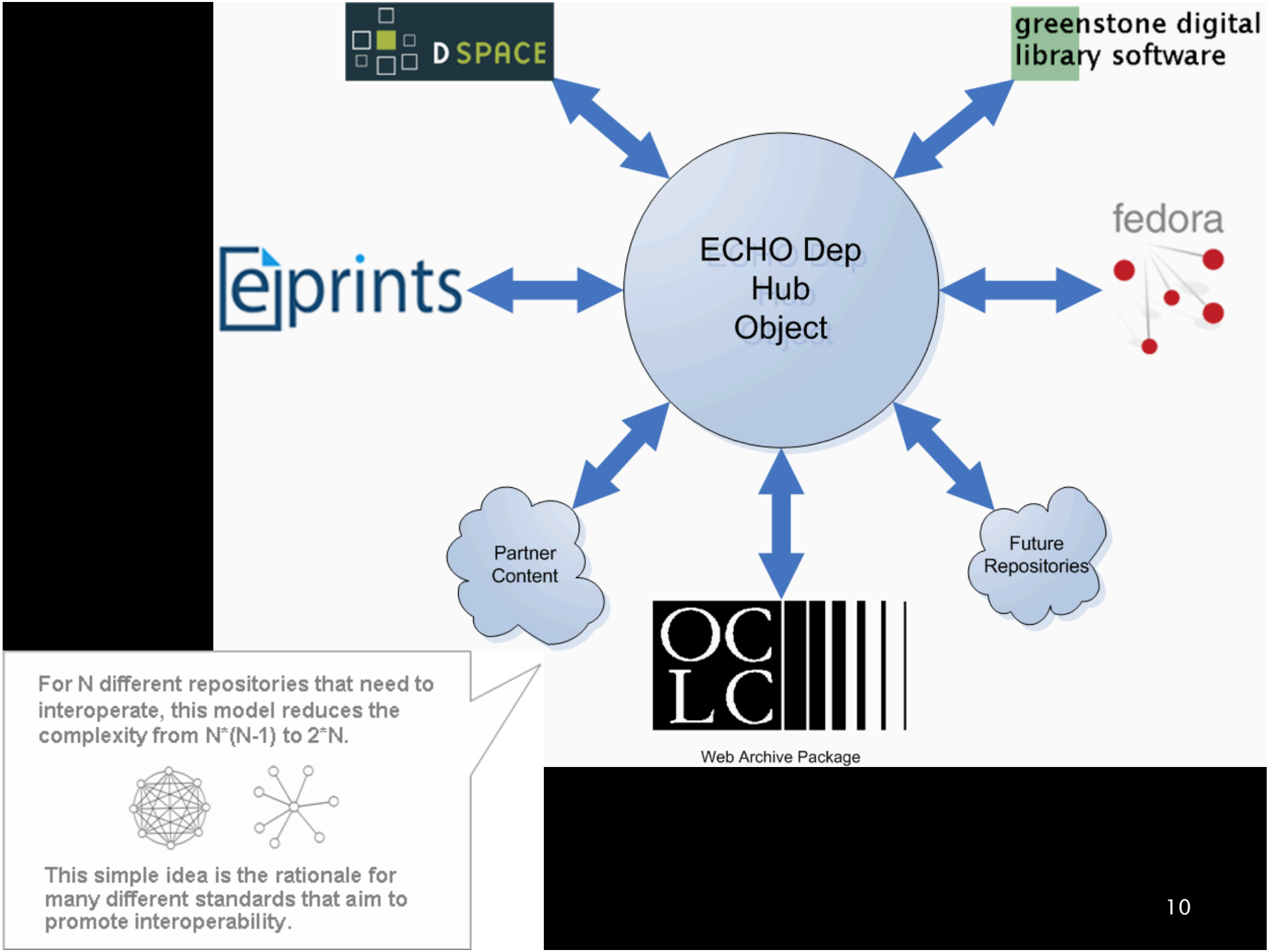
What We Have…

- Plethora of repositories
- Overabundance of data sources
- Current integration solutions are local and ad hoc
- Not much preservation

A Solution: The Hub & Spoke…

- 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-1) to 2*N.

This simple idea is the rationale for many different standards that aim to promote interoperability.

Hub and Spoke METS Profiles

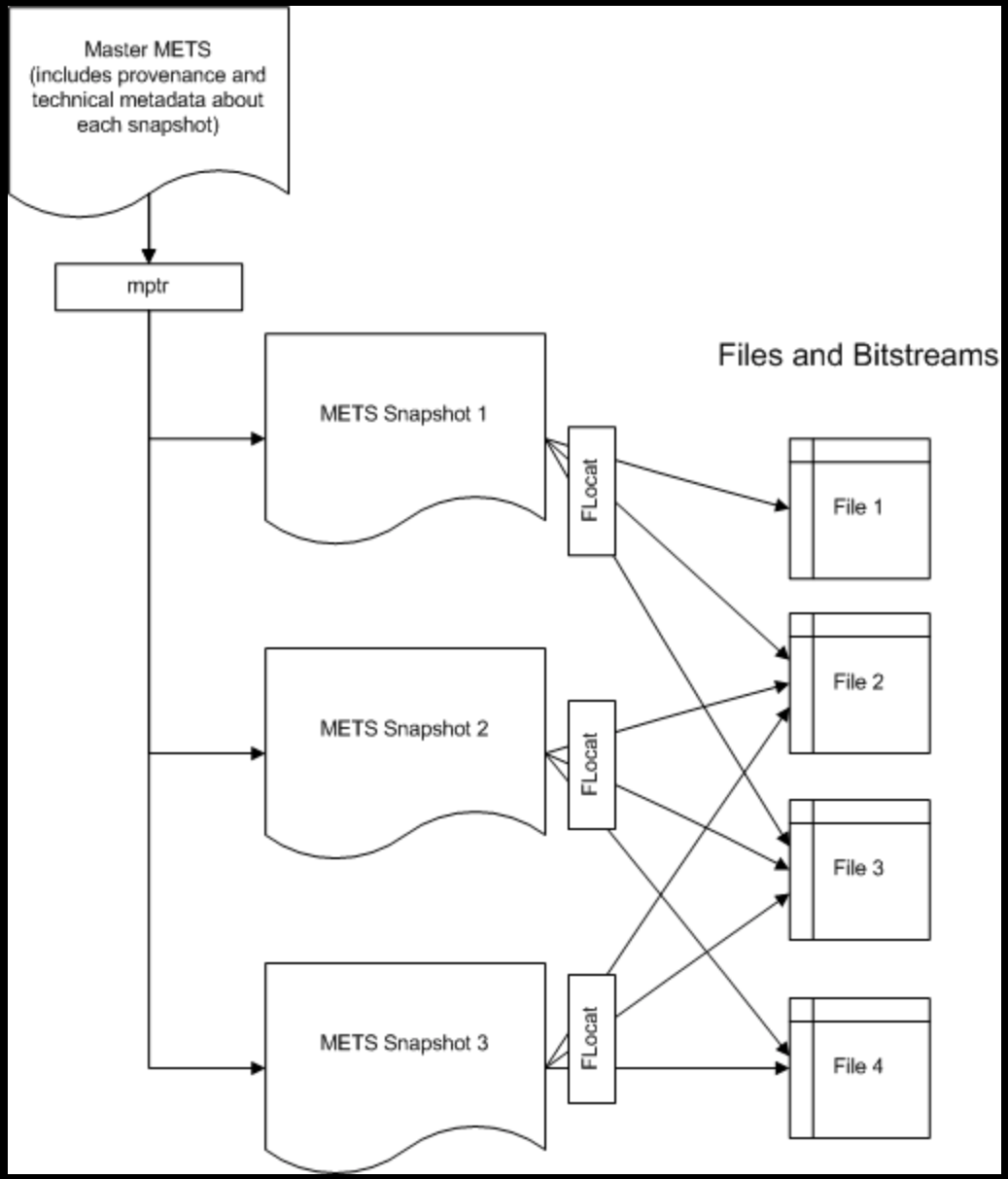
Preserving the Representations of Intellectual Entities

- Representations are encapsulated in:
  - structMap
  - structLink
  - dmdSec
  - fileSec
  - behaviorSec
- Metadata about representations:
  - metsHdr
  - amdSec
  - Various attributes of the structMap, structLink, dmdSec, fileSec, and bahaviorSec

Hub and Spoke METS Profiles

- 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



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

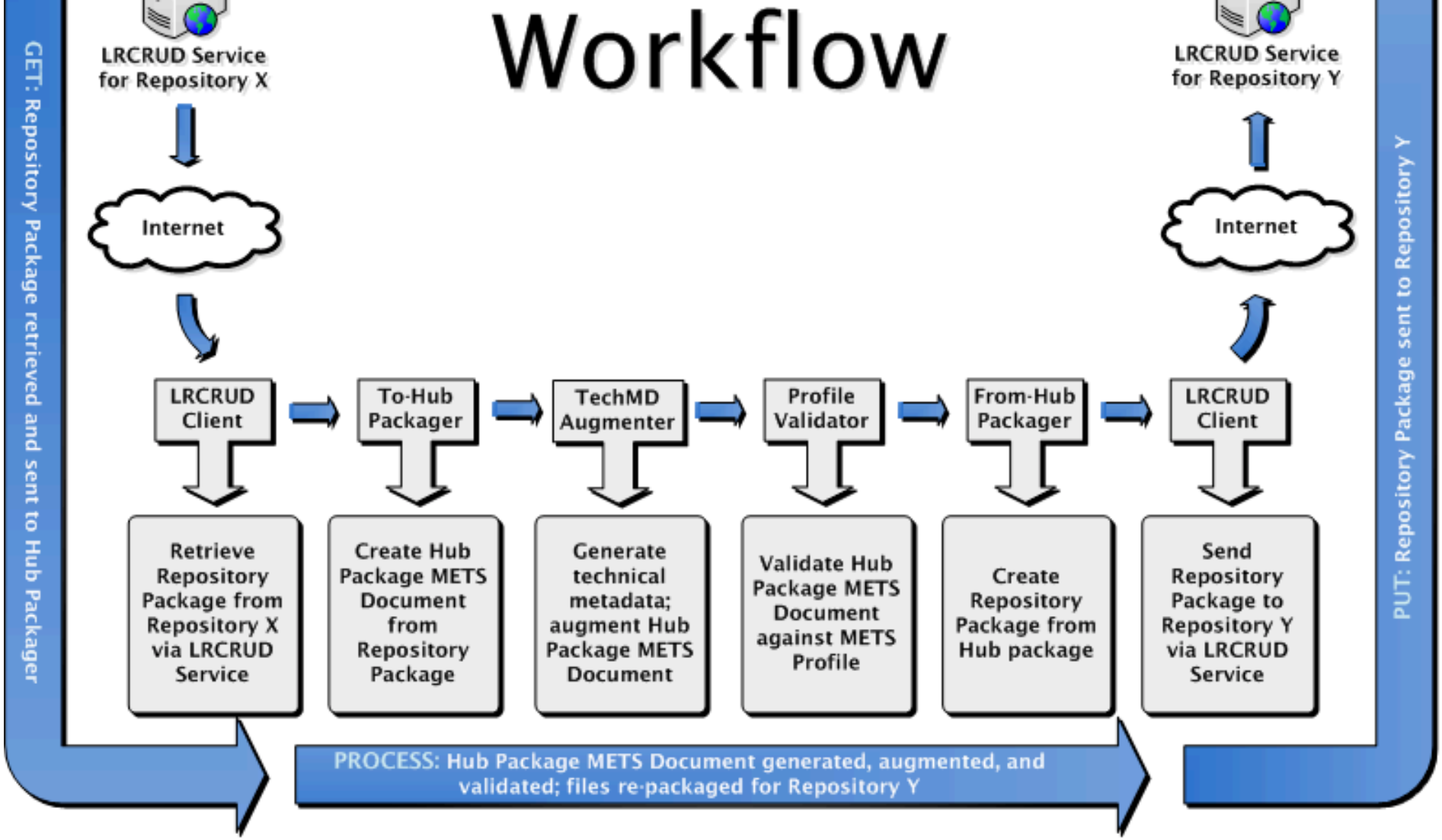
- Open Source
- Java
- XMLBeans
- Download http://sourceforge.net/projects/echodep/

METS Profiles

- Generic
  - http://www.loc.gov/standards/mets/profiles/00000015.xml
- Web Capture
  - http://www.loc.gov/standards/mets/profiles/00000016.xml
- Master METS – (not yet registered)
  - http://dli.grainger.uiuc.edu/echodep/METS/DRAFTS/MasterMETSProfile.xml

Technical Overview

Workflow


LRCRUD Service for Repository X
Internet
LRCRUD Client
To-Hub Packager
TechMD Augmenter
Profile Validator
From-Hub Packager
LRCRUD Client
LRCRUD Service for Repository Y
Internet
Retrieve Repository Package from Repository X via LRCRUD Service
Create Hub Package METS Document from Repository Package
Generate technical metadata; augment Hub Package METS Document
Validate Hub Package METS Document against METS Profile
Create Repository Package from Hub package
Send Repository Package to Repository Y via LRCRUD Service
GET: Repository Package retrieved and sent to Hub Packager
PUT: Repository Package sent to Repository Y
PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y

LRCRUD

Service

Lightweight
Repository
Create
Retrieve
Update
Delete

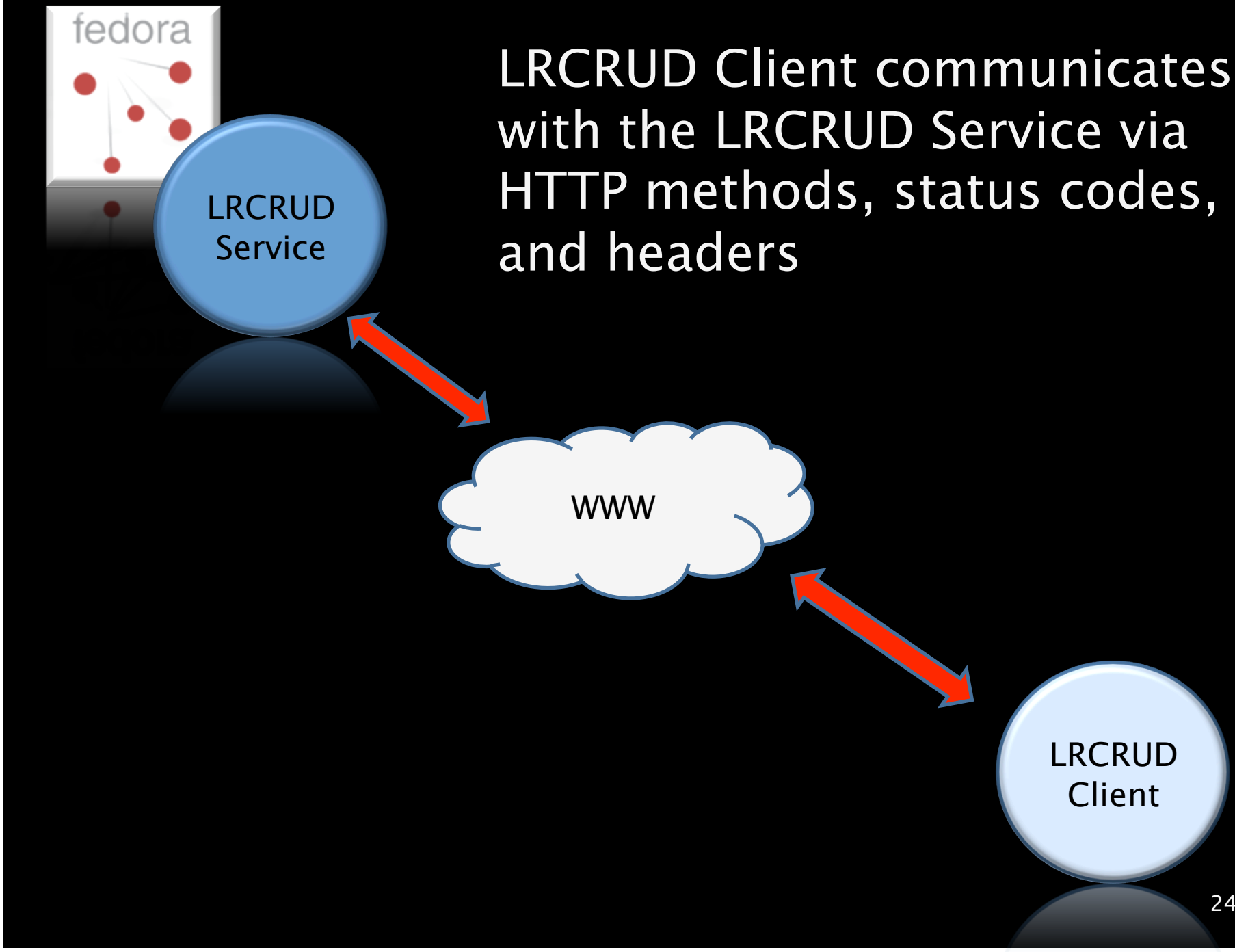
REST

Representational State Transfer

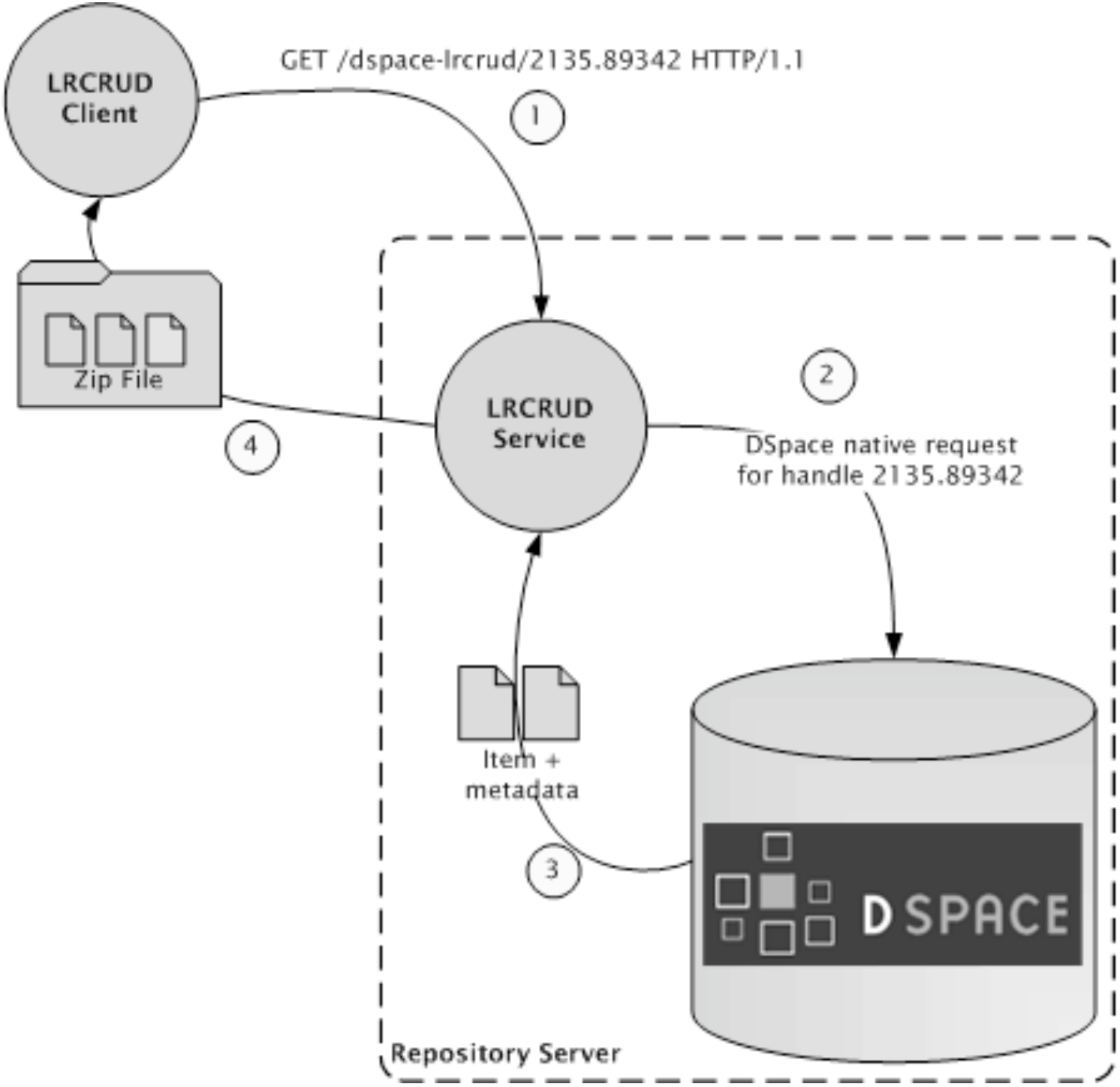
LRCRUD Web Service HTTP Methods

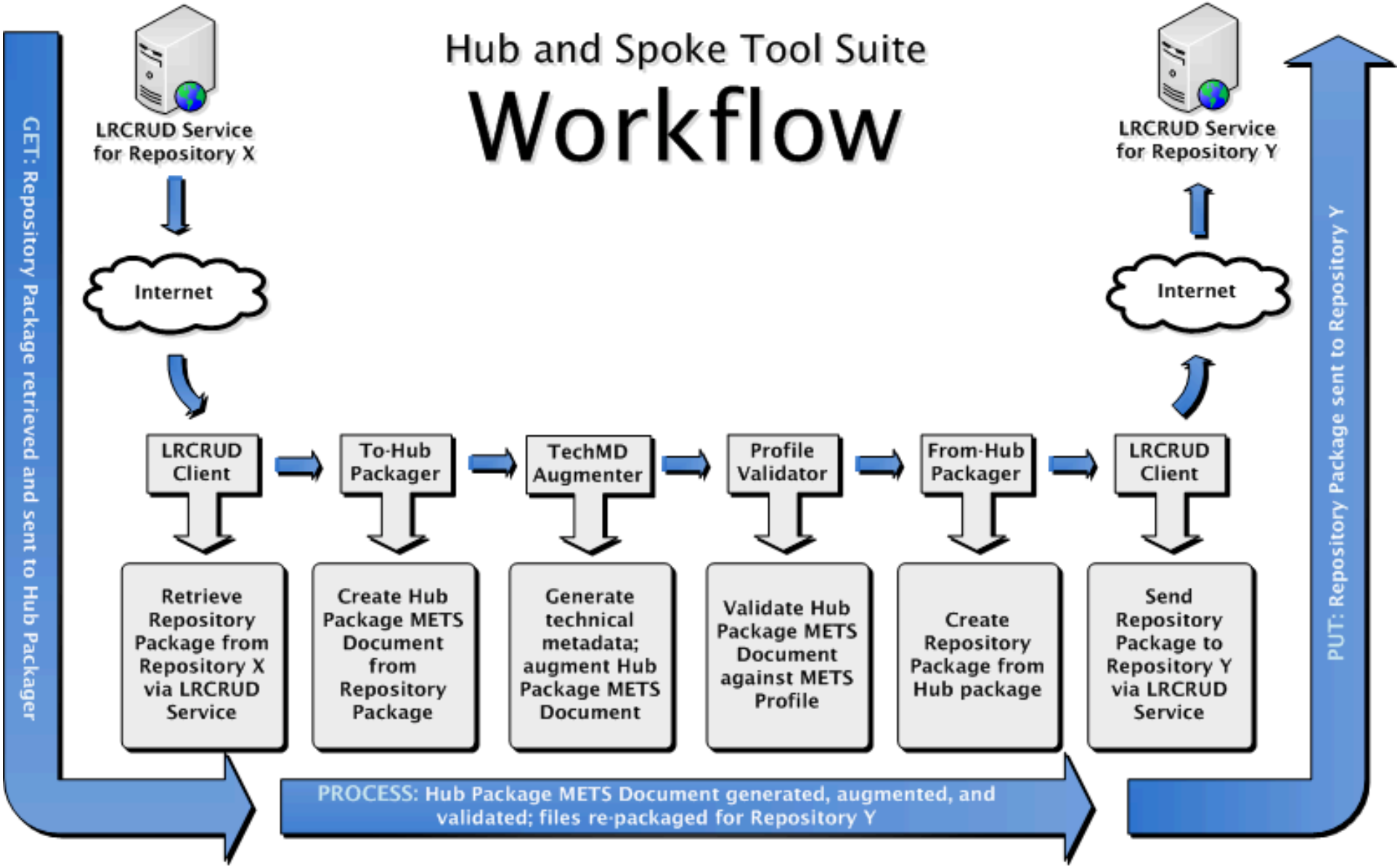
CRUD Action	HTTP Method	URL Path
Create	POST	/collection_id
Retrieve	GET	/item_id
Update	PUT	/item_id
Delete	DELETE	/item_id

LRCRUD
Client



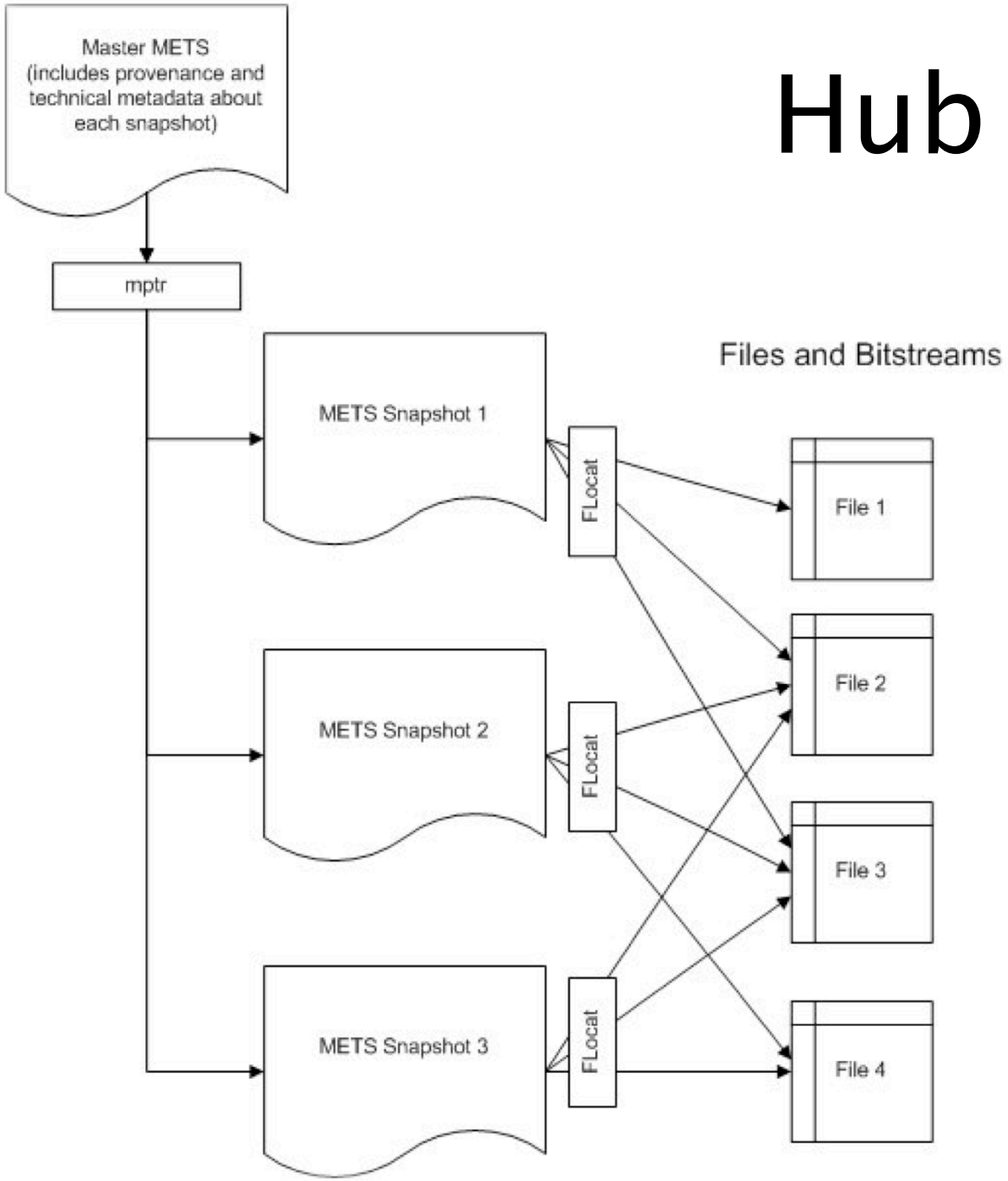
LRCRUD Client communicates with the LRCRUD Service via HTTP methods, status codes, and headers


LRCRUD Client
GET /dspace-lrcrud/2135.89342 HTTP/1.1
1
Zip File
LRCRUD Service
2
4
DSpace native request for handle 2135.89342
Item + metadata
3
DSPACE
Repository Server


Hub and Spoke Tool Suite
Workflow
LRCRUD Service for Repository X
Internet
LRCRUD Client
To-Hub Packager
TechMD Augmenter
Profile Validator
From-Hub Packager
LRCRUD Client
LRCRUD Service for Repository Y
Internet
Retrieve Repository Package from Repository X via LRCRUD Service
Create Hub Package METS Document from Repository Package
Generate technical metadata; augment Hub Package METS Document
Validate Hub Package METS Document against METS Profile
Create Repository Package from Hub package
Send Repository Package to Repository Y via LRCRUD Service
GET: Repository Package retrieved and sent to Hub Packager
PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
PUT: Repository Package sent to Repository Y

To-Hub Packager

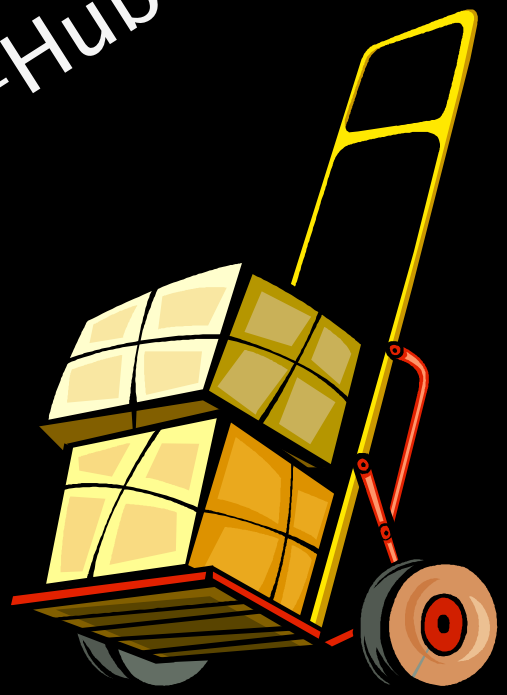
Hub Package


Master METS
(includes provenance and technical metadata about each snapshot)
mptr
Files and Bitstreams
METS Snapshot 1
METS Snapshot 2
METS Snapshot 3
FLocat
FLocat
FLocat
File 1
File 2
File 3
File 4

To-Hub Packagers

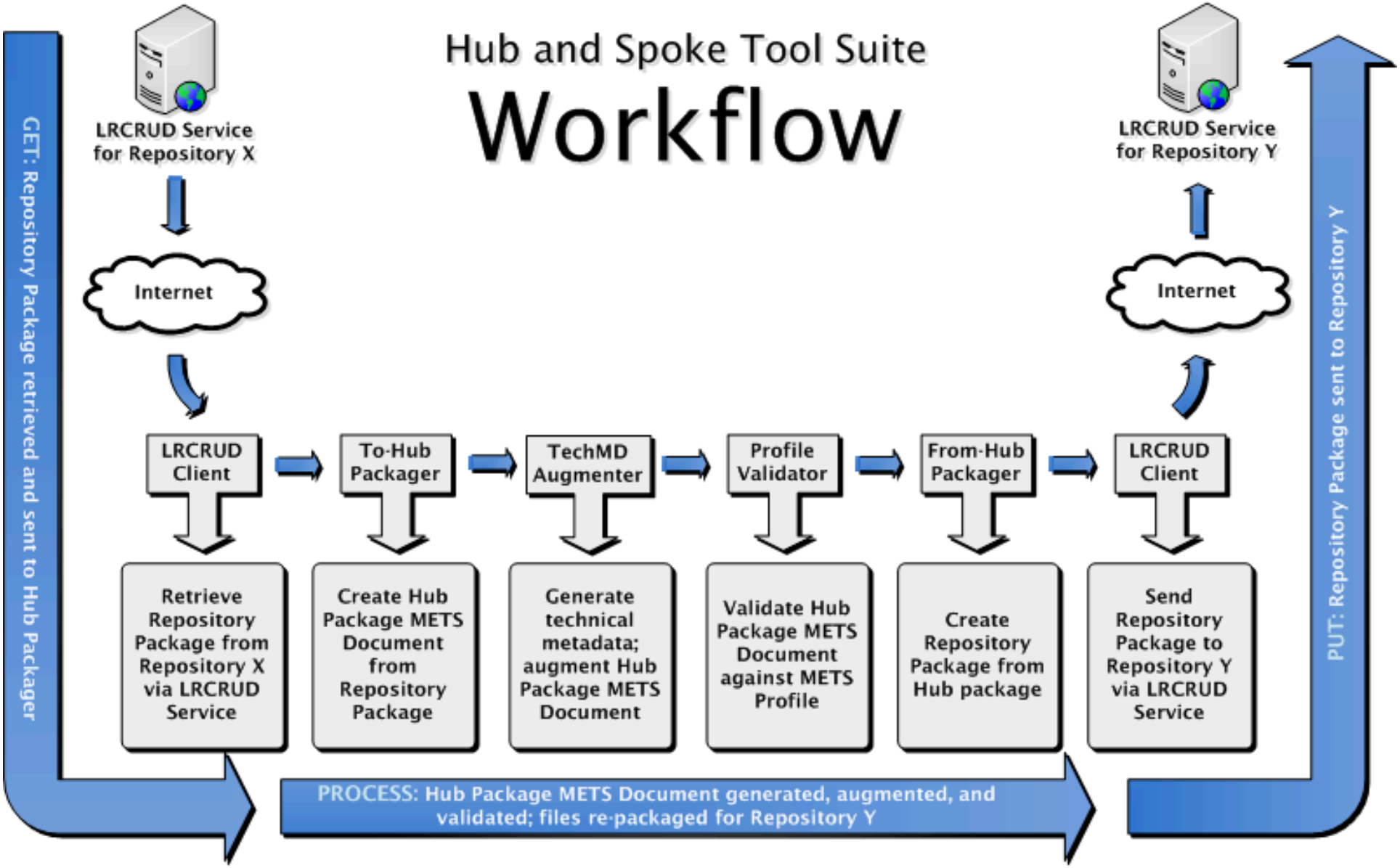
DSpace-to-Hub

EPrints-to-Hub



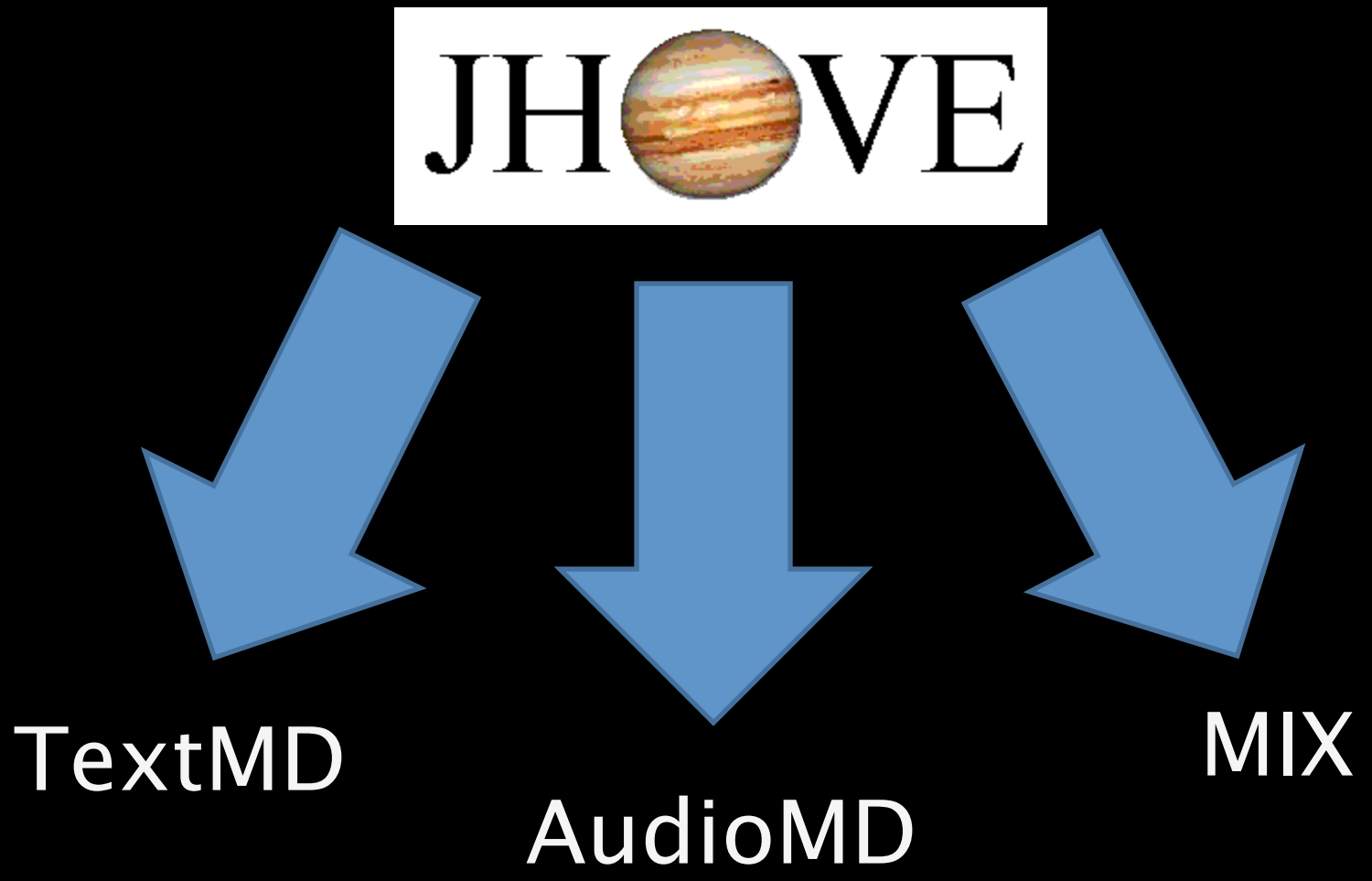
Fedora-to-Hub

Directory-to-Hub

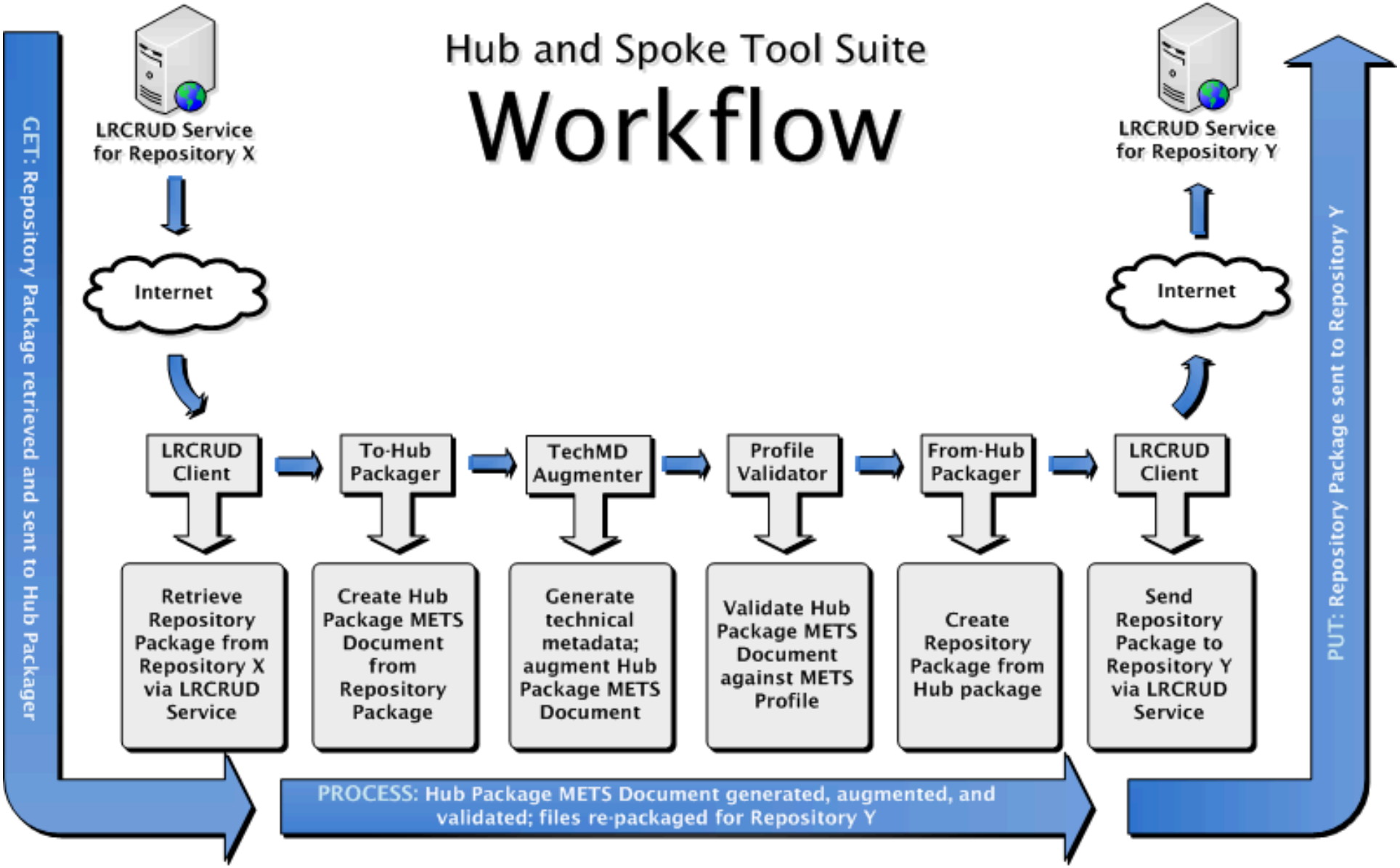

Hub and Spoke Tool Suite
Workflow
LRCRUD Service for Repository X
Internet
LRCRUD Client
To-Hub Packager
TechMD Augmenter
Profile Validator
From-Hub Packager
LRCRUD Client
LRCRUD Service for Repository Y
Internet
Retrieve Repository Package from Repository X via LRCRUD Service
Create Hub Package METS Document from Repository Package
Generate technical metadata; augment Hub Package METS Document
Validate Hub Package METS Document against METS Profile
Create Repository Package from Hub package
Send Repository Package to Repository Y via LRCRUD Service
GET: Repository Package retrieved and sent to Hub Packager
PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
PUT: Repository Package sent to Repository Y

TechMD Augmenter

JSTOR/Harvard Object Validation Environment (JHOVE)



Saved in METS as PREMIS object technical metadata


Hub and Spoke Tool Suite
Workflow
LRCRUD Service for Repository X
Internet
LRCRUD Client
To-Hub Packager
TechMD Augmenter
Profile Validator
From-Hub Packager
LRCRUD Client
Internet
LRCRUD Service for Repository Y
Retrieve Repository Package from Repository X via LRCRUD Service
Create Hub Package METS Document from Repository Package
Generate technical metadata; augment Hub Package METS Document
Validate Hub Package METS Document against METS Profile
Create Repository Package from Hub package
Send Repository Package to Repository Y via LRCRUD Service
GET: Repository Package retrieved and sent to Hub Packager
PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
PUT: Repository Package sent to Repository Y

Profile Validator

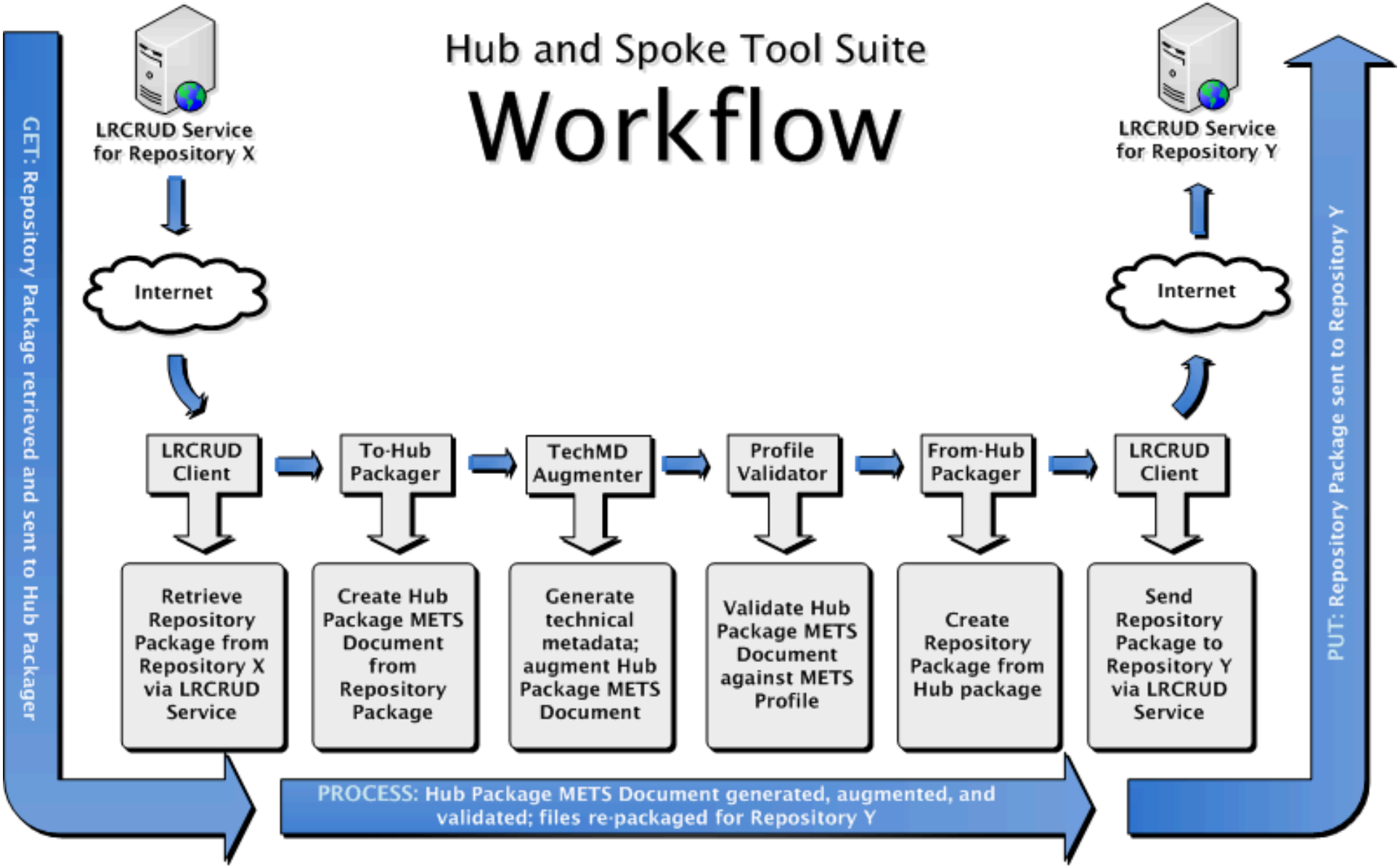
Valid Hub & Spoke METS files:

Aquifer MODS as primary descriptive metadata

PREMIS object technical metadata for each file

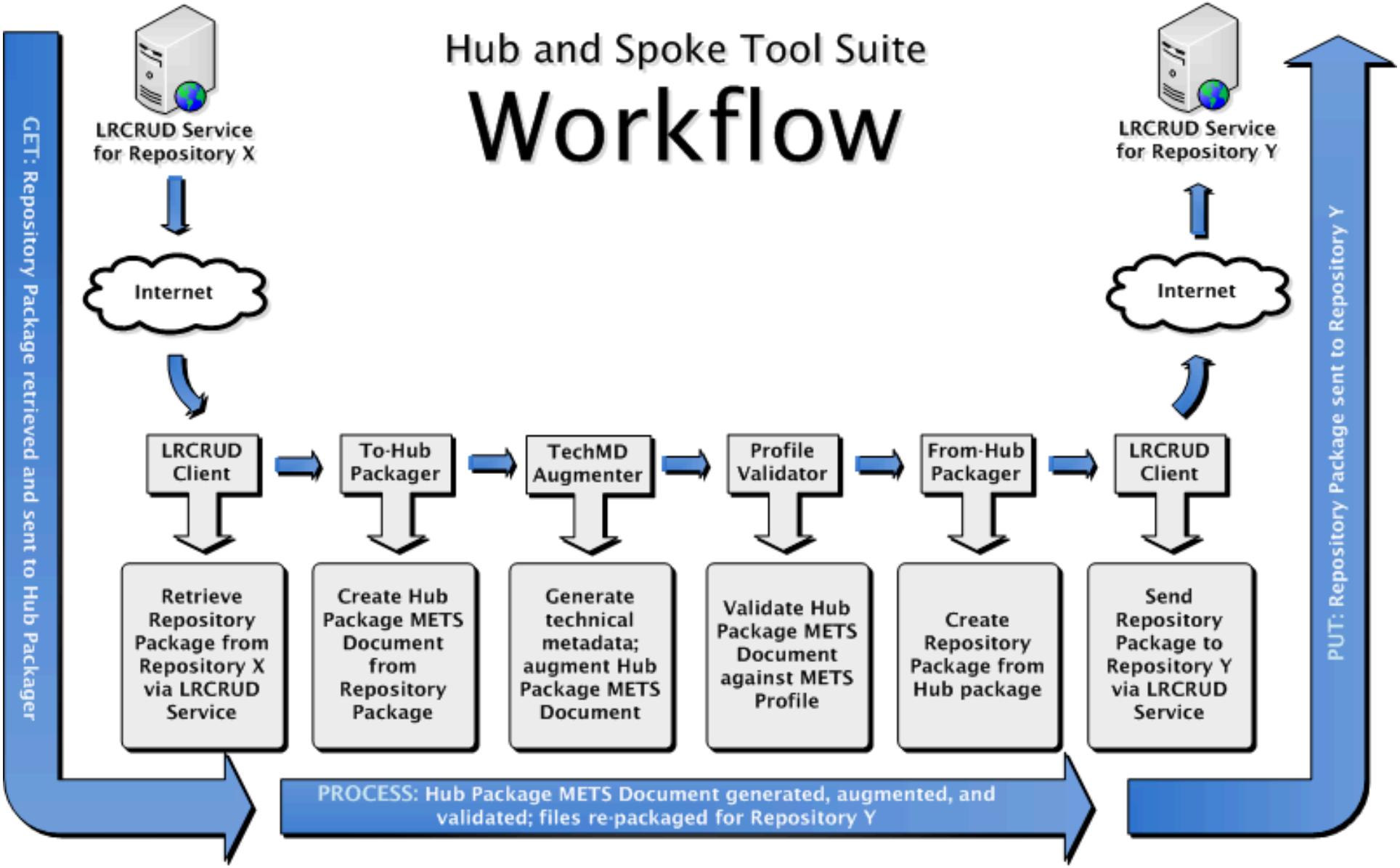
Valid PREMIS event elements for provenance metadata

Every referenced file is present and has correct checksum, file-size, and mime-type values


Hub and Spoke Tool Suite
Workflow
LRCRUD Service for Repository X
Internet
LRCRUD Client
To-Hub Packager
TechMD Augmenter
Profile Validator
From-Hub Packager
LRCRUD Client
LRCRUD Service for Repository Y
Internet
Retrieve Repository Package from Repository X via LRCRUD Service
Create Hub Package METS Document from Repository Package
Generate technical metadata; augment Hub Package METS Document
Validate Hub Package METS Document against METS Profile
Create Repository Package from Hub package
Send Repository Package to Repository Y via LRCRUD Service
GET: Repository Package retrieved and sent to Hub Packager
PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
PUT: Repository Package sent to Repository Y

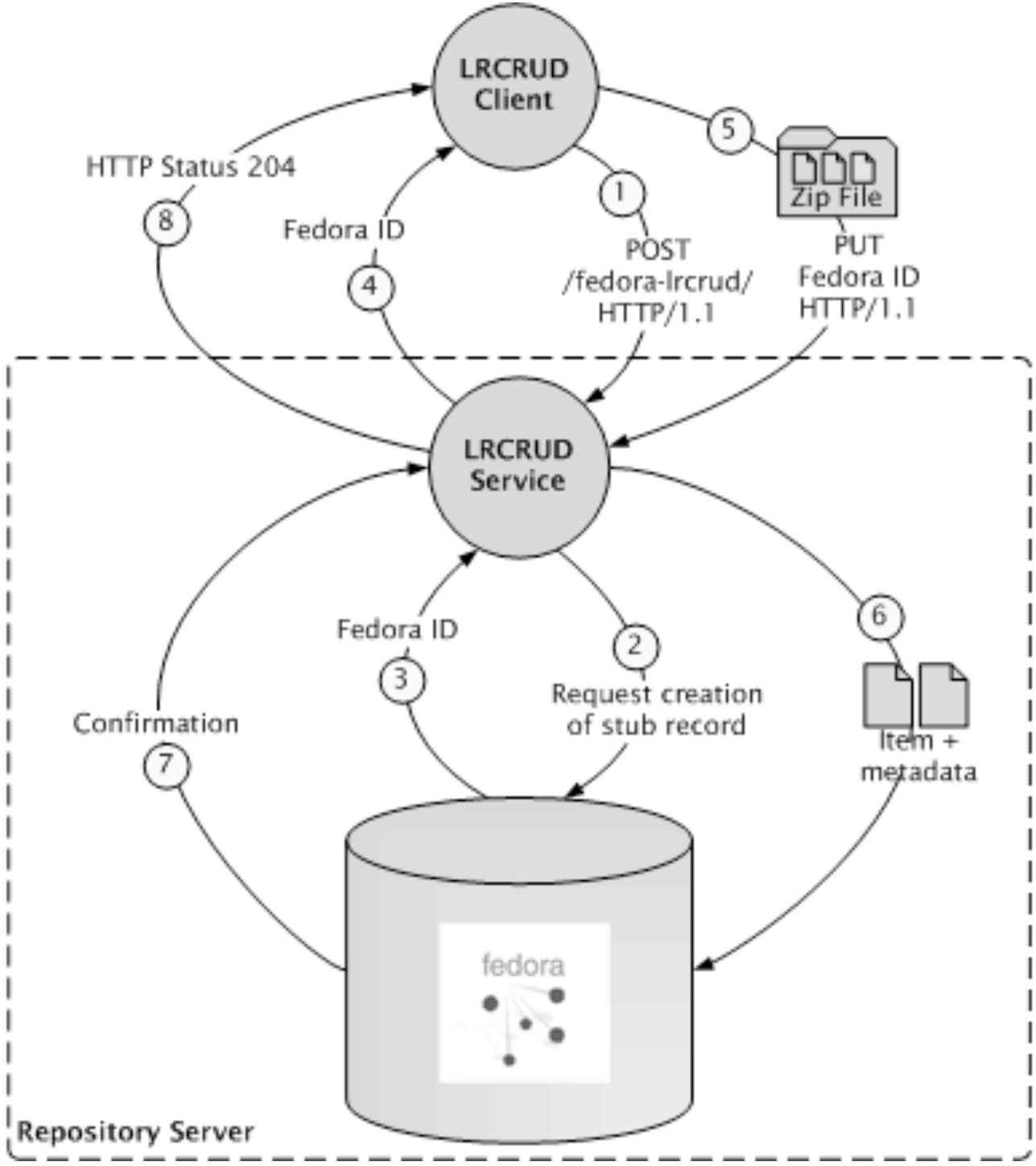
From-Hub Packager

From-Hub Packagers


Hub and Spoke Tool Suite
Workflow
LRCRUD Service for Repository X
Internet
LRCRUD Client
To-Hub Packager
TechMD Augmenter
Profile Validator
From-Hub Packager
LRCRUD Client
Internet
LRCRUD Service for Repository Y
Retrieve Repository Package from Repository X via LRCRUD Service
Create Hub Package METS Document from Repository Package
Generate technical metadata; augment Hub Package METS Document
Validate Hub Package METS Document against METS Profile
Create Repository Package from Hub package
Send Repository Package to Repository Y via LRCRUD Service
GET: Repository Package retrieved and sent to Hub Packager
PROCESS: Hub Package METS Document generated, augmented, and validated; files re-packaged for Repository Y
PUT: Repository Package sent to Repository Y

LRCRUD

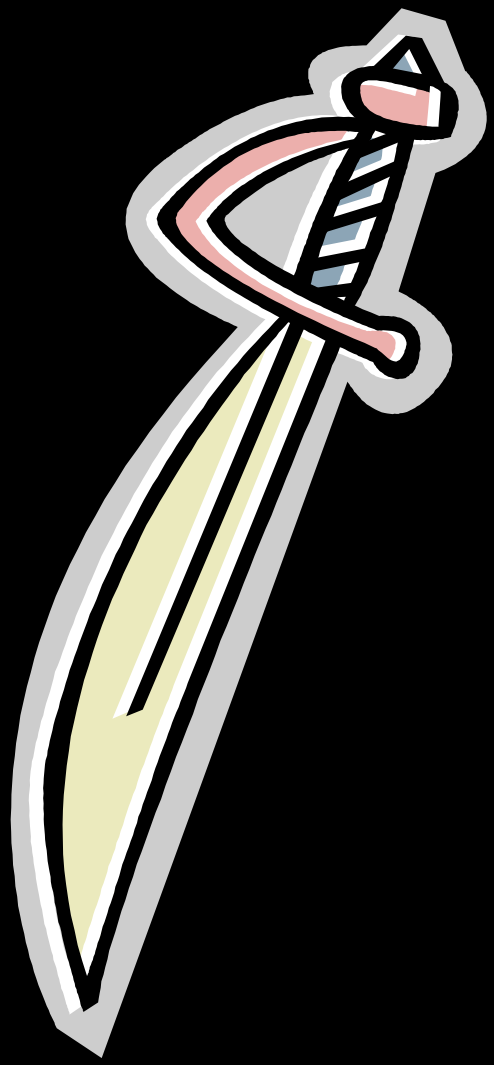
(again)


LRCRUD Client
HTTP Status 204
8
Fedora ID
4
1
POST /fedora-lrcrud/ HTTP/1.1
5
Zip File
PUT Fedora ID HTTP/1.1
LRCRUD Service
Fedora ID
3
2
Request creation of stub record
6
Item + metadata
Confirmation
7
fedora
Repository Server

DEMO

SWORD

Simple
Web-service
Offering
Repository
Deposit



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/)

SWAP Application Model

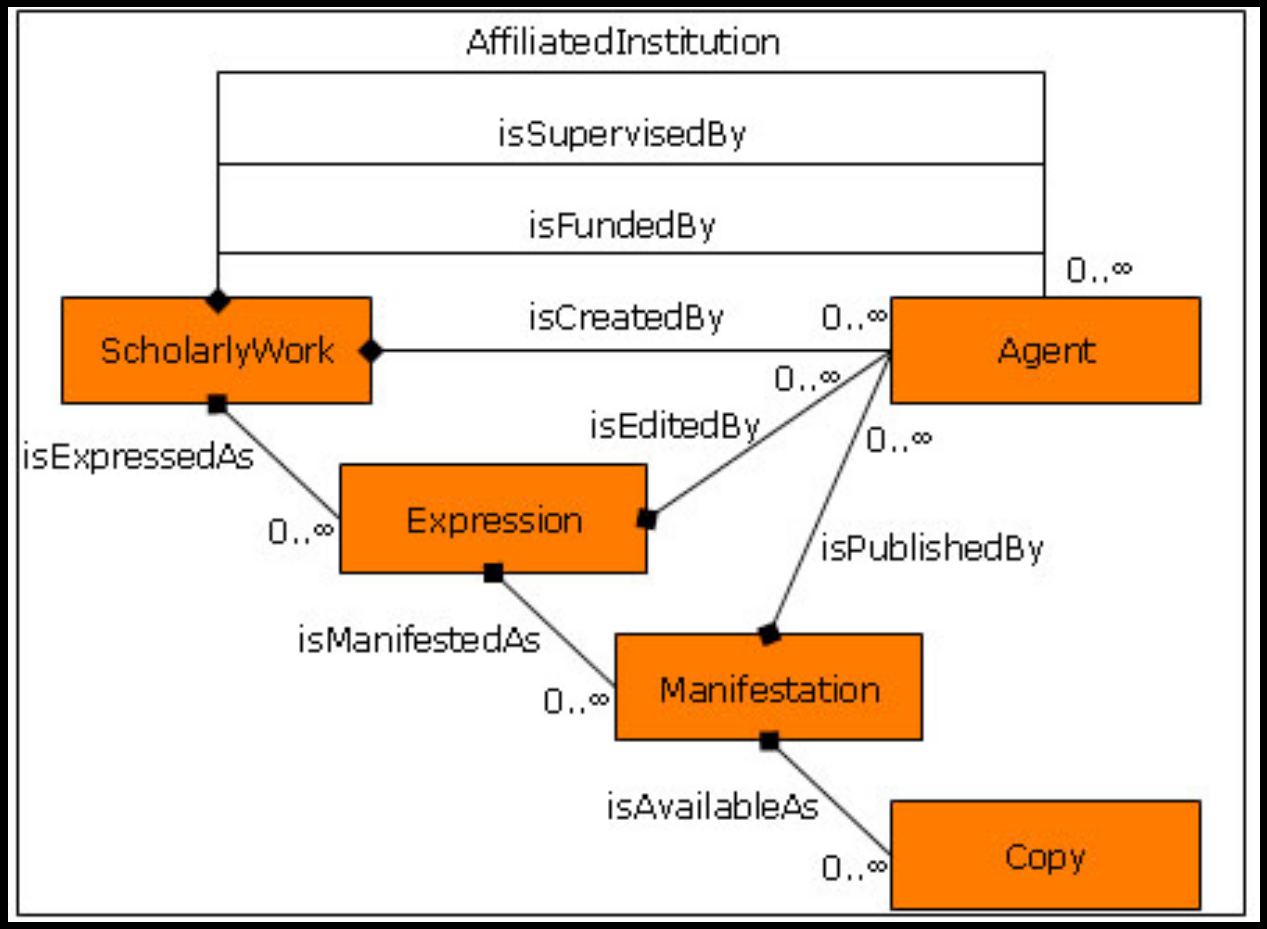
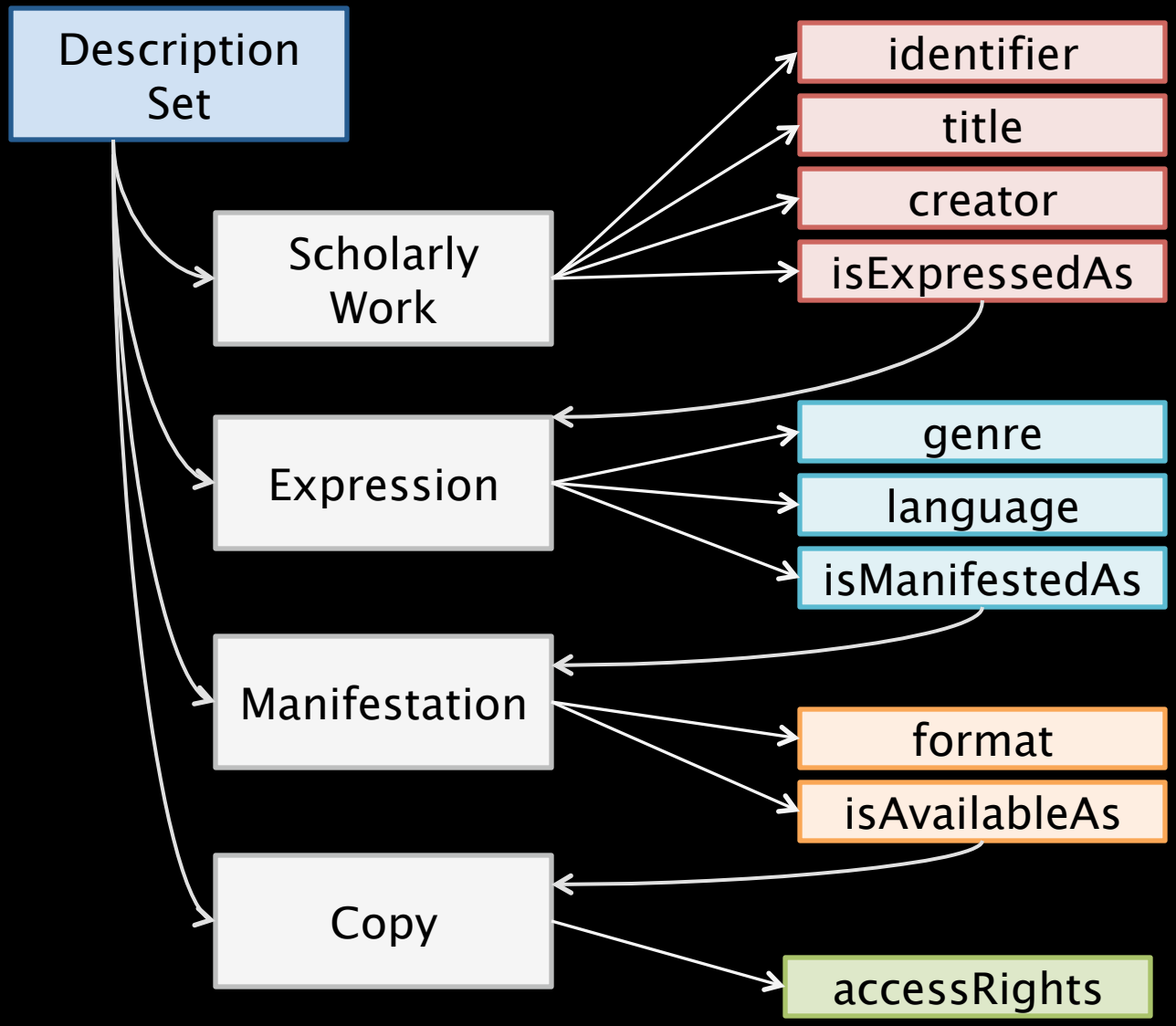


Diagram from: J. Allinson, P. Johnston & A. Powell. (2007). "A Dublin Core application profile for scholarly works." *Ariadne* (50). Retrieved April 26, 2009 from http://www.ariadne.ac.uk/issue50/allinson-et-al/.

Negotiating a Crosswalk

MODS → SWAP

SWAP Visualized



MODS to SWAP

MODS	DC/QDC	FRBR Level
<titleinfo> <title>	<title>	Work
<subject> <topic> <temporal> <geographical>	<subject>	Work
<genre>	<type>	Expression
<mimeType>	<format>	Manifestation
<accessCondition>	<accessRights>	Expression (but maybe Item)

Element Crosswalk Examples

Observations Post-Crosswalk

- 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.

Conclusion

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

Conclusion

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

Hub and Spoke (HandS) Project Team

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