Upgradation batch collection ingest using fedora4

Eric James

Fedora Users Group Presentation: May 11, 2014

Current Fedora 3 implementation

Content Model types (defined in ladybird table)

- Collection (object representing each of 39 collections, loaded/updated by 30 second rake task)
- Simple (child of collection, no further hierarchy)
- ComplexParent (child of collection, can have children)
- ComplexChild (child of complexParent or ComplexChild,can have children)
- ComplexChildUnstruct (like complexChild 1 object…)

Current Fedora 3 implementation

Content Model datastreams (some required)

- Tif/jp2/jpg/
- access/rights/desc
- Ocr
- Pdf
- Undefined (one file, for research data special use case)

Current Fedora 3 implementation

Hydra_publish table

Acts as a queue

Actions=insert,update,activate,inactivate,delete,purge

Timestamps

Identifiers-CM,collection,ladybirdID (OID),hydraID (PID)

Tracking properties (servers, queue, priorities, attempts)

Special properties on objects (viewOpt,hierarchy level,digitalchildren)

Other

Current Fedora 3 implementation

Hydra_publish_path table

- Datastreams (tif/jp2/jpg/ocr/pdf/undefined/access/rights/desc)
- Ladybird currently responsible for creation and staging of these (upload,staging,derivatives generation,assembling metadata into files
- Tables has http and path UNC locations, checksum, and other properties for fedora 3 ingest (control group, mimetype)

Current Fedora 3 implementation

Ingest script as rake task (yulhy14:ingest)

- A script – can be run concurrently for performance (7 instance sweet spot)
- Need for stored procedure that does an insert swap rather than update (SQL server limitation) interfacing w/ hydra_publish table
- Reads for hydra_publish tables, triages via action (update,insert,delete,etc)
- Gets datastreams from hydra_publish_path
- Assembles a properties datastream
- handles

Current Fedora 3 implementation

Hydra

- Uses ActiveFedora API for ingest commands
- Model objects created for xml datastreams, using OM as handles to elements and attributes, and solrizer to map to solr doc
- Bulk of work has been maintaining these mappings (LB fdid ->LB data-> XML serializations, OM definition->solr document)

Fedora 4 implementation?

Legacy CM replacement by PCDM

- https://wiki.duraspace.org/display/FF/Portland+Common+Data+Model
- Collection -> pcdm:collection
- Simple,ComplexParent,ComplexChild,ComplexChildUnstruct ->pcdm:object
- Binaries -> under pcdm:File
- Introduction of proxy objects for ordering
- Xml metadata files -> RDF

Fedora 4 implementation?

upgradation

- hydra_publish_4 table rows related 1:1 with hydra_publish
- Use this hydra_publish_4 association to hydra_publish and hydra_publish datastreams to retrieve binaries from fedora3 to swap space to stage for fedora4 upload
- Once legacy migration remove the “training wheels” and continue to use hydra_publish_4 fresh

Fedora 4 implementation?

- Metadata organized in ladybird in data tables with mapping classes to MODS XML, accessconditions XML, and table properties
- Replace this with serialized RDF for SPARQL update to fedora 4 nodes
- OR write these to a RDB or triple store of RDF statements read rows into AF statement or SPAQL update

SQLQuery6.sql - blues.library.yale.edu.pamoja (YALE\ermadmix (82))* - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

New Query | Execute | Debug | pamoja

Object Explorer

Connect

- blues.library.yale.edu (SQL Server 10.50.
  - Databases
  - Security
  - Server Objects
  - Replication
  - Management
  - SQL Server Agent

SQLQuery6.sql - bl...ALE\ermadmix (82))* | SQLQuery5.sql - bl...ALE\ermadmix (59))* | SQLQuery4.sql - bl...ALE\ermadmix (71))* | SQLQuery3.sql - bl...ALE\ermadmix (61))* | SQLQuery2.sql - bl...ALE\ermadmix (67))* | SQLQuery1.sql - bl...LE\ermadmix (128))*

```
select * from field_definition
```

100 %

Results | Messages

	fdid	frid	date	creator	ftid	acid	isMultiValued	handle	readOnly	style	faid	tooltip	displayField	textboxRows	exportField	template	autoNumber	digitalCollections	digitalCollectionsSize	exif	lockAcidValues	required
1	51	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Cataloger	0	nv	NULL	NULL	1	1	1	1	0	0	NULL	0	n	n
2	52	3	2010-08-19 16:08:39.857	NULL	1	NULL	0	Record source	1	nv	NULL	NULL	0	1	0	0	0	0	NULL	0	n	n
3	53	3	2010-08-19 16:08:39.857	NULL	1	NULL	0	Record date	1	nv	NULL	NULL	0	1	0	0	0	0	NULL	0	n	n
4	54	3	2010-08-19 16:08:39.857	NULL	1	NULL	0	Record modified date	1	nv	NULL	NULL	1	1	0	0	0	0	NULL	0	n	n
5	55	3	2010-08-19 16:08:39.857	NULL	1	NULL	0	Record ID	0	nv	NULL	NULL	1	1	0	0	0	1	500	0	n	n
6	56	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Local Record Number	0	nv	NULL		1	1	1	0	0	1	500	0	n	n
7	57	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Local record ID, other	0	nv	NULL	Use this to keep track of a locally useful number not...	1	1	1	0	0	1	500	0	n	n
8	58	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Call number	0	nv	NULL	Call number as found in Orbis	1	1	1	1	0	1	500	0	n	n
9	59	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Accession number	0	nv	NULL	Accession number, or another number used to identi...	1	1	1	1	0	1	500	0	n	n
10	60	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Box	0	nv	NULL	Archival box number	1	1	1	1	0	1	500	0	n	n
11	61	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Folder	0	nv	NULL	Archival folder number	1	1	1	1	0	1	500	0	n	n
12	62	3	2010-08-19 16:08:39.857	NULL	1	69	1	Source Creator	0	nv	NULL	Author or creator of the host item: either the author ...	1	1	1	1	0	1	1500	0	n	n
13	63	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Source Title	0	nv	NULL	Title of the host item: the title of the archival collecti...	1	1	1	1	0	1	1500	0	n	n
14	64	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Source place of creation	0	nv	NULL	Place of publication for the host	1	1	1	1	0	1	500	0	n	n
15	65	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Source publisher	0	nv	NULL	Publisher of the host item	1	1	1	1	0	1	500	0	n	n
16	66	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Source date	0	nv	NULL	Date of the host item	1	1	1	1	0	1	500	0	n	n
17	67	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Source edition	0	nv	NULL	Edition, if noted, of the host item	1	1	1	1	0	1	500	0	n	n
18	68	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Source note	0	nv	NULL	Note related to the host item	1	1	1	1	0	1	max	0	n	n
19	69	3	2010-08-19 16:08:39.857	NULL	1	69	1	Creator	0	nv	151	Creator of the object being digitized which might be ...	1	1	1	1	0	1	max	0	n	m
20	70	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Title	0	nv	NULL	Title of the object being digitized, i.e. the title of a pl...	1	2	1	1	0	1	500	0	n	y
21	71	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Variant Titles	0	nv	84	Any and all alternative titles	1	2	1	1	0	1	max	0	n	n
22	73	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Number	0	nv	NULL	Any needed numbering to designate negatives, volu...	1	1	1	1	0	1	500	0	n	n
23	74	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Caption	0	nv	NULL	Caption text if the caption is in addition to an item's t...	1	1	1	1	0	1	max	0	n	n
24	75	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Parts scanned	0	nv	NULL	An optional, descriptive listing of the components th...	1	1	1	1	0	1	500	0	n	n
25	76	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Edition	0	nv	NULL	Edition or version, if listed, transcribe from imprint if a...	1	1	1	1	0	1	500	0	n	n
26	77	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Place of origin	0	nv	NULL	Name of a place associated with the issuing, public...	1	1	1	1	0	1	500	0	n	n
27	78	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Publisher	0	nv	NULL	The name of the entity that published, printed, distri...	1	1	1	1	0	1	500	0	n	n
28	79	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Date, created	0	nv	NULL	Record the date of the thing depicted, not the date ...	1	1	1	1	0	1	500	0	n	m
29	80	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Date, depicted	0	nv	NULL	Record the date of the carrier (media) not the thing ...	1	1	1	1	0	1	500	0	n	n
30	81	3	2010-08-19 16:08:39.857	NULL	5	81	1	Date, key	0	nv	NULL	Choose one or more date ranges	1	1	1	1	0	1	500	0	n	n
31	82	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Physical description	0	nv	NULL	A narrative statement of the number and specific ma...	1	1	1	1	0	1	max	0	n	n
32	83	3	2010-08-19 16:08:39.857	NULL	1	83	1	Materials	0	nv	NULL	Narrative description of the physical form or medium ...	1	1	1	1	0	1	500	0	n	n
33	84	3	2010-08-19 16:08:39.857	NULL	5	84	1	Language	0	nv	NULL	Language of the item	1	1	1	1	0	1	500	0	n	m
34	85	3	2010-08-19 16:08:39.857	NULL	5	84	1	Language of cataloging	0	nv	NULL	Language of the majority of the cataloging	1	1	1	1	0	1	500	0	n	n
35	86	3	2010-08-19 16:08:39.857	NULL	2	NULL	1	Notes	0	mx	NULL	General textual information	1	NULL	1	1	0	1	max	0	n	n
36	87	3	2010-08-19 16:08:39.857	NULL	2	NULL	1	Abstract	0	mx	NULL	A succinct summary of some aspect of the content ...	1	NULL	1	1	0	1	max	0	n	n
37	88	3	2010-08-19 16:08:39.857	NULL	1	69	1	Associated Names	0	nv	151	A name used as a subject	1	1	1	1	0	1	1500	0	n	n
38	90	3	2010-08-19 16:08:39.857	NULL	1	90	1	Subject, topic	0	nv	NULL	Topical subjects	1	1	1	1	0	1	max	0	n	m
39	91	3	2010-08-19 16:08:39.857	NULL	1	91	1	Subject, geographic	0	nv	279	Geographic subject terms	1	1	1	1	0	1	1500	0	n	n
40	92	3	2010-08-19 16:08:39.857	NULL	1	92	1	Subject, geographic c...	0	nv	NULL	A geographic area code associated with the object	1	1	1	1	0	1	1500	0	n	n
41	93	3	2010-08-19 16:08:39.857	NULL	1	93	1	Period/Style	0	nv	NULL	Record such terms as Renaissance, Baroque, 'Got...	1	1	1	1	0	1	500	0	n	n
42	94	3	2010-08-19 16:08:39.857	NULL	1	94	1	Culture	0	nv	NULL	Mainly for VRC use, define the a sovereign state or ...	1	1	1	1	0	1	500	0	n	n
43	95	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Scale	0	nv	NULL	NULL	1	1	1	1	0	1	500	0	n	n
44	96	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Projection	0	nv	NULL	Used for cartographic description	1	1	1	1	0	1	500	0	n	n
45	97	3	2010-08-19 16:08:39.857	NULL	1	NULL	1	Coordinates	0	nv	NULL	Used for cartographic description	1	1	1	1	0	1	500	0	n	n
46	98	3	2010-08-19 16:08:39.857	NULL	1	98	1	Genre	0	nv	NULL	Used to characterize the content of the resource rat...	1	1	1	1	0	1	500	0	n	n

Query executed successfully. | blues.library.yale.edu (10.... | YALE\ermadmix (82) | pamoja | 00:00:00 | 254 rows

Output

Show output from:

Properties

Current connection parameters

- Aggregate Status
  - Connection failur
  - Elapsed time 00:00:00.436
  - Finish time 5/8/2015 11:18:01 AM
  - Name blues.library.yale.edu
  - Rows returned 254
  - Start time 5/8/2015 11:18:00 AM
  - State Open
- Connection
  - Connection name blues.library.yale.edu (Y
- Connection Details
  - Connection elaps 00:00:00.436
  - Connection finish 5/8/2015 11:18:01 AM
  - Connection rows 254
  - Connection start t 5/8/2015 11:18:00 AM
  - Connection state Open
  - Display name blues.library.yale.edu
  - Login name YALE\ermadmix
  - Server name blues.library.yale.edu
  - Server version 10.50.1617
  - Session Tracing ID
  - SPID 82

Name
The name of the connection.

Ready | Ln 1 | Col 31 | Ch 31 | INS

Fedora 4 implementation?

- Work with catalogers to map fdid handles to predicates
- Change current use of fdid values from strings and ACID (controlled vocab) to URIs where possible

Fedora 4 Implementation?

Solr indexing

- Option 1 – Use hydra/ActiveFedora, define RDF models
- Option 2 – Use fcrepo4 REST API for ingest and use camel route to transform to a solr document
- resolving URI labels*
- Dealing with nested linked data*

Fedora 4 Implementation?

Other Loose Ends

- Ladybird does derivatives and metadata, could push this functionality into fedora 4
- Using projections for certain collections (large AV)
- Policy driven storage
- Integration with preservation actions (build dynamic workflow in hydra? Hybrid with vendor solution?)
- Asynch goodness (triplestore, audit statements, file serialization, push to systemX)

Content

- Special collections (~10) and institutional IR content

Limit your search

- Creator
- Date
- Content Type
- Topic
- Language
- Digital Collection
  - Henry A. Kissinger Papers 16169
  - Lewis Walpole Library 7703
  - Day Missions Periodicals 3088
  - Israel Sack Furniture Archive 2682
  - Day Missions Annual Reports 1723
  - Yale Indian Papers 1464
  - Drama School Collection 850
  - Sanborn Fire Insurance Maps 416
  - Maurice Durand Han Nom 187
  - Arabic and Persian Medical Books and Manuscripts 72
  - Persian Philological Texts 22
- Access Restrictions

Yale University Library Digital Collections

Welcome to the new Yale University Library Digital Collections repository.
This interface will ultimately replace the many individual digital collections now available from the Yale University Library.
During this period of development, we will add both new collections and new features and are very interested in your feedback.

These collections are currently in the repository:

Arabic and Persian Medical Books and Manuscripts
Arabic and Persian manuscripts and books, as well as early translations of Arabic and Persian works dating from 1300 to 1921, from the Medical Historical Library, Cushing/Whitney Medical Library.

Connecticut Sanborn Fire Insurance Maps
Over 6,500 maps in 470 atlases for Connecticut towns dating from 1880 to 1970 at 1:600 scale. These highly detailed maps provide building, outbuilding, and property footprints, labeled streets, addresses, and information about building materials and construction features for urban and residential areas of Connecticut.

Day Missions Collection: Annual Reports
Annual reports of mission agencies and institutions document educational, medical, and religious work worldwide. Dating primarily from the 1830s through the 1930s, these reports form part of the Yale Divinity Library's Day Missions Collection, the preeminent North American collection for documentation of the history of missions and world Christianity.

Day Missions Collection: Periodicals
Periodicals published by mission agencies and worldwide religious organizations provide documentation of educational, medical, and religious work, and eyewitness accounts of events and conditions. These periodicals form part of the Yale Divinity Library's Day Missions Collection, the preeminent North American collection for documentation of the history of missions and world Christianity.

Israel Sack Furniture Archive
The Israel Sack Furniture Archive is a comprehensive digital database of American decorative arts, consisting of over 7,000 records cataloging the material objects bought and sold by the firm of Israel Sack, Inc. As the premier antiques firm representing early American furniture for much of the twentieth century, Israel Sack, Inc. was an influential corporation, operating from 1905 to 2002 in Boston and then New York City. The Israel Sack Furniture Archive will provide students, scholars and decorative arts enthusiasts alike with an unparalleled resource of comparative material for the study of early American furniture.

Lewis Walpole Digital Images Collection
Highlights from the Lewis Walpole Library's eighteenth-century British collections: visual caricatures and satires; Horace Walpole and Strawberry Hill--views, contents, and extra-illustrated copies of Walpole's "Description of the Villa"; selected topographical, portrait, and historical prints and drawings; ephemera such as trade cards, advertisements, bookplates, and playbills, and more.

Maurice Durand Han Nom Handwritten and Woodblock Manuscripts
Original woodblock or brush ink texts and translations of Han Nom texts into modern Romanized Vietnamese collected by Maurice Durand, a prominent Vietnamese/French scholar of Han Nom from the mid-20th century. Han Nom script uses classical Chinese characters to represent Sino-Vietnamese vocabulary and some native Vietnamese words, while other words are represented using locally created characters based on the Chinese model.

Persian Philological Texts
Selected older Persian philology texts which originate from South Asia, are rare European translations, or are reprints. All of the selected items are in the "endangered" preservation category. Many have extensive marginalia from scholars, or are from the personal libraries of Yale scholars such as Edward Salisbury.

Yale Indian Papers
Letters, committee reports, covenants and maps and other materials documenting the history of New England Native Americans. The Yale Indian Papers Project is a documentary editing endeavor and collaborative research initiative with the mission to advance scholarship on the history and culture of New England Native Americans.

Note - currently this resource works best with Chrome, Safari, Firefox (latest two versions), IE 10, and Safari for iPhone and iPad.

Thanks!