From Ingest To Access: A Day In The Life Of A HathiTrust Digital Object

Jeremy York (University of Michigan)

University of MIchigan

Hardware | Process | User Interface/Tools

http://www.hathitrust.org/papers

1 Bibliographic Ingest

- Metadata requirements are validated
- A process identifies bibliographic records that already exist in the repository. If there is an existing record, new holdings and item records are added to that record. If there is not, a new bibliographic record is created.

2 Bibliographic Data Store (Aleph)

3 Rights Determinations

- Automatic using bibliographic data
- Public Domain if
  - US federal government documents, or
  - published in the US prior to 1923, or
  - published outside of the US before 1869
- Public Domain in the United States if
  - published outside of the US between 1868 and 1909
- Otherwise in-copyright
- Bib-determined statuses may be overridden by corrections to bibliographic records, copyright review, out-of-print and brittle determinations, and rights holder permissions.

8 Processes check for updated information in bib records and in the rights database daily to be sure all rights information is current.

9 Bibliographic Search Solr Index

12 Content Ingest (TIFF, JPEG2000)

13 Server | Server | Server (Red Hat Linux) | Server | Server

17 SLIP (Solr Large-Scale Indexing Processer)

To create Solr documents, SLIP

1. Receives a list of ids to be indexed from the Shadow Rights Database.
2. Queries the repository, opens the zip file, unzips the OCR
3. Queries VuFind Solr index for Solr metadata fields
4. Queries Rights Database for rights information

After building a document, SLIP sends it to one of the Large scale search Solr servers in round robin fashion for indexing. There are N number of processes for creating documents in each SLIP instance on the ingest servers.

14 GROOVE

Content Validation

- Luhn validation on barcodes
- Fixity check on JPEG2000, TIFF, UTF-8 using MD5
- Well-formedness and embedded metadata check on JPEG2000, TIFF, UTF-8 using JHOVE

METS file created

1. metsHdr with an ID, creator of METS document, and creation date
2. dmdSec with marcxml
3. dmdSec with a reference to ILS record
4. amdSec containing one techMD with PREMIS metadata
5. fileSec with four fileGrps (zip archive, images, OCR, and coordinate OCR)
6. Physical structMap connecting files with metadata (pg. numbers or features)

Quality Review Sample Created | Handle Assigned | Image and OCR Zip Package Created

15 Quality Review

- 20 consecutive pages from each volume are set aside
- Approximately 1% of all content is reviewed manually for blur, cleaning, warp, crop, obscured, and colorization errors, as well as errors relating to TIFF thresholding (thick or broken).

18 Large Scale Search Servers

Building Solr Instances
These Solr instances build one Solr shard index each. Once a day, snapshots of the shards are mounted and served by the Serving Solr Instances. There are a total of 10 shards on 5 machines, handled by 20 Solr instances

Serving Solr Instances
These instances receive queries from the LSS application and send them to the shards on all of the other machines. They merge results from the shards and send the results back to the application.

Large Scale Search Solr Index

Unzip OCR

16 Content Data Store

Pairtree directory structure

ZIP file
/sdr/mdp/pairtree_root/39/01/50/04/07/96/56/39015015004965 6/390150049656.zip

OCR
00000001.txt
00000002.txt
...
Page Images
00000001.tiff
00000002.jp2
...
Coordinate OCR
00000001.html
00000002.html
...
Google METS file
UOM_390150049656.xml
HathiTrust METS file
390150049656.mets.xml

Isilon Storage

- Built for disaster recovery. High redundancy.
- Divided into nodes (CPUs + storage)
  - Data is distributed across all nodes
  - N+3 parity protection = 3 nodes can completely fail and all data will still be available
- Sync IQ replication software
- MediaScan scans blocks in the system for bad disk sectors. If it finds one, it uses parity information to rebuild the necessary data and rewrite a block somewhere else on the drive. Content is migrated to a block in a new node and re-balanced across the system, incorporating the new node.
- OneFS 5.0 operating system release will allow checks of system data and metadata via associated checksums.

4 MySQL Database Server (Debian)

- **5** Rights Database
- **6** Shadow Rights Database
- **7** Geo IP Database

METS File

Unzip — Page Images and OCR

22 Collection Builder Solr Index

Rights Database

- All determinations are saved in the database with an id, rights attribute (e.g. pd, pdus, ic) reason code, source (e.g. Google, Internet Archive, local digitization), user, and timestamp.

Attributes

1	pd	public domain
2	ic	in copyright
3	opb	out of print and brittle
4	orph	copyright orphaned
5	und	undetermined
6	umall	available to all UM affiliates and walk-in patrons
7	world	available to everyone
8	nobody	blocked for all users
9	pdus	public domain only when viewed in the U.S.

Reasons

1	bib	derived automatically from bibliographic data
2	ncn	no printed copyright notice
3	con	contractual agreement with rights holder on file
4	ddd	due diligence documentation on file
5	man	manual access control override
6	pvt	private personal information visible
7	ren	Copyright renewal research conducted
8	nfi	needs further investigation
9	cdpp	title page or verso have copyright date and/or publication information not in bib record
10	cip	condition review and in-print status research conducted
11	unp	unpublished work

Sources

id	name	dscr
1	google	Google
2	lit-dlps-dc	Library IT, Digital Library Production Service, Digital Conversion

Rights Database (example)

id	attr	reason	source	user	time	note
mpd.39015054477651	1	1	1	root	2006-01-12 11:34:26	
mpd.39015017678577	1	1	1	root	2006-01-12 11:34:27	
mdp.39015017678577	4	4	1	sooty	2006-02-08 15:18:24	determined by jaheim as in-copyright, but orphaned
mdp.39015034781842	2	1	1	root	2006-01-12 11:34:28	
mdp.39015034781842	6	5	1	pwillett	2006-03-08 09:12:45	agreement reached with publisher for campus access

19 Web Servers (Redhat Linux)

23 PageTurner

1. Retrieves bib data from Aleph
2. Queries Rights Database for attribute and source
3. Loads METS
4. Gets/transforms images, gets OCR

21 Collection Builder

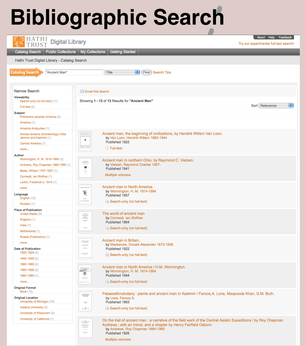
1. Limited metadata for collection items (title, publisher, date) is saved in a database table with the collection to which it belongs.
2. When new items are added, the volume OCR is retrieved from the repository and indexed (or reindexed) along with other volumes from that collection.

20 Large Scale Search Application

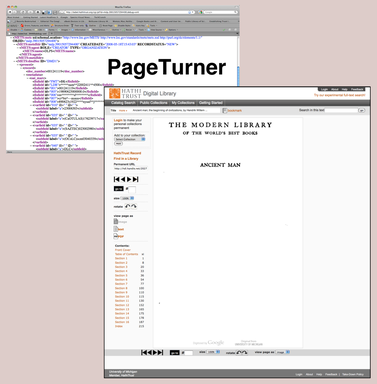
1. Sends a query to Serving Solr Instances
2. Receives and displays results

10 Queries Bibliographic Solr Index

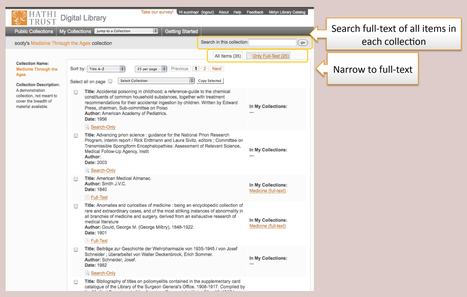
11 **Bibliographic Search**



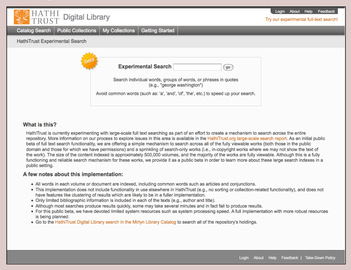
PageTurner



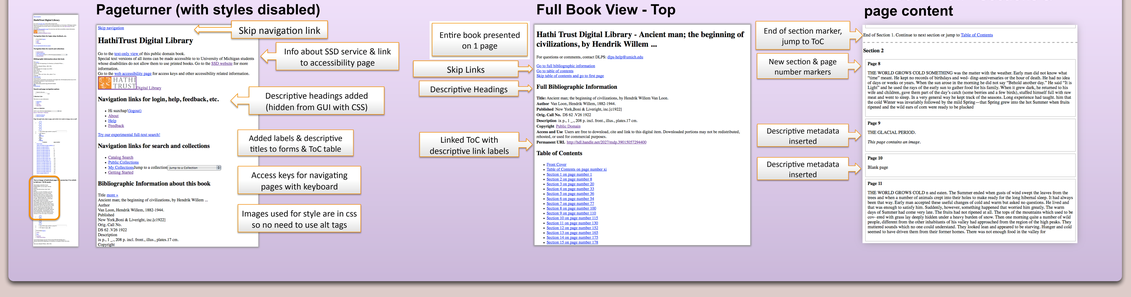
Collection Builder



Large Scale Search



24 **Accessible Interface**



Pageturner (with styles disabled)

Skip navigation link
Info about SSD service & link to accessibility page
Descriptive headings added (hidden from GUI with CSS)
Added labels & descriptive titles to forms & ToC table
Access keys for navigating pages with keyboard
Images used for style are in css so no need to use alt tags

Full Book View - Top

Entire book presented on 1 page
Skip Links
Descriptive Headings
Linked ToC with descriptive link labels

Full Book View - Sections & page content

End of section marker, jump to ToC
New section & page number markers
Descriptive metadata inserted
Descriptive metadata inserted

25 **Computational Research**

Three methods:

1. Data Distribution (for public domain volumes)
2. Protocol-based method (e.g. SEASR)
3. Research Center

Access: Indiana University & University of Michigan