Metadata Migration Managed: Fixing Metadata That Was Up to No Good

Anna Neatrour
Jeremy Myntti

Brian McBride
Alan Witkowski
Matt Brunsvik

J. Willard Marriott Library
THE UNIVERSITY OF UTAH

Overview of Marriott Library Digital Collections

263 Collections

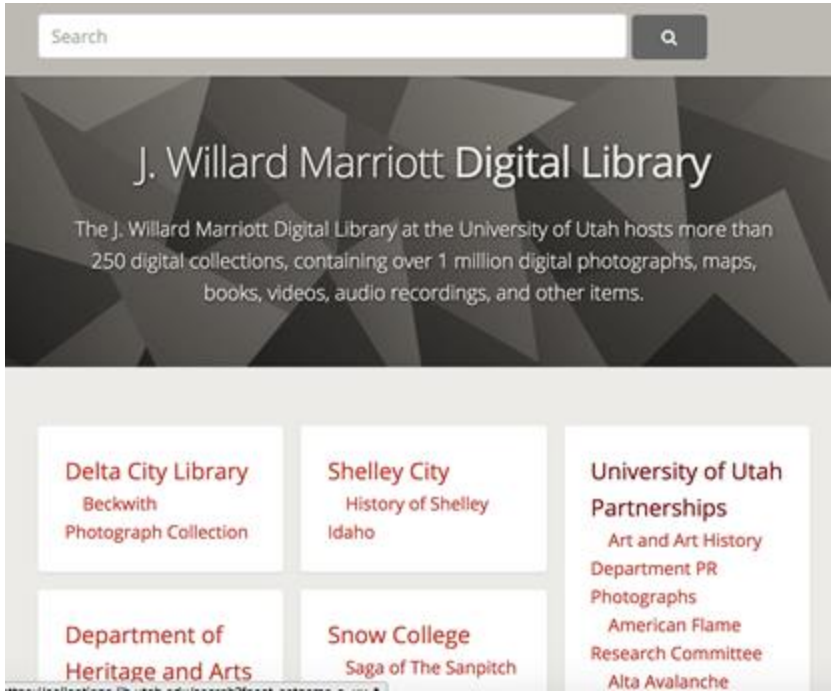
691,154 IE's (Items)

133,289 Compound Objects

13 file types (images,videos,pdfs...)

2.8 TB storage

Over 50 Partners (internal and external) hosting collections with the Marriott Library



https://collections.lib.utah.edu/

Utah Digital Newspapers (UDN)

20,486,348 articles

1,823,561 pages

135 Newspaper titles

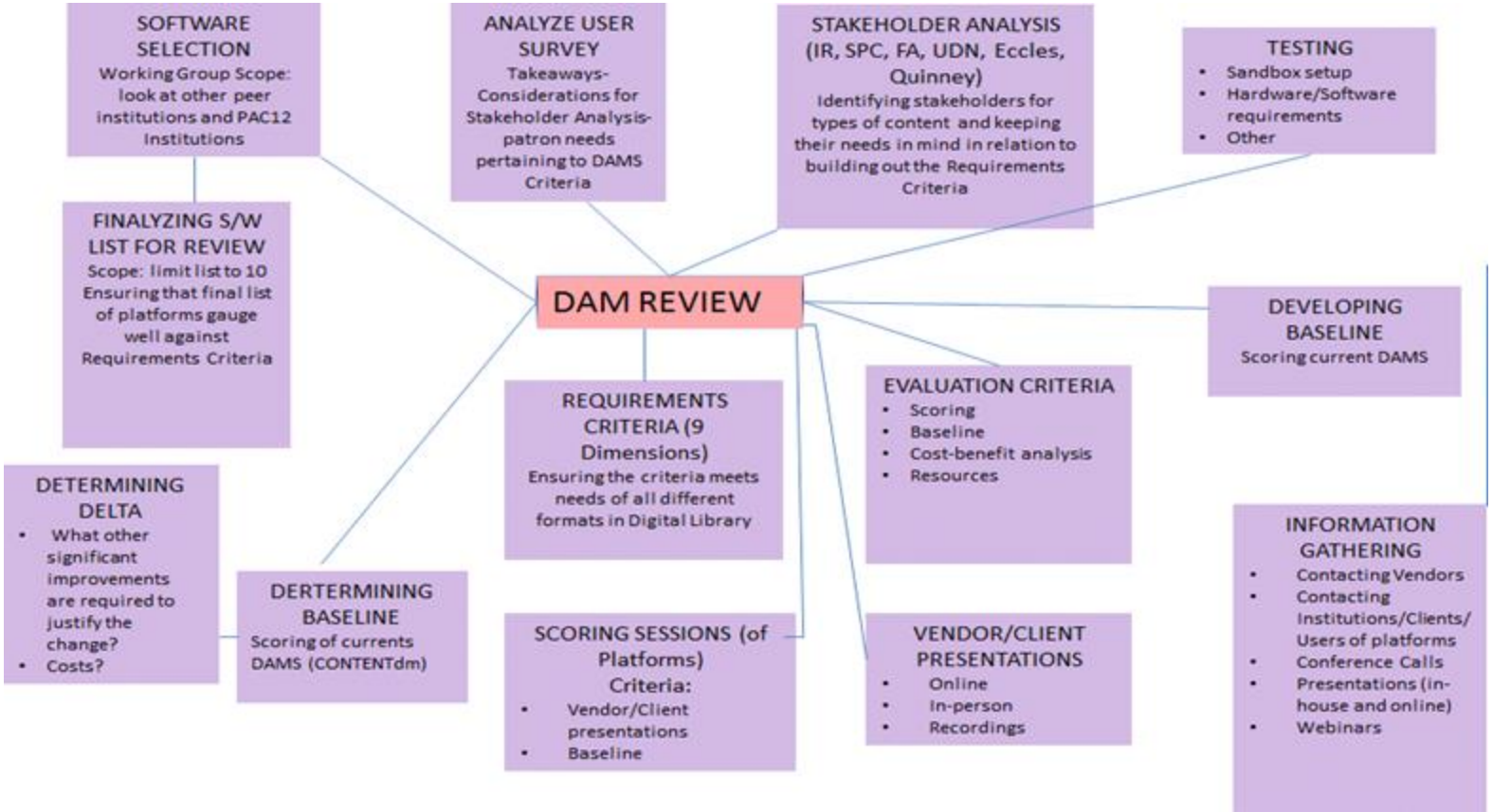
Originally setup for article level data in newspapers, not page level

Some newspapers digitized through NDNP, still have many unique papers not in Chronicling America

https://digitalnewspapers.org/

DAMS review in 2013

http://mwdl.org/events/DAMS_options.php



CONTENTdm set-up vs Open source set-up

CONTENTdm (locally hosted)

2 servers (digital library and UDN)

CDM (2x12 Core 3.0 Ghz, 64 Gb mem) UDN (2x8 Core 2.4 Ghz, 96 Gb mem)

Hosting for partners

Scalability issues

Limited customizations

Open Source

- 3 VMs (indexing services, Utah Digital Newspapers and Digital Collections)
- VMs avg. 2 cores, 3 Gb mem
- Hosting for partners
- Extremely small footprint (hardware and power savings)
- Highly scalable

Benefits of new system

Apache Solr (indexer), NGINX (webserver), and phalcon (PHP framework)

- No longer bound by CONTENTdm scalability limitations
- In-house expertise and support
- Student/end-user and partner convenience (faster response times, advanced search-ability functions, front-end metadata editing, customized interface...)
- Large cost/power/hardware savings for hosting
- Improved workflows for all teams
- Staff time savings -- 1 less FTE supporting servers
- Demonstrating and sharing open source solutions with the public as well as

Speed and indexing improvements

UDN

55 GB of metadata

21,648,462 records

CONTENTdm indexing = ~1440 hrs

Solr indexing = 2 hrs

Digital Library

6 GB of metadata

2,248,922 records

CONTENTdm indexing = ~144 hrs

Solr indexing = 17 min

SIMP Tool for metadata management

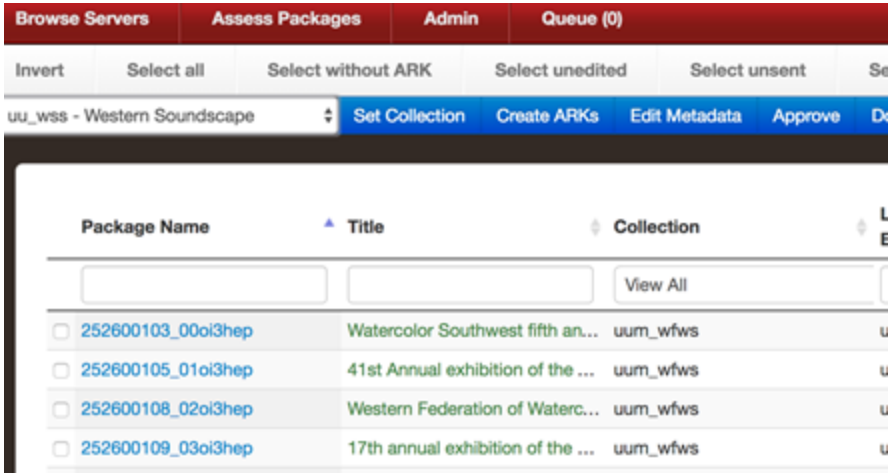
Initially designed to facilitate management between CONTENTdm and Rosetta, our preservation system

Platform agnostic and modular, provides ability to use other DAMs

Options for automatically extracting some metadata (format, OCR text)

Implementing controlled vocabularies (starting with type field)

Templates saved with static data



More details available in D-Lib article:
http://www.dlib.org/dlib/july14/neatrour/07neatrour.html

Metadata template customization in CDM

Fields gone wild!

735 unique field names before migration

441 unique field names after migration, with most of the fields mapped to a DC field

Local field labels like:

Subject 1, Subject 2, Subject 3, and Hidden Description

Orphaned collections with very little metadata

Collections that were 10 years old and hadn't been updated since they were digitized

Legacy collections created during an era where CONTENTdm was used for

Metadata issues we had to correct during migration

Metadata Standardization

Establish core fields needed for faceting and queries in new DAM, and change varying fields to the following:

- Creator
- Date
- Subject
- Rights (To be added after migration for faceting)
- Coverage-Spatial
- Type

Standardize Field Names to port over

- Capitalize first letter of all field names
- Remove periods and other unnecessary characters
- Delete all empty metadata fields
- Combine fields when necessary and use semicolons as separators to remove redundancy
- Change collection templates to match new standards

UDN migration vs. digital library migration

Utah Digital Newspapers

Newspapers metadata more standardized, but not perfect

Inconsistencies with titles, dates, paper names, and type fields

Clean-up for nonstandard characters, newspaper names issues

Previous architecture had one newspaper title split into multiple collections (scalability issues)

Consistent field names and metadata meant that

Digital Library

More communication needed with both internal and external partners during migration

Problem or missing metadata values in older collections not fixed (yet)

Field label standardization

Standardization of type and format

Lots of clean-up work still to do in the

Future plans for metadata fixes

Have been doing assessment and fixes prior to migration with some work completed by student workers, but scale of the problem means that even though some collections were fixed, there is a great deal of work left to do

Rights statement remediation for older collections - transition to rightsstatement.org for newer collections, but still need to assess and fix older, inaccurate rights statements (currently over 12,000 unique statements).

Assess field values against MWDL application profile. Add data to missing required fields.

Duplicate images

Duplicate titles

Thank you! Please contact us if you have any questions.

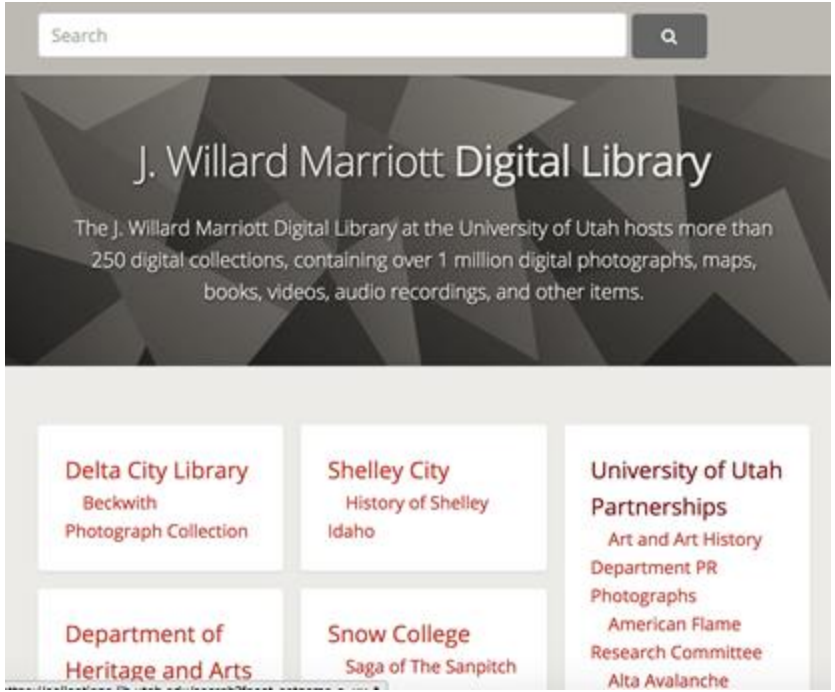
Anna Neatrour (anna.neatrour@utah.edu)

Jeremy Myntti (jeremy.myntti@utah.edu)

Brian McBride (brian.mcbride@utah.edu)

Alan Witkowski (alan.witkowski@utah.edu)

Matt Brunsvik (matt.brunsvik@utah.edu)



https://collections.lib.utah.edu/