Supercharged Digital Collections:
Moving to the fast lane with scalable open source infrastructure

Photos from University of Utah, Multimedia Archives Photo Collection, Gary Dean Brown
Presenters

Anna Neatrour
Brian McBride

Matt Brunsvik
Harish Maringanti

Other Presenters/contributors:

Jeremy Myntti
Alan Witkowski
What we’ll cover today

▸ History
▸ SIMP demo
▸ Migration
▸ Lessons learned
History of UU Marriott Digital Library

2001

Launched ContentDM
History of UU Marriott Digital Library

2002

Launched Utah Digital Newspapers (UDN)
History of UU Marriott Digital Library

2003

UDN receives IMLS NLG grant
History of UU Marriott Digital Library

2010

UDN reaches 1 Million pages milestone
History of UU Marriott Digital Library

- 2001
  - Purchased Rosetta for long term archiving
  - SIMP development started

- 2012
History of UU Marriott Digital Library

- 2001: SIMP launched
- 2013: DAMS review Committee report
History of UU Marriott Digital Library

- 2001
- 2015

Upgraded Server & Storage infrastructure
History of UU Marriott Digital Library

2001

UDN migrated (Jan – June)

All Other digital Collections migrated (June – Dec)

2016
Marriott Library Digital Collections

- 269 collections
- 717,858 IEs (items)
- 139,583 compound objects
- 12 file types (images, videos, pdfs...)
- 3 TB storage
- Over 50 partners (internal and external) hosting collections with the Marriott Library

https://collections.lib.utah.edu/
Utah Digital Newspapers (UDN)

- ~1.926 Million pages
- ~21.64 Million articles
- ~138 Newspaper titles
- Article level newspaper data
- Many unique papers not in Chronicling America

https://digitalnewspapers.org/
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/Preservation Systems

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

- Updated in 2016 to support Solphal

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

Patron ease of use
• End-user experience
• Search options and accuracy - discoverability of the content
• Handle all formats in similar fashion for all browsers
• ADA compliance

Types of content supported
• Major content categories (e.g. IR, newspapers, EAD)
• File types and datasets
• Tiers of content (both object and item level)
• Media streaming

Issues related to ingest/conversion/exit
• Workflows for adding new content
• Converting existing collection data
• Barrier(s) to exit

Collection administration
• Interface design
• User permissions
• Website configuration
• Statistics/Reporting/Logs

Metadata administration
• Metadata creation, editing, and maintenance options
• Faceting
• Copyright/Embargo
• Search Engine Optimization (SEO)

Integration with other library platforms
• Discovery layer (e.g. Primo)
• Digital preservation system (e.g. Rosetta)
• Locally developed tools (e.g. SIMP Tool)

Technical infrastructure/administration
• Scalability of collections/content
• Hardware/Server(s) specs
• API support
• System and server installation, configuration, and upgrades
• Internal resources required to run the system

Technical support
• Training
• Help desk

Future and strategic directions
DAMS Review Outcome - Solphal
Hardware Requirements - CONTENTdm vs Solphal

CONTENTdm

• 2 servers (UDN & Digital Collections)

• Digital Collections (2x12 Core 3.0 Ghz, 64 Gb mem) UDN (2x8 Core 2.4 Ghz, 96 Gb mem)

• Scalability issues

• Limited customizations

Solphal

• 3 VMs (indexing services, UDN, and Digital Collections)

• VMs avg. 2 cores, 3 Gb mem

• Extremely small footprint (hardware and power savings)

• Highly scalable and sustainable

• Fully customizable

• Easy replication and cloud-centric
Benefits of Solphal

• In-house expertise and support

• Student, patron, and partner convenience
  • faster response times
  • advanced search functions
  • front-end metadata editing
  • customized interface

• Substantial power, hardware, and licensing cost savings

• Improved workflows and QA/QC for all teams

• Staff time savings = ~1 less FTE supporting servers

• Demonstrating and sharing open source solutions with other institutions
# Indexing Improvements

## Utah Digital Newspapers

- 55 GB of metadata
- 21,648,462 records
- CONTENTdm
  - Full re-index ~ 1440 hrs (site is online)
  - Full re-index ~ 240 hrs (site is offline)
- Apache Solr indexing
  - Initial index ~ 2.75 hrs
  - Incremental index ~ instant

## Digital Collections

- 6 GB of metadata
- 2,248,922 records
- CONTENTdm
  - Full re-index ~ 144 hrs (site is online)
  - Full re-index ~ 72 hrs (site is offline)
- Apache Solr indexing
  - Initial index ~ 25 min
  - Incremental index ~ instant
User Interface Improvements

**Utah Digital Newspapers**
- CONTENTdm
  - Page load times (avg) 10s
  - Pdf display issues
- Solphal
  - Page load times (avg) under 500ms

**Digital Collections**
- CONTENTdm
  - Page load times (avg) 5s
  - Pdf display issues
- Solphal
  - Custom HTML5 image viewer
  - HTML response times (avg) under 50ms
  - Complete page load times (avg) under 500ms
Sample Solr Ingest File

```xml
<add>
  <doc>
    <field name="title_t">U Archives D. Art 1890-1939 fd 1 #2</field>
    <field name="form_t">photograph</field>
    <field name="source_t">University of Utah Archives</field>
    <field name="work_id_t">101</field>
    <field name="date_digital_t">4/20/2010</field>
    <field name="rights_t">Digital Image Copyright 2010, University of Utah. All Rights Reserved.</field>
    <field name="type_t">Image</field>
    <field name="imaging_technician_t">Amelia Walsh</field>
    <field name="metadata_cataloger_t">Amelia Walsh</field>
    <field name="filename_t">101.tiff</field>
    <field name="ark_t">ark:/81728/s65b7cp8</field>
    <field name="archival_file_t">Volume1/U Archives D. Art 1890-1939 fd 1 #2.tiff</field>
    <field name="created_tdc">2010-04-19T00:00:00Z</field>
    <field name="modified_tdc">2010-04-28T00:00:00Z</field>
    <field name="oldid_t">Art2 0</field>
    <field name="id">15729</field>
    <field name="parent_id">0</field>
    <field name="thumb_s">68/84/6884a28421c31d96dd9ec466f47feebb9af700.jpg</field>
    <field name="file_s">5c/bc/55bc0793020d24aba7eb6a44d3d0c65749705c57.jpg</field>
    <field name="format_t">image/jpeg</field>
    <field name="setname_s">uu_sah_artp</field>
    <field name="restricted_t">0</field>
  </doc>
</add>
```
UDN migration vs. Digital Collections 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 migration happened sooner and faster

• No external partners to contact

Digital Collections

• 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
## Metadata Standardization

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

- Creator
- Date
- Subject
- Rights
- Spatial Coverage
- Type
- Title

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Mr. J. Holman Waters: Recollections of Samuel Newhouse, Salt Lake City, Utah: an interview by Hynda Rudd, July 16, 1977</td>
</tr>
<tr>
<td>Creator</td>
<td>Waters, J. Holman, 1912-1995</td>
</tr>
<tr>
<td>Date</td>
<td>1977-07-16</td>
</tr>
<tr>
<td>Spatial Coverage</td>
<td>Salt Lake City, Salt Lake County, Utah, United States</td>
</tr>
<tr>
<td>Subject</td>
<td>Newhouse, Samuel, 1853-1930; Businessmen--Utah--Salt Lake City--Interviews; Jews--Utah--Salt Lake City</td>
</tr>
<tr>
<td>Type</td>
<td>Text</td>
</tr>
<tr>
<td>Rights</td>
<td>Digital Image © 2016 University of Utah. All Rights Reserved. Contact the J. Willard Marriott Library Manuscripts Division for use and restrictions. 801-585-3082</td>
</tr>
</tbody>
</table>
Metadata fields gone wild!
## Variations in title field name

<table>
<thead>
<tr>
<th>Collection Alias</th>
<th>Local Field Name</th>
<th>CONTENTdm nickname</th>
</tr>
</thead>
<tbody>
<tr>
<td>col/ehsl-wht</td>
<td>Title</td>
<td>title</td>
</tr>
<tr>
<td>col/epubs</td>
<td>Article Title</td>
<td>articl</td>
</tr>
<tr>
<td>col/fhwa</td>
<td>Document Title</td>
<td>title</td>
</tr>
<tr>
<td>col/fitz</td>
<td>Title</td>
<td>title</td>
</tr>
<tr>
<td>col/flwright-jp2</td>
<td>Alternative Title</td>
<td>latin</td>
</tr>
<tr>
<td>col/flwright-jp2</td>
<td>Title</td>
<td>title</td>
</tr>
<tr>
<td>col/GBA_Photos</td>
<td>Caption</td>
<td>title</td>
</tr>
<tr>
<td>col/GBM</td>
<td>Caption</td>
<td>title</td>
</tr>
<tr>
<td>col/gis</td>
<td>Title</td>
<td>title</td>
</tr>
</tbody>
</table>
## Controlled Vocabularies

<table>
<thead>
<tr>
<th>Type (DCMI Type vocab)</th>
<th>Format (Internet Media Type)</th>
<th>Language (ISO 639-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td>application/msword</td>
<td>ara</td>
</tr>
<tr>
<td>Dataset</td>
<td>application/pdf</td>
<td>chi</td>
</tr>
<tr>
<td>Event</td>
<td>application/vnd.google-earth.kmz</td>
<td>eng</td>
</tr>
<tr>
<td>Image</td>
<td>application/vnd.ms-powerpoint</td>
<td>fre</td>
</tr>
<tr>
<td>Image/MovingImage</td>
<td>application/xml</td>
<td>ger</td>
</tr>
<tr>
<td>Image/StillImage</td>
<td>application/x-shockwave-flash</td>
<td>heb</td>
</tr>
<tr>
<td>InteractiveResource</td>
<td>application/zip</td>
<td>ita</td>
</tr>
<tr>
<td>PhysicalObject</td>
<td>audio/mpeg</td>
<td>kor</td>
</tr>
<tr>
<td>Service</td>
<td>image/jpeg</td>
<td>lat</td>
</tr>
<tr>
<td>Software</td>
<td>image/png</td>
<td>nav</td>
</tr>
<tr>
<td>Sound</td>
<td>text/html</td>
<td>rus</td>
</tr>
<tr>
<td>Text</td>
<td>video/mp4</td>
<td>spa</td>
</tr>
</tbody>
</table>
Delete Unneeded Legacy Collections

Please take a moment to fill out our survey

- 100 THINGS TO DO BEFORE U JOIN THE REAL WORLD
  How do you get a job without experience and get experience without a job? It's the question virtually every college student or recent graduate faces. In Getting from College to Career, Lindsey Pollak offers the first definitive guide to building...

- 10 THINGS EMPLOYERS WANT YOU TO LEARN IN COLLEGE
  A college diploma can open many doors, but it's just an expensive piece of paper if you don't have what employers seek. Long-time professor and student advisor Bill Coplin surveyed the nation's top companies and identified the essential skills...

- 100 BEST CAREERS FOR WRITERS AND ARTISTS
  What and where the jobs are for people with creative talents! For all the aspiring writers and artists who want to put their creative talents to good use, here's the guide that can show them how! Compiled by one of America's leading career...
Future metadata corrections - next steps

Review 441 fields manually, noted issues (field used in only one collection, field synonyms)

Develop spreadsheet to cluster possible fixes by partner, write up metadata recommendations - ongoing process

To do:

• Explore possibilities of assessment scripts against solr to review metadata values
• Explore visualization tools
<table>
<thead>
<tr>
<th>Number</th>
<th>Rights Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>40867</td>
<td>This material may be protected by copyright. Permission required for use in any form. For further information please contact the Multimedia Archives, Special Collections, J. Willard Marriott Library, University of Utah.</td>
</tr>
<tr>
<td>20207</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/1.0/">http://creativecommons.org/licenses/by-nc-sa/1.0/</a></td>
</tr>
<tr>
<td>4325</td>
<td>Digital Image Copyright 2010. University of Utah. All Rights Reserved.</td>
</tr>
<tr>
<td>3597</td>
<td>Copyright owner unknown</td>
</tr>
<tr>
<td>3563</td>
<td>Digital image copyright 2005. University of Utah. All rights reserved.</td>
</tr>
<tr>
<td>3488</td>
<td>Digital image. copyright 2003 Uintah County Library</td>
</tr>
<tr>
<td>3405</td>
<td>This material is protected by copyright. Express permission is required for use in any form. For further information please contact the Multimedia Archivist, Special Collections, J. Willard Marriott Library, University of Utah.</td>
</tr>
<tr>
<td>3033</td>
<td>This material may be protected by copyright. Permission required for use in any form. Please contact the Multimedia Archives, Special Collections, J. Willard Marriott Library, University of Utah.</td>
</tr>
<tr>
<td>2975</td>
<td>This material is protected by copyright. Permission required for use in any form. For further information please contact the Multimedia Archives, Special Collections, J. Willard Marriott Library, University of Utah.</td>
</tr>
<tr>
<td>title</td>
<td>date</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>National Ski Association officers and directors</td>
<td>January 13 - 15, 1961</td>
</tr>
<tr>
<td>Group Picture in the Orderville Ward</td>
<td></td>
</tr>
<tr>
<td>Saltair Beach Pavilion and bathers, Saltair Resort.</td>
<td>1893-1925 destroyed by fire</td>
</tr>
<tr>
<td>Interior of the Dancing Pavilion, Saltair Resort.</td>
<td>1893-1925 destroyed by fire</td>
</tr>
<tr>
<td>Saltair Beach Pavilion, Saltair Resort.</td>
<td>1893-1925 destroyed by fire</td>
</tr>
<tr>
<td>Group of ski tourers high in the Wasatch mountains</td>
<td>A group photo of ski tourers in the Wasatch Mountains</td>
</tr>
<tr>
<td>Painting of soldiers.</td>
<td></td>
</tr>
<tr>
<td>George Thompson in England.</td>
<td></td>
</tr>
<tr>
<td>George Thompson with second wife Katherine.</td>
<td></td>
</tr>
<tr>
<td>William Jennings Bryan at the radio station.</td>
<td></td>
</tr>
<tr>
<td>Nels Just with Emma Bennett Just</td>
<td></td>
</tr>
<tr>
<td>Emma Bennett Just's son Fred Bennett.</td>
<td></td>
</tr>
<tr>
<td>Charley and James Just</td>
<td></td>
</tr>
<tr>
<td>Nels Just</td>
<td></td>
</tr>
<tr>
<td>Thomas Moran</td>
<td></td>
</tr>
<tr>
<td>Just family by their ranch house</td>
<td></td>
</tr>
<tr>
<td>Francis and Will Just</td>
<td></td>
</tr>
<tr>
<td>Emma Bennett Just at age 16.</td>
<td></td>
</tr>
</tbody>
</table>
## Future metadata corrections - Subject faceting

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>American newspapers</td>
<td>1</td>
</tr>
<tr>
<td>Antisemitism</td>
<td>12</td>
</tr>
<tr>
<td>Auschwitz (Concentration camp)</td>
<td>1</td>
</tr>
<tr>
<td>Bamberger Railroad Company</td>
<td>2</td>
</tr>
<tr>
<td>Bar mitzvah</td>
<td>1</td>
</tr>
<tr>
<td>Cantors (Judaism)</td>
<td>1</td>
</tr>
<tr>
<td>Conservative Judaism</td>
<td>1</td>
</tr>
<tr>
<td>Eastern Front (World War (1914-1918))</td>
<td>1</td>
</tr>
<tr>
<td>Emigration and immigration</td>
<td>3</td>
</tr>
<tr>
<td>Ethnic groups</td>
<td>1</td>
</tr>
<tr>
<td>Evacuation and relocation of Japanese Americans (United States : 1942-1945)</td>
<td>1</td>
</tr>
</tbody>
</table>
Roadmap for future development

**Solphal (Front end) development**
- International Image Interoperability Framework (IIIF)
- RDFa and schema.org support

**SIMP Tool enhancements**
- Workflow automation
- Large updates monthly, small updates as needed

**Large projects**
- IMLS Grants
  - Newspapers in Hydra (LG-70-17-0043-17)
  - Western Name Authority File (LG-72-16-0002-16)
- Born digital content
- Partner projects
- Data Repository
Lessons Learned

Wild wild west of metadata wrangling

Don’t keep shoveling while it is still snowing!

Shooting of Patrick Coughlin
https://collections.lib.utah.edu/details?id=962118

Unidentified man holding snow shovel in front of his house, (family, friend or acquaintance of Norman D. Nevills).
https://collections.lib.utah.edu/details?id=984034
Thank You!

Contact us at:

Jeremy Myntti (jeremy.myntti@utah.edu)

Anna Neatrour (anna.neatrour@utah.edu)

Harish Maringanti (harish.maringanti@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/
https://newspapers.lib.utah.edu/