IIIF, Metadata Aggregation, and Structured Transformation
The Implications for Improving Discovery

Shane Huddleston, Product Manager
[Jeff Mixter, Software Engineer]
[Bruce Washburn, Software Engineer]
What is CONTENTdm?

- OCLC’s digital repository cloud service
- 2600+ libraries worldwide
- 65+ million digital objects
- Diverse descriptive metadata
- IIIF Image & Presentation APIs
What is IIIF?

- Community-focused and growing
- Defines APIs for access to digital content
- Encourages application development
- Favors real-world use & developer happiness
- Supports research and scholarship
What are the APIs?

- **Image** – image properties and transformation
- **Presentation** – structural info and metadata
- **Search** – user queries within a digital object
- **Authentication** – interaction pattern only
- **Change Discovery** – activity notification feed
Five Hypotheses

- W3C Activity Streams can be created from CONTENTdm collections
- An API can make Activity Streams accessible
- Data from the API can drive web crawling of CONTENTdm IIIF Manifests
- CONTENTdm discovery can be improved by indexing useful metadata
- Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core
HYPOTHESIS 1:
W3C Activity Streams can be created from CONTENTdm collections

TRUE:
✓ List of CONTENTdm manifests & creation dates built as Activity Streams data
✓ A manual process to assemble the list of manifests
✓ Crawling is automated & repeated monthly

However...
- Only image records represented
- Item context pulled from other sources (collection & organization descriptions)
HYPOTHESIS 2:
An API can make Activity Streams accessible

TRUE:
- ✓ IIIF Change Discovery API similar to OAI-PMH and ResourceSync
- ✓ JSON service with list of records created/updated/deleted sorted in reverse chronological order

However...
- o The experimental API has not yet been exercised by external users
- o Operational support and deployment is provided by developers, not as an OCLC production service
HYPOTHESIS 3:
Data from the API can drive web crawling of CONTENTdm IIIF Manifests

TRUE:
✓ IIIF outlines a processing algorithm for the Activity Streams API
✓ We harvested our experimental API endpoint to index 13 million Manifests

However...
○ Harvest rate limited to prevent potential abuse
○ Testing revealed some manifest issues, which are being remedied
HYPOTHESIS 4:

CONTENTdm discovery can be improved by indexing useful metadata

We think so:

✓ An aggregated index across all collections provides one-stop keyword searching
✓ We are finding unexpected things in unexpected places

However...

○ Metadata sometimes describes the digitized item, sometimes the physical
○ Discovery expectations shaped by Europeana and DPLA in our domain cannot be met
HYPOTHESIS 5:
Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

Not exactly:
- Fields in CONTENTdm can be mapped to any Dublin Core element
- We looked closely at DC Type, Format, Medium, Temporal, Spatial, and Audience

But...
- Mapping practices are inconsistent
- Automated reconciliation is strongly dependent on source data quality
- Remediation requires attention upstream and domain expertise
**HYPOTHESIS 5:** Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

<table>
<thead>
<tr>
<th>Original field string</th>
<th>DC Type</th>
<th>LC TGM Term</th>
<th>Getty AAT Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>black-and-white negatives</td>
<td>Image</td>
<td>Negatives</td>
<td>black-and-white negatives</td>
</tr>
<tr>
<td>9 1/2 x 7 pen &amp; ink drawing</td>
<td>Image</td>
<td>Drawings</td>
<td>pen and ink drawings</td>
</tr>
<tr>
<td>programs (documents)</td>
<td>Text</td>
<td>Documents</td>
<td>programmes (documents)</td>
</tr>
<tr>
<td>1 letter (2 p.)</td>
<td>Text</td>
<td>Correspondence</td>
<td>letters (correspondence)</td>
</tr>
</tbody>
</table>
HYPOTHESIS 5: Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

<table>
<thead>
<tr>
<th>Original Audience field string</th>
<th>LC Demographic Term</th>
<th>Audience category</th>
</tr>
</thead>
<tbody>
<tr>
<td>genealogists and local history researchers</td>
<td>Genealogists</td>
<td>Occupation</td>
</tr>
<tr>
<td>graduate</td>
<td>Graduate students</td>
<td>Education level</td>
</tr>
<tr>
<td>elementary k-8</td>
<td>School children</td>
<td>Education level</td>
</tr>
<tr>
<td>française</td>
<td>French speakers</td>
<td>Spoken language</td>
</tr>
<tr>
<td>american indian/navajo</td>
<td>Navajo (North American people)</td>
<td>Nationality</td>
</tr>
<tr>
<td>children ages 2-5 years</td>
<td>Children</td>
<td>Demographic</td>
</tr>
</tbody>
</table>
**HYPOTHESIS 5:** Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

<table>
<thead>
<tr>
<th>Original date string</th>
<th>Start date</th>
<th>End date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940/1965-01-21</td>
<td>1940-01-01</td>
<td>1965-01-21</td>
</tr>
<tr>
<td>twentieth century, c. e.</td>
<td>1900-01-01</td>
<td>1999-12-31</td>
</tr>
<tr>
<td>1960s</td>
<td>1960-01-01</td>
<td>1969-12-31</td>
</tr>
<tr>
<td>(1789-1820) north carolina's early statehood</td>
<td>1789-01-01</td>
<td>1820-12-31</td>
</tr>
<tr>
<td>deuxième guerre mondiale</td>
<td>1939-09-01</td>
<td>1945-09-02</td>
</tr>
</tbody>
</table>
HYPOTHESIS 5:
Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core.
HYPOTHESIS 5: Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

Geo-Coordinates

39.768074,-86.16198
39.762053,-86.151266
39.768628,-86.15603
**HYPOTHESIS 5:** Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

<table>
<thead>
<tr>
<th>Local CONTENTdm Field</th>
<th>Dublin Core Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic subject (street address)</td>
<td>Spatial</td>
</tr>
<tr>
<td>Geographic subject (city or populated place)</td>
<td>Spatial</td>
</tr>
<tr>
<td>Geographic subject (county)</td>
<td>Spatial</td>
</tr>
<tr>
<td>Geographic subject (state/province)</td>
<td>Spatial</td>
</tr>
<tr>
<td>Geographic subject (country)</td>
<td>Spatial</td>
</tr>
<tr>
<td>Geographic subject (other)</td>
<td>Spatial</td>
</tr>
<tr>
<td>Geographic coordinates</td>
<td>Spatial</td>
</tr>
</tbody>
</table>
Five Findings

- Activity Streams and the IIIF Change Discovery API is a sound and stable syndication architecture
- Aggregation adds value
- Structured data can be reconciled to provide authority control for searching
- CONTENTdm data is too varied and incomplete to support downstream reconciliation
- The potential for deep and meaningful discovery can be realized if data is provided as structured, linked data at the source
Next steps: Applying what we learned

✓ Expanding support for IIIF APIs in CONTENTdm

✓ Evaluate whether reconciliation tools can be effective and feasible at scale

✓ Give domain experts those tools to produce reconciled structured data