OCLC Open Collections Symposium – June 19, 2019

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…

- The experimental API has not yet been exercised by external users
- 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

Original field string	DC Type	LC TGM Term	Getty AAT Term
black-and-white negatives	Image	Negatives	black-and-white negatives
9 1/2 x 7 pen & ink drawing	Image	Drawings	pen and ink drawings
programs (documents)	Text	Documents	programmes (documents)
1 letter (2 p.)	Text	Correspondence	letters (correspondence)

HYPOTHESIS 5: Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

Original Audience field string	LC Demographic Term	Audience category
genealogists and local history researchers	Genealogists	Occupation
graduate	Graduate students	Education level
elementary k-8	School children	Education level
française	French speakers	Spoken language
american indian/navajo	Navajo (North American people)	Nationality
children ages 2-5 years	Children	Demographic

HYPOTHESIS 5: Structured, linked data can be derived from CONTENTdm fields mapped to Dublin Core

Original date string	Start date	End date
1940/1965-01-21	1940-01-01	1965-01-21
twentieth century, c. e.	1900-01-01	1999-12-31
1960s	1960-01-01	1969-12-31
(1789-1820) north carolina's early statehood	1789-01-01	1820-12-31
deuxième guerre mondiale	1939-09-01	1945-09-02

Organization

Audience

unspecified (48,917)

Nationality (6)

Type

Dates

Filter by Date

Start date: 1800

End date: 2019

search for manifests

type.id:http://vocab.getty.edu/aat/300026879

Search

48,923 search results for : type.id:http://vocab.getty.edu/aat/300026879

Letters (correspondence)

Free Public Library of Newark, New Jersey

Newark Eagles

View Add

Letter, W. N. Mitchell to R. C. Mitchell, Camp DuBois, January 22, [18]62

Missouri State Library

Confluence & Crossroads - The Civil War in the American Heartland

View Add

Paul Henderson letter to Elizabeth C. Clarke - Febraury 1, 1923

Missouri State Library

Over There - Missouri and the Great War

View Add

S.B. Gorham letter to Mrs. D.B. Brady - December 22, 1917

Missouri State Library

Over There - Missouri and the Great War

View Add

Robert Kirk Brady Letter to Folks - December 18, 1918

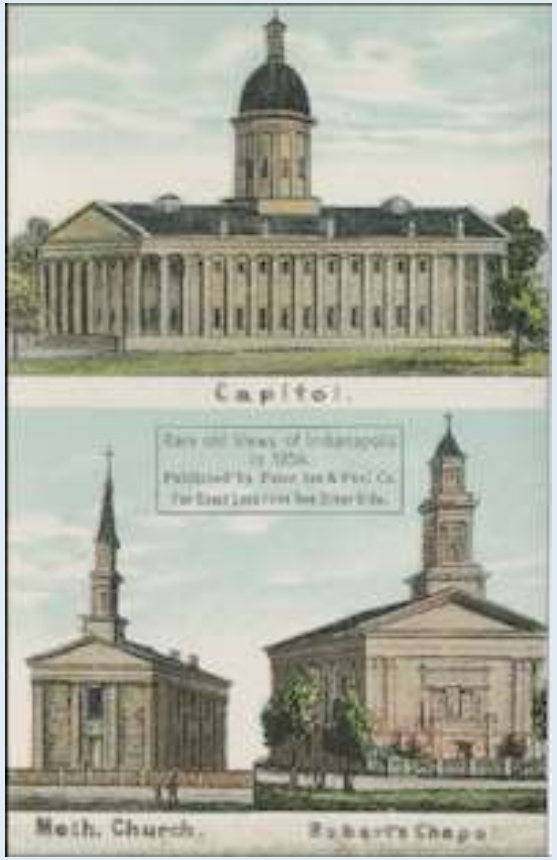
Eugene V. Debs letter to Frank P. O'Hare - February 22, 1918

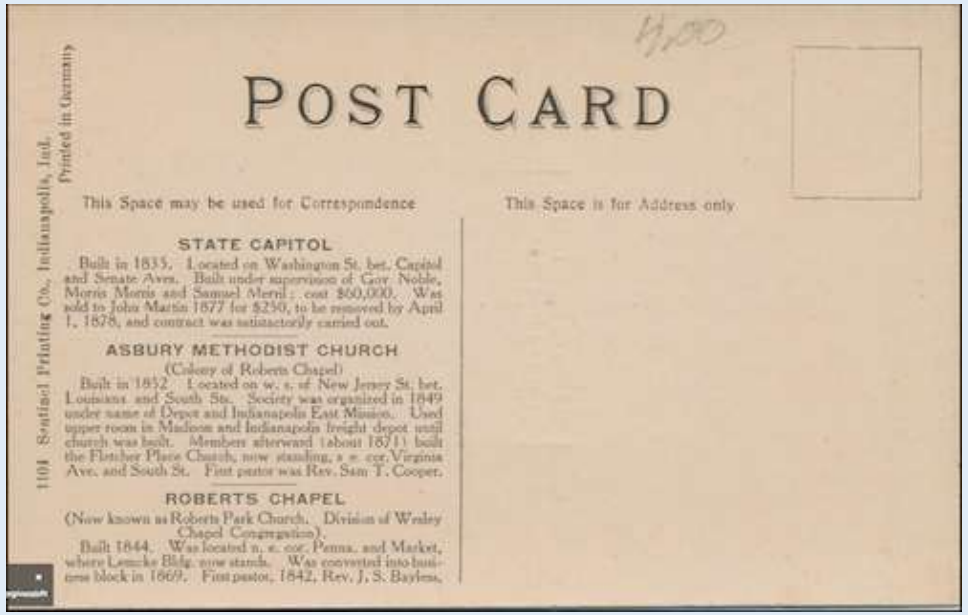
J. Bird Humphrey letter to Will Carleton 1871-10-16

Letters (correspondence)

Free Public Library of Newark, New

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

Local CONTENTdm Field	Dublin Core Field
Geographic subject (street address)	Spatial
Geographic subject (city or populated place)	Spatial
Geographic subject (county)	Spatial
Geographic subject (state/province)	Spatial
Geographic subject (country)	Spatial
Geographic subject (other)	Spatial
Geographic coordinates	Spatial

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**