OCLC RESEARCH DISCUSSION SERIES

Next Generation of Metadata

#OCLCmetadataseries
Welcome and Introduction

Rachel Frick
Executive Director Research Library Partnership, OCLC
Housekeeping rules

• You are currently in 'listening only' mode.
• If you experience any technical difficulties, please contact the WebEx host via the Chat, look for the “Host” option in the drop-down menu.
• If you have any questions, please put them in the Chat. Look for the “Host and Panelist” option. All questions will be addressed during the Q&A.
• Today’s webinar will be recorded. The recording will be published online afterwards.
OCLC RESEARCH DISCUSSION SERIES

Next Generation of Metadata

#OCLCmetadata series
After a decade, it is okay to keep asking, “Why linked data?”
What is OCLC doing to help libraries prepare for next generation metadata?

1. **Cultivating understanding** of this “next generation” metadata ecosystem

2. **Experimenting** with new data models, semantic web technologies, workflows, methods, and tools

3. **Building** a “Shared Entity Management Infrastructure”
This report synthesizes six years of OCLC Research Library Partners Metadata Managers Focus Group discussions.

oc.lc/nextgen-metadata-report

#OCLCmetadata
Experimenting and Building

- Publish linked data - FAST, VIAF, WorldCat (2009 - )
- EntityJS Research Project (2013)
- Person Entity Lookup Pilot (2014)
- CONTENTdm Metadata Refinery (2015-16)
- Project Passage (2017-18)
- CONTENTdm Linked Data Pilot (2019-20)
- Shared Entity Management Infrastructure (2020-21)
VIAF, FAST, and WorldCat: Publish linked data on the web with a UI, API, and downloadable datasets
2019-2021 and next steps

- CONTENTdm Linked Data Pilot
- Shared Entity Management Infrastructure
- More Research
- More convening, more understanding, more sharing

#OCLCmetadataSeries
Cultivating understanding: the Metadata Series

OPENING PLENARY WEBINAR
Tuesday 23 February 2021,

INTERACTIVE ROUND TABLE
During the first two weeks of March 2021.

CLOSING PLENARY WEBINAR
Tuesday 13 April 2021,

#OCLCmetadataseries
Opening Plenary • 23 February 2021

Transitioning to the Next Generation of Metadata

Dr. Annette Dortmund
Senior Product Manager & Research Consultant, OCLC

@libsun
https://orcid.org/0000-0003-1588-9749

#OCLCmetadataseries
Karen Smith-Yoshimura
OCLC Research Senior Program Officer (retired Nov 30, 2020)
• The Transition to Linked Data and Identifiers
• Describing “Inside-Out” and “Facilitated” Collections
• Evolution of “Metadata as a Service”
• Preparing for Future Staffing Requirements

https://doi.org/10.25333/rggd-b343
Format-specific metadata management based on curated text strings in bibliographic records understood only by library systems is nearing obsolescence, both conceptually and technically.

In short, the metadata could be better, there is not enough of it, and the metadata that does exist is not used widely outside the library domain.
Transition to Linked Data & Identifiers

"Persistent identifiers were viewed as crucial to transitioning from current metadata to future applications."
"Identity management poses a change in focus...to describing entities ...and the relationships among them."
“Addressing language issues is important as libraries seek to develop relationships and build trust with marginalized communities.”
Describing “Inside-Out” and “Facilitated” Collections

- Archival collections
- Archived websites
- Audio and video collections
- Image collections
- Research data

https://researchworks.oclc.org/iiif-explorer/

“Metadata underlies all discovery regardless of format, now and in the future…”
Libraries’ expertise in metadata standards, identifiers, linked data, and data sharing systems as well as technical systems can be invaluable to the research life cycle.
Evolution of “Metadata as a Service”

New applications

Plus:
- Metrics
- Consultancy
- Semantic indexing

Bibliometrics
Preparing for Future Staffing Requirements

A culture shift is needed: from pride in production alone to valuing opportunities to learn, explore, and try new approaches to metadata work.
Conclusion

The next generation of metadata will become even more focused on entities rather than record-based descriptions of an institution’s collections.

Good linked data requires good metadata.
• The Transition to Linked Data and Identifiers
• Describing “Inside-Out” and “Facilitated” Collections
• Evolution of “Metadata as a Service”
• Preparing for Future Staffing Requirements

QUESTIONS?
Thank you!

Dr. Annette Dortmund
Sr. Product Manager & Research Consultant, OCLC
annette.dortmund@oclc.org

@libsun
https://orcid.org/0000-0003-1588-9749

Because what is known must be shared.
Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project

Titia van der Werf
Senior Program Officer, OCLC Research
Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project

Greta Bahnemann
Minnesota Digital Library

Michael Carroll
Temple University Libraries

Paul Clough
University of Miami Libraries

Mario Einaudi
The Huntington Library, Art Museum, and Botanical Gardens

Chatham Ewing
Cleveland Public Library

Jeff Mixter
OCLC Research

Jason Roy
Minnesota Digital Library

Holly Tomren
Temple University Libraries

Bruce Washburn
OCLC Research

Elliot Williams
University of Miami Libraries

#OCLCmetadataseries
The CONTENTdm Linked Data Pilot questions

1. Divergent practice and collection assessment
2. Shared data models for diverse collections and institutions
3. Machine learning and human intervention
4. Tools for subject matter experts
5. Discovery tools
6. The paradigm shift
The CONTENTdm Linked Data Pilot

- Manually reviewed, mapped and reconciled the metadata
- Imported the data into Wikibase
- Used Wikibase as a sandbox
- Involved the community to co-create and learn together
- Tested tools and workflows
The CONTENTdm Linked Data Pilot

New applications:

1. The Field Analyzer
2. The Image Annotator
3. The Retriever
4. The Describer
5. The Explorer
Findings

- It takes a lot of human effort to create the structured data
- Wikibase is a powerful and flexible infrastructure for creating, managing, and curating structured data
- There is a lot of potential for enhancing existing metadata about cultural heritage items
REFLECTION: Rethink the systematic cleanup of our legacy metadata

“the Field Analyzer, proved so useful that it stands above all the others. This tool enabled us to **review all our collections systematically and plan cleanup more effectively.** (…) 

We will use the knowledge gained from this project to **rethink our workflows and our descriptive metadata with an eye toward the promise of linked data.”**
REFLECTION: Reimagining data curation

“An overarching question driving the linked data project was, for a paradigm shift of this magnitude, **how can the foundational changes be made more scalable, affordable, and sustainable?**

The project showed that **the scope and magnitude of the effort required** to completely analyze, transform, and reconcile all current descriptive metadata into consistently modeled linked data **is beyond the reach of a single centralized agency.**

It will require substantial and shared resource **commitments from a decentralized community of practitioners** who will need to be supplied with easily accessible **tools and workflows** for carrying out the transition.”
REFLECTION: Enhancing discovery beyond collections

“One of the most important value propositions of working with linked data is for entities to link to other related things in other systems, leveraging the network to obtain more contextual data “on the fly” instead of duplicating data across systems.”
“By bringing and storing ‘national’ data into our local systems we are taking away some of the power of linked data; power that comes in the form of networked vocabularies that work best in a layer above our localized instances.

Linked data is powerful, in part because it is not tied to any one system, but rather, integrates content across collections, thereby creating user-discoverable connections across collections and, more importantly, repositories.”
QUESTIONS?
OCLC’s Entity Management Infrastructure

John Chapman
Senior Product Manager, Metadata Strategy & Operations

https://orcid.org/0000-0002-5388-5063
WHY A
“METADATA INFRASTRUCTURE?”
Feedback from OCLC member libraries

- Provide persistent identifiers relevant to library workflows
- Enable the creation of new identifiers within metadata management workflows
- Provide interfaces and ecosystem to create native linked data descriptions
- Seed the web with persistent identifiers
- Provide broad reconciliation across vocabularies and ontologies

oc.lc/passagereport

#OCLCmetadataSeries
Our goals

• Address infrastructure needs identified by libraries
  • Stand behind entity URIs
  • Provide ID creation services to help “at the point of need”
  • Expand on “native” metadata management
  • Link library data to non-library data… and shared data to local data
• Operate at a large scale – and be sustainable
• Complement other efforts
• Deliver products and services December 2021
Timeline of activities

2020

- Wikibase: Millions of entities
- Simple search & read

2021

- Ramp up: 10s of millions of entities
- Creation and editing tools

- Add more data sources
- Advanced querying, new UI tools

- Scale and refine
- Prepare for release

#OCLCmetadataseries
WHAT IS IT?

Shared Entity Management Infrastructure (2020-21)

#OCLCmetadataseries
What is the “Infrastructure”? 

• Community-curated Knowledge Graph
• Integration of facts from library data from around the world
  – Seeded from the knowledge contained in bibliographic authority files, WorldCat creative works, and controlled vocabularies
• Provenance and context of the knowledge claims as the facts come from a variety of heterogeneous sources
• Published following Linked Data Principles, a set of APIs and query endpoints
1. **Knowledge Creation**: Integration of heterogeneous data sources, through ‘Semantic lifting’.
2. **Knowledge Hosting**: Storage of the knowledge in a suitable way (e.g., semantic repository, a graph database, triple store).
3. **Knowledge Curation**: Make sure that the correctness and completeness of the Knowledge Graph satisfy ongoing needs.
4. **Knowledge Deployment**: Applications, APIs use the graph.

---

Knowledge Graphs - Methodology, Tools and Selected Use Cases. Springer (2020)
Done in 2020

• Entity pipeline
  – Extracted, transformed, loaded multiple sources to graph
  – Studied the landscape (probabilistic/fuzzy matching, gazetteer)

• Stable, repeatable Knowledge Hosting
  – Continued the learning with Wikibase
  – Focused on Loading at scale

• Creation/curation at scale
  – Measures, models, tools
Next steps on architecture, systems

- Multilingual approaches
- Moving beyond the Wikibase structure
- Integrating input on data models
- Building out curation support
WORKFLOWS AND LINKING
Example Workflow

Metadata

- Determine that record needs authority information
- Add persistent identifier that links to Entity Manager entity

Entity Manager

- Search for the entity
- Identify the correct entity (disambiguate)
- Open the entity
- Create a new entity
- Update the entity, as necessary
How entities are built from WorldCat data
Multiple links for multiple entity types
Next steps on workflows

• API and UI testing
• Wireframe review with Advisory Group (ongoing)
• Prioritization and scheduling of UI features
DATA ACTIVITIES
Processes

• Staff focused on two areas: modeling and quality
• For modeling work, documenting:
  – MVE description
  – SPARQL queries to validate MVE model
  – Data selection – sources, and logic used to select data
Works modeling

• Based on WorldCat clustering
  – Elements of Work and Expression cluster together
  – Manifestation elements from the record, Item elements in WMS or local system

• Supported by models and ontology to support tracking of provenance
Minimum Viable Entity (MVE): Work

• Some elements from Wikibase: label, description, also known as (when applicable)
• Remaining elements based on LRM and BIBFRAME, i.e., a combination of Work and Expression elements: instance of, title, agent, realization date (often based on publication date for first known realization), content type, “exemplar identifier” (points to a thing in WorldCat)
Scale

- Refining processes for ontology definition and data loading
- >90M entities
  - Roughly 80% works, 20% persons (<.01% places)
Quality Composite

Accurancy

- Consistency
- Currency
- Precision
- Relevancy
- Timeliness

Disambiguation

Confidence

- Completeness
- Accessibility
- Provenance

Structural

- Definition
- Syntax
Data in 2021

- Continue to build out data models and entity description
- Further work on Quality Composite
- Ontology development
- Broader testing
WHAT HAVE WE LEARNED?
What we have learned so far

• Need to increase capabilities for monitoring quality, breadth, depth
• APIs, machines as “users”
• Need redundancy, multiple environments, and robust testing capabilities
• Need to engineer loading and ingest technologies
Thank you!

John Chapman
Senior Product Manager,
Metadata Strategy & Operations

chapmanj@oclc.org

https://orcid.org/0000-0002-5388-5063
QUESTIONS?
The Discussion Series: towards a shared perspective of the future metadata landscape

Titia van der Werf
Senior Program Officer, OCLC Research
Emerging trends

1. Promoting the re-use of library data

2. The shift from Dublin Core metadata to structured heritage data

3. The shift from “authority control” to “entity management”
Promoting the re-use of library data

source: International Linked Data Surveys for Implementers (2014-2018)
The shift to structured heritage data

• CONTENTdm linked data pilot project
• Europeana
• Wikidata GLAM projects
• DERA – Digital Heritage Reference Architecture
From authority control to entity management

- OCLC’s Shared Entity Management Infrastructure (SEMI)
- French National Entities File (FNE)
- Wikidata/CrossRef/ORCID/ISNI/etc.
- Ecosystem of Wikibase instances
- Project HERCULES
Main question for the discussions

How do we make the transition to the Next Generation of Metadata happen at the **right scale** and in a **sustainable manner**, building an **interconnected ecosystem**, not a garden of silos?
QUESTIONS?
Don't miss the Closing Plenary Session!

Tuesday 13 April 2021
15:00 (CET)

Register at:
oc.lc/metadata-series