OCLC and Linked Data: The transition to contextual metadata

Shane Huddleston
Product Manager, CONTENTdm

John Chapman
Senior Product Manager, Metadata Strategy and Operations

Nathan Putnam
Director, Metadata Quality
Shane Huddleston
Product Manager, CONTENTdm

John Chapman
Senior Product Manager, Metadata Strategy and Operations
CONTENTdm Linked Data Pilot: Transforming metadata and improving discoverability

Shane Huddleston
Product Manager, CONTENTdm
Building on our experience
Building on our experience
Building on our experience

Publish linked data - FAST, VIAF, WorldCat (2009 - )

EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)
Building on our experience

Publish linked data - FAST, VIAF, WorldCat (2009 - )

EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)

CONTENTdm Metadata Refinery (2015-16)
Building on our experience

- 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)
Building on our experience

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)
Entity Management Infrastructure (2020-21)
Project Passage (2017-18)
CONTENTdm Linked Data Pilot (2019-20)
Entity Management Infrastructure (2020-21)
Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Entity Management Infrastructure (2020-21)
Overview of the pilot

• Three-phase, one-year project
• Leveraging what we’ve learned from past linked data projects
• Unique digital items are well-suited to entity-based description
• Working with real data in partnership with libraries
The report

- Published in 2021
- Summarizes the activity of the project over the year
- Key findings and conclusions

oc.lc/transform-linked-data
KEY FINDINGS
Benefits of a linked data environment

• Manage richer metadata with greater efficiency
• Add contextual information that better reflects knowledge in the real world
• Help researchers achieve a fuller understanding of collection materials to increase engagement
Potential to develop a shared data model

- Headings are associated with linked data entities and reused
- Relationships between entities support connecting and aggregating related items
- Connecting items can make management more efficient and discovery more intuitive
Challenges

• Converting existing text headings will be difficult to do at scale
• Analyzing, transforming, and reconciling is beyond the reach of a single central agency
• Data transformation needs to be shared and workflows decentralized
Success through cooperation

• Linked data transformation is a paradigm shift requiring long-term strategies
• Working partnerships represent strength in numbers
• Sharing practices and expertise are critical
Shared Entity Management Infrastructure: Where we are

John Chapman
Senior Product Manager, Metadata Strategy and Operations
Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Entity Management Infrastructure (2020-21)
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 & ontologies

oc.lc/passagereport
Project overview

- Two-year, $2.436M grant, matched by OCLC
- Production infrastructure for Work and Person entities
- Support for multiple descriptive and encoding standards
- Use of persistent identifiers
- Most importantly: a collaboration with the library community
Timeline of activities

2020

January - June

- Wikibase: Millions of entities
- Simple search & read

July - December

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

2021

January - June

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

July - December

- Scale and refine
- Prepare for release

Production release
Advisory group members
WHAT IS IT?
Entity Management Infrastructure (2020-21)
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
Done in 2020

- Explored an entity pipeline
  - Extracted, transformed, loaded multiple sources to graph
  - Studied the landscape (probabilistic/fuzzy matching, gazetteer)

- Established stable, repeatable knowledge hosting
  - Continued the learning with Wikibase
  - Focused on Loading at scale

- Explored iterative 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
How entities are built from WorldCat data

WorldCat

WorldCat Record A
Tom Sawyer (Print book)

WorldCat Record B
Tom Sawyer (Audiobook)

WorldCat Record C
Tom Sawyer (Large Print)

Automated Process
System remembers where the facts were extracted

Entity Manager

Tom Sawyer
- Name
- Alternate Names
- Description
- Language
- Work Entity Type
- Significant Dates
- Related External Sources
- Related Entities

Similar process for authority files or other data sets
Multiple links for multiple entity types
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
An overview of OCLC’s work model within the entity management infrastructure

Nathan Putnam
Director, Metadata Quality
**Model comparisons**

**IFLA-LRM**
- **Work**
  - Is realized through **Expression**
  - **Manifestation**
    - Is embodied in
    - Is exemplified by **Item**

**BIBFRAME**
- **Hub**
  - **bf:Work**
    - **bf:hasExpression**
    - **bf:hasInstance**
    - **Instance**
      - **Item**

**Share-VDE**
- **Super Work**
  - **svde:Work**
    - **bf:hasExpression**
    - **bf:hasInstance**
    - **Instance**
      - **Item**

---

**IFLA-LRM**
- IFLA’s Library Reference Model is a conceptual entity—relationship model developed by the International Federation of Library Associations and Institutions that expresses the “logical structure of bibliographic information”.

**BIBFRAME**
- BIBFRAME is a data model for bibliographic description. BIBFRAME was designed to replace the MARC standards, and to use linked data principles to make bibliographic data more useful both within and outside the library community.

**Share-VDE**
- Share-VDE is a library-driven initiative which brings together the bibliographic catalogues and authority files of a community of libraries in a shared discovery environment based on linked data.
Model comparisons
Tom Sawyer example

“Work (LRM)”
Tom Sawyer by Mark Twain created in 1876

manifestation of
“Expression (LRM)”
French Translation (Simon), 1984
translation of
“Manifestation (LRM)”
extent: 186 pages
publ: Nathan
Year: 1984
OCN: 70397484

reading of
“Expression (LRM)”
audiobook narr. by Hagon, 1996
“Manifestation (LRM)”
extent: 7 CDs
publ: Naxos Audiobooks
Year: 2020
OCN: 1144935849

manifestation of
“Expression (LRM)”
Norwegian translation (Hop), 1949
translation of
“Expression (LRM)”
Norwegian audiobook narr. by Eckhoff, 2003
manifestation of
“Manifestation (LRM)”
extent: 4 compact tapes
publ: Billingstad Fono
Year: 2003
OCN: 1028431700

manifestation of
“Expression (LRM)”
extent: 274 pages
publ: American Publ.
Year: 1876
OCN: 24991049
Questions?

Shane Huddleston

John Chapman

Nathan Putnam

LinkedData@oclc.org