OCLC RESEARCH DISCUSSION SERIES

Next Generation of Metadata

#OCLCmetadataseries

Opening Plenary • 23 February 2021

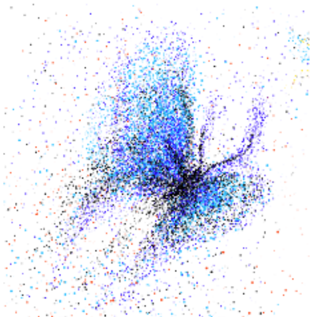
Welcome and Introduction

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.



#OCLCmetadataseries

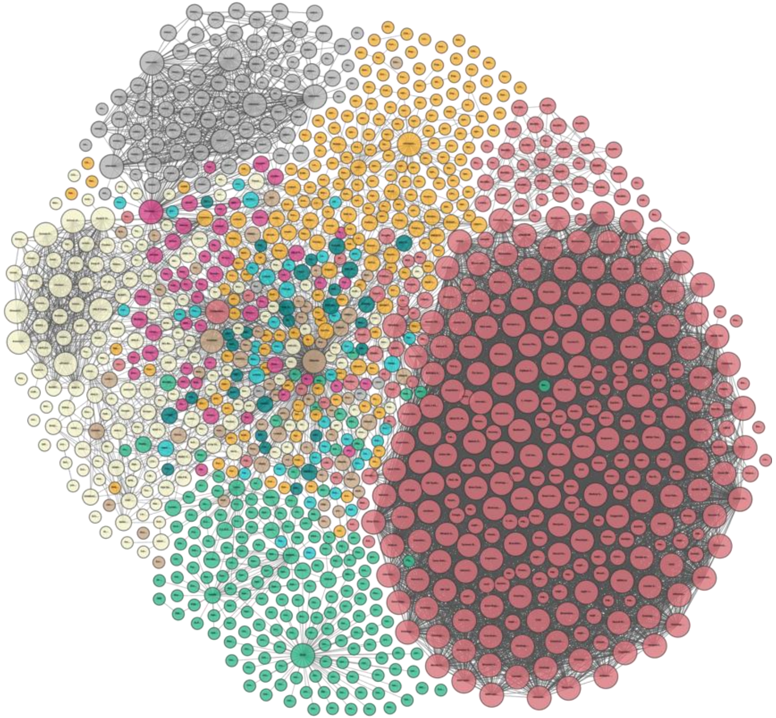
OCLC RESEARCH DISCUSSION SERIES

Next Generation of Metadata

#OCLCmetadataseries

OCLC

After a decade, it is okay to keep asking, *“Why linked data?”*



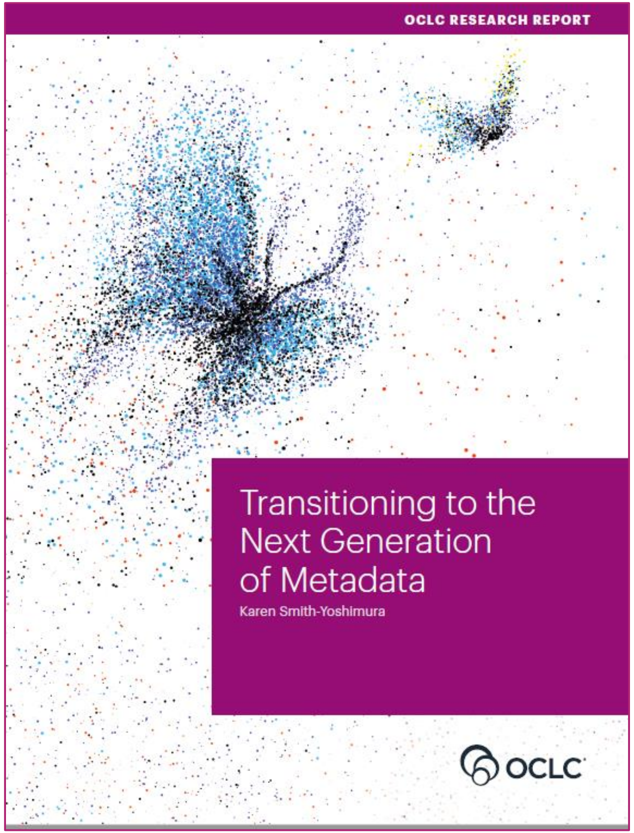
#OCLCmetadataseries

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”

#OCLCmetadataseries

Cultivating understanding



OCLC RESEARCH REPORT

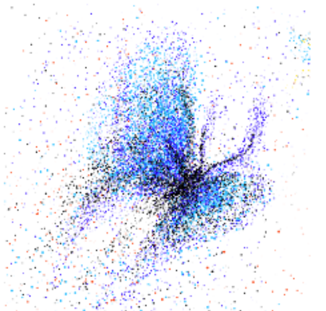
Transitioning to the Next Generation of Metadata

Karen Smith-Yoshimura

OCLC

oc.lc/nextgen-metadata-report

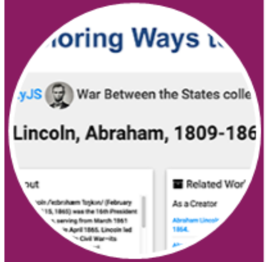
This report synthesizes six years of OCLC Research Library Partners Metadata Managers Focus Group discussions.



#OCLCmetadataseries

Experimenting and Building

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


oring Ways t

yJS War Between the States colle

Lincoln, Abraham, 1809-186

out

Related Wor

As a Creator

EntityJS Research Project (2013)

WorldCat®
Person Entity Lookup Pilot

Person Entity Lookup Pilot (2014)

Digital Maryland
Building and Exhibitions
Museum of Art

N PROFILE | ITEM METADATA | CLEA

nciliation with Contr

CONTENTdm Metadata Refinery (2015-16)

Project Passage (2017-18)

CONTENTdm
Linked Data Pilot

CONTENTdm Linked Data Pilot (2019-20)

Shared Entity Management Infrastructure (2020-21)

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

oring Ways t

yJS War Between the States colle

Lincoln, Abraham, 1809-186

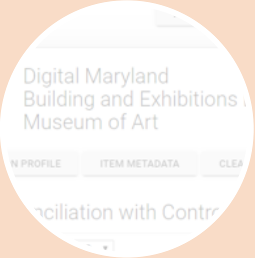
Related Wor

As a Creator

EntityJS Research Project (2013)

WorldCat®
Person Entity Lookup Pilot

Person Entity Lookup Pilot (2014)



Digital Maryland
Building and Exhibitions
Museum of Art

N PROFILE ITEM METADATA CLEA

nciliation with Contr

CONTENTdm Metadata Refinery (2015-16)

Project Passage (2017-18)

CONTENTdm
Linked Data Pilot

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 (2019-20)

Shared Entity Management Infrastructure (2020-21)

- 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

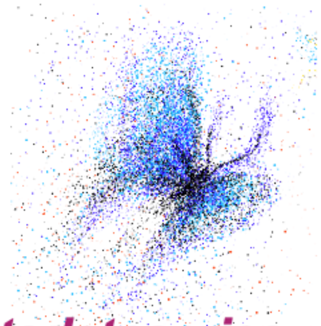
Tuesday 23 February 2021,

INTERACTIVE ROUND TABLE

During the first two weeks of March 2021.

CLOSING PLENARY WEBINAR

Tuesday 13 April 2021,



#OCLCmetadataseries

OCLC

Opening Plenary • 23 February 2021

Transitioning to the Next Generation of Metadata

Dr. Annette Dortmund

Senior Product Manager & Research Consultant, OCLC

@libsun

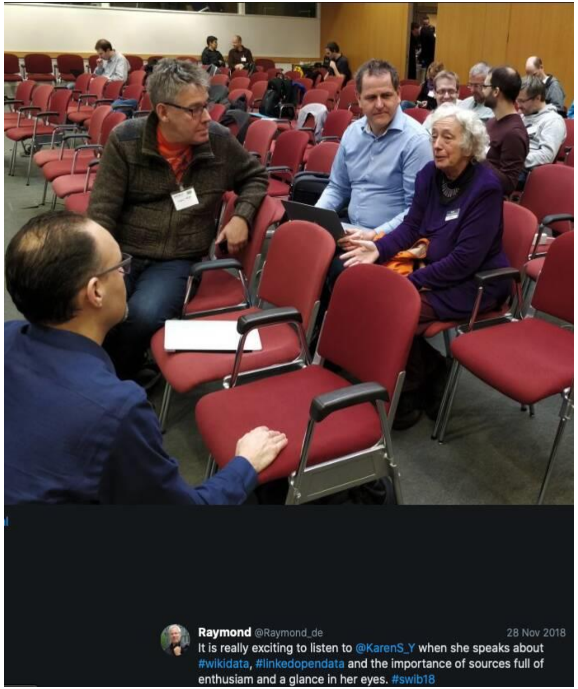
https://orcid.org/0000-0003-1588-9749

#OCLCmetadataseries

OCLC

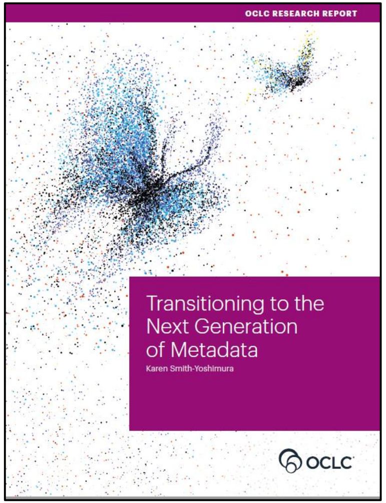
Karen Smith-Yoshimura

OCLC Research Senior Program Officer (retired Nov 30, 2020)



Raymond @Raymond_de 28 Nov 2018

It is really exciting to listen to @KarenS_Y when she speaks about #wikidata, #linkedopendata and the importance of sources full of enthusiam and a glance in her eyes. #swib18



OCLC RESEARCH REPORT

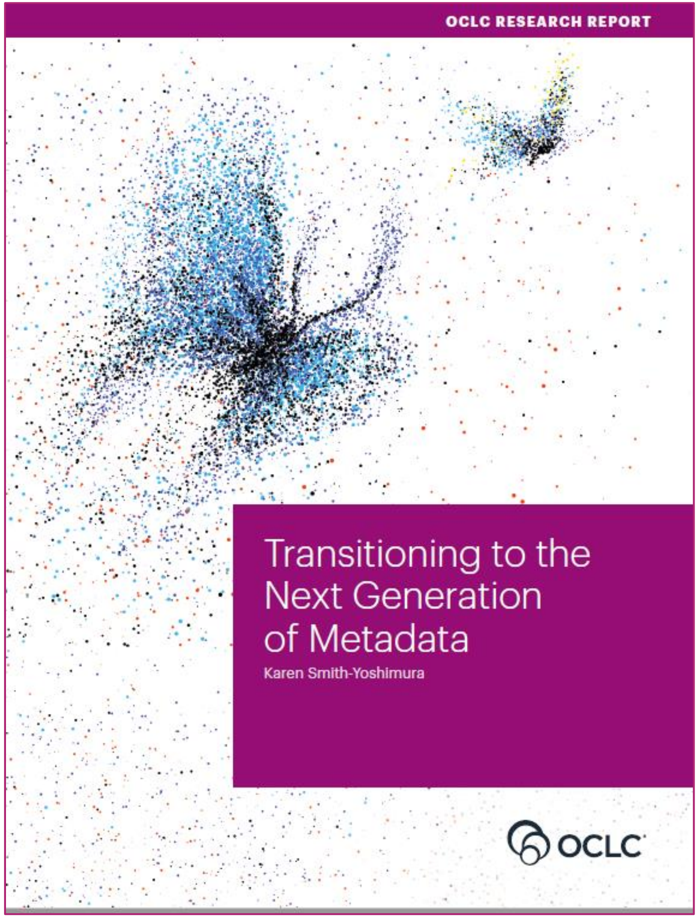
Transitioning to the Next Generation of Metadata

Karen Smith-Yoshimura

OCLC

- The Transition to Linked Data and Identifiers
- Describing “Inside-Out” and “Facilitated” Collections
- Evolution of “Metadata as a Service”
- Preparing for Future Staffing Requirements

Smith-Yoshimura, Karen. 2020. *Transitioning to the Next Generation of Metadata*. Dublin, OH: OCLC Research. https://doi.org/10.25333/rqgd-b343



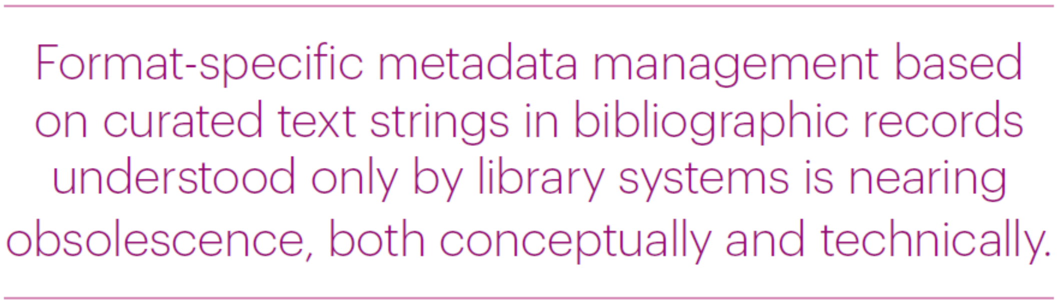
OCLC RESEARCH REPORT

Transitioning to the Next Generation of Metadata

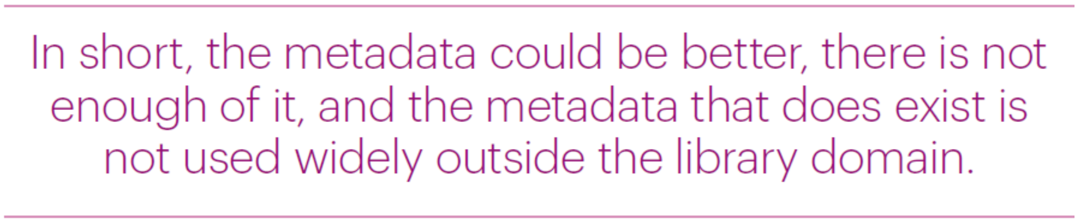
Karen Smith-Yoshimura

OCLC

Context



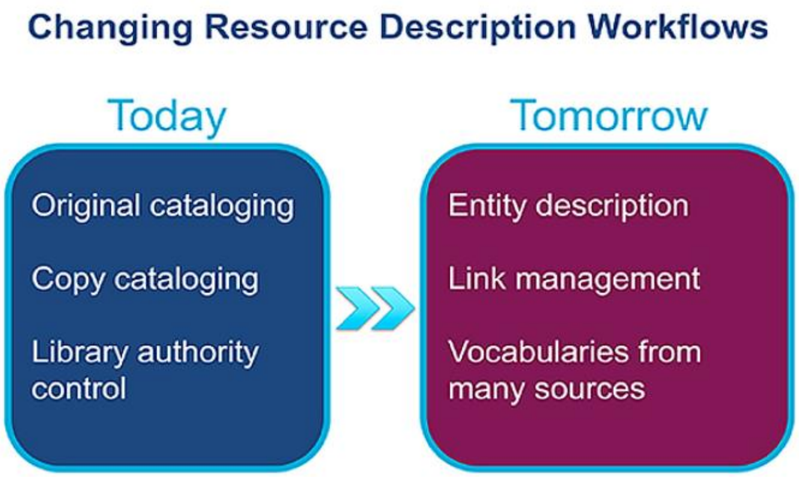
Format-specific metadata management based on curated text strings in bibliographic records understood only by library systems is nearing obsolescence, both conceptually and technically.



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

Transition to Linked Data & Identifiers



Changing Resource Description Workflows

Today

- Original cataloging
- Copy cataloging
- Library authority control

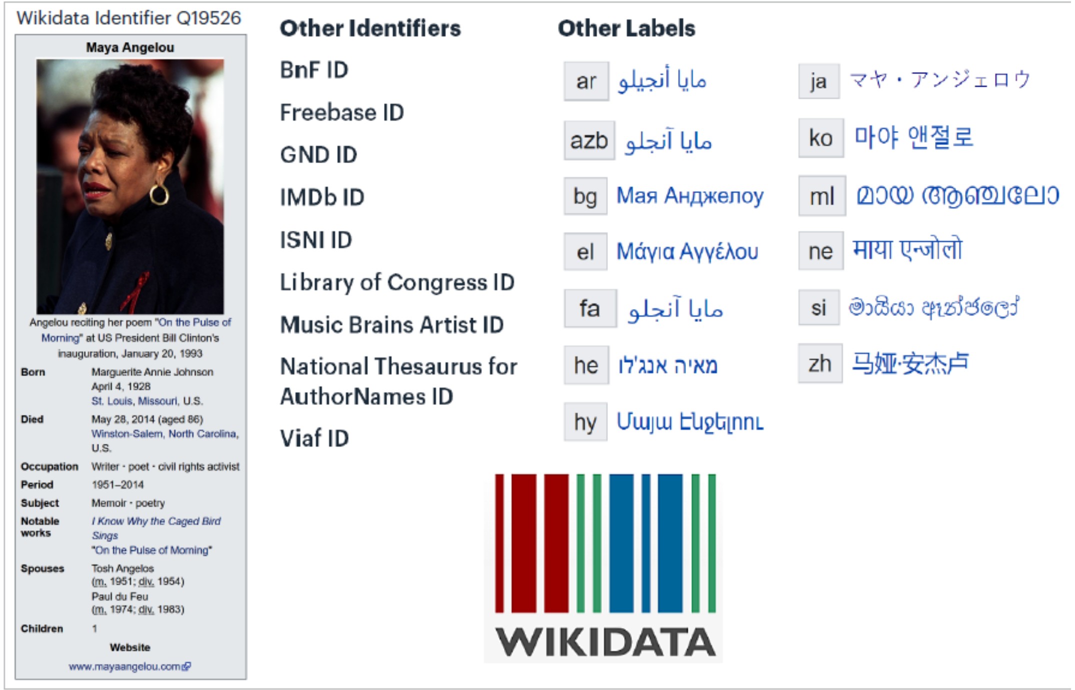
Tomorrow

- Entity description
- Link management
- Vocabularies from many sources

FIGURE 1. "Changing Resource Description Workflows" by OCLC Research[15]

"Persistent identifiers were viewed as crucial to transitioning from current metadata to future applications."

One Wikidata Identifier Links to Other Identifiers and Labels in Different Languages



Wikidata Identifier Q19526

Maya Angelou

Angelou reciting her poem "On the Pulse of Morning" at US President Bill Clinton's inauguration, January 20, 1993

Born	Marguerite Annie Johnson April 4, 1928 St. Louis, Missouri, U.S.
Died	May 28, 2014 (aged 86) Winston-Salem, North Carolina, U.S.
Occupation	Writer · poet · civil rights activist
Period	1951–2014
Subject	Memoir · poetry
Notable works	*I Know Why the Caged Bird Sings* "On the Pulse of Morning"
Spouses	Tosh Angelos (m. 1951; div. 1954) Paul du Feu (m. 1974; div. 1983)
Children	1

Website

www.mayaangelou.com

Other Identifiers

BnF ID

Freebase ID

GND ID

IMDb ID

ISNI ID

Library of Congress ID

Music Brains Artist ID

National Thesaurus for AuthorNames ID

Viaf ID

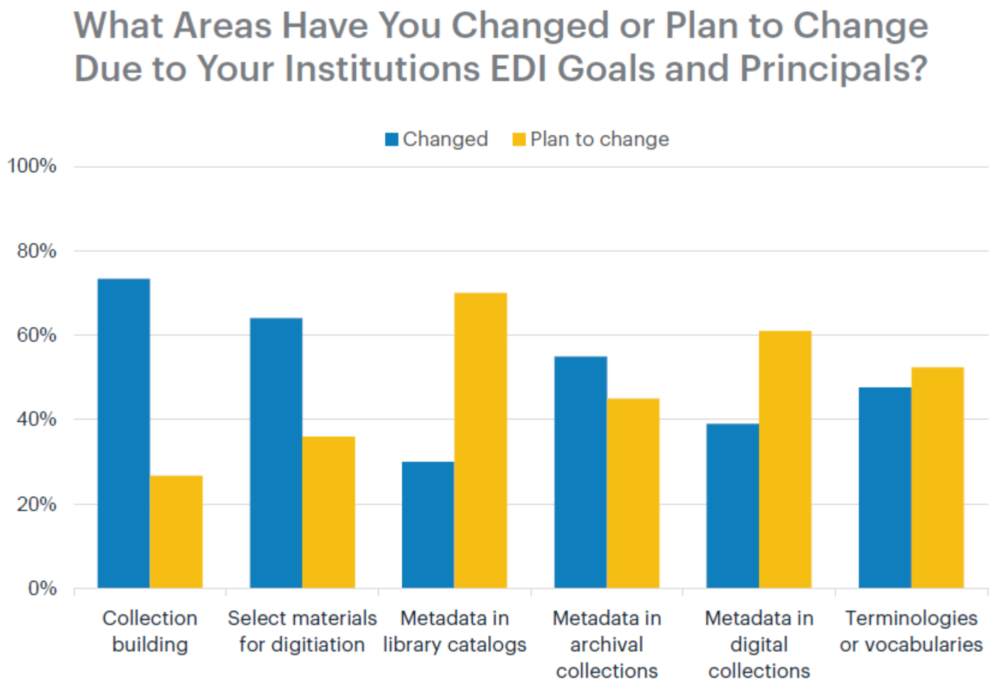
Other Labels

ar	مايا أنجيلو	ja	マヤ・アンジェロウ
azb	مایا آنجلو	ko	마야 앤절로
bg	Мая Анджелоу	ml	മായ ആഞ്ചലോ
el	Μάγια Αγγέλου	ne	माया एन्जोलो
fa	مایا آنجلو	si	මායා ඇන්ජලෝ
he	מאיה אנג'לו	zh	马娅·安杰卢
hy	Մայա Էնջելոու		

FIGURE 4. One Wikidata identifier links to other identifiers and labels in different languages

“Identity management poses a change in focus…to describing entities …and the relationships among them.”

Transition to Linked Data & Identifiers


What Areas Have You Changed or Plan to Change Due to Your Institutions EDI Goals and Principals?

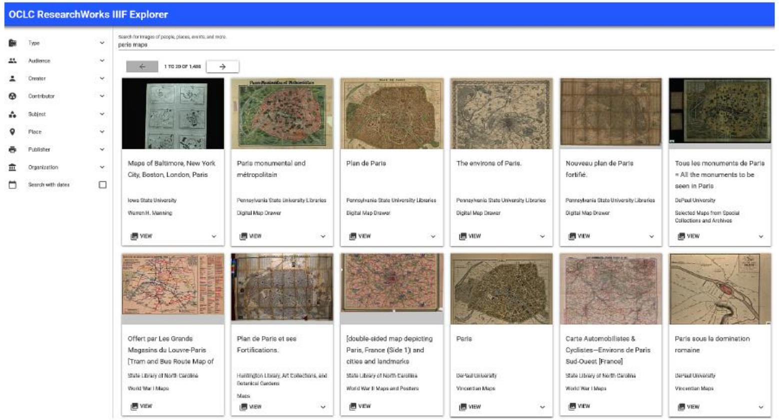
“Addressing language issues is important as libraries seek to develop relationships and build trust with marginalized communities.”

„Inside-Out“ & „Facilitated“ Collections

Describing “Inside-Out” and “Facilitated” Collections

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

The OCLC ResearchWorks IIIF Explorer Retrieves Images about “Paris Maps” across CONTENTdm Collections



Search for images of people, places, events, and more.
paris maps

1 TO 20 OF 1,488

Maps of Baltimore, New York City, Boston, London, Paris
Iowa State University
Warren H. Manning

Paris monumental and métropolitain
Pennsylvania State University Libraries
Digital Map Drawer

Plan de Paris
Pennsylvania State University Libraries
Digital Map Drawer

The environs of Paris.
Pennsylvania State University Libraries
Digital Map Drawer

Nouveau plan de Paris fortifié.
Pennsylvania State University Libraries
Digital Map Drawer

Tous les monuments de Paris = All the monuments to be seen in Paris
DePaul University
Selected Maps from Special Collections and Archives

Offert par Les Grands Magasins du Louvre-Paris [Tram and Bus Route Map of
State Library of North Carolina
World War I Maps

Plan de Paris et ses Fortifications.
Huntington Library, Art Collections, and Botanical Gardens
Maps

[double-sided map depicting Paris, France (Side 1) and cities and landmarks
State Library of North Carolina
World War II Maps and Posters

Paris
DePaul University
Vincentian Maps

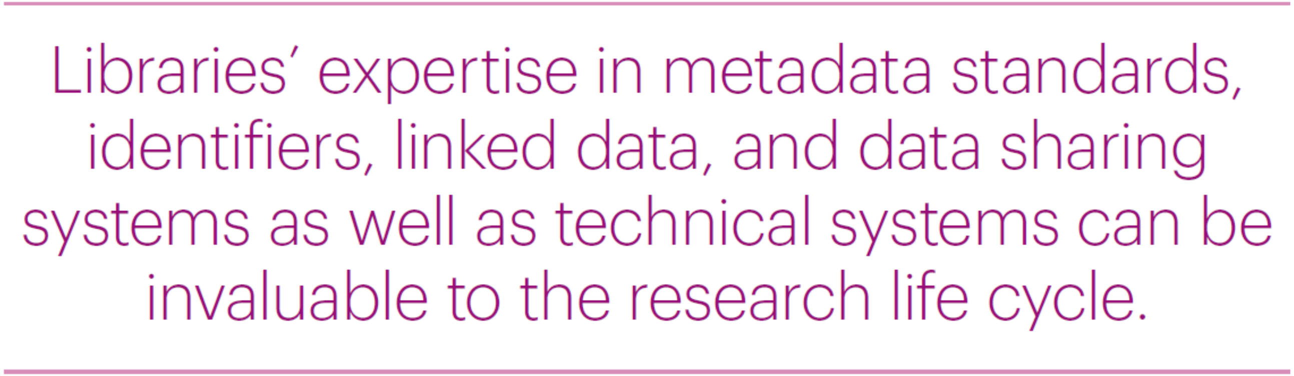
Carte Automobilistes & Cyclistes—Environs de Paris Sud-Ouest [France]
State Library of North Carolina
World War I Maps

Paris sous la domination romaine
DePaul University
Vincentian Maps

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

“Metadata underlies all discovery regardless of format, now and in the future…”

„Inside-Out“ & „Facilitated“ Collections

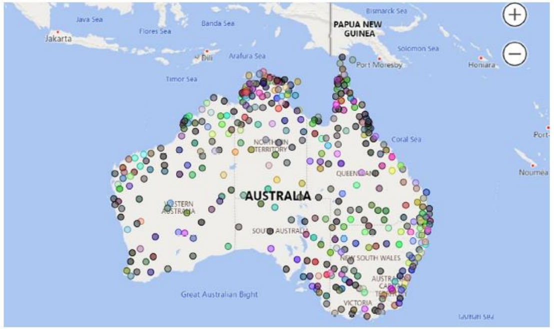


Libraries’ expertise in metadata standards, identifiers, linked data, and data sharing systems as well as technical systems can be invaluable to the research life cycle.

Metadata as a Service

Evolution of “Metadata as a Service”

Distribution of 465 Indigenous Language Codes in the Australian National Bibliographic Database



New applications

UK Hatchette’s “River of Authors” Generated from the British Library’s Catalog Metadata

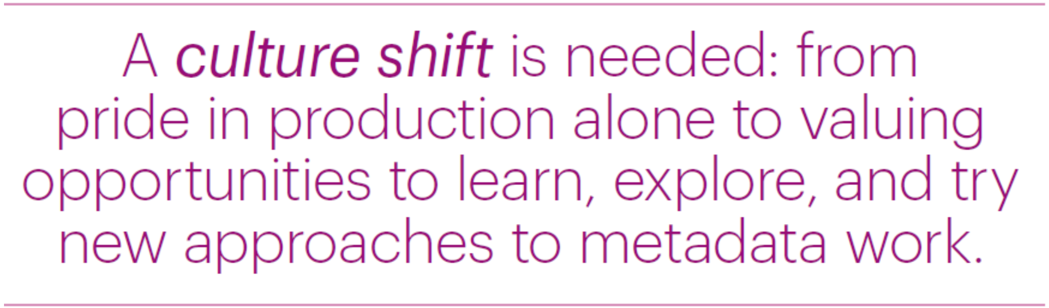
Bibliometrics

Plus:

- Metrics
- Consultancy
- Semantic indexing

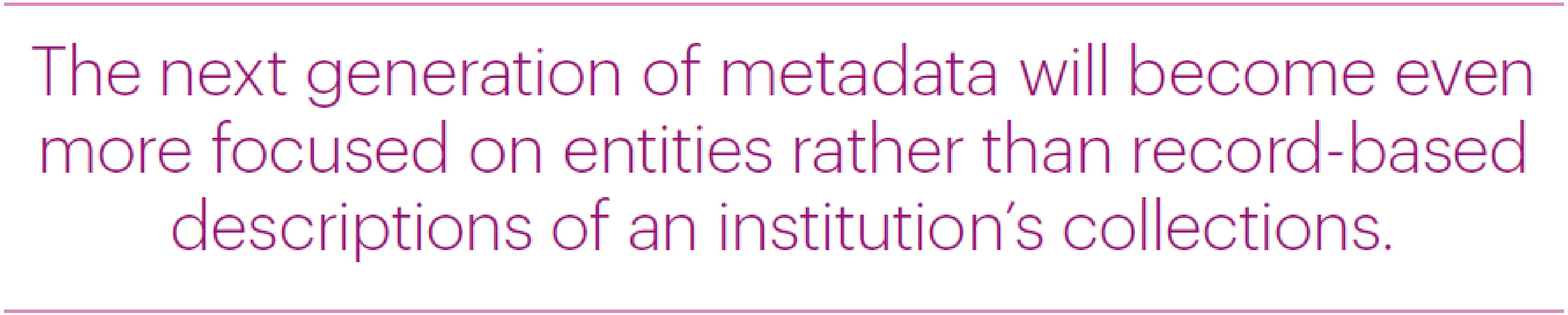
Future Staffing Requirements

Preparing for Future Staffing Requirements

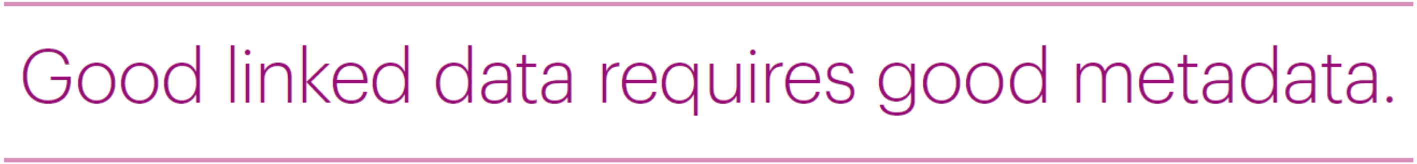


A *culture shift* is needed: from pride in production alone to valuing opportunities to learn, explore, and try new approaches to metadata work.

Conclusion



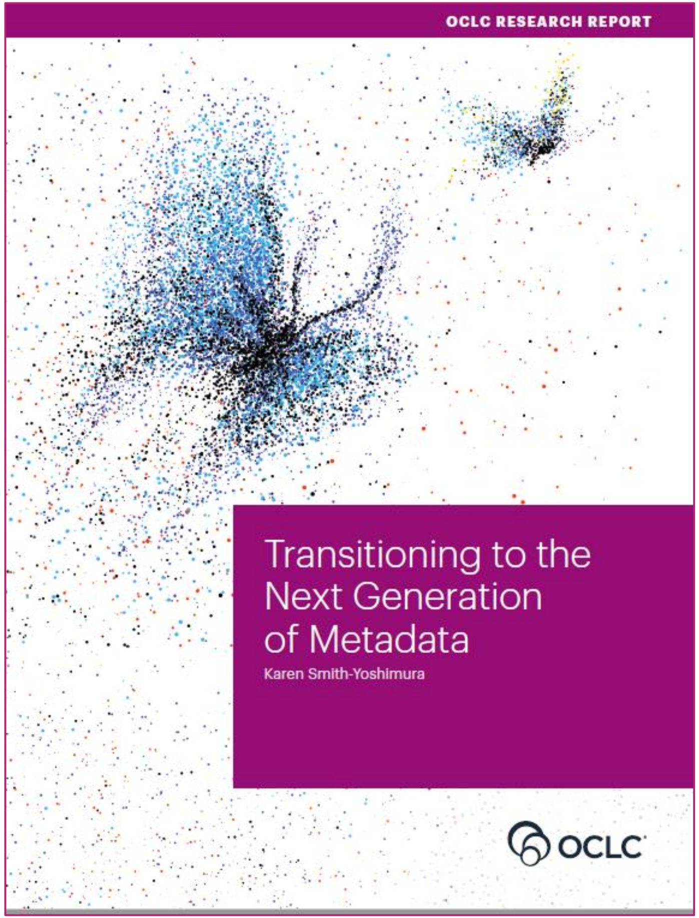
The next generation of metadata will become even more focused on entities rather than record-based descriptions of an institution's collections.



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

Smith-Yoshimura, Karen. 2020. *Transitioning to the Next Generation of Metadata*. Dublin, OH: OCLC Research. https://doi.org/10.25333/rqgd-b343


OCLC RESEARCH REPORT

Transitioning to the Next Generation of Metadata

Karen Smith-Yoshimura

OCLC

QUESTIONS?

Thank you!

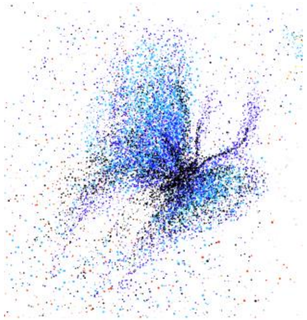
Dr. Annette Dortmund

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.®

Opening Plenary • 23 February 2021

Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project

Titia van der Werf

Senior Program Officer, OCLC Research

#OCLCmetadataseries

OCLC

OCLC RESEARCH REPORT

Transforming metadata into linked data to improve digital collection discoverability:

A CONTENTdm Pilot Project

OCLC

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

The CONTENTdm Linked Data Pilot questions

CONTENTdm Linked Data Pilot (2019-20)

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

CONTENTdm
Linked Data Pilot

CONTENTdm Linked Data Pilot (2019-20)

- 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

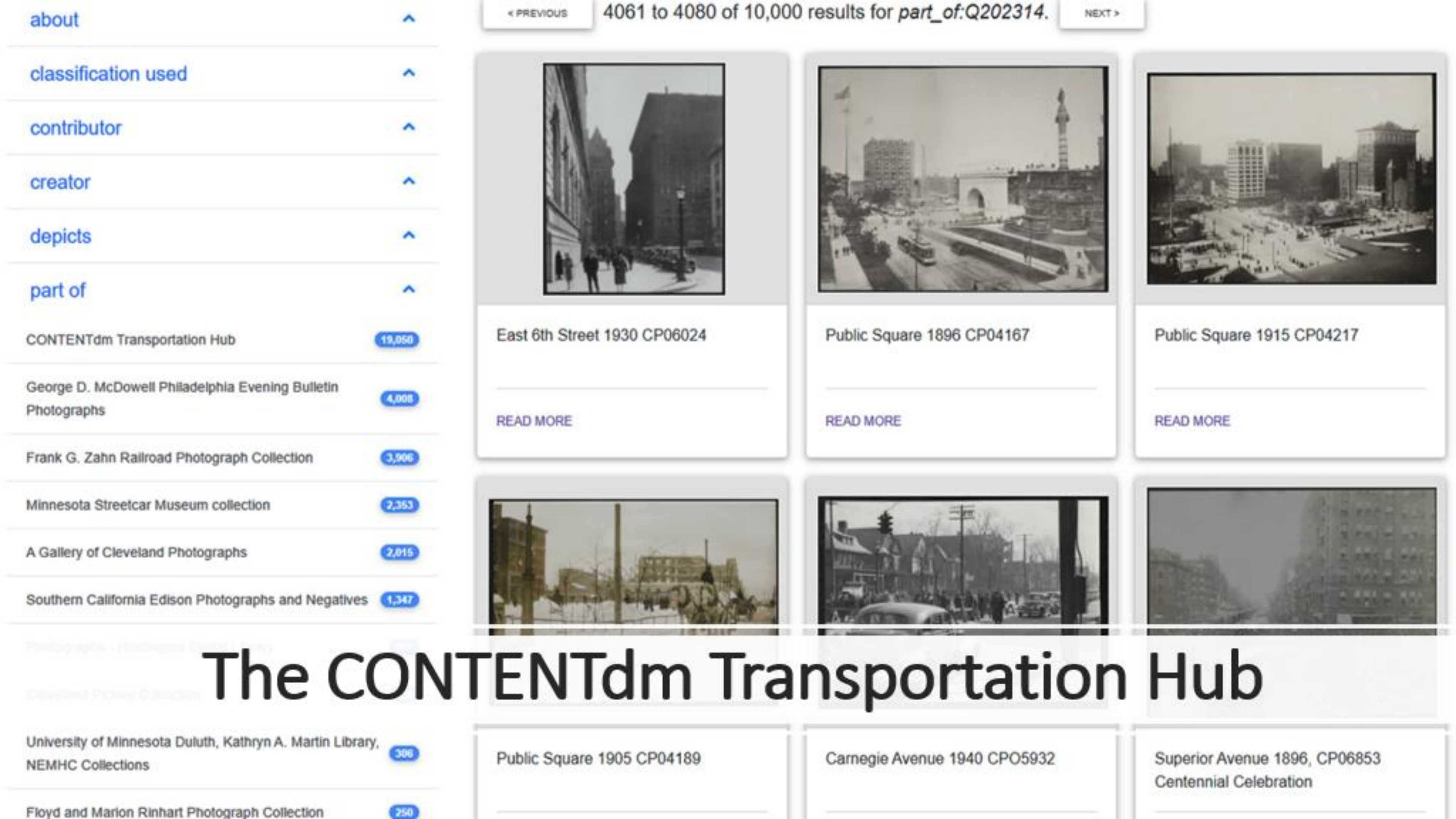
The CONTENTdm Linked Data Pilot

CONTENTdm
Linked Data Pilot

CONTENTdm Linked Data Pilot (2019-20)

New applications:

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



The CONTENTdm Transportation Hub

< PREVIOUS 4061 to 4080 of 10,000 results for *part_of:Q202314*. NEXT >

about

classification used

contributor

creator

depicts

part of

CONTENTdm Transportation Hub	19,050
George D. McDowell Philadelphia Evening Bulletin Photographs	4,008
Frank G. Zahn Railroad Photograph Collection	3,906
Minnesota Streetcar Museum collection	2,353
A Gallery of Cleveland Photographs	2,015
Southern California Edison Photographs and Negatives	1,347
Photographs - Washington Digital Library	[illegible]
Cleveland Picture Collection	[illegible]
University of Minnesota Duluth, Kathryn A. Martin Library, NEMHC Collections	306
Floyd and Marion Rinhart Photograph Collection	250

East 6th Street 1930 CP06024

READ MORE

Public Square 1896 CP04167

READ MORE

Public Square 1915 CP04217

READ MORE

Public Square 1905 CP04189

Carnegie Avenue 1940 CPO5932

Superior Avenue 1896, CP06853 Centennial Celebration

Findings

CONTENTdm Linked Data Pilot

CONTENTdm Linked Data Pilot (2019-20)

- 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.”

REFLECTION: Leveraging the power of linked data

“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.”

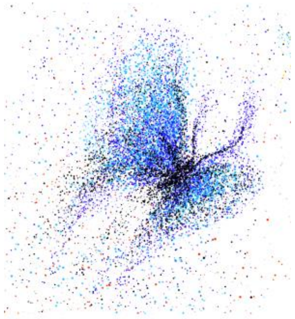
QUESTIONS?

Opening Plenary • 23 February 2021

OCLC’s Entity Management Infrastructure

John Chapman

Senior Product Manager, Metadata Strategy & Operations

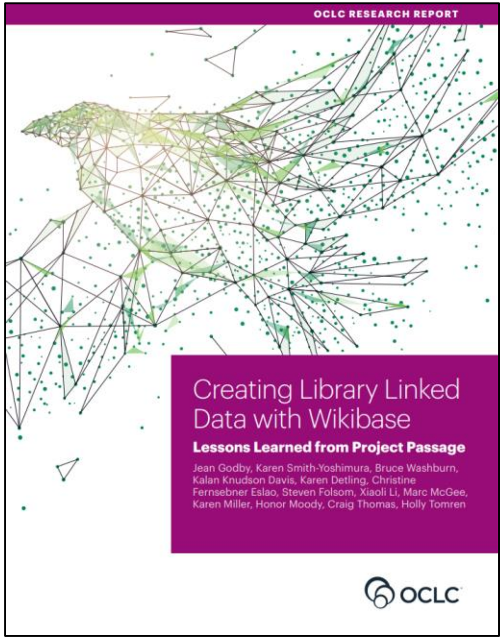


https://orcid.org/0000-0002-5388-5063

#OCLCmetadataseries

WHY A “METADATA INFRASTRUCTURE?”

Feedback from OCLC member libraries



OCLC RESEARCH REPORT

Creating Library Linked Data with Wikibase

Lessons Learned from Project Passage

Jean Godby, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Craig Thomas, Holly Tomren

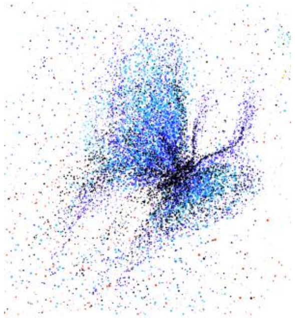
OCLC

oc.lc/passagereport

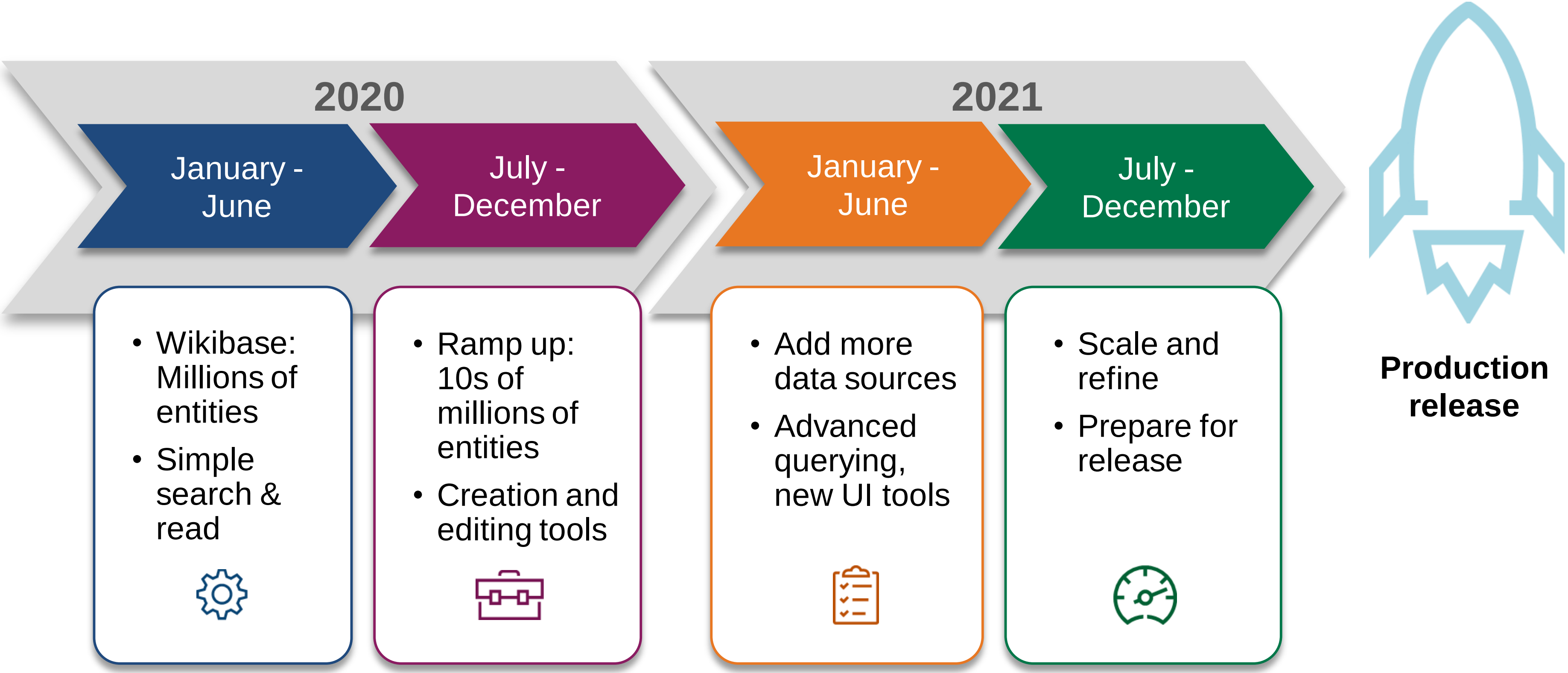
- Provide persistent identifiers relevant to library workflows
- Enable the creation of new identifiers within metadata management workflows
- Provide interfaces and ecosystem to create native linked data descriptions
- Seed the web with persistent identifiers
- Provide broad reconciliation across vocabularies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

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

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

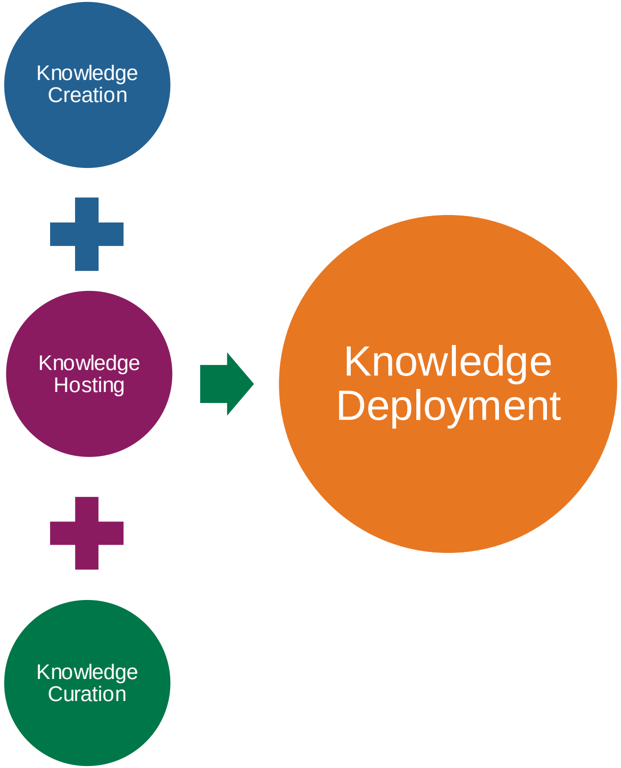
Production release

WHAT IS IT?

What is the “Infrastructure”?

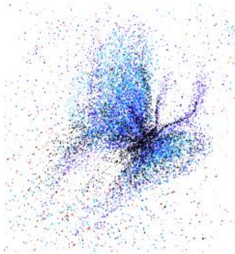
- Community-curated Knowledge Graph
- Integration of facts from library data from around the world
  - Seeded from the knowledge contained in bibliographic authority files, WorldCat creative works, and controlled vocabularies
- Provenance and context of the knowledge claims as the facts come from a variety of heterogeneous sources
- Published following Linked Data Principles, a set of APIs and query endpoints

Knowledge Graph - processes



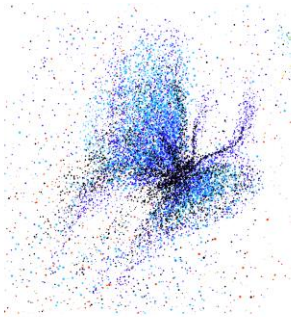
1. **Knowledge Creation**: Integration of heterogeneous data sources, through ‘Semantic lifting’.
2. **Knowledge Hosting**: Storage of the knowledge in a suitable way (e.g., semantic repository, a graph database, triple store).
3. **Knowledge Curation**: Make sure that the correctness and completeness 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)]



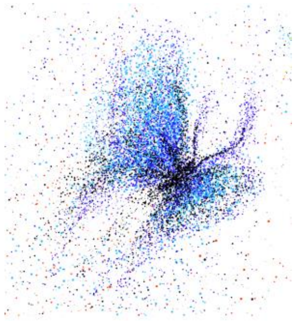
Done in 2020

- 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



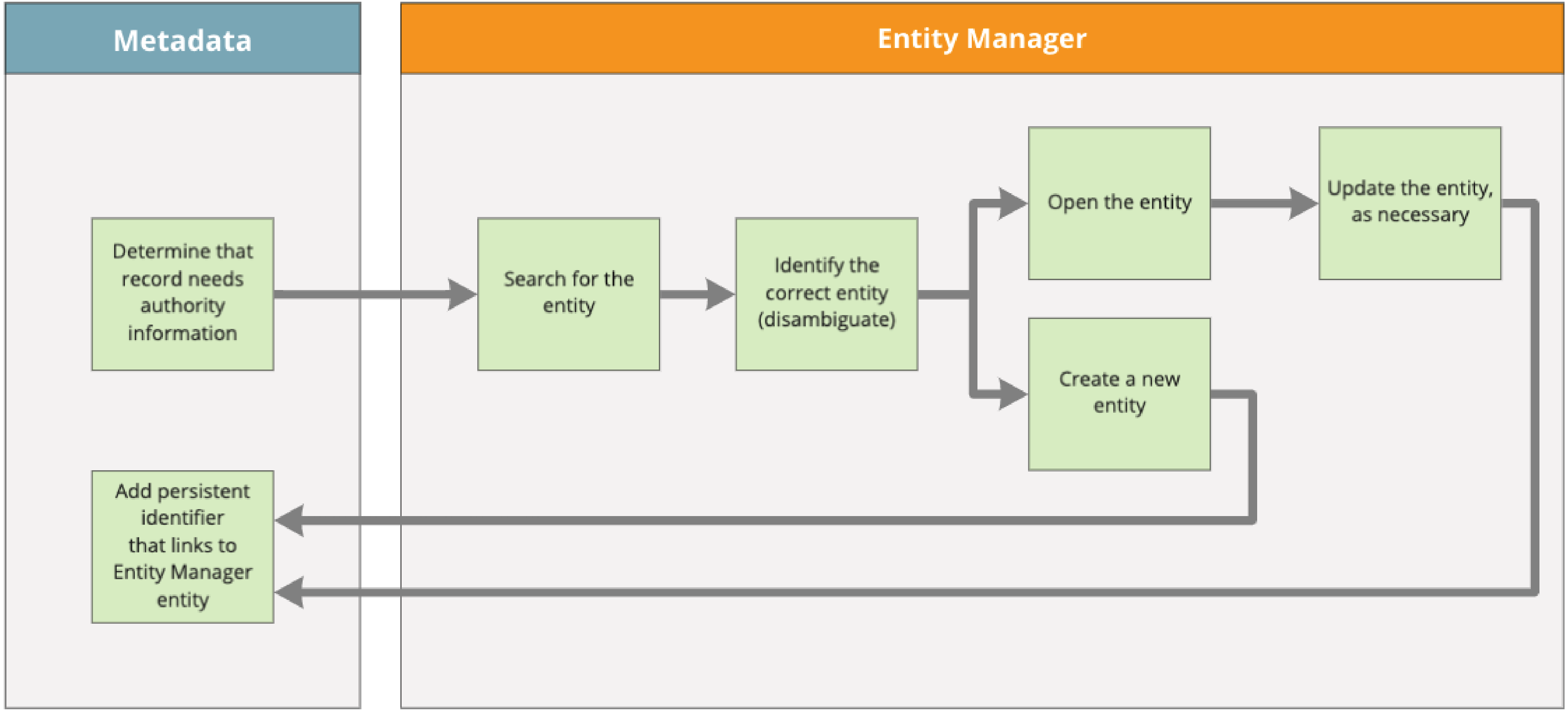
Next steps on architecture, systems

- Multilingual approaches
- Moving beyond the Wikibase structure
- Integrating input on data models
- Building out curation support



WORKFLOWS AND LINKING

Example Workflow



Metadata

Entity Manager

Determine that record needs authority information

Search for the entity

Identify the correct entity (disambiguate)

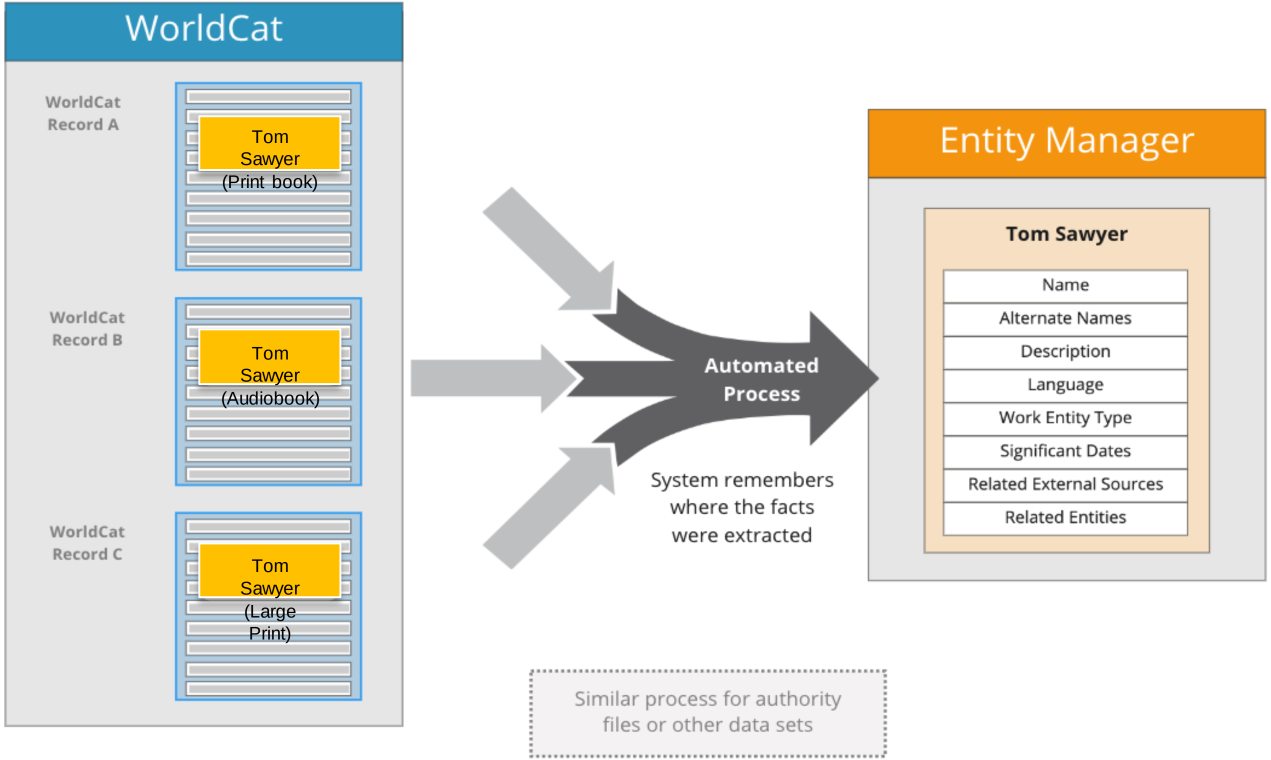
Open the entity

Update the entity, as necessary

Create a new entity

Add persistent identifier that links to Entity Manager entity

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

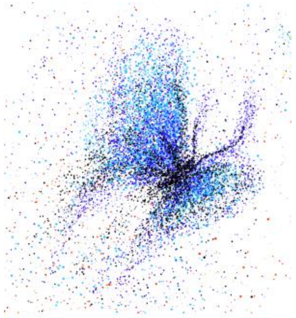
System remembers where the facts were extracted

Entity Manager

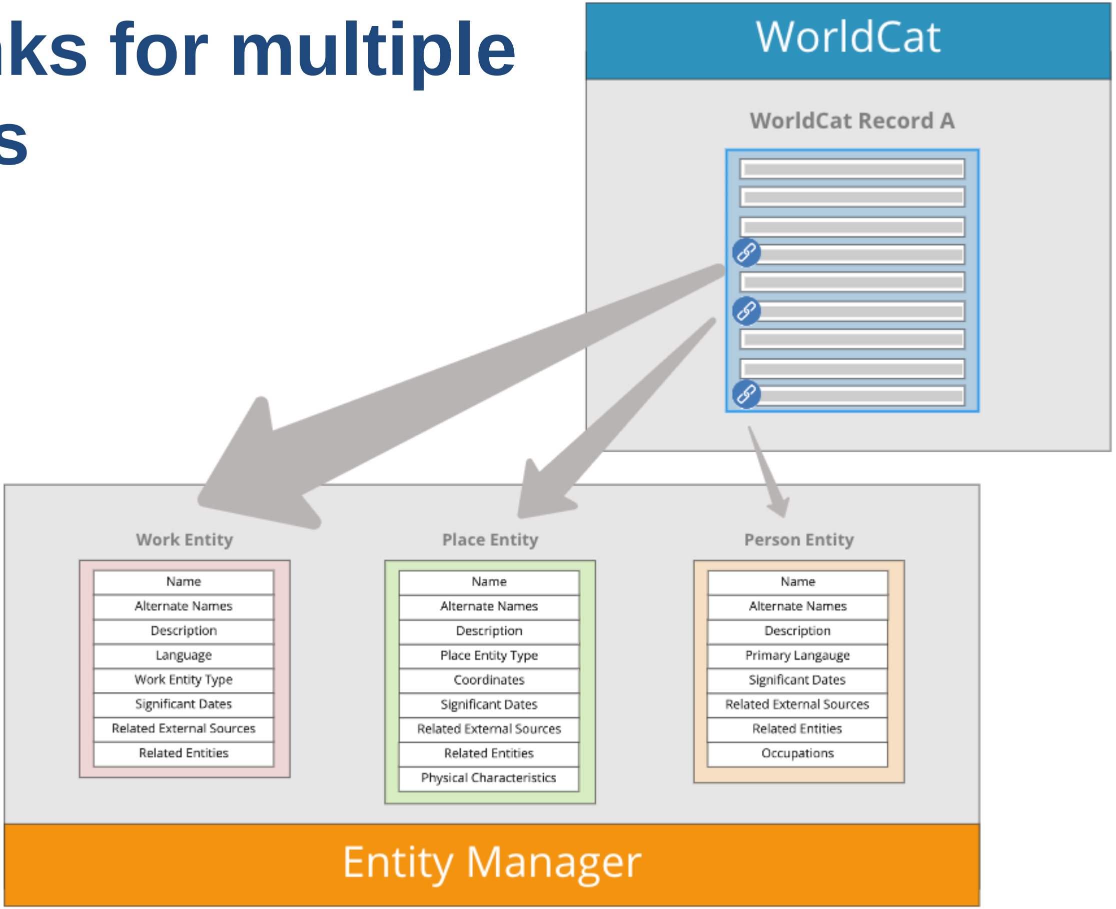
Tom Sawyer

- 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



WorldCat

WorldCat Record A

Work Entity

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

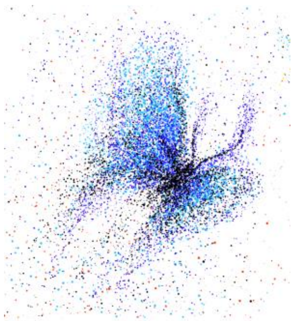
Place Entity

- Name
- Alternate Names
- Description
- Place Entity Type
- Coordinates
- Significant Dates
- Related External Sources
- Related Entities
- Physical Characteristics

Person Entity

- Name
- Alternate Names
- Description
- Primary Langauge
- Significant Dates
- Related External Sources
- Related Entities
- Occupations

Entity Manager



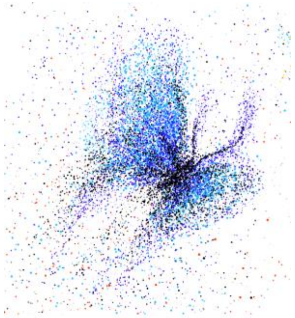
Next steps on workflows

- API and UI testing
- Wireframe review with Advisory Group (ongoing)
- Prioritization and scheduling of UI features

DATA ACTIVITIES

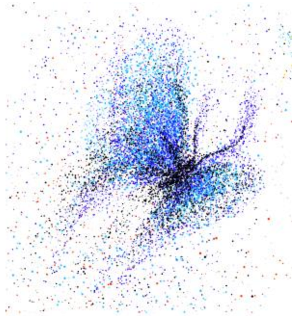
Processes

- 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



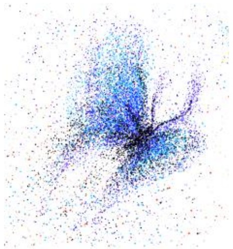
Works modeling

- 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



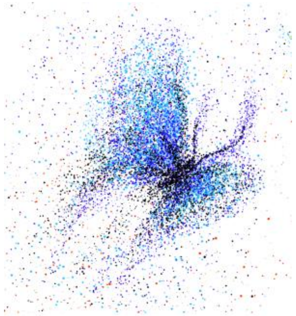
Minimum Viable Entity (MVE): Work

- 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)

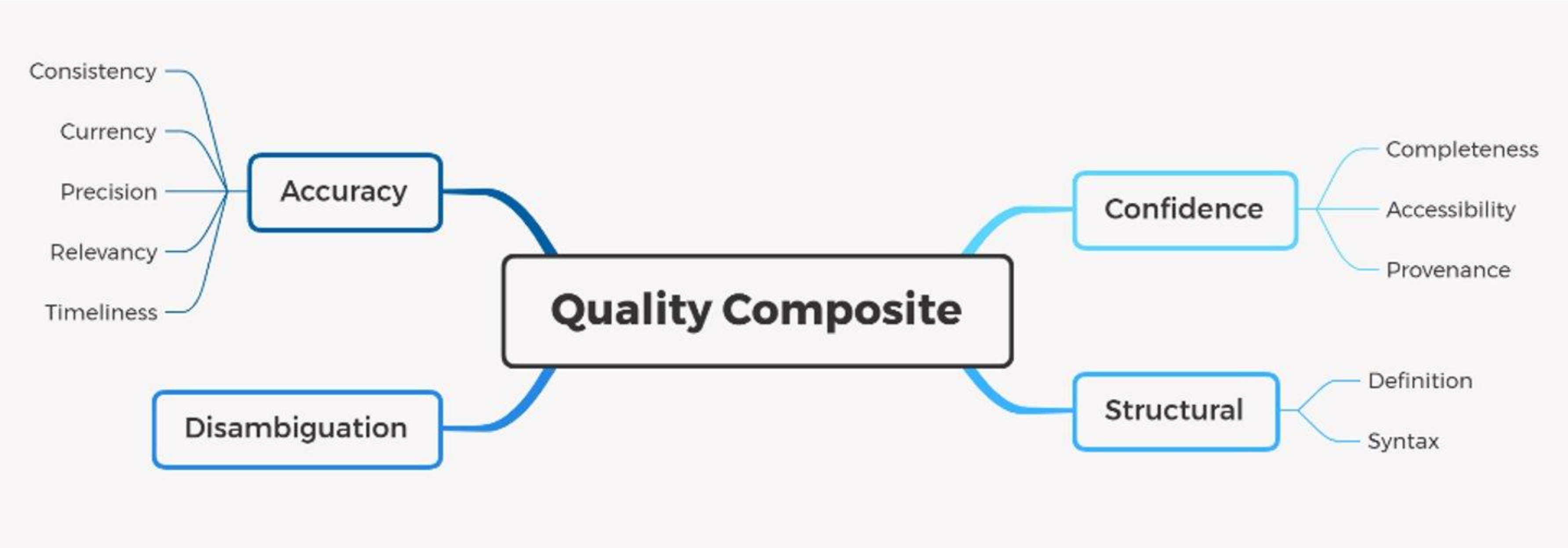


Scale

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

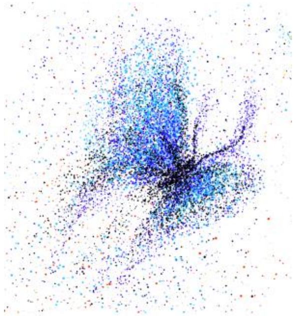


Quality Composite



Data in 2021

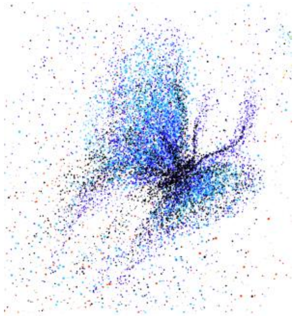
- Continue to build out data models and entity description
- Further work on Quality Composite
- Ontology development
- Broader testing



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



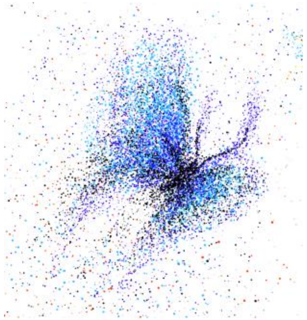
Thank you!

John Chapman

Senior Product Manager,
Metadata Strategy & Operations

chapmanj@oclc.org

https://orcid.org/0000-0002-5388-5063



Because what is known must be shared.®

QUESTIONS?

Opening Plenary • 23 February 2021

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

nomisma.org

NDL Search

BnF Bibliothèque nationale de France data.bnf.fr

hbz lobid

europeana collections

OCLC

VIAF
Virtual International Authority File

LIBRARY OF CONGRESS

LC Linked Data Service
Authorities and Vocabularies

OCLC WorldCat

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