

2020 DCMI
17 September 2020

Deep semantic representation from metadata descriptions: a linked data perspective

Andrew K. Pace

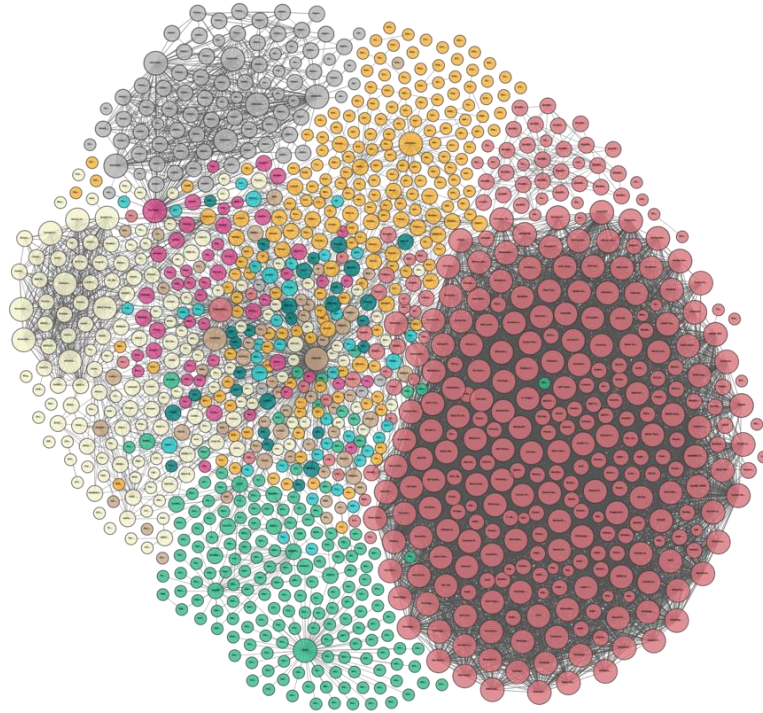
Executive Director, Technical Research
OCLC

Agenda

- Research and Findings: a decade of linked data research
- A shared Entity Management Infrastructure
- Applying a linked data infrastructure to distinctive collections
- Shared Infrastructure & Local Entity Management: a juxtaposition

Acknowledgements: OCLC Colleagues Jeff Mixter, Bruce Washburn, Shane Huddleston, Jeff Young, John Chapman, *et al*; and the dozens of libraries who have participated in OCLC Research prototypes, experiments, and research efforts.

Why Linked Data?



A decade with Linked Data


oclc.org/linkeddataresearch



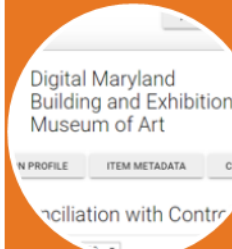
Publish linked data - FAST, VIAF, WorldCat (2009 -)



EntityJS Research Project (2013)




WorldCat®
Person Entity Lookup Pilot




Digital Maryland Building and Exhibitions Metadata Refinery (2015-16)



Project Passage (2017-18)



CONTENTdm
Linked Data Pilot (2019-20)



Shared Entity Management Infrastructure (2020-21)





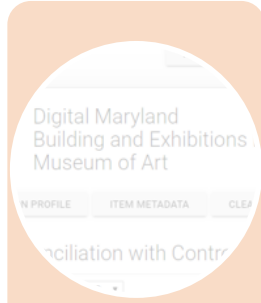
Publish linked data -
FAST, VIAF,
WorldCat (2009 -)



EntityJS Research
Project (2013)



Person Entity Lookup
Pilot (2014)



CONTENTdm
Metadata Refinery
(2015-16)



Project Passage
(2017-18)



CONTENTdm Linked
Data Pilot (2019-20)

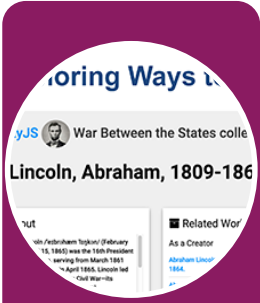


Shared Entity
Management
Infrastructure
(2020-21)

VIAF and FAST: Publish Linked Data on the web with a UI, API, and downloadable datasets



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)

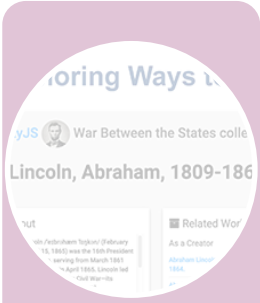


Shared Entity
Management
Infrastructure
(2020-21)

EntityJS: Explore how Linked Data maximizes the discovery potential for sets of related entities (related by an event, a literature domain, etc.)



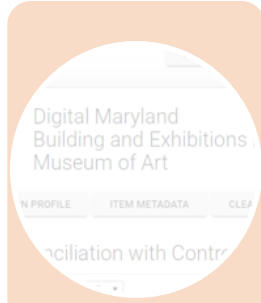
Publish linked data -
FAST, VIAF,
WorldCat (2009 -)



EntityJS Research
Project (2013)



Person Entity Lookup
Pilot (2014)



CONTENTdm
Metadata Refinery
(2015-16)



Project Passage
(2017-18)



CONTENTdm Linked
Data Pilot (2019-20)

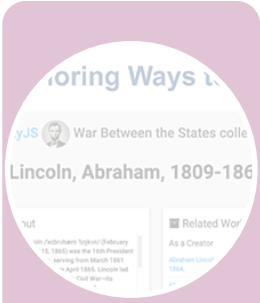


Shared Entity
Management
Infrastructure
(2020-21)

Person Entity Lookup Pilot: Test use cases and client interoperability for Linked Data as a web service



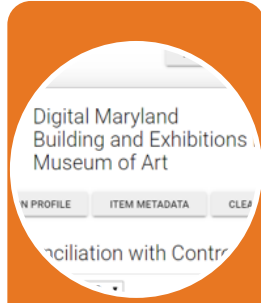
Publish linked data -
FAST, VIAF,
WorldCat (2009 -)



EntityJS Research
Project (2013)



Person Entity Lookup
Pilot (2014)



CONTENTdm
Metadata Refinery
(2015-16)



Project Passage
(2017-18)

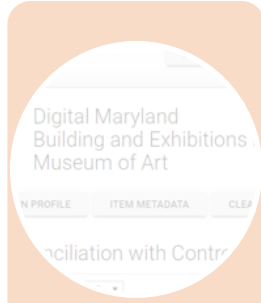
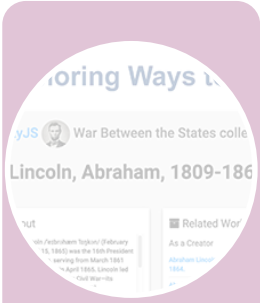


CONTENTdm Linked
Data Pilot (2019-20)



Shared Entity
Management
Infrastructure
(2020-21)

Metadata Refinery: Evaluate shared tools that help institutions take control of the Linked Data creation workflow



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)

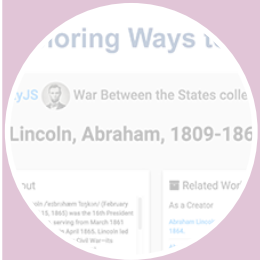
CONTENTdm Linked
Data Pilot (2019-20)

Shared Entity
Management
Infrastructure
(2020-21)

Project Passage: Think big... Build a complete system based on Linked Data, and see how workflows change



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)

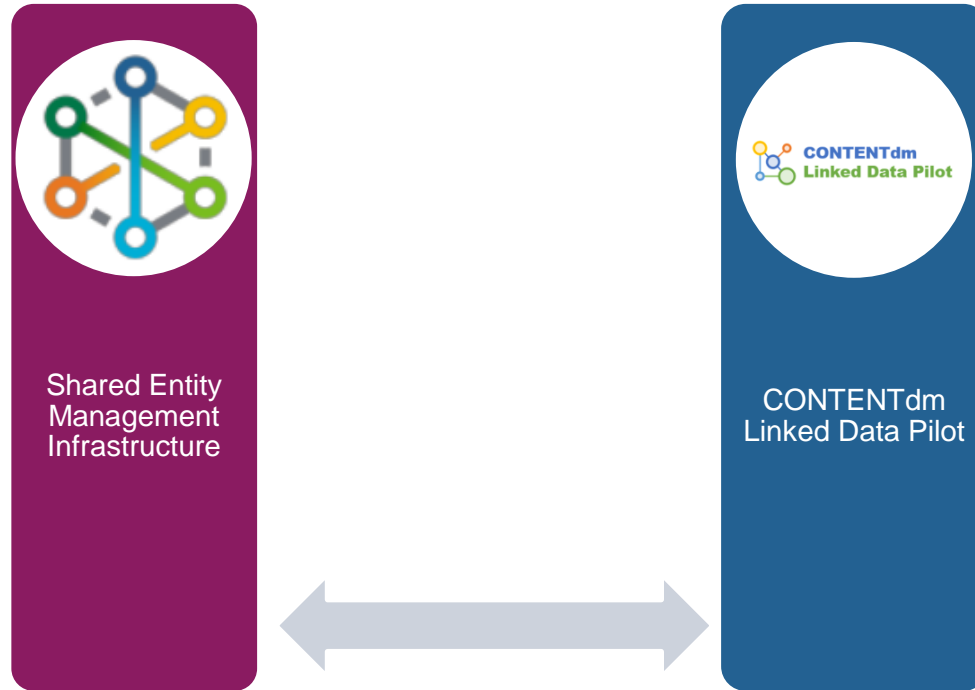


CONTENTdm Linked
Data Pilot (2019-20)



Shared Entity
Management
Infrastructure
(2020-21)

Current Events with Linked Data



SHARED ENTITY MANAGEMENT INFRASTRUCTURE



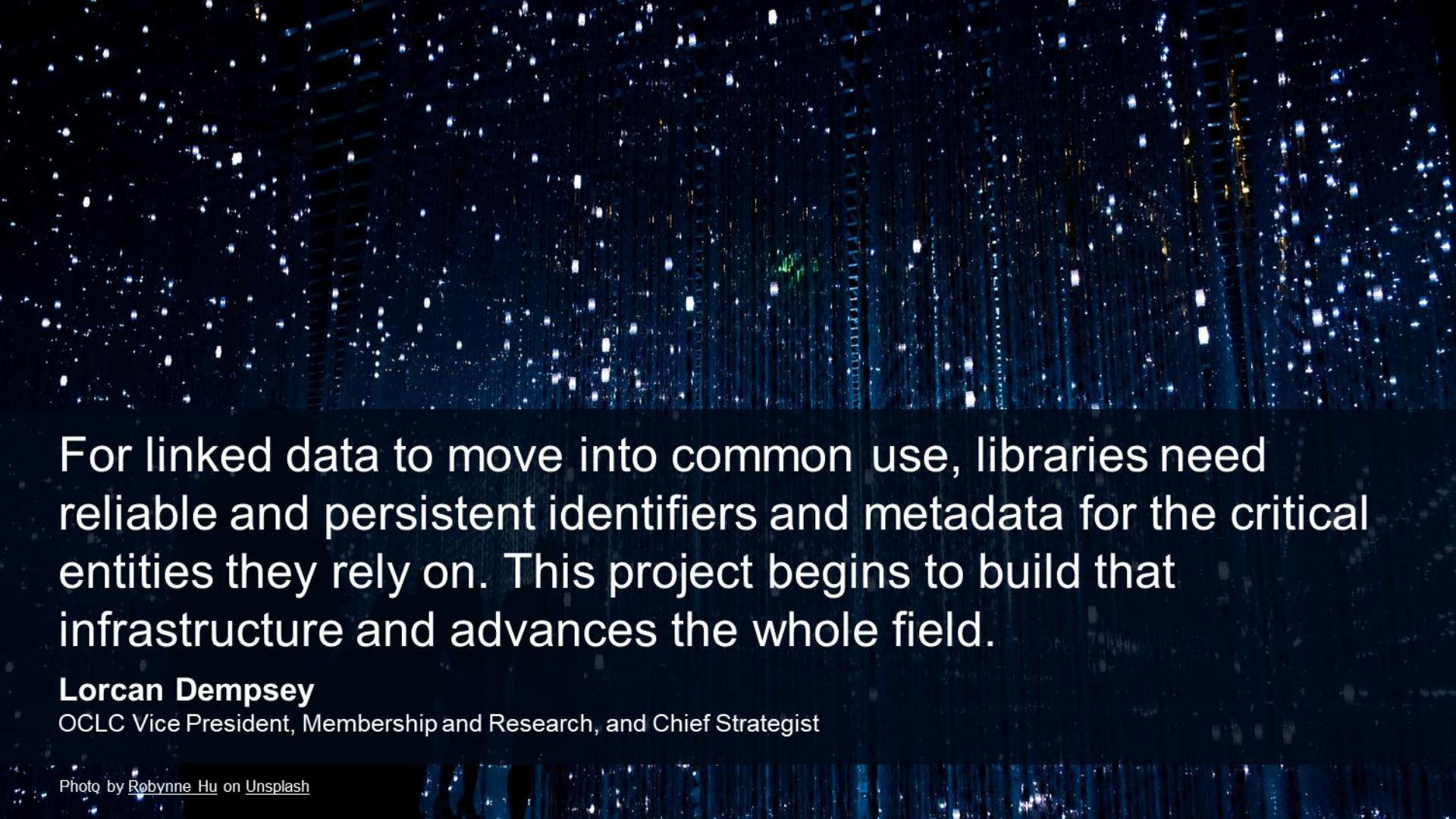
OCLC awarded Mellon Foundation grant to develop infrastructure to support linked data management initiatives

'Entity Management Infrastructure' will advance use of linked data and ultimately improve discoverability of scholarly materials on the web

DUBLIN, Ohio, 9 January 2020—[OCLC](#) has been awarded a grant from The Andrew W. Mellon Foundation to develop a shared "Entity Management Infrastructure" that will support linked data management initiatives underway in the library and scholarly communications community. When complete, this infrastructure will be jointly curated by the community and OCLC, and will ultimately make scholarly materials more connected and discoverable on the web.

The two-year grant, for \$2.436 million, will support work on the project that will run from January 2020 to December 2021. The Mellon grant funding represents approximately half of the total cost of the Entity Management Infrastructure project. OCLC is contributing the remaining half of the required investment.

"OCLC has been a leader in library linked data research for years, and we have developed prototypes, innovative pilot programs and partnerships that continue to inform our work," said Skip Prichard, OCLC President and CEO. "OCLC enables libraries to work together to achieve economies, efficiencies, and consistency in metadata creation. We're grateful to The



For linked data to move into common use, libraries need reliable and persistent identifiers and metadata for the critical entities they rely on. This project begins to build that infrastructure and advances the whole field.

Lorcan Dempsey

OCLC Vice President, Membership and Research, and Chief Strategist

Entity Management



Shared Entity
Management
Infrastructure
(2020-21)

Project goals

- Address infrastructure needs identified by libraries
 - Expand on “native” metadata management
 - Link library data to non-library data... and shared data to local data
 - Provide ID creation services to help “at the point of need”
 - Persistent and maintained entity URIs
- Operate at a large scale – and be sustainable
- Complement other efforts—LD4, PCC, DCMI, etc.

Entity Management



Shared Entity
Management
Infrastructure
(2020-21)

Methods

- 24-month project, six-month increments
- Leverage Wikibase for 12+ months
- Multiple communication channels for input and iteration
- Division-spanning project including staff from engineering, UX research, architecture, systems, and technical research
- Multiple “workstreams” represent coherent teams

Entity Management



Shared Entity
Management
Infrastructure
(2020-21)

Communication channels

- Ad-hoc with libraries, groups (ex: PCC)
- Presentations and reports
- Ongoing with LD4P
- Entity Management Advisory Group
 - Monthly meetings
 - “Breakouts” / focus groups
 - Testing

Advisory Group Members



Shared Entity
Management
Infrastructure
(2020-21)

{BnF | Bibliothèque
nationale de France

TEMPLE
UNIVERSITY

LIBRARY
HSILITIB

abes
agence bibliographique
de l'enseignement supérieur

Penn
UNIVERSITY of PENNSYLVANIA

Yale

CMU
CENTRAL MICHIGAN
UNIVERSITY

CLEVELAND PUBLIC LIBRARY

UCDAVIS

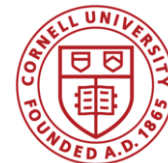


NIH
U.S. National Library
of Medicine



PRINCETON
UNIVERSITY

University
of Victoria



Biblioteca
nazionale
centrale
di Roma

BYU
BRIGHAM YOUNG
UNIVERSITY

NYU

HARVARD
UNIVERSITY

UNIVERSITY OF
OXFORD

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
UNIVERSITY LIBRARIES

UNIVERSITY OF MINNESOTA

NLB
National Library Board
Singapore

DEUTSCHE
NATIONAL
BIBLIOTHEK

Smithsonian

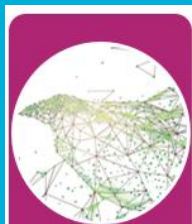
Entity Management



Shared Entity
Management
Infrastructure
(2020-21)

- First increment was recently completed
 - Basic functionality
 - API and UI
 - Process, procedures, cadence
- “Findings” so far
 - Need focus: creative works and persons
 - Internal communication (especially now) takes effort
 - Scaling is a challenge

LINKED DATA FOR DISTINCTIVE COLLECTIONS



Project Passage
(2017-18)

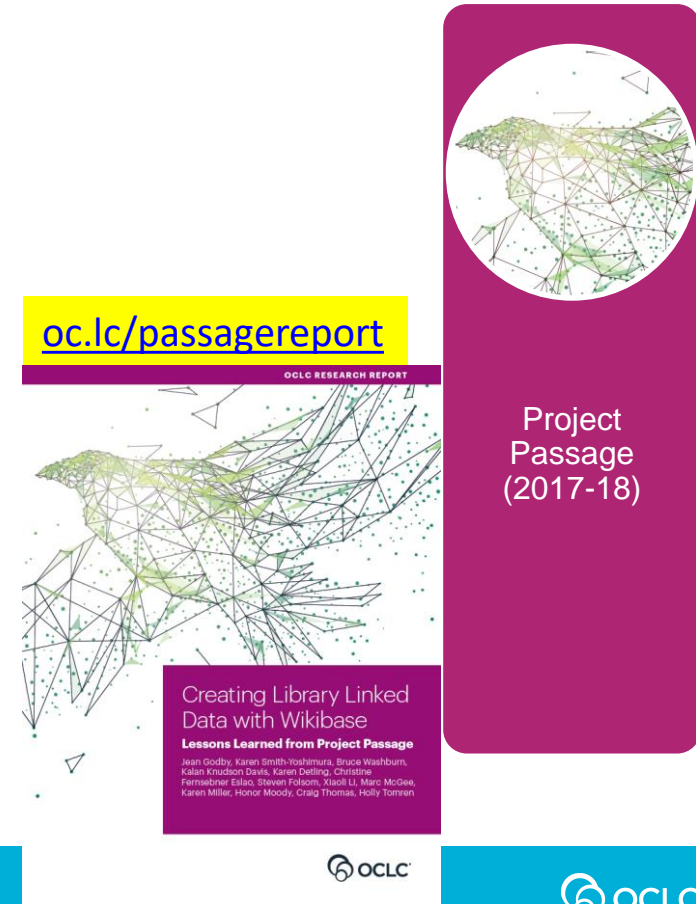


CONTENTdm Linked
Data Pilot (2019-20)

Project Passage: Linked Data Wikibase Prototype

Project goals

- Evaluate a framework for reconciling, creating, and managing bibliographic and authority data as linked data entities and relationships.
- Build a community of users who could create and curate data in the ecosystem and imagine or propose future workflows.
- Not originally planned: Evaluate Wikibase and Wikidata as a technical platform



Project
Passage
(2017-18)

Project Passage: Linked Data Wikibase Prototype

Findings (Part 1)

- Wikibase can be used to create structured data with a precision that exceeds current library standards.
- The platform enables user-driven ontology design but raises concerns about how to manage and maintain ontologies.
- The platform, supplemented with OCLC's enhancements and stand-alone utilities, enables librarians to see the results of their effort in a discovery interface without leaving the metadata-creation workflow.

oclc.org/linkeddata/wikibase

The screenshot shows a search for 'abraham' in Wikibase. The results list 'Abraham' (Biblical patriarch), 'Abraham Lincoln' (16th President of the United States), and 'Abrahamic religion'. A callout points to the 'Abraham Lincoln' entry, stating 'Information for disambiguation'. Below, a table shows 'Abraham Lincoln' in multiple languages. A callout points to the 'URIs' column. Another callout points to the 'Statements' section, listing 'farmer', 'politician', and 'lawyer' with their respective reference counts. A final callout points to the 'Identifiers for VIAF, FAST, LCNAF, WikiData' section.

Language	Label	Description	Also known as
English	Abraham Lincoln	16th President of the United States	Honest Abe Lincoln Abe Lincoln
Spanish	Abraham Lincoln	decimosexto presidente de los Estados Unidos	
Chinese	亞伯拉罕·林肯	第16任美國總統	林肯

Statements

occupation	farmer	0 references
	politician	0 references
	lawyer	

Project Passage: Linked Data Wikibase Prototype

Findings (Part 2)

- Robust tools are required for local data management.
- To populate knowledge graphs with library metadata, tools that facilitate the import and enhancement of data created elsewhere are recommended.
- The pilot underscored the need for interoperability between data sources, both for ingest and export.
- The traditional distinction between authority and bibliographic data can disappear in a linked data description.

The screenshot shows a search for 'abraham' in Wikibase. The results list 'Abraham' (Biblical patriarch), 'Abraham Lincoln' (16th President of the United States), and 'Abrahamic religion'. The 'Abraham Lincoln' entry is selected, showing a disambiguation table and a list of statements.

Information for disambiguation

Language	Label	Description	Also known as
English	Abraham Lincoln	16th President of the United States	Honest Abe Lincoln Abe Lincoln
Spanish	Abraham Lincoln	decimosexto presidente de los Estados Unidos	
Chinese	亞伯拉罕·林肯	第16任美國總統	林肯

URIs

Occupation, place of birth, type, sex or gender, place of death, spouse, child

Statement	Value	References
occupation	farmer	0 references
	politician	0 references
	lawyer	

Identifiers for VIAF, FAST, LCNAF, WikiData

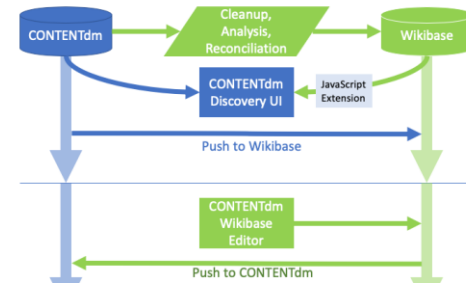
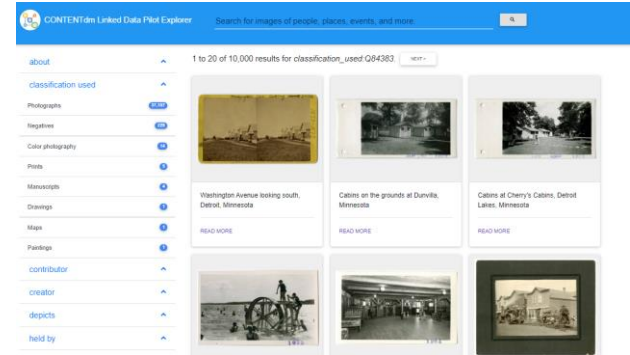
CONTENTdm Linked Data Pilot

CONTENTdm
Linked Data Pilot

CONTENTdm
Linked Data
Pilot (2019-20)

Project goals

- Developing the scalable methods and approaches needed to produce richer, state-of-the-art machine representations of entities and relationships to make visible connections that were formerly invisible.
- Prototype an application for library staff to:
 - convert existing record-based metadata into linked data by replacing strings of characters with identifiers from known authority files and local library-defined vocabularies
 - manage and publish the resulting entities and relationships



Phase 1:

- Try to map CONTENTdm data to entities
- Engage CONTENTdm users to help
- Encourage best practices for CONTENTdm metadata curation

Phase 2:

- Provide UI to create and maintain CONTENTdm data in Wikibase
- Handle unreconciled data

MASS AGGREGATION

Data Harvest & Explorer

- Developed a prototype of the IIF Change Discovery API for 13 Million CONTENTdm items
- Harvested the metadata
- Reconciled string headings to linked data URIs
 - Only looked at strings that were associated with Dublin Core fields
 - Limited those strings to ones that occurred more than 2000 times
- Built a prototype application to search and explore the 13 Million harvested CONTENTdm records
 - <https://researchworks.oclc.org/iiif-explorer/>

- Type ▼
- Audience ▼
- Creator ▼
- Contributor ▼
- Subject ▼
- Place ▼
- Publisher ▼
- Organization ▲
- Oberlin College
72
- Missouri State Library
50
- Houston Public Library
34
- Hamilton College
32
- Illinois State Library
30
- Michigan Department of Natural...
27
- University of Nevada, Las Vegas
24
- Indiana Historical Society
19
- Gonzaga University
16
- Digital Horizons - North Dakota ...
13
- Ohio History Connection
12
- Denver Public Library
12

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

Louis Armstrong

← 1 TO 20 OF 517 →



Louis Armstrong

Houston Public Library
Image Collections

VIEW ▼



Louis Armstrong

Temple University
John W. Mosley Photographs

VIEW ▼



Louis Armstrong

CSUC (Consorc de Serveis
Universitaris de Catalunya)
Fons fotogràfic del Palau de la Música
Catalana (Orfeo Català)

VIEW ▼



Louis Armstrong

Indiana Historical Society
Duncan Schiedt Collection

VIEW ▼



[Louis Armstrong]

Pikes Peak Library District
African Americans In Colorado Springs

VIEW ▼



Louis Armstrong



Louis Armstrong



Louis Armstrong



Louis Armstrong



Louis Armstrong

☰ Louis Armstrong



1 of 1 • Louis Armstrong

Louis Armstrong

Subject

- [singer](#)
- [musician](#)
- [jazz musician](#)
- [African Americans](#)
- [United States of America](#)
- [jazz](#)

Type

- [Image](#)

Organization

- [Temple University](#)

Collection

- [John W. Mosley Photographs](#)

More Like This



Findings

- It was exciting to find unexpected things in unexpected places
- Reconciliation was limited based on
 - Scale of the data
 - Heterogenous nature of metadata
 - Algorithmic approach to matching
- Exercise in harvesting and mapping record-based data for discovery

LINKED DATA FROM SCRATCH

Data Harvest & Explorer (part 2)

- Worked with 5 CONTENTdm users and selected 3 collections from each
- Manually reviewed, mapped, and reconciled metadata
- Imported the data into a Wikibase instance for management
- Built a new prototype application to search and explore the data ingested into the Wikibase instance

about



Jazz

9

African American men

9

Philadelphia

9

Musicians

9

African Americans

8

Composers

8

Entertainers

8

African American pioneers

8

African American entertainers

8

Jazz musicians

8

African American icons

8

classification used



depicts



part of



photographer



1 to 10 of 10 results for *louis armstrong*.



Louis Armstrong

[READ MORE](#)



Louis Armstrong

[READ MORE](#)



Louis Armstrong

[READ MORE](#)



Louis Armstrong



Louis Armstrong



Louis Armstrong and Jake Armstrong



☰ Louis Armstrong



Satch," was widely recognized as a founding father of jazz. His influence, as an artist and cultural icon, is universal, unmatched, and very much alive today.

date created

1944

height

8 inch

width

10 inch

part of

[John W. Mosley Photograph Collection](#)

classification used

[Photographs](#)

process or format

[Black and white prints](#)

about

[Philadelphia](#) | [Entertainers](#) | [Musicians](#) | [Singers](#) | [Jazz](#) | [African Americans](#) | [Composers](#) | [African American entertainers](#) | [Jazz musicians](#) | [African American men](#) | [African American pioneers](#) | [African American icons](#) | [Composers--United States](#) | [Innovators](#)

depicts

[Louis Armstrong](#)

Armstrong was a musician, composer, jazz trumpeter, film star, and comedian. Considered one of the most artists in jazz history, Armstrong is known for songs like "Blue," "La Vie En Rose" and "What a Wonderful World." In 1971, the man known around the world as "Ambassador" was recognized as a founding father of jazz. His influence, as an artist and cultural icon, is universal, unmatched, and very much alive today.

[John W. Mosley Photograph Collection](#)

[Philadelphia](#) | [Entertainers](#) | [Musicians](#) | [Singers](#) | [Jazz](#) | [African Americans](#) | [Composers](#) | [African American entertainers](#) | [Jazz musicians](#) | [African American men](#) | [African American pioneers](#) | [African American icons](#) | [Composers--United States](#) | [Innovators](#)

View in [CONTENTdm](#) | View in the [Pilot Wikibase](#)
IIIF Presentation Manifest viewer provided by [Project M](#)
<http://rightsstatements.org/vocab/InC/1.0/>



[about](#) ^

[classification used](#) ^

Photographs **8**

[depicts](#) ^

[part of](#) ^

[photographer](#) ^

[process or format](#) ^

1 to 8 of 8 results for *depicts:Q161624*.



Louis Armstrong

[READ MORE](#)



Louis Armstrong

[READ MORE](#)



Louis Armstrong

[READ MORE](#)



Louis Armstrong and Jake Armstrong

[READ MORE](#)



Louis Armstrong Outside of the Pyramid Club

[READ MORE](#)



Louis Armstrong

[READ MORE](#)



Louis Armstrong and Jake Armstrong

[READ MORE](#)



Louis Armstrong

[READ MORE](#)

Item Discussion

Read

View history

More ▾

Search CONTENTdm Linked Data Pilot



Louis Armstrong (Q161624)

American jazz trumpeter, composer and singer

Satchmo | Pops | Armstrong, Louis

edit

◄ In more languages Configure

Language	Label	Description	Also known as
English	Louis Armstrong	American jazz trumpeter, composer and singer	Satchmo Pops Armstrong, Louis
español	Louis Armstrong	trompetista y cantante estadounidense de jazz	Satchmo Pops

All entered languages

Constraint Violation Report View

Statements

type	person ↔ edit
	► 1 reference
	+ add value

website	http://www.louisarmstronghouse.org/ edit
	► 1 reference
	+ add value

sex or gender	male ↔ edit
	► 1 reference
	+ add value

birth date	4 August 1901 <i>Gregorian</i> edit
	► 1 reference
	+ add value

Context and Background

Louis Armstrong (August 4, 1901 – July 6, 1971), nicknamed Satchmo or Pops, was an American trumpeter, composer, singer and occasional actor who was one of the most influential figures in jazz. His career spanned five decades, from the 1920s to the 1960s, and different eras in jazz. Coming to prominence in the 1920s as an "inventive" trumpet and cornet player, Armstrong was a foundational influence in jazz, shifting the focus of the music from collective improvisation to solo performance. With his instantly recognizable gravelly voice, Armstrong was also an influential singer, demonstrating great dexterity as an improviser, bending the lyrics and melody of a song for expressive purposes. He was also skilled at scat singing. Renowned for his charismatic stage presence and voice almost as much as for his trumpet-playing, Armstrong's influence extends well beyond jazz music, and by the end of his career in the 1960s, he was widely regarded as a profound influence on popular music in general. Armstrong was one of the first truly popular African-American entertainers to "cross over", whose skin color was secondary to his music in an America that was extremely racially divided. He rarely publicly politicized his race, often to the dismay of fellow African-Americans, but took a well-publicized stand for desegregation in the Little Rock Crisis. His artistry and personality allowed him socially acceptable access to the upper echelons of American society which were highly restricted for black men of his era.

Sources: DBpedia Wikipedia Wikimedia Commons

Main page
Recent changes
Random page
Help about MediaWiki

Tools

What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information
Concept URI

In other languages

Add links



UPDATE THE ENTITY

- About Philadelphia [X]
- Depicts _____
- About Entertainers [X]
- Depicts _____
- About Musicians [X]
- Depicts _____
- About Singers [X]
- Depicts _____
- About Jazz [X]
- Depicts _____
- About African Americans [X]
- Depicts _____
- About Composers [X]
- Depicts _____
- About African American entertainers [X]
- Depicts _____
- About Jazz musicians [X]
- Depicts _____
- About African American men [X]
- Depicts _____



UPDATE THE ENTITY

- About Philadelphia [X]
- Depicts Philadelphia [X]
- About Entertainers [X]
- Depicts Entertainers [X]
- About Musicians [X]
- Depicts Musicians [X]
- About Singers [X]
- Depicts Singers [X]
- About Jazz [X]
- Depicts Jazz [X]
- About African Americans [X]
- Depicts African Americans [X]
- About Composers [X]
- Depicts Composers [X]
- About African American entertainers [X]
- Depicts African American entertainers [X]
- About Jazz musicians [X]
- Depicts Jazz musicians [X]
- About African American men [X]
- Depicts African American men [X]



UPDATE THE ENTITY

Depicts

About Jazz musicians

Depicts

About African American men

Depicts

About African American pioneers

Depicts

About African American icons

Depicts

About Composers--United States


Depicts

About Innovators

Depicts

About Louis Armstrong

Depicts



+ ADD A DEPICTION



UPDATE THE ENTITY

- About African American pioneers ✖


Depicts
- About African American icons ✖

Depicts
- About Composers--United States ✖


Depicts
- About Innovators ✖

Depicts
- About Louis Armstrong 📷 ✖

Depicts


- About Trumpet 📷 ✖

Depicts



+ ADD A DEPICTION

depicts	<p> African American entertainers </p> <p>▼ 0 references</p> <p>+ add reference</p>	 edit
	<p> Jazz musicians </p> <p>▼ 0 references</p> <p>+ add reference</p>	 edit
	<p> African American men </p> <p>▼ 0 references</p> <p>+ add reference</p>	 edit
	<p> Louis Armstrong </p> <p>digital representation URL https://cdm16002.contentdm.oclc.org/digital/iiif/p15037coll17/60/2029_969_2712_3385/full/0/default.jpg</p>  <p>▼ 0 references</p> <p>+ add reference</p>	 edit
	<p> Trumpet </p> <p>digital representation URL https://cdm16002.contentdm.oclc.org/digital/iiif/p15037coll17/60/3092_1682_1649_1119/full/0/default.jpg</p>  <p>▼ 0 references</p> <p>+ add reference</p>	 edit
	+ add value	

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

A semantic continuum



Shared Entity
Management
Infrastructure
(2020-21)



CONTENTdm
Linked Data
Pilot (2019-20)

Shared, homogeneous, and centralized entities...

...accounting for the reality of localized, heterogeneous, de-centralized collections

Machine-matching, highly automated reconciliation...

...with tools for hand-matching, semi-automated reconciliation

**Well-accepted context:
Persons & Works**

**Granular context:
About, Depicts, Annotations, Notes**

Blurs the line between bib and authority work

Blurs the line between object and context description

Custom applications and interfaces needed

This is the new Knowledge Work

Next Steps

- **Shared entity management infrastructure**
 - Targeting millions of entities in the infrastructure by December 2020
 - Complete the project! — December 2021
- **CONTENTdm Linked Data Pilot**
 - Evaluate how to better balance algorithmic record conversion with domain knowledge expertise
 - Determine how to pull apart contextual metadata and descriptive metadata
 - Explore how to leverage the new contextual metadata in end-user applications
 - OCLC Research Report forthcoming later this year
- **More Research**
 - More contextual linked data being added to the infrastructure (e.g. Places)
 - Linked data for Concepts and Subjects
 - Federated Linked Data and Local Subject description: Linking and Localization
 - Entity Alignment: knowledge graph integration from multiple sources
 - Archives & Special Collections Linked Data: moving forward from CONTENTdm prototype and the findings of OCLC's recent report, "Archives and Special Collections Linked Data: Navigating between Notes and Nodes"

Thank you!

Andrew K. Pace

Executive Director, Technical Research

pacea@oclc.org

[@andrewkpace](https://twitter.com/andrewkpace)

<https://www.oclc.org/research/people/pace-andrew.html>

**Because
what is
known must
be shared.®**