2020 LD4 Conference, July 2020

OCLC Linked Data: Research, experimental applications, and shared infrastructure

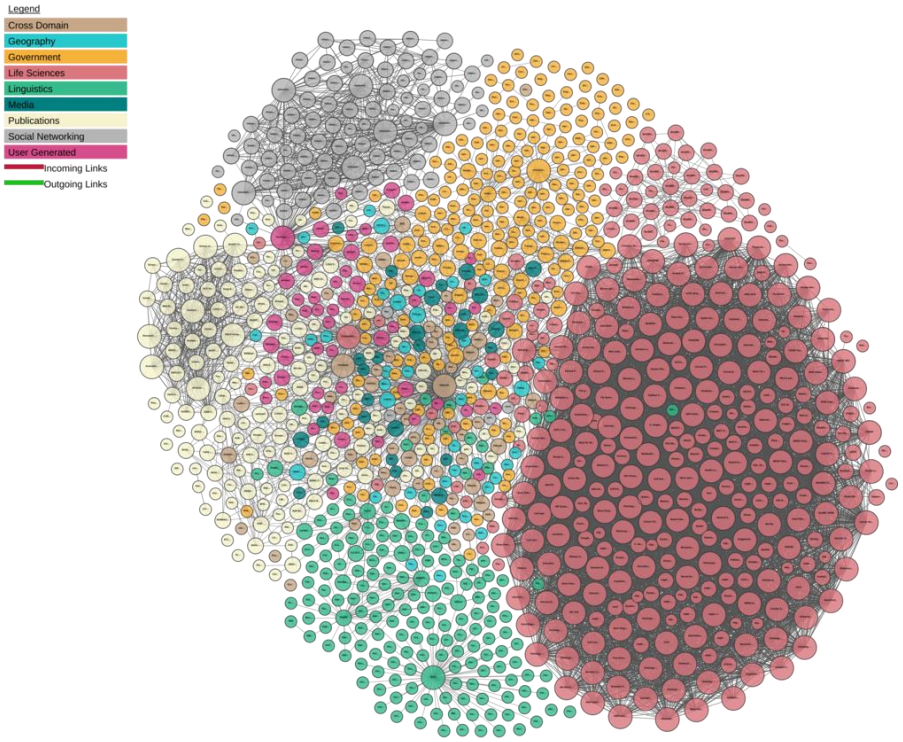
Andrew K. Pace
Executive Director,
Technical Research
OCLC

John Chapman
Senior Product Manager,
Metadata Strategy & Operations
OCLC

Agenda

- Why linked data?
- 5 Habits of successful pilots and prototypes
- Research and Findings: a decade of linked data research
- A shared Entity Management Infrastructure

Why linked data?



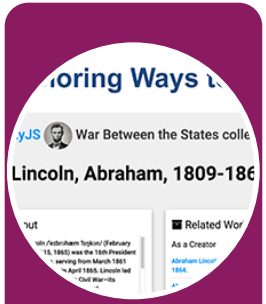
5 Habits of successful pilots and prototypes

- **Vision statement** – set out what you want to prove or disprove
- **Justification** – a succinct business-focused view that helps determine the amount of effort needed before you start
- **Partners** – find real users who can evaluate tools, workflows, data, and models based on their real-life use cases
- **Expectations** – demand participation, expect resistance, set an end date, hope for something that is different than initially imagined
- **Acceptance** – not every idea is a winner; prototypes and pilots will shift and change focus; document your process and findings

A decade with Linked Data

oc.lc/linkeddataresearch

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)

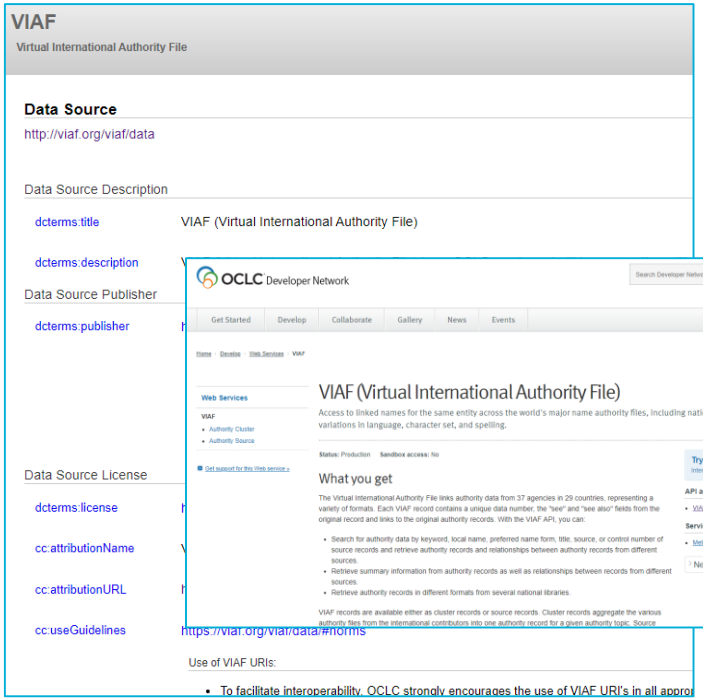
CONTENTdm Linked Data Pilot (2019-20)

Shared Entity Management Infrastructure (2020-21)

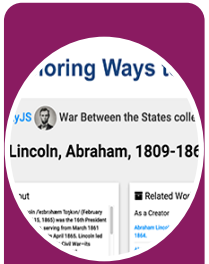
VIAF and FAST: early additions to the LD landscape

Virtual International Authority File (VIAF)

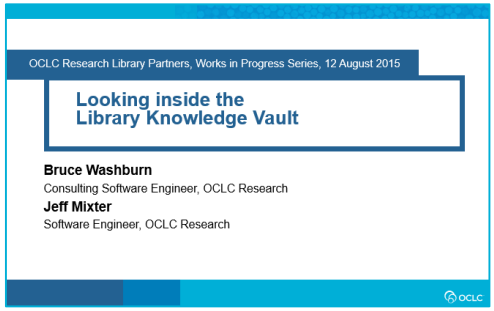
Faceted Application of Subject Terminology (FAST)

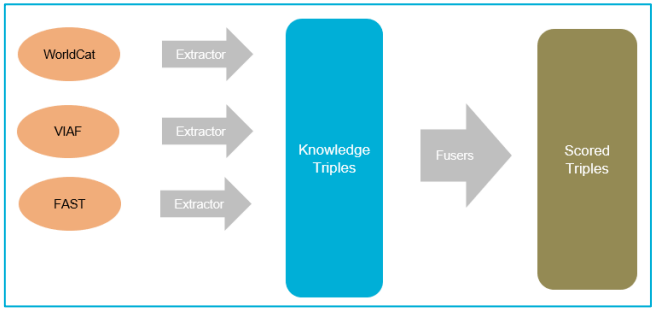


EntityJS: entities & their relationships



EntityJS Research Project (2013)





Project Goals

- Prototype an application that runs in a browser and uses RDF data sources from OCLC and elsewhere
- Search across entities and show relationships of one entity to others
- Examine questions around user-contributed improvements to entity relationships

Findings

- Co-occurrence of entities mentioned in descriptions of creative works shows important relationships; aggregation adds value.

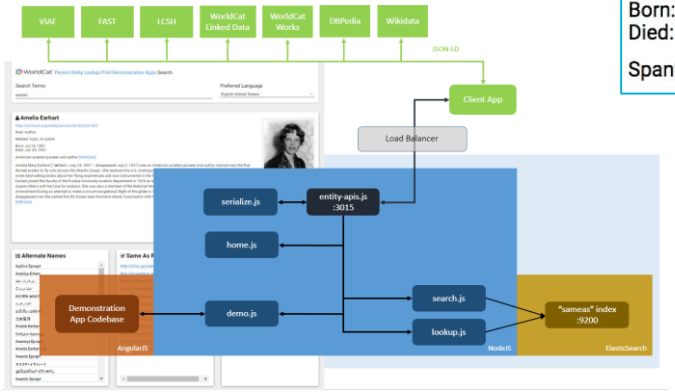
Person Entity Lookup Pilot

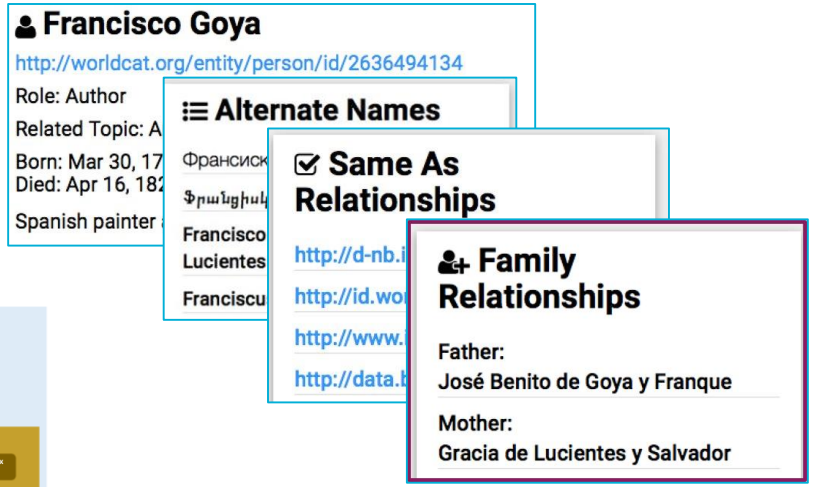
Project goals

- improve access to entities via “API First” services
- Determine changes needed in indexing, data, workflow to improve metadata creation and Improve discovery outcomes

Findings

- Many sources available
- Data Aggregation is crucial
- Workflow is the cataloger’s delight

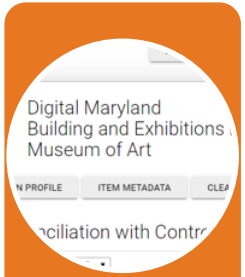




WorldCat
Person Entity Lookup Pilot

Person Entity Lookup Pilot (2014)

CONTENTdm Metadata Refinery



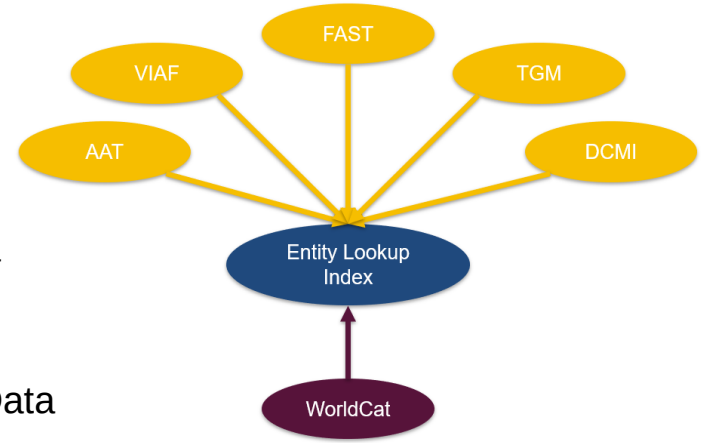
CONTENTdm Metadata Refinery (2015-16)

Starting points: Distinctive Collections

Project Goals:

- Building a web app to help CONTENTdm sites create linked data from scratch
  - **CLEAN UP** the data,
  - **MAP** the local fields to a common schema
  - **RECONCILE** field values against shared vocabularies to get persistent identifiers
  - **TRANSFORM** the data into RDF Linked Data

Metadata Refinery Entity Lookup Index



Findings:

- Aggregation adds value
- Centralize the web app tools
- Decentralize the work of cleanup, mapping, and refining/correcting entity lookup results

Project Passage: Linked Data Wikibase Prototype

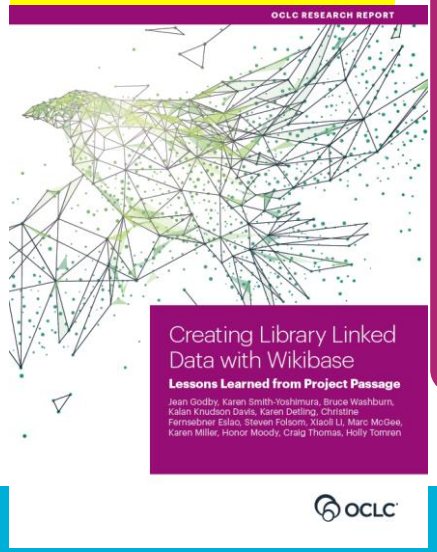
Project goals

- 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.
- [Evaluate Wikibase and Wikidata as a technical platform]

Method

- A Wikibase/Wikidata sandbox in which librarians from 16 US institutions could experiment with creating linked data to describe resources—without requiring knowledge of the technical machinery of linked data.
- Use cases where pilot participants created metadata for resources in various formats and languages using the Wikibase editing interface.

oc.lc/passagereport

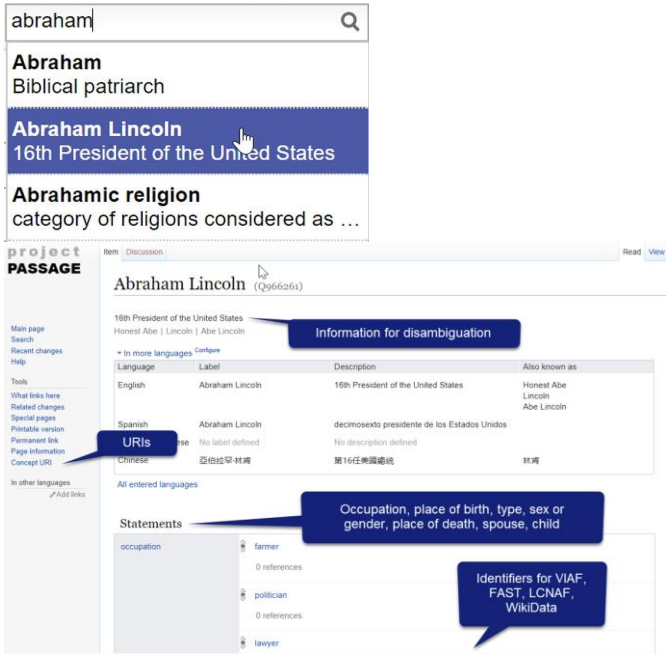


Project Passage (2017-18)

Project Passage: Linked Data Wikibase Prototype

Findings

- 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.
- 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.

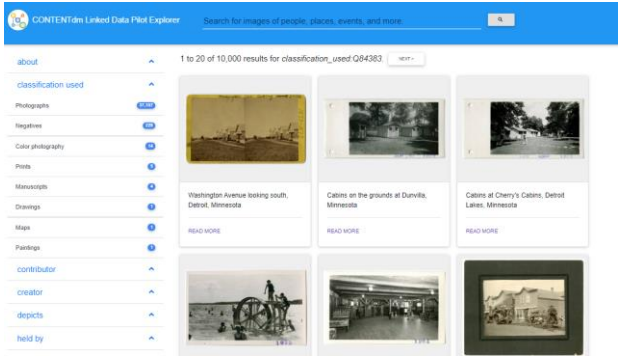


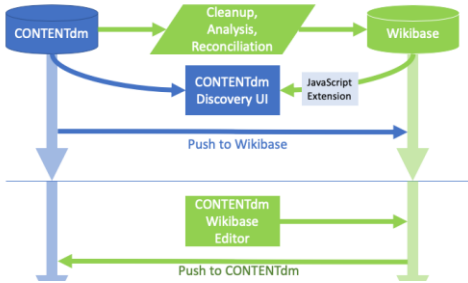
CONTENTdm Linked Data Pilot

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

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”
    - Stand behind entity URIs
  - Operate at a large scale – and be sustainable
  - Complement other efforts (including LD4P!)

Entity Management

- 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

Shared Entity Management Infrastructure

(2020-21)

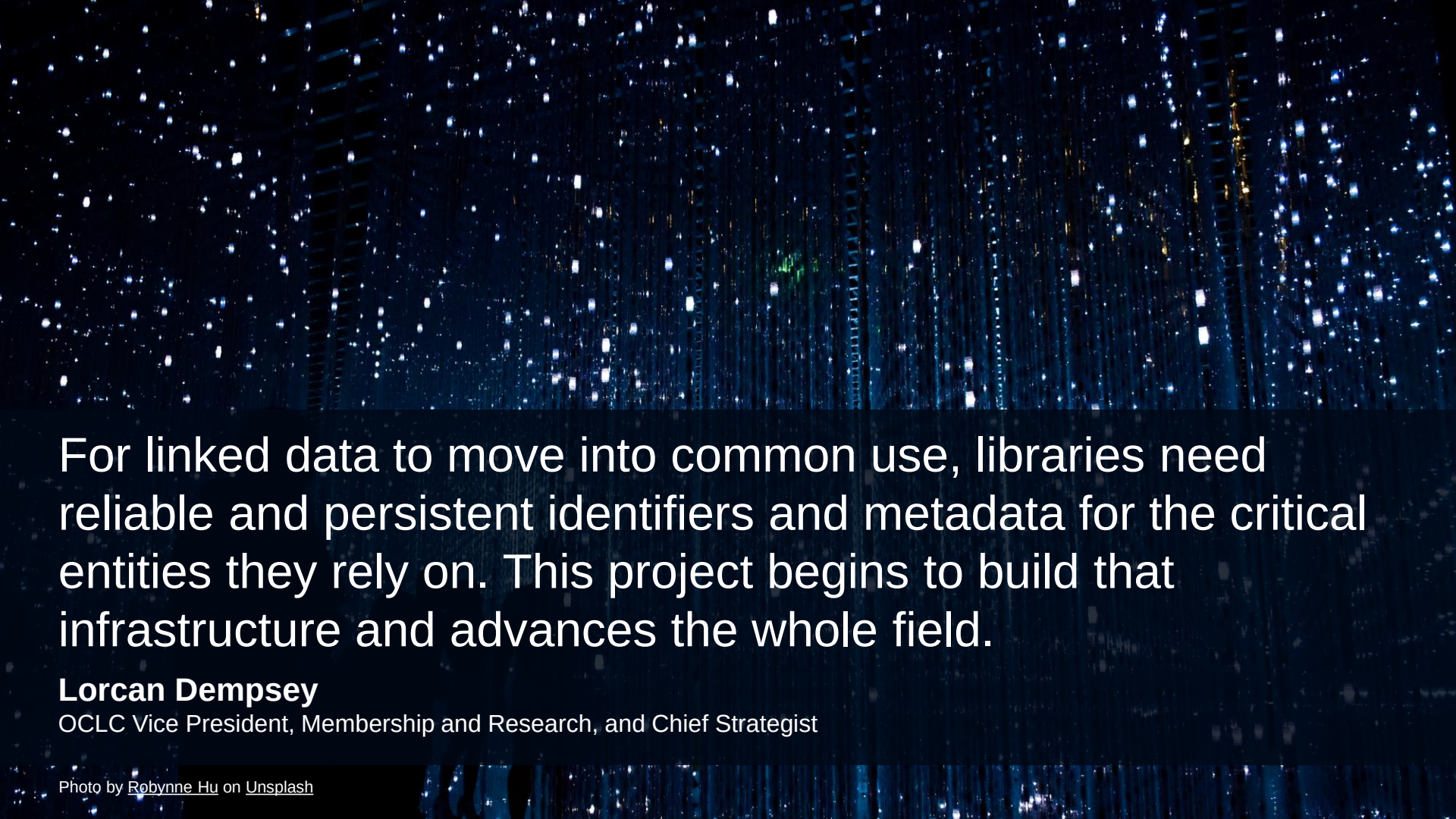
OCLC awarded Mellon Foundation grant to develop infrastructure to support linked data management initiatives

'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 Mellon Foundation for their generous support for this project. And we're eager to apply our knowledge and expertise to develop this infrastructure on behalf of libraries and the scholarly communications community."



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)

- 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

BnF Bibliothèque nationale de France

Alma Mater Studiorum Università di Bologna A.D. 1088

Princeton University

Penn University of Pennsylvania

British Library

NLB National Library Board Singapore

Deutsche National Bibliothek

Yale

Smithsonian

Cleveland Public Library

abes agence bibliographique de l'enseignement supérieur

BYU Brigham Young University

Biblioteca nazionale centrale di Roma

NIH U.S. National Library of Medicine

University of Oxford

CMU Central Michigan University

University of Minnesota

University of Victoria

Cornell University Founded A.D. 1865

NYU

Temple University

Harvard University

The University of Tennessee Knoxville University Libraries

UC Davis

Entity Management

Shared Entity Management Infrastructure

(2020-21)

- Currently in testing phase for first increment
  - 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

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

EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)

CONTENTdm Metadata Refinery (2015-16)

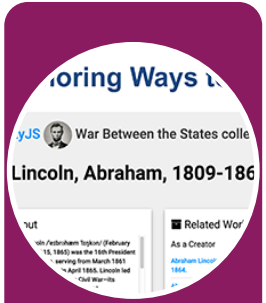
Project Passage (2017-18)

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)

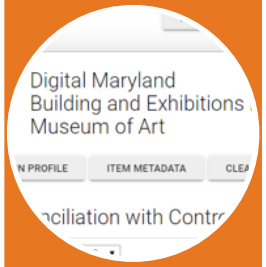
Shared Entity Management Infrastructure (2020-21)

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)

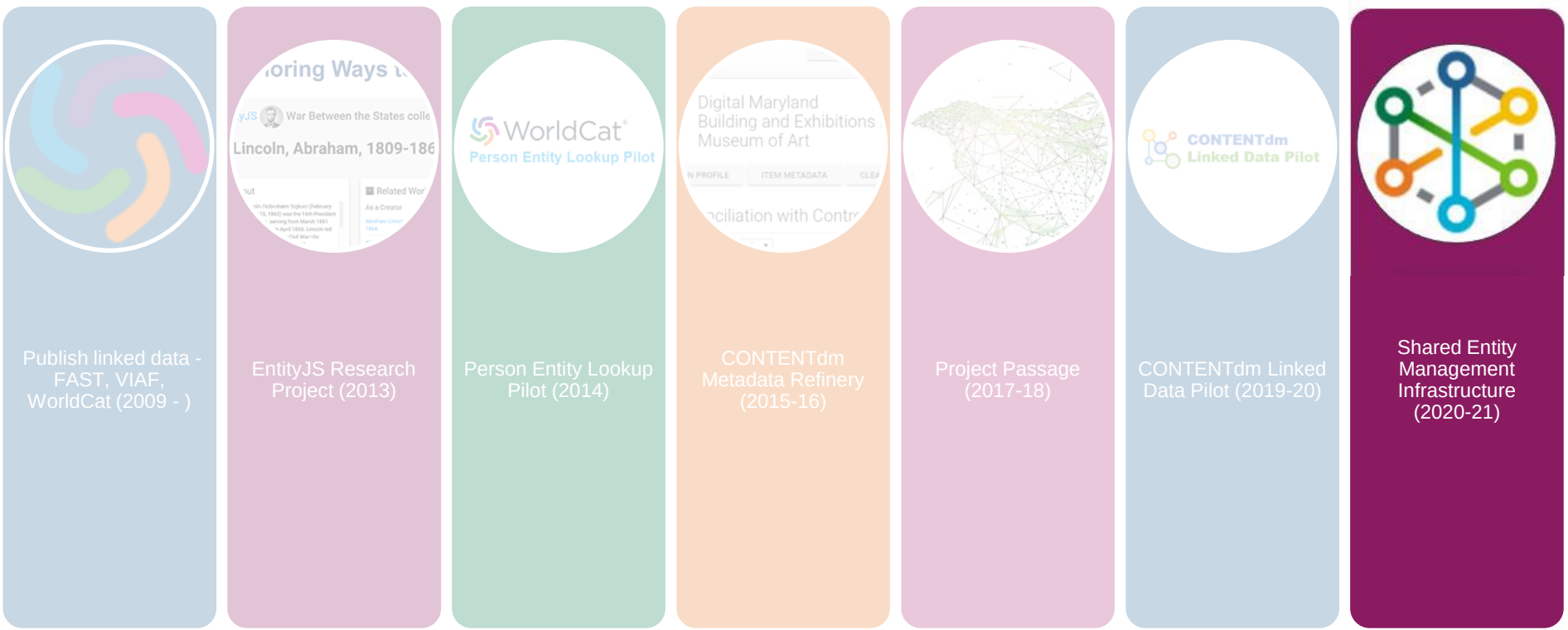
CONTENTdm Metadata Refinery (2015-16)

Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Shared Entity Management Infrastructure (2020-21)

CONTENTdm Linked Data Pilot: Think "long tail". Attend to the issues around the rare, local, and unique.

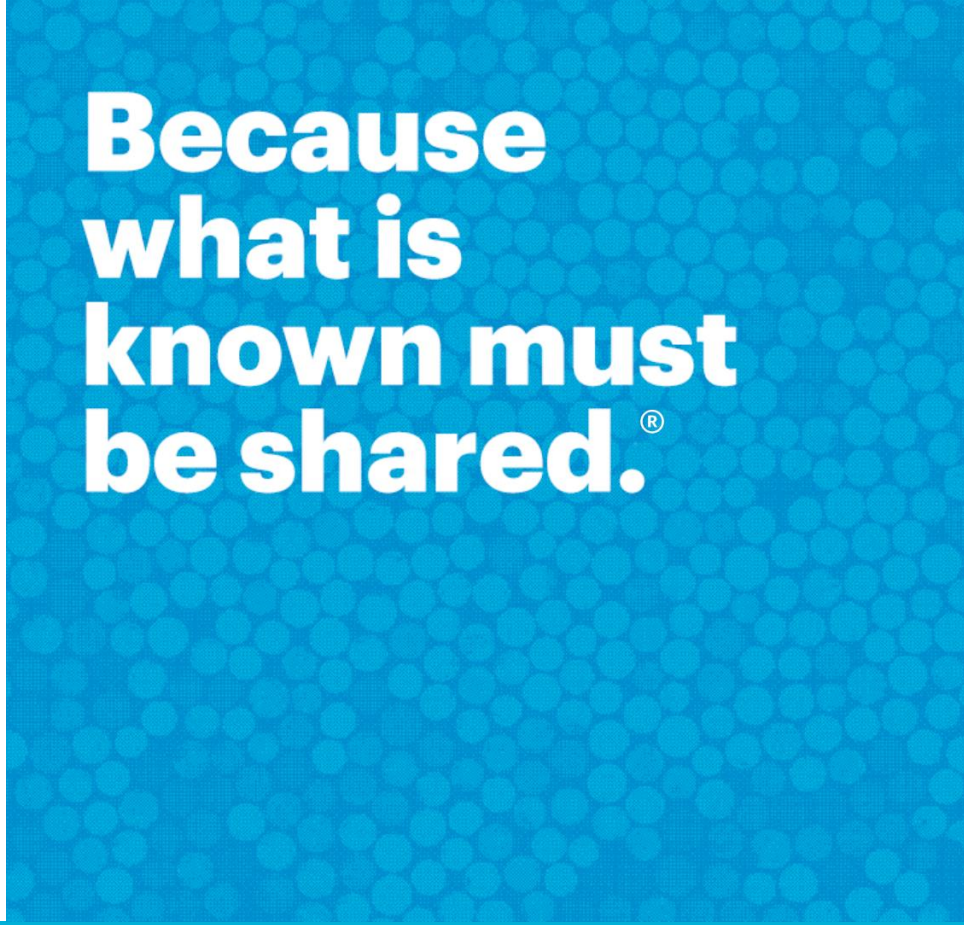


Entity Management: The Future is Now. Given our deep experience, build production entity management data and services at a global scale.

Thank you!

Andrew K. Pace
Executive Director, Technical Research
pacea@oclc.org
@andrewkpace
https://www.oclc.org/research/people/pace-andrew.html

John Chapman
Senior Product Manager,
Metadata Strategy and Operations
chapmanj@oclc.org



OCLC