OCLC and Linked Data: An update on infrastructure testing and linked data quality

Anne Washington  
OCLC

Laura Ramsey  
OCLC

Charlene Chou  
NYU Libraries
Agenda

• Welcome
• Presentations
• Questions/discussion
Shared Entity Management Infrastructure

- Project overview and updates
- Testing insights
- Findings from data workshops
Anne Washington
Product Analyst
OCLC

Shared Entity Management Infrastructure

- Project overview and updates
- Testing insights
- Findings from data workshops

Laura Ramsey
Senior Metadata Operations Manager
OCLC

The OCLC Quality Metric
Anne Washington
Product Analyst
OCLC

Shared Entity Management Infrastructure
• Project overview and updates
• Testing insights
• Findings from data workshops

Laura Ramsey
Senior Metadata Operations Manager
OCLC

The OCLC Quality Metric

Charlene Chou
Head, Knowledge Access
NYU Libraries

Is "Minimum Viable Entity Description" "viable" for clustering and discovery?: SEMI (Shared Entity Management Infrastructure) test outcomes and observations
Charlene Chou
Division of Libraries, New York University
OCLC and Linked Data Virtual Meeting
October 27, 2022
Shared Entity Management Infrastructure

- Project overview and updates
- Testing insights
- Findings from data workshops

Anne Washington
Product Analyst
Metadata Services
washinga@oclc.org
OVERVIEW AND UPDATES
Project overview

• Two-year, $2.436M Mellon grant, matched by OCLC
  – December 2019 - December 2021

• Production infrastructure for Person and Work entities
  – Persistent identifiers for entities in aggregation
  – Create and edit entities, with new IDs and changes reflected immediately
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 project, with the remaining funding coming from OCLC and other libraries.

“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.”
Current focus

- Refine methods for building entities
- Entity view in UI and Read API
- Create and edit entities via UI and API
- Language support
- SPARQL
**Jackie Robinson**

- **Entity ID:** Q561707
- **Entity Type:** Person
- **Description:** American baseball player

### Key Information

<table>
<thead>
<tr>
<th>Language of Content</th>
<th>Entity Label</th>
<th>Description</th>
<th>Aliases</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Jackie Robinson</td>
<td>American baseball player</td>
<td>3</td>
</tr>
<tr>
<td>German</td>
<td>Jackie Robinson</td>
<td>enser afroamerikanischer Baseball-Spieler in der MLB</td>
<td>1</td>
</tr>
<tr>
<td>Spanish</td>
<td>Jackie Robinson</td>
<td>Primer beisbolista afroamericano en integrar un equipo de las ligas m...</td>
<td>2</td>
</tr>
<tr>
<td>French</td>
<td>Jackie Robinson</td>
<td>joueur de baseball américain</td>
<td>2</td>
</tr>
<tr>
<td>Dutch</td>
<td>Jackie Robinson</td>
<td>Amerikaans honkballer (1919-1972)</td>
<td>1</td>
</tr>
</tbody>
</table>
**Jackie Robinson**

- **Entity ID:** Q561707
- **Entity Type:** Person
- **Description:** American baseball player

**Key Information**

<table>
<thead>
<tr>
<th>Language of Content</th>
<th>Entity Label</th>
<th>Description</th>
<th>Aliases</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Jackie Robinson</td>
<td>American baseball player</td>
<td>3</td>
</tr>
<tr>
<td>German</td>
<td>Jackie Robinson</td>
<td>erster afroamerikanischer Baseball-Spieler in der MLB</td>
<td>1</td>
</tr>
<tr>
<td>Spanish</td>
<td>Jackie Robinson</td>
<td>Primer beisbolista afroamericano en integrar un equipo de las ligas m...</td>
<td>2</td>
</tr>
<tr>
<td>French</td>
<td>Jackie Robinson</td>
<td>joueur de baseball américain</td>
<td>2</td>
</tr>
<tr>
<td>Dutch</td>
<td>Jackie Robinson</td>
<td>Amerikaans honkballer (1919-1972)</td>
<td>1</td>
</tr>
</tbody>
</table>
# View

## Jackie Robinson

**Entity ID:** Q561707  
**Entity Type:** Person  
**Description:** American baseball player

## Key Information

<table>
<thead>
<tr>
<th>Language of Content</th>
<th>Entity Label</th>
<th>Description</th>
<th>Aliases</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Jackie Robinson</td>
<td>American baseball player</td>
<td>3</td>
</tr>
<tr>
<td>German</td>
<td>Jackie Robinson</td>
<td>erster afroamerikanischer Baseball-Spieler in der MLB</td>
<td>1</td>
</tr>
<tr>
<td>Spanish</td>
<td>Jackie Robinson</td>
<td>Primer beisbolista afroamericano en integrar un equipo de las ligas m...</td>
<td>2</td>
</tr>
<tr>
<td>French</td>
<td>Jackie Robinson</td>
<td>joueur de baseball américain</td>
<td>2</td>
</tr>
<tr>
<td>Dutch</td>
<td>Jackie Robinson</td>
<td>Amerikaans honkballer (1919-1972)</td>
<td>1</td>
</tr>
<tr>
<td>Language</td>
<td>Jackie Robinson</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>----------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>Jackie Robinson</td>
<td>Hungarian</td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>Jackie Robinson</td>
<td>Indonesian</td>
<td></td>
</tr>
<tr>
<td>Irish</td>
<td>Jackie Robinson</td>
<td>Irish</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>Jackie Robinson</td>
<td>Italian</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>Jackie Robinson</td>
<td>Japanese</td>
<td></td>
</tr>
</tbody>
</table>
Names

<table>
<thead>
<tr>
<th>Name</th>
<th>Name Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack Roosevelt Robinson</td>
<td>Birth Name</td>
</tr>
</tbody>
</table>

External Sources

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Source ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDL Auth ID</td>
<td>00454404</td>
</tr>
<tr>
<td>IM/Do ID</td>
<td>nm0732697</td>
</tr>
<tr>
<td>BNE ID</td>
<td>XX5484585</td>
</tr>
<tr>
<td>Library of Congress</td>
<td>n79141359</td>
</tr>
<tr>
<td>GND ID</td>
<td>11954248X</td>
</tr>
</tbody>
</table>

Significant Dates

- **01/31/1919**
  - Date of Birth

- **10/24/1972**
  - Date of Death
### External Sources

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Source ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDL Auth ID</td>
<td>00454404</td>
</tr>
<tr>
<td>OMDo ID</td>
<td>nm0732697</td>
</tr>
<tr>
<td>BNE ID</td>
<td>0005484585</td>
</tr>
<tr>
<td>Library of Congress Name Auth ID</td>
<td>n79141357</td>
</tr>
<tr>
<td>GND ID</td>
<td>11954248X</td>
</tr>
</tbody>
</table>

### Significant Dates

- **01/31/1919**: Date of Birth
- **10/24/1972**: Date of Death

2 references

2 references
TESTING INSIGHTS
Advisory group members
Our goals

• Learn about experience with new UI
• Gather feedback on data
• Identify needs for future iterations
Areas of focus

- User interface
- API
- Entity data
  - Completeness of entities varies; impacts disambiguation
  - Additional details in description will help with disambiguation, especially for works
  - Need more connections between works and persons
What next?

• Continue testing with Advisory Group members
• Feedback helps inform priorities
• Plan for future community input
DATA WORKSHOPS
Themes explored

- Shifting from records to entities
- Data models & cataloging workflows
- Representing complex resources
- Disambiguation
Going forward

• Analyze and share results
• Incorporate ideas in post-grant infrastructure
• Shape our communication
STAY TUNED
Thank you!

LinkedData@oclc.org

Because what is known must be shared.
The OCLC Quality Metric

Laura Ramsey
Senior Metadata Operations Manager
ramseyl@oclc.org
Quality metric goals

1. Calculate fit-for-purpose of the entities
2. Enable workflows based on the quality score
3. Provide comprehensive view of WorldCat entity quality
Scope

WorldCat entities

Scores for individual entities
Scope

WorldCat entities

Scores for individual entities
Scope

WorldCat entities

Scores for individual entities

Wikidata entities

Scores in relation to all entities
Scope

WorldCat entities

Scores for individual entities

Wikidata entities

Scores in relation to all entities
Metric components

- Completeness
- Confidence
- Disambiguation
- Structure
Metric components

- Completeness
- Adherence to MVED
- Confidence
- Disambiguation
- Structure
# Minimum Viable Entity Description

<table>
<thead>
<tr>
<th>Person</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>Emily Brontë</td>
</tr>
<tr>
<td>Description</td>
<td>English novelist and poet</td>
</tr>
<tr>
<td>Alias</td>
<td>Emily Jane Brontë</td>
</tr>
<tr>
<td>Instance of</td>
<td>person</td>
</tr>
<tr>
<td>Date of birth</td>
<td>07/30/1818</td>
</tr>
<tr>
<td>Date of death</td>
<td>12/19/1848</td>
</tr>
<tr>
<td>External identifier</td>
<td>VIAF, LCNAF, FAST, Wikidata, etc.</td>
</tr>
<tr>
<td>Work</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Label</td>
<td>Jane Eyre</td>
</tr>
<tr>
<td>Description</td>
<td>A novel by English writer Charlotte Brontë</td>
</tr>
<tr>
<td>Alias</td>
<td>Jane Eyre: an autobiography</td>
</tr>
<tr>
<td>Instance of</td>
<td>text</td>
</tr>
<tr>
<td>Title</td>
<td>Jane Eyre</td>
</tr>
<tr>
<td>Agent</td>
<td>Charlotte Brontë</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Creation date</td>
<td>1847</td>
</tr>
<tr>
<td>OCLC number</td>
<td>1113914124</td>
</tr>
</tbody>
</table>
Metric components

- Completeness
  - Adherence to MVED
- Confidence
  - Presence of reference information
- Disambiguation
- Structure
References

Significant Dates

Date of Birth
02/18/1931

Date of Death
08/05/2019

Rank
Normal

References

VIAF ID
109406177

Ingest ID
VIAF_20210616_121513
Metric components

- Completeness
- Adherence to MVED
- Confidence
- Presence of reference information
- Disambiguation
- Distinguishing similar entities
- Structure
Inbound & Outbound edges

Inbound edges
What links here

Vienna → Beethoven
Scoring entity = “Vienna”
One link, Vienna is linked to by Beethoven

Outbound edges
What links elsewhere

Beethoven → Vienna
Scoring entity = “Beethoven”
One link, links to Vienna
Metric components

- Completeness
  - Adherence to MVED
- Confidence
  - Presence of reference information
- Disambiguation
  - Distinguishing similar entities
- Structure
  - Conformance to property rules & constraints
Structure (constraint examples)

**Single value constraint**
Birthdate: 18 February 1931
Birthdate: 18 February 1931, 12 May 1945

**Distinct value constraint**
Toni Morrison VIAF ID = 109406177
Charlotte Brontë VIAF ID = 109406177

**Domain includes constraint**
Toni Morrison is author of *Beloved* (work)
Toni Morrison is author of Platteville, WI (place)

**Range includes constraint**
*Beloved* is authored by Toni Morrison (person)
*Beloved* is authored by Platteville, WI (place)
Quality metric workflow
The OCLC Quality Metric

Quality metric goals

1. Calculate
2. Enable
3. Provide

Metric components

Quality metric workflow

View

Maya Pergolesi

meta data
person

Description: American author, 1970 (birth) - 2020 (death); lived in dellere for most of her life. Died of cancer. Two of her books were selected in the list. Information on her was found in the OCLC database.

Key Information

Language of Content: English

English: Joe Q. Author

Spanish: Angel Perez

Description: A novel about love and loss in the 21st century. 2020 (death)

Quality Score: 75
Is “Minimum Viable Entity Description” “viable” for clustering and discovery?: SEMI (Shared Entity Management Infrastructure) test outcomes and observations

Charlene Chou
Division of Libraries, New York University

OCLC and Linked Data Virtual Meeting
October 27, 2021
Agenda

- Observations from recently concluded testing
- Perspectives on OCLC’s data modeling
- Skills and competencies that will be important to metadata workers
- Challenges that NYU hopes to address with linked data technologies
SEMI: Work & Person Entity

- Work: the definition of a work combines parts of WEMI work and expression, and uses the Minimum Viable Entity Description (MVED) properties OCLC has defined
- Person: include both human being and fictitious character
## Minimum Viable Entity Description (MVED)

### Person

<table>
<thead>
<tr>
<th>Label</th>
<th>Emily Brontë</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>English novelist and poet</td>
</tr>
<tr>
<td>Alias</td>
<td>Emily Jane Brontë</td>
</tr>
<tr>
<td>Instance of</td>
<td>person</td>
</tr>
<tr>
<td>Date of birth</td>
<td>07/30/1818</td>
</tr>
<tr>
<td>Date of death</td>
<td>12/19/1848</td>
</tr>
<tr>
<td>External identifier</td>
<td>VIAF, LCNAF, FAST, Wikidata, etc.</td>
</tr>
</tbody>
</table>

### Work

<table>
<thead>
<tr>
<th>Label</th>
<th>Jane Eyre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>A novel by English writer Charlotte Brontë</td>
</tr>
<tr>
<td>Alias</td>
<td>Jane Eyre: an autobiography</td>
</tr>
<tr>
<td>Instance of</td>
<td>text</td>
</tr>
<tr>
<td>Title</td>
<td>Jane Eyre</td>
</tr>
<tr>
<td>Agent</td>
<td>Charlotte Brontë</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Creation date</td>
<td>1847</td>
</tr>
<tr>
<td>OCLC number</td>
<td>1113914124</td>
</tr>
</tbody>
</table>
Test examples, outcomes & observations

- **API tests**: we did API search smoothly but there was only broad keyword search and Q number search.
- **UI test examples** focused on classic or popular works with multiple expressions and multilingual resources’ discoverability.
- **Pride and Prejudice**: related works/expressions in NYU collections mostly
  - Original text in English: print book published in 1813
  - Spanish translation published in 2006
  - German translation published in 1980
  - 2006 DVD for 1940 Motion picture
  - 1995 music CD for 1995 TV drama
  - 1995 TV drama: music score published in 2011
  - A serial thematic issue for Pride and prejudice (Volume 1, no. 2)
- **Black lives matter**
- **Zhi qing sui yue**
Pride and Prejudice

- **979 results** for Pride and prejudice
  - Work (977) & Person (2) [Person: Mr Darcy & Elizabeth Bennet, without Jane Austen]
  - Searched with its subtitle (*a novel in three volumes*) but still 5 entities; no sufficient information to disambiguate
- Spanish translation published in 2006 ([Work ID](https://www.wikidata.org/wiki/Work:4862607))
  - Work (46) under *Orgullo y prejuicio*: very difficult to search; no label defined in English
- German translation published in 1980
  - *Stolz und Vorurteil: Roman*: very easy to find due to its subtitle
- 2006 DVD for 1940 Motion picture: very easy due to entity label with 1940 release
- 1995 music CD for 1995 TV drama (18 results for performed music)
  - Easier if adding related work: Pride and prejudice (Television program: 1995) and music composer and conductor--Carl Davis.
- 1995 TV drama: music score published in 2011: easy with subtitle & notated music
- A serial thematic issue for Pride and prejudice (in 246): none in SEMI
Correct Description: crucial for identifying a correct entity

<table>
<thead>
<tr>
<th>Entity Label</th>
<th>Entity Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pride and prejudice</td>
<td>work</td>
<td>Fiction</td>
</tr>
<tr>
<td>Jane Austen's Pride and prejudice</td>
<td>work</td>
<td>notated music</td>
</tr>
<tr>
<td>Pride and prejudice : suite for small orchestra</td>
<td>work</td>
<td>notated music</td>
</tr>
<tr>
<td>Pride and prejudice : for violin and piano : theme from the TV series</td>
<td>work</td>
<td>Songs and music</td>
</tr>
<tr>
<td>Jane Austen : the music : themes from Sense and sensibility and Pride and prejudice</td>
<td>work</td>
<td>Songs and music</td>
</tr>
<tr>
<td>Jane Austen's world : evocative music from the classic feature films Pride &amp; prejudice, Sense &amp; sensibility, Emma, and Persuasion</td>
<td>work</td>
<td>notated music</td>
</tr>
</tbody>
</table>
## Filtered search results

Showing 1 - 4 of 4 results for pride and prejudice 1940

### Entity Type: work

<table>
<thead>
<tr>
<th>Entity Label</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pride and Prejudice</td>
<td>work</td>
<td>Drama</td>
</tr>
<tr>
<td>Pride and prejudice (1940 release)</td>
<td>work</td>
<td>Drama</td>
</tr>
<tr>
<td>Pride and prejudice 1940</td>
<td>work</td>
<td>Film adaptations</td>
</tr>
<tr>
<td>Pride and prejudice : production material ca. 1940</td>
<td>work</td>
<td>Fiction</td>
</tr>
</tbody>
</table>

### Additional resources:
- Streaming video
- DVD
- Italian VHS
- Archives/manuscripts

Showing 1 - 2 of 2 results for pride and prejudice 1940 moving image

<table>
<thead>
<tr>
<th>Entity Label</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pride and prejudice (1940 release)</td>
<td>work</td>
<td>Drama</td>
</tr>
<tr>
<td>Pride and prejudice 1940</td>
<td>work</td>
<td>Film adaptations</td>
</tr>
</tbody>
</table>
Missing “description” for archives/manuscripts materials

SEMI:

Pride and prejudice : production material, ca. 1940

Entity ID: Q24752702
Entity Type: Work
Description: Fiction

OCLC record linked to the entity:

Author: Helen Sewell; Jane Austen
Edition/Format: Archival material : English
Publication: Papers
Summary: Ink illustration.

Genre/Form:

Document Type: Archival Material
All Authors / Contributors: Helen Sewell; Jane Austen
OCLC Number: 62444196
In: Sewell, Helen
Description:

Illustrations: 1 item
### History vs. Moving Image

#### Black Lives Matter Entity
- **Entity ID:** Q24880174
- **Entity Type:** Work
- **Description:** History

#### Table: Entity Data
<table>
<thead>
<tr>
<th>Entity Label</th>
<th>Type</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black lives matter</td>
<td>work</td>
<td>moving image</td>
</tr>
<tr>
<td>Black Lives Matter</td>
<td>work</td>
<td>creative work</td>
</tr>
</tbody>
</table>

**Image:**
- A visual representation of the Black Lives Matter entity with its entity ID, type, and description.
- A table showing the entity's type and associated data.

**Source:** NYU Libraries
Title in different Romanized forms (zhiqing vs. zhi qing)

2 names from Wikidata (not inverted names)
Romanized title is findable, but Chinese title is not findable.
Observations summary

- UI design is great and clear.
- A generic or popular title without a subtitle or correct description information: much harder to identify or find the entity.
- Description under Key Information
  - No consistent search results for "Description" such as History, text, juvenile works, posters, biographies, drama, two-dimensional moving image or adaptations, in the list of search results.
  - "Description" with different terms due to sources, e.g. MARC, VIAF and Wikidata
  - When the Work has more than one content type/genre/form such as moving image, creative works or history, which one should be chosen for "Description"?
- Subtitle: an easiest way to find the work entity
- Original creation date: helpful if data available in the entity
Perspectives on OCLC’s data modeling (WEMI)

- OCLC# linked under External Sources
  - Assumptions:
    - A bibliographic record (OCLC#) should be defined as a manifestation or instance of a Work entity in SEMI. An OCLC# is not equal to a Work ID.
    - Linking an instance/manifestation to the Work entity as a “related” instance/manifestation: could be multiple instances
  - OCLC#: is not clearly defined as a related “instance” under External Sources while Misc. includes multiple “instance of” from Wikidata or other sources
  - If only one OCLC# is linked to each Work entity, what criteria used to select the OCLC# when there is more than one bib record (algorithm vs. quality)?
- The definition of “Work” may not be consistent for certain entities.
Required skills and competencies

- Terms & properties
  - RDA content type terms such as performed music, two-dimensional moving image, notated music, text, etc. in MARC 336 field
  - Wikidata properties

- Preferred titles: qualifying elements

- Data models
  - WEMI presented in LRM, BIBFRAME, Share-VDE and OCLC SEMI

- Data mapping and crosswalks
  - MARC, BIBFRAME, Wikidata, etc.

- Authorities and identity management
  - NAF, VIAF, ISNI, ORCID, etc.

- Linked data platforms
  - SEMI, Share-VDE, Wikidata, etc.
Challenges

- Like Wikidata, data quality has been a longstanding concern.
  - Wikidata: duplicate items, vandalism, items changed without consistent reasons or guidelines, items with inconsistent labels in different wiki platforms
  - SEMI: inconsistent description or content type terms and multilingual issues
- Maintenance and quality control are crucial for discovery!
  - Incorrect linking, incorrect clustering and change management
- We need a linked data platform to create new entities or authority records for multilingual resources (Unicode, ISO codes, etc.) and can rely on smart batch matching and loading.
- In addition to linked data/semantic web/knowledge graphs, shall we consider other technologies such as AI/NLP (Natural language processing) to assist in quality control or algorithm improvement?
Thank You!

For questions please contact: Charlene Chou: charlene.chou@nyu.edu
Questions?
Thank you to our presenters

Anne Washington
OCLC

Laura Ramsey
OCLC

Charlene Chou
NYU Libraries
Because what is known must be shared.