

Transforming metadata
into linked data
to improve digital
collection discoverability:
A CONTENTdm Pilot Project

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



© 2021 OCLC.

This work is licensed under a Creative Commons Attribution 4.0 International License.
<http://creativecommons.org/licenses/by/4.0/>



January 2021

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-185-9

DOI: 10.25333/fzcv-0851

OCLC Control Number: 1230259668

ORCID iDs

Greta Bahnemann, Minnesota Digital Library  <https://orcid.org/0000-0002-5823-7217>

Michael Carroll, Temple University Libraries  <https://orcid.org/0000-0003-3736-0678>

Paul Clough, University of Miami Libraries  <https://orcid.org/0000-0001-6939-2805>

Mario Einaudi, The Huntington Library, Art Museum, and Botanical Gardens  <https://orcid.org/0000-0002-6859-594X>

Chatham Ewing, Cleveland Public Library  <https://orcid.org/0000-0002-8402-0652>

Jeff Mixter, OCLC Research  <https://orcid.org/0000-0002-8411-2952>

Jason Roy, Minnesota Digital Library  <https://orcid.org/0000-0002-3644-1970>

Holly Tomren, Temple University Libraries  <https://orcid.org/0000-0002-6062-1138>

Bruce Washburn, OCLC Research  <http://orcid.org/0000-0003-4396-7345>

Elliot Williams, University of Miami Libraries  <https://orcid.org/0000-0001-6925-7144>

Please direct correspondence to:

OCLC Research
oclcresearch@oclc.org

Suggested citation:

Bahnemann, Greta, Michael Carroll, Paul Clough, Mario Einaudi, Chatham Ewing, Jeff Mixter, Jason Roy, Holly Tomren, Bruce Washburn, and Elliot Williams. 2021. *Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project*. Dublin, OH: OCLC Research.
<https://doi.org/10.25333/fzcv-0851>.

CONTENTS

| | |
|---|-------------|
| Acknowledgments | viii |
| Executive Summary | ix |
| Introduction | 11 |
| Three-Phase Project Plan | 13 |
| Phase 1: Mapping textual metadata to entities | 15 |
| Phase 2: Tools for managing metadata in Wikibase | 15 |
| Phase 3: Wikibase entities drive discovery | 15 |
| The Wikibase Environment | 16 |
| Developing A Data Model | 17 |
| Describing the “type” of a creative work at three levels | 18 |
| Distinguishing between instances of concepts and ontological classes | 19 |
| Managing the data model in Wikibase | 20 |
| Managing source metadata outside of the data model | 21 |
| Gathering and Transforming Metadata | 22 |
| Selecting and analyzing collections from pilot partner CONTENTdm sites | 23 |
| Optimizing tools and workflows for reconciliation and transformation | 24 |
| Adding related entities to the Contentdm Wikibase from external sources | 25 |
| Creating entities in advance for anticipated matches | 26 |
| Testing an alternative openrefine reconciliation endpoint | 26 |
| Creating placeholder entities for things that could not be reconciled | 27 |
| Representing Compound Objects | 28 |

Syndicating Data in Standard Schemas 29

Wikibase Ecosystem Advantages..... 29

- Implementing authority control29
- Decreasing cataloging inefficiencies, increasing descriptive quality 30
- Generating data visualizations32

User Interface Extensions 33

- MediaWiki gadgets.....33
 - Adding the Mirador viewer33
 - Showing contextual information from Wikidata.....33
 - Contextual Data and Image from DBPedia and Wikimedia Commons
Embedded in the Wikibase User Interface34
 - Revealing constraint violations34
- CONTENTdm custom pages35
 - Embedding Schema.org JSON-LD in CONTENTdm pages 36
 - Showing contextual information for headings based on Wikibase data..... 37

New Applications..... 39

- The Image Annotator 39
 - User study results42
- The Retriever 43
- The Describer 46
- The Explorer and the Transportation Hub..... 47
- The Field Analyzer53

Cohort Communication 55

Partner Reflections 56

- Cleveland Public Library56

| | |
|---|-----------|
| The Huntington Library, Art Museum, And Botanical Gardens | 58 |
| Minnesota Digital Library | 59 |
| Invitation | 59 |
| Development of three tools by OCLC..... | 59 |
| Leveraging the power of linked data | 60 |
| Concluding thoughts | 61 |
| Temple University Libraries..... | 61 |
| University of Miami Libraries | 63 |
| Key Findings and Conclusions..... | 64 |
| Testing the linked data value proposition | 64 |
| Evaluating a shared data model | 64 |
| Selecting and transforming metadata | 65 |
| Continuing the journey to linked data | 65 |
| Working partnerships represent strength in numbers | 66 |
| Notes..... | 67 |

FIGURES

| | | |
|------------------|--|----|
| FIGURE 1 | Planned project phases | 14 |
| FIGURE 2 | The Wikibase Ecosystem | 16 |
| FIGURE 3 | A CONTENTdm class hierarchy data model..... | 17 |
| FIGURE 4 | Example type, classification used, and process or format properties and values for a description of a postcard | 18 |
| FIGURE 5 | A depicts statement for the concept of “Dogs” | 19 |
| FIGURE 6 | A type classification of “dog” for a specific dog | 19 |
| FIGURE 7 | The “dog” class is defined by the concept of “Dogs” | 20 |
| FIGURE 8 | Wikibase templates for proposing new properties | 21 |
| FIGURE 9 | Unmapped CONTENTdm metadata displayed in the Wikibase user interface using a Gadget extension..... | 22 |
| FIGURE 10 | Wikibase Discussion page for a collection review..... | 23 |
| FIGURE 12 | A “placeholder” entity for a person without an established identity..... | 27 |
| FIGURE 13 | Example “has creative work part” statements and sequencing for the first four parts of an album | 28 |
| FIGURE 14 | Other names associated with the Los Angeles Dodgers entity | 30 |
| FIGURE 15 | First parts of the description of Jasper Wood | 31 |
| FIGURE 16 | SPARQL Query map visualization of places depicted in works from a collection | 32 |
| FIGURE 17 | Mirador image viewer embedded in the Wikibase user interface | 33 |
| FIGURE 18 | Contextual data and image from DBPedia and Wikimedia Commons embedded in the Wikibase user interface | 34 |
| FIGURE 19 | A constraint violation indicating that the “occupation” property should only be used for instances of the type “person” | 35 |
| FIGURE 20 | Schema.org data evaluated using Google’s Structured Data Testing Tool | 37 |
| FIGURE 21 | Additional contextual information displayed in CONTENTdm based on entity descriptions in the pilot Wikibase | 38 |
| FIGURE 22 | Image Annotator initial view of an image and subjects | 40 |
| FIGURE 23 | Image Annotator cropping an image of a person | 40 |
| FIGURE 24 | Image Annotator after adding more depicted subjects | 41 |
| FIGURE 25 | Wikibase item updated with illustrated depicts statements | 42 |
| FIGURE 26 | Retriever search results from Wikidata, VIAF, and FAST for “Lake Vermilion” | 44 |
| FIGURE 27 | Retriever entity editor | 45 |
| FIGURE 28 | Wikibase entity created by the Retriever | 45 |

| | | |
|------------------|--|----|
| FIGURE 29 | Editing essential details for an entity in the Describer | 46 |
| FIGURE 30 | Explorer home page | 48 |
| FIGURE 31 | Explorer Transportation Hub and related collections | 49 |
| FIGURE 32 | Explorer search results for “strike” | 50 |
| FIGURE 33 | Explorer view of a truck bringing employees home during a PTC walkout | 51 |
| FIGURE 34 | Explorer view of a protest against the Philadelphia Transportation Company | 51 |
| FIGURE 35 | Explorer view of an 1899 Cleveland transit strike in Public Square | 52 |
| FIGURE 36 | Explorer view of streetcars parked on the street during a transit strike | 53 |
| FIGURE 37 | Field Analyzer field usage chart | 54 |
| FIGURE 38 | Field Analyzer list of field values | 55 |

ACKNOWLEDGMENTS

The OCLC CONTENTdm Linked Data Pilot project team consisted of the following OCLC staff: Hanning Chen, Eric Childress, Shane Huddleston, Jeff Mixter, Mercy Procaccini, and Bruce Washburn.

The Linked Data project team wishes to thank the project partners who enthusiastically and generously collaborated with us in this endeavor. Your vision for and commitment to a linked data future have been illuminating and inspiring. OCLC particularly appreciates the efforts of those who contributed to or co-authored this report:

- Cleveland Public Library: Chatham Ewing, Rachel Senese, Amia Wheatley
- The Huntington Library, Art Museum, and Botanical Gardens: Mario Einaudi
- Minnesota Digital Library: Greta Bahnemann, Jolie Graybill, Jason Roy
- Temple University Libraries: Michael Carroll, Stefanie Ramsay, Holly Tomren
- University of Miami Libraries: Paul Clough, Elliot Williams

The team also acknowledges the consultation, guidance, and support provided by our OCLC colleagues: Dave Collins, Rachel Frick, Marti Heyman, Erik Mayer, Carolyn Morgan, Andrew Pace, Taylor Surface, and Diane Vizine-Goetz.

Thank you to Jeanette McNicol for the excellent design of this report and to Erica Melko for her skillful editing.

EXECUTIVE SUMMARY

In the CONTENTdm Linked Data Pilot project, OCLC partnered with institutions that manage their digital collections with OCLC's CONTENTdm service to investigate methods for—and the feasibility of—transforming metadata into linked data to improve the discoverability and management of digitized cultural materials and their descriptions. This report, *Transforming Metadata into Linked Data to Improve Digital Collection Discoverability*, describes the course of the project and its primary areas of investigation and summarizes key findings and conclusions generated by the collaborative study.

The project was designed to help the OCLC team and the pilot participants better understand the following questions:

- How divergent are the descriptive data practices across the institutions using CONTENTdm, and what tools are needed to make that assessment?
- Can a shared and extensible data model be developed to support the differing needs and demands for a range of material types and institution types?
- What is the right mix of human attention and automation to effectively reconcile metadata headings to linked data entities?
- What types of tools can help extend the description of cultural materials to subject matter experts?
- After metadata from different institutions and collections is transformed, are there new discovery tools that can help researchers find new—or previously hidden—connections through a centralized discovery system?
- What are the institutional and individual interests in the paradigm shift of moving to linked data?

Over the course of the pilot, the project team and partners observed improved metadata management and discovery in action...

Five organizations representing a cross-section of different types of institutions—The Huntington Library, Art Museum, and Botanical Gardens; the Cleveland Public Library; the Minnesota Digital Library; Temple University Libraries; and University of Miami Libraries—participated in the project.

The pilot focused on developing efficient workflows for transforming metadata, evaluating existing interfaces to leverage linked data, and testing applications built in the Wikibase environment for managing the newly created linked data.

Over the course of the pilot, the project team and partners observed improved metadata management and discovery in action and reflected on the potential benefits: higher-quality and richer metadata can be managed with greater efficiency by staff, and linked data can be used to add contextual information and to create a network of connections that better reflects knowledge in the real world. This context and these connections can help researchers achieve a fuller understanding of collection materials, inviting increased engagement and use by community members.

Higher-quality and richer metadata can be managed with greater efficiency by staff, and linked data can be used to add contextual information and to create a network of connections that better reflects knowledge in the real world.

While the pilot project findings are based on a limited set of institutions and collections, they strongly suggest that there is significant potential for improved discovery and more efficient data management when the materials that have been digitized are described using a shared data model, where headings are associated with linked data entities and relationships, and when the entities and relationships are brought together into a single aggregation.

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. Evidence gathered during the project and detailed in this report about data modeling, metadata reconciliation, and data analysis provides new knowledge about how these tools and workflows could be designed and used.

INTRODUCTION

The CONTENTdm¹ Linked Data Pilot project (also referred to throughout this report as the “Linked Data project”) is the latest (as of 2020) in a series of investigations² that OCLC has organized and led over several years in the interest of developing a shared understanding how libraries, archives, and museums can make the transition to linked data. OCLC works in partnership with these institutions to increase researchers’ ability to discover, evaluate, and use digitized cultural materials, principally through its support of the CONTENTdm service for building, preserving, and showcasing a library’s unique digital collections.

This Linked Data project was focused on envisioning and evaluating scalable and affordable systems and workflows that will be needed to produce rich linked data representations of entities and relationships, which will then help to make visible connections that were formerly invisible. The project was grounded in the context of the linked data value proposition, which states that these best practices for publishing structured data on the web—using URIs (Uniform Resource Identifiers) as names for things, using HTTP URIs so that people can look up those names, providing useful information using standards when someone looks up a URI, and including links to other URIs so that people can discover more things—lead to an interconnected global network of data that can serve both developers and researchers.³

Five organizations representing a cross-section of different types of institutions—The Huntington Library, Art Museum, and Botanical Gardens; the Cleveland Public Library; the Minnesota Digital Library; Temple University Libraries; and University of Miami Libraries—participated as partners in the project. The pilot participants collaborated with OCLC on a range of focused studies, including developing efficient workflows for transforming source metadata into linked data, evaluating CONTENTdm interface customizations to leverage linked data for discovery and syndication, and testing new applications built in the Wikibase environment for data retrieval, image annotation, editing, metadata analysis, and discovery.

This report describes the course of the CONTENTdm Linked Data Pilot project and its primary areas of investigation, shares the experiences of the five participating partner institutions, and summarizes key findings and conclusions generated by that collaborative study.

The Linked Data project’s focus on sustainability and scalability posed many questions to pursue, including:

How divergent are the descriptive data practices across the institutions using CONTENTdm, and what tools are needed to make that assessment?

The large volume of cultural material descriptive metadata stored in CONTENTdm offered an excellent test bed for evaluating a large-scale transition to linked data. Additionally, the outcomes

and findings from OCLC’s Metadata Refinery project completed in 2016 and its Project Passage⁴ linked data prototype completed in 2018 provided important insights into how to implement a system to facilitate the mapping, reconciliation, storage, and retrieval of structured data for unique digital materials. This pilot project built on those insights and successes. The sections below that describe the Wikibase⁵ environment, the steps for gathering and transforming metadata, and the prototype “Field Analyzer” application highlight the challenges of applying this work at scale.

Can a shared and extensible data model be developed to support the differing needs and demands for a range of material types and institution types?

The wide variety of data models and descriptive practices currently used across CONTENTdm could be significantly easier for staff to manage if there was a shared data model available, and if that shared model could also support rich discovery for researchers in a single, aggregated discovery system. This project set out to develop a shared data model, building on existing standards but allowing for extensions as evidence surfaced in the source metadata for additional classes and relationships. The section below on developing the data model provides an overview and examples of the results of that work.

What is the right mix of human attention and automation to effectively reconcile metadata headings to linked data entities?

The project spent substantial time and effort on testing reconciliation workflows and prototyping new tools to make this work more efficient while maintaining quality. Prototyping a new metadata reconciliation endpoint helped us understand the potential for improving the performance of what can be a time-consuming automated process. The development of the “Retriever” web application for finding related entities in other systems and transforming them into new Wikibase entities addressed a cataloger workflow stumbling block. Both prototypes are described below.

What types of tools can help extend the description of cultural materials to subject matter experts?

The project team developed—and the participants tested—an “Image Annotator” prototype application that could be used by either library staff or subject matter experts from outside the library to associate subject headings with depicted entities in images, envisioning how the transformed data along with new tools could open the door to more and richer descriptions from an engaged community. The description below of the Image Annotator includes a summary of its usability test results.

After metadata from different institutions and collections is transformed, are there new discovery tools that can help researchers find new, or previously hidden, connections through a centralized discovery system?

The “Explorer” prototype application, developed during the project and described below, demonstrated the ability to search across data from a range of repositories, with searching and faceting powered by entities derived from authority files and from vocabularies created by librarians. And the “Transportation Hub” virtual collection included in the Explorer gave the project team and participants a way to test linked data discovery in action, working with thematically related item descriptions that were supplied by a cross-section of institutions and collections and transformed into separate entities and relationships.

What are the institutional and individual interests in the paradigm shift of moving to linked data?

The close collaboration between OCLC and the pilot project partners was one of the most rewarding aspects of the project. Given that most of the project was carried out as people and the organizations they work for were experiencing transformative disruptions to their lives and work as the 2020 COVID-19 pandemic began and unfolded, it was unclear at first what relative priority and attention the pilot could receive. But attention and participation from the participants—and support from OCLC—never wavered, and we mutually benefited greatly from the endeavor. Look to the following sections on cohort communication and the partner reflections for more insights and perspectives on the impact of this project and the partners’ first-hand views on the implications for our shared futures.

The findings of the project—detailed in this report—about data modeling, metadata reconciliation, and data analysis provide new knowledge about how these tools and workflows could be designed and used, which we anticipate will inform future linked data investigations and developments from the library, archives, and museum communities.

The CONTENTdm Linked Data Pilot project is another stage in a growing body of linked data research and development that OCLC has undertaken over the past decade. The findings of the project—detailed in this report—about data modeling, metadata reconciliation, and data analysis provide new knowledge about how these tools and workflows could be designed and used, which we anticipate will inform future linked data investigations and developments from the library, archives, and museum communities.

Three-Phase Project Plan

The pilot project was planned as a one-year effort to be carried out in three phases (figure 1) so that the project could address the most pressing questions first and allow for reconsideration and adjustments to the plan as it progressed:

- **Phase 1:** Concentrated on mapping metadata for digital collections to descriptions of related entities: works, people, organizations, places, concepts, and events. Three partner institutions joined the project in Phase 1: The Huntington Library, Art Museum, and Botanical Gardens; the Cleveland Public Library; and the Minnesota Digital Library.

- **Phase 2:** Focused on a needs assessment and prototypes for managing metadata in the Wikibase environment. Two more partner institutions joined the project in Phase 2 after the OCLC team had developed a better understanding of the institutional support requirements, and to expand representation of materials from academic research libraries: Temple University Libraries and University of Miami Libraries.
- **Phase 3:** Anticipated testing an end-user discovery experience based entirely on the data and tools developed within the Wikibase environment.

CONTENTdm Linked Data Planned Project Phases

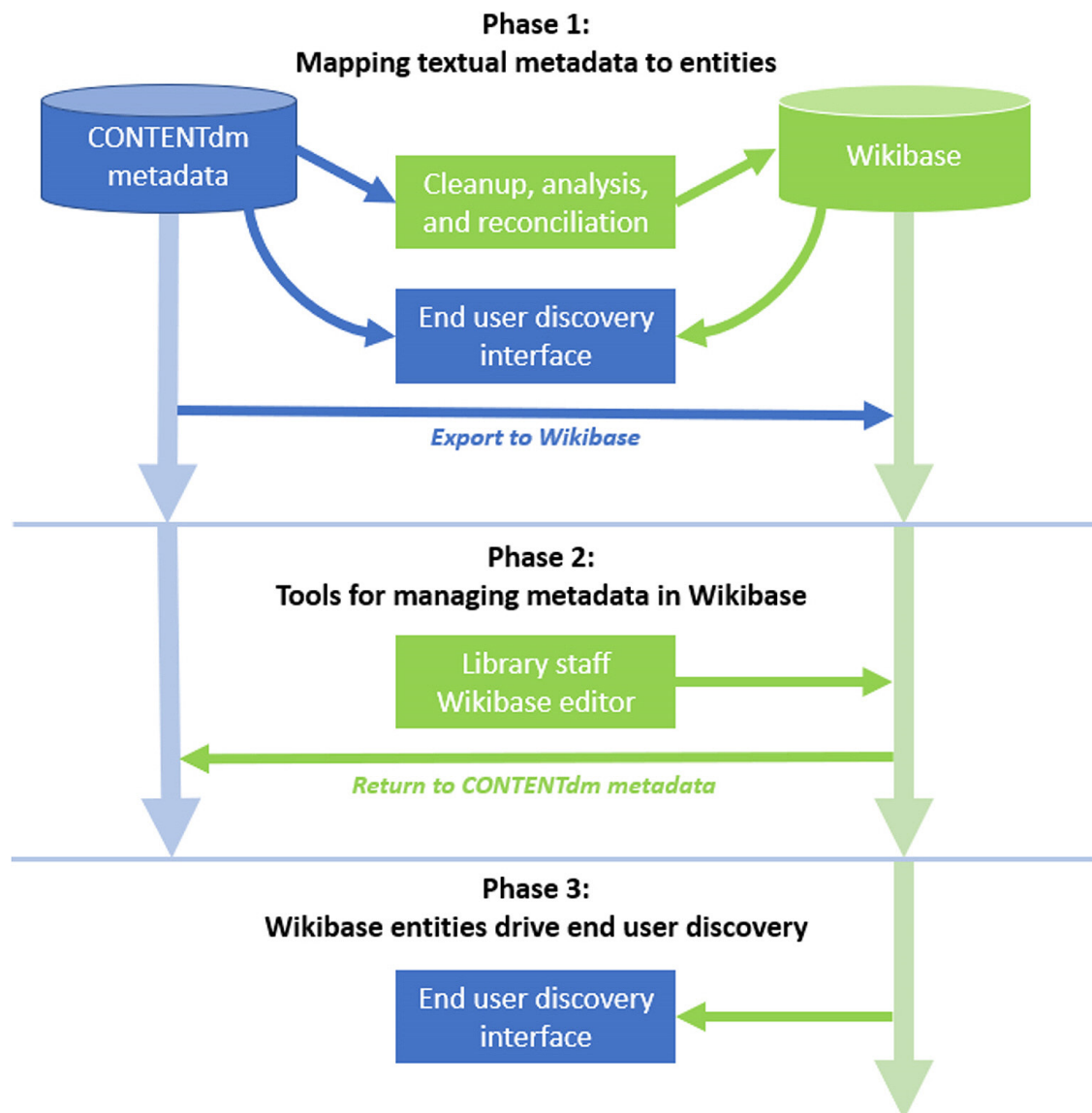


FIGURE 1. Planned project phases.⁶ View a larger image [online](#).

PHASE 1: MAPPING TEXTUAL METADATA TO ENTITIES

In the first phase, the plan was to focus on the systems and workflows needed to clean up, analyze, and reconcile CONTENTdm metadata for input into a linked data environment.

Building on the project team's prior experience with the environment in OCLC's earlier linked data pilot, Project Passage, the Wikibase extension to the MediaWiki platform was selected as the project's linked data environment. The Wikibase environment of related databases, indexes, and services is described in fuller detail below.

In this phase, linked data was expected to be shown in the CONTENTdm interface, delivering data from the pilot project Wikibase using CONTENTdm's custom Javascript feature.

PHASE 2: TOOLS FOR MANAGING METADATA IN WIKIBASE

In the second phase, the work was expected to focus on the Wikibase editing interface and on supplementary tools that could be used to extend that environment. These tools would help bridge the gap between CONTENTdm staff user expectations and the features and limitations of the Wikibase environment.

The design and development of mechanisms for returning data from the Wikibase to the production CONTENTdm environment were also expected to be part of this phase.

PHASE 3: WIKIBASE ENTITIES DRIVE DISCOVERY

The focus of phase three was intended to be on a discovery interface that relied solely on data within the Wikibase to evaluate the features that could be part of a redesigned CONTENTdm discovery system.

As the project unfolded, the project team made adjustments to the original plan, responding to new findings from its early phases. For example, work on some staff tools for editing Wikibase data began during Phase 1 (planned for Phase 2). On the other hand, the initial plan included the prototyping of a user interface for entity editing as an alternative to the Wikibase user interface but the team did not completely build and test that prototype before the project ended.

The Phase 2 work anticipated that the project would encourage loading data from the Wikibase back into the CONTENTdm system using its "Catcher"⁷ web service that can add and edit metadata using a standard XML-based method. But given the conditions of the pilot project, the project partners could not be sure if the modified headings would conflict with their ongoing data management work.

The first two phases concentrated on building and evaluating workflows for analyzing, transforming, and reconciling CONTENTdm metadata into Wikibase Linked Data with as complete and lossless a result as was feasible. In the third phase, a new course was charted to see how much linked data could be generated from CONTENTdm with minimal human intervention and evaluate the results in a front-end discovery application to more clearly demonstrate the linked data value proposition.

These types of adjustments to the project plan are expected in a research-oriented pilot, where a full understanding of the issues and questions that will naturally surface over time are not defined at the outset.

The Wikibase Environment

To proceed through the planned phases, the project needed an effective and proven platform for working with linked data. Based on the successful results from Project Passage (2018), the CONTENTdm Linked Data Pilot project used the Wikibase environment, which includes several interrelated APIs, databases, indexes, and services (figure 2):

- MediaWiki is the primary software platform, the same software on which the Wikipedia⁸ encyclopedia and other “wikis” operate.
- To handle structured data, the Wikibase extension to MediaWiki is used, which is the same software that supports the Wikidata⁹ knowledge base. Together, MediaWiki and Wikibase provide both a user interface for searching and editing and a range of APIs for access to authentication and editing services.
- But to support linked data, a parallel system is synchronized with the Wikibase data, including its own linked database or triplestore that can be accessed using a linked data query language called SPARQL.¹⁰ A SPARQL Query service user interface is also provided. These powerful tools are the product of years of open source software development and support provided by the Wikidata and Wikimedia communities.

The Wikibase Ecosystem

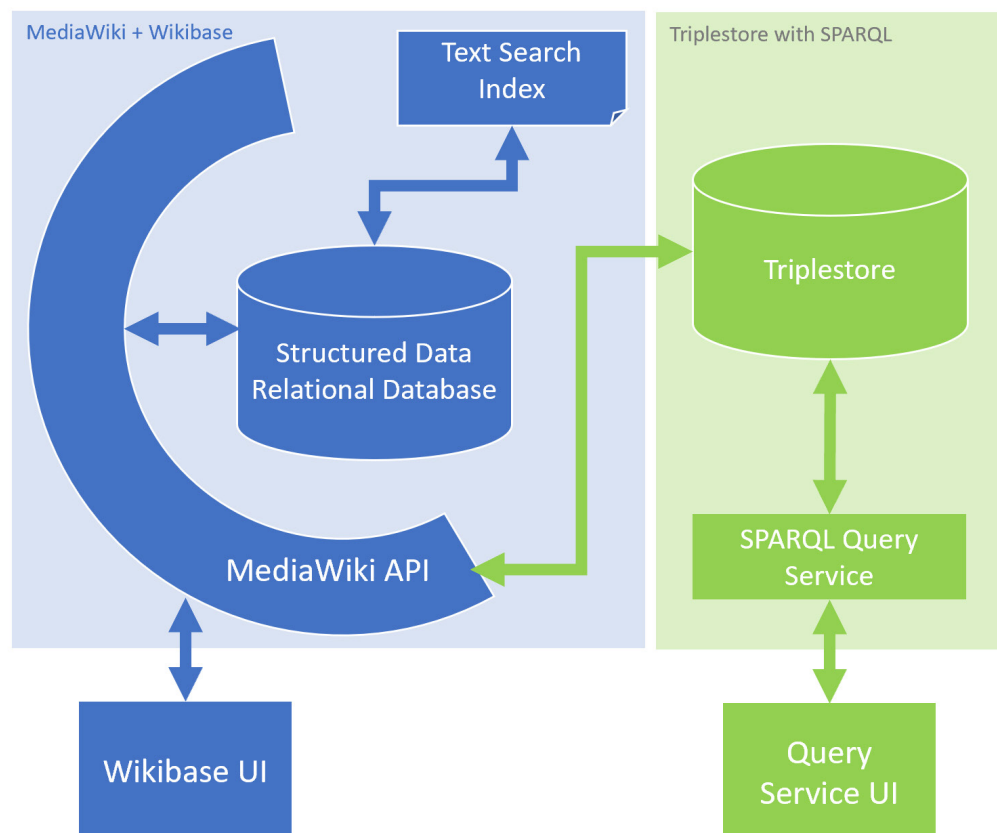


FIGURE 2. The Wikibase Ecosystem.¹¹ View a larger image [online](#).

Developing A Data Model

CONTENTdm repositories employ a wide range of vocabularies and institution-specific data dictionaries. Some institutions apply patterns to their data descriptions that are consistent across all their collections, while others use different patterns for different collections, either due to evolution of their institutional preferences over time and the effort required to maintain and revise “legacy” patterns in previously described collections, or to account for special characteristics in the data and use cases associated with specific collections.

For the Linked Data project Wikibase, a single data model was needed that could reflect the variations seen in the metadata across CONTENTdm sites. Rather than selecting an existing data model to which we’d force CONTENTdm metadata to conform, the pilot project tested the theory that, through sampling current metadata and looking for general patterns, a model could be developed that was driven by data and that avoided speculation. Where appropriate, the properties and classes defined in the project data model were linked to equivalent properties and classes in other ontologies and vocabularies.

CONTENTdm Class Hierarchy Data Model

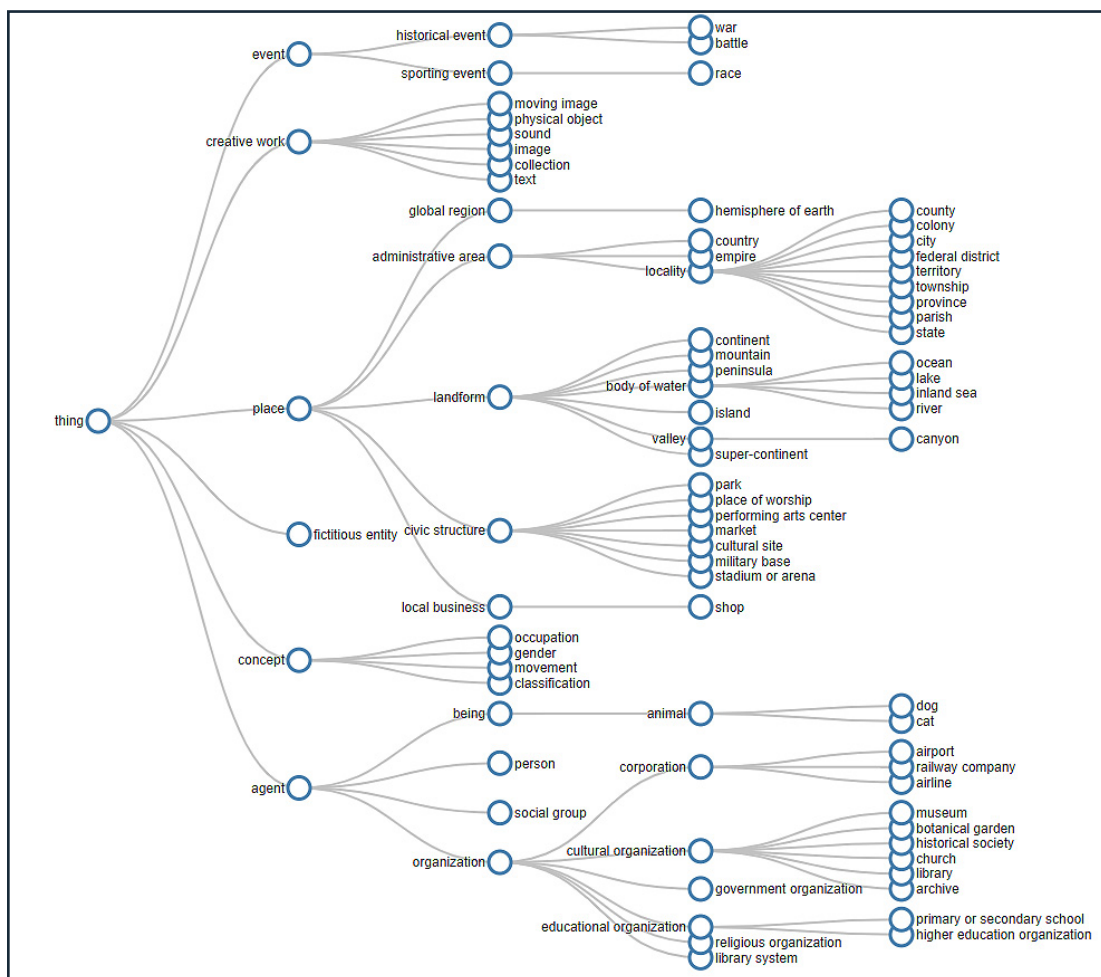


FIGURE 3. A CONTENTdm class hierarchy data model.¹² View a larger image [online](#).

This work began by looking across an inventory of CONTENTdm metadata for the most common metadata practices, leveraging CONTENTdm’s ability to help institutions relate their local vocabulary terms to the Dublin Core¹³ element set and associated controlled vocabularies. This step identified the classes and relationships that would be encountered most frequently in the pilot participants’ data and gave a starting point for building the pilot project data model. A field analysis survey was conducted for about 13 million records, selected from all CONTENTdm sites, that evaluated the most frequently used fields to identify important properties for creative works. From that same CONTENTdm survey, the most frequently used terms were extracted to build an initial class taxonomy for creative works. This method was later revised based on conversations with partners and colleagues. The class hierarchy from the project’s data model is illustrated in figure 3.


















DESCRIBING THE “TYPE” OF A CREATIVE WORK AT THREE LEVELS

As analysis of the pilot project participants’ data began, new classes and relationships were encountered and were evaluated as possible extensions to the data model. One part of the model that changed substantially was the Creative Work taxonomical branch. It was originally populated with the “types” of creative works based on how they were described in the source metadata, but that resulted in a large and unstructured list of classes.

After consulting with the pilot partners and with colleagues at the J. Paul Getty Trust, the team decided to revise the model using a three-level approach. At the top level, creative work “type” classes were mapped to the Dublin Core DCMI Type¹⁴ terms. An immediate benefit of that decision was the ability to neatly facet results across the different DCMI Types, a common way of providing a high-level filter for search and retrieval of digital collections.

To refine the top level DCMI Type classes, a second level “classification used” property was created that was associated with 25 “classification” entities. The set of classification entities was developed based on work done in the Linked.Art¹⁵ project as well as through consultation with colleagues at the pilot partner Minnesota Digital Library. If more detail was needed, a third level for the “process or format” property could be used to connect the item to any conceptual entity. An example of this revision to the data model is illustrated in figure 4 for a postcard, which is a type of “image,” uses the classification “Prints,” and adds a “process or format” of “Postcards.”

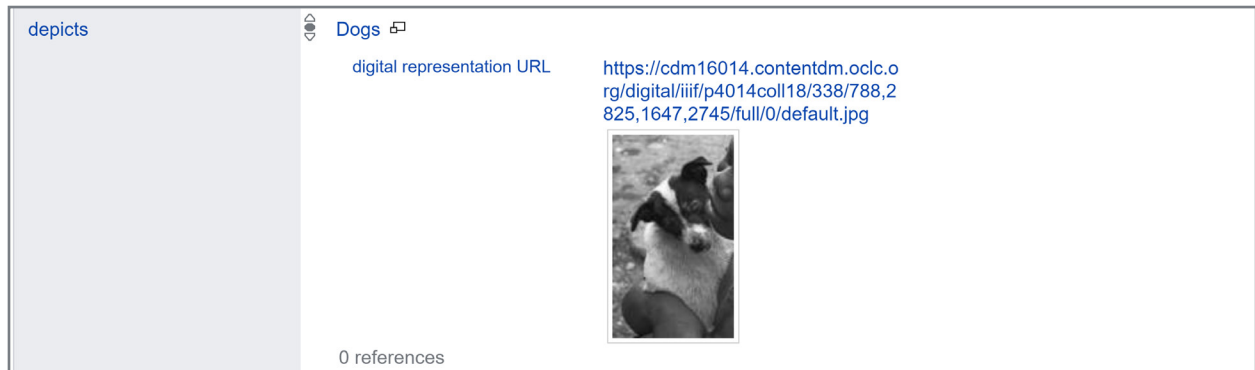
Example of Mapped DCMI Data Levels for a Postcard

| | |
|------|--|
| type |                  |
|------|--|

DISTINGUISHING BETWEEN INSTANCES OF CONCEPTS AND ONTOLOGICAL CLASSES

Distinguishing between instances of concepts and ontological classes presented a data modeling challenge. This challenge is related to how, in the library domain, controlled vocabularies have been developed and translated to ontology-based systems. Concepts derived from a controlled vocabulary can be used both as conceptual entities for subject headings and as ontological classifications for a specific instance of the subject. A good example of this dual use is the concept entity of “Dogs.” As a concept it can be used to describe what a photograph depicts as seen in figure 5.

“Depicts” Statement for the Concept of “Dogs”

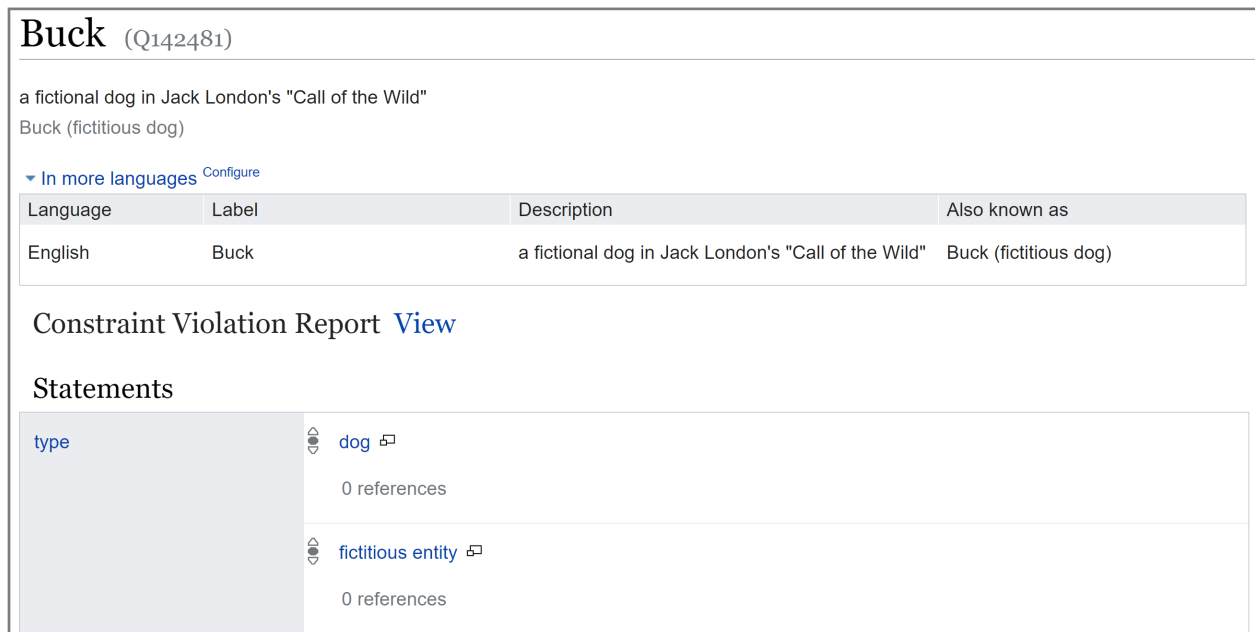


The screenshot shows a 'depicts' statement in a software interface. On the left, the word 'depicts' is highlighted in blue. To its right, the concept 'Dogs' is listed with a small icon. Below 'Dogs', the text 'digital representation URL' is followed by a long URL: 'https://cdm16014.contentdm.oclc.org/digital/iiif/p4014coll18/338/788,2825,1647,2745/full/0/default.jpg'. Below the URL is a small thumbnail image of a dog. At the bottom of the interface, it says '0 references'.

FIGURE 5. A depicts statement for the concept of “Dogs.”¹⁷ View a larger image [online](#).

But “dog” can also be used as an ontological class to describe specific dogs, such as the dog named “Buck” who appears in a photograph (figure 6).

Type Classification of “Dog” for a Specific Dog



The screenshot shows the Wikidata entry for 'Buck' (Q142481). At the top, it says 'Buck (Q142481)'. Below that, it describes 'a fictional dog in Jack London's "Call of the Wild"' and 'Buck (fictitious dog)'. There is a link for 'In more languages' and a 'Configure' link. Below this is a table with four columns: 'Language', 'Label', 'Description', and 'Also known as'. The table has one row for 'English' with the label 'Buck', the description 'a fictional dog in Jack London's "Call of the Wild"', and 'Buck (fictitious dog)'. Below the table is a 'Constraint Violation Report' link. At the bottom, there is a 'Statements' section with two entries: 'type' with the value 'dog' and '0 references', and 'type' with the value 'fictitious entity' and '0 references'.

| Language | Label | Description | Also known as |
|----------|-------|---|-----------------------|
| English | Buck | a fictional dog in Jack London's "Call of the Wild" | Buck (fictitious dog) |

FIGURE 6. A type classification of “dog” for a specific dog.¹⁸ View a larger image [online](#).

To distinguish the conceptual entity of “Dogs” from the ontological class “dog” in the pilot data model, an “is defined by” property was created, based on the property “isDefinedBy,” which is found in the linked data modeling vocabulary RDF Schema¹⁹ to connect the class to the conceptual entity that describes it (figure 7).

The “Dog” Class “isDefinedBy” the Concept of “Dogs”

The screenshot shows the CONTENTdm linked data interface for the class 'dog (Q73829)'. The page is titled 'dog (Q73829)' and is categorized as a 'domestic animal'. It lists labels in English ('dog') and other languages ('Canis lupus familiaris', 'Canis familiaris', 'domestic dog'). The 'Statements' section shows three entries: 'subclass of animal', 'is defined by Dogs', and 'type class'. The 'is defined by' statement is highlighted, indicating the relationship between the class and the concept of 'Dogs'.

FIGURE 7. The “dog” class is defined by the concept of “Dogs.”²⁰ View a larger image [online](#).

MANAGING THE DATA MODEL IN WIKIBASE

OCLC staff took advantage of the components built into the Wikibase infrastructure to manage the process of developing the data model, using a template form to submit and review proposals for new properties and classes. This approach helped illustrate the expected advantages that these additions to the model would bring and provided a history to look back on as the project proceeded.

OCLC staff found that these templates and the proposal/review/acceptance workflow were an effective way for a small but distributed team to manage the process and recommends this approach to others who are building a system using the Wikibase software platform (figure 8).

Wikibase Templates for Proposing New Properties

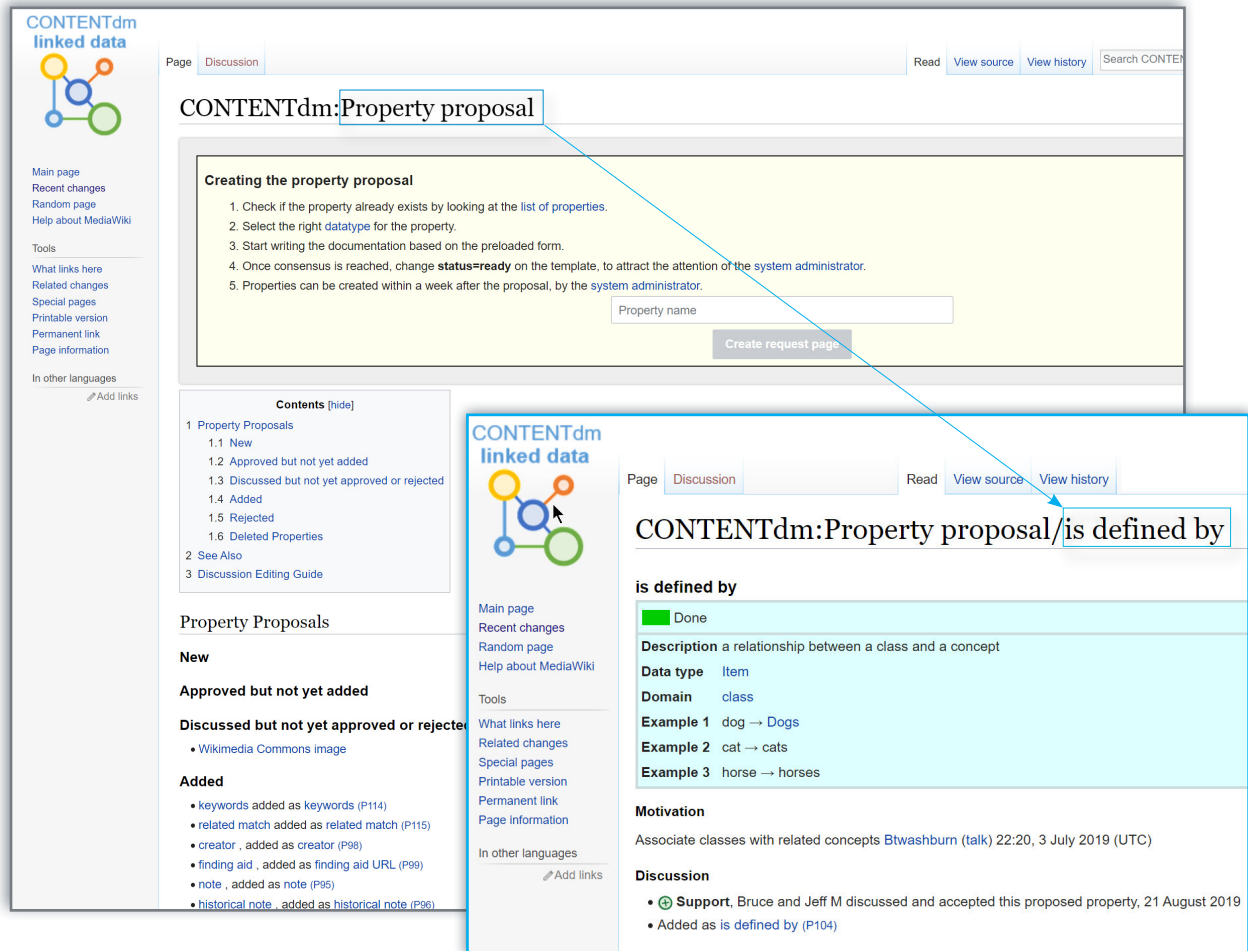


FIGURE 8. Wikibase templates for proposing new properties.²¹ View the larger Wikibase images [property proposal](#) and [property proposal/is defined by](#) online.

MANAGING SOURCE METADATA OUTSIDE OF THE DATA MODEL

Some of the CONTENTdm source metadata fell outside of the information that was expected to be accounted for in the data model. The data model was intended to support the description of cultural materials, but the source metadata also included technical information about their digital representations and administrative data associated with the cataloging process.

To prevent this additional information from being lost in the transformation process, the associated fields and values were indexed in a system separate from the Wikibase. The indexed data included the identifier for the associated entity in the project Wikibase. This allowed unmapped elements to be displayed in the Wikibase user interface (illustrated in figure 9) without disrupting the data model with entities and relationships that were administrative or technical in nature.

Unmapped CONTENTdm Metadata Displayed in the Wikibase User Interface Using a Gadget Extension

| Field | Value |
|-------------------------|--|
| Local Identifier | 2519 |
| CONTENTdm file name | 40.jp2 |
| Scanning Center | Minnesota Historical Society |
| Date created | 2013-02-14 |
| Rights Management | This image may not be reproduced for any reason without the express written consent of the Becker County Historical Society. |
| MDL Identifier | mhs24636 |
| Contact Information | Becker County Historical Society, 714 Summit Avenue, Detroit Lakes, Minnesota 56501; www.beckercountyhistory.org |
| Master File Software | Adobe Photoshop CS |
| Master File Compression | none |
| Date modified | 2018-11-19 |
| Master File Hardware | Epson 10000XL scanner |
| Master File System | Windows XP |
| Project Affiliation | Minnesota Reflections 2011-12; |
| Master File Size | 29619690 |
| Date Digital | 2012-07-12 8:12 |
| Master File Resolution | 450 |
| Master File Format | image/tiff |
| CONTENTdm number | 39 |
| Item Digital Format | image/jp2 |
| Master File Width | 3740 |
| file extension | jp2 |
| Reference URL | http://cdm16022.contentdm.oclc.org/cdm/ref/collection/becker/id/39 |
| IIIF manifest URL | https://cdm16022.contentdm.oclc.org/iiif/info/becker/39/manifest.json |
| Master File Checksum | d02864f0c44f6513460436b16cfd24a1 |
| Master File Height | 2638 |
| Full resolution | Volume11/mhs24636.tif |
| Type | Still Image |
| Master File Bit Depth | 24 |
| Fiscal Sponsor | Funding provided to the Minnesota Digital Library through the Minnesota Arts and Cultural Heritage Fund, a component of the Minnesota Clean Water, Land and Legacy constitutional amendment, ratified by Minnesota voters in 2008. |
| CONTENTdm file path | /becker/image/40.jp2 |

FIGURE 9. Unmapped CONTENTdm metadata displayed in the Wikibase user interface using a Gadget extension.²² View a larger image [online](#).


Gathering and Transforming Metadata

The primary focus of the first phase of the linked data project involved assembling metadata describing digitized cultural materials and transforming it to descriptions of related entities. The following notes provide a detailed view of that work, including how metadata was selected for inclusion and analyzed, the development of tools and workflows to manage the transformation, and how the database was enriched to build more connections between entities.

SELECTING AND ANALYZING COLLECTIONS FROM PILOT PARTNER CONTENTDM SITES

Pilot project participants were asked to suggest a small group of CONTENTdm collections that they wanted to work with. OCLC suggested working with collections of varying sizes and content types but emphasized that the described materials should be primarily visual (photographs, prints, maps, etc.) rather than finding aids or PDF documents. In some cases, for very large collections, OCLC chose to represent a subset of the entire collection, given the pilot project's resource constraints.²³

Wikibase Discussion Page for a Collection Review



CONTENTdm
linked data

English [Log in](#)

Item Discussion Read View source View history

Item talk:Q148309

John W. Mosley Photograph Collection

OpenRefine Project

<http://18.217.24.229/project?project=1664881014972>

OpenRefine transformation notes

| CONTENTdm source field | Wikibase data element | Transformation notes | Example value |
|------------------------|---|---|---|
| Title | label | | 1946 Cleveland Buckeyes |
| Title | title (P124) | Decided on using the newly created title property as staff can't easily confirm if titles are cataloger-supplied or original without access to and evaluation of each of the physical materials | 1946 Cleveland Buckeyes |
| Date | date created (P21) | Use for single dates. | "1964-06-11", "1968, September, 09", "August 1954" |
| Date | approximate date (P123) | Use for single circa dates. | "circa 1965" |
| Date | earliest date (P87) | Use when there is a range of dates. | "1940-1949" |
| Date | latest date (P88) | Use when there is a range of dates. | "1940-1949" |
| Photographer | photographer (P60) John W. Mosley (Q148310) | Same value in all records. | Mosley, John W. |
| Subject | about (P2) | Split concatenated values, reconcile, and verify that these are generally "about" relationships rather than depictions | Schools; Segregation; Demonstrations; Civil rights; Girard College |
| Organization-Building | about (P2) | Split concatenated values, reconcile, and verify that these are generally "about" relationships rather than depictions | Girard College; State Office Building (Philadelphia, Pa.) |
| Intersection | | unmapped | Broad and Spring Garden Streets (Philadelphia, Pa.); |
| Geographic Subject | about (P2) | Reconcile, and verify that these are generally "about" relationships rather than depictions | Philadelphia (Pa.) |
| Personal Names | depicts (P24) | Split concatenated values, reconcile, and verify that these are depicted people | Goodman, Benny, 1909-1986; Hampton, Lionel, 1908-2002; Wilson, Teddy, 1912-1986; Krupa, Gene, 1909-1973 |
| Description | description | Use the first sentence | Cat Anderson (left) with "Dizzy" Gillespie. John Birks "Dizzy" Gillespie (1917–1993) jazz trumpeter, composer and bandleader, moved to Philadelphia with his family at 18 years old and joined Frankie Fairfax's band before moving to New York City. |
| Description | description (P26) | Use an OpenRefine recipe to transform lengthy strings into multiple statements | Cat Anderson (left) with "Dizzy" Gillespie. John Birks "Dizzy" Gillespie (1917–1993) jazz trumpeter, composer and bandleader, moved to Philadelphia with his family at 18 years old and joined Frankie Fairfax's band before moving to New York City. |

FIGURE 10. Wikibase Discussion page for a collection review.²⁴ View a larger image [online](#).

OCLC staff exported CONTENTdm metadata for each suggested collection and created an entity description for it in the CONTENTdm Wikibase and used the Wikibase “Discussion Page” feature to develop a metadata crosswalk, analyzing fields used in the collection and mapping them to Wikibase properties and classes. After OCLC staff created the initial crosswalk, individual meetings were held with each pilot site to review the initial analysis and address questions.

This process highlighted the importance of domain expertise when thinking through the metadata transformation process, as institution-specific, and sometimes collection-specific, cataloging practices cannot always be discerned by others outside the institution.

OPTIMIZING TOOLS AND WORKFLOWS FOR RECONCILIATION AND TRANSFORMATION

After analyzing collection fields and reviewing the analysis with the pilot participants, OCLC created a project for each collection in the program OpenRefine²⁵ (figure 11), which provides tools for data analysis, cleanup, and reconciliation. OpenRefine has a significant learning curve but is a tool OCLC has used frequently for metadata analysis; it was a natural fit for this project and proved to be an effective platform.

CONTENTdm Collection Metadata in an OpenRefine Project

| All | Title | Creator | Contributors | Subject | Description | Date (Alpha) | Date | Type |
|-----|------------------------|-------------------------|--------------|-------------------------------|---|---------------|--|-------------|
| ☆ | 1. Hulett Unloader | Gray, Arthur, 1884-1976 | | Cleveland (Ohio)--Photographs | Arthur Gray worked for Standard Oil, photographing the progress of industry from 1929 to 1949. | ca. 1930-1939 | 1930; 1931; 1932; 1933; 1934; 1935; 1936; 1937; 1938; 1939 | Still Image |
| ☆ | 2. Williamson Building | Gray, Arthur, 1884-1976 | | Cleveland (Ohio)--Photographs | Arthur Gray worked for Standard Oil, photographing the progress of industry from 1929 to 1949. The Plain Dealer building can be seen illuminated on the left. Rosenblum's on the right. | ca. 1930-1939 | 1930; 1931; 1932; 1933; 1934; 1935; 1936; 1937; 1938; 1939 | Still Image |
| ☆ | 3. West Side Market | Gray, Arthur, 1884-1976 | | Cleveland (Ohio)--Photographs | Arthur Gray worked for Standard Oil, photographing the progress of industry from 1929 to 1949. | ca. 1930-1939 | 1930; 1931; 1932; 1933; 1934; 1935; 1936; 1937; 1938; 1939 | Still Image |
| ☆ | 4. Terminal Tower | Gray, Arthur, 1884-1976 | | Cleveland (Ohio)--Photographs | Arthur Gray worked for Standard Oil, photographing the progress of industry from 1929 to 1949. | ca. 1930-1939 | 1930; 1931; 1932; 1933; 1934; 1935; 1936; 1937; 1938; 1939 | Still Image |
| ☆ | 5. Weber's Restaurant | Gray, Arthur, 1884-1976 | | Cleveland (Ohio)--Photographs | Arthur Gray worked for Standard Oil, photographing the progress of industry from 1929 to 1949. | ca. 1930-1939 | 1930; 1931; 1932; 1933; 1934; 1935; 1936; 1937; 1938; 1939 | Still Image |

FIGURE 11. CONTENTdm collection metadata in an OpenRefine project.²⁶ View a larger image [online](#).

As OCLC staff gained more experience with CONTENTdm metadata, reusable OpenRefine recipes were developed for carrying out generic data transformation tasks, which helped speed up the data processing for OCLC staff. For example, a recipe was developed for looking up an item's Wikibase identifier using its IIF²⁷ Manifest URL ("IIF" is an image interoperability standard, and it defines a "manifest" that represent the digital content associated with a collection or item) and retrieving data from the pilot project's linked data "triplestore"²⁸ database, a recipe for converting personal names from indirect order to direct order, and a recipe to extract and format individual height and width values and corresponding unit from extent data text strings. The code for each recipe was documented and stored in a Wikibase Help page for sharing and reuse by OCLC staff.

An important advantage of the OpenRefine platform is its ability to reconcile strings of text against external vocabularies to obtain a persistent identifier for the thing that the text string describes. The reconciliation feature is built into OpenRefine and can be configured to compare strings against external OpenRefine-compatible reconciliation endpoints. OCLC staff worked with the OpenRefine reconciliation endpoint software²⁹ developed for the Wikidata community and reconfigured it as an endpoint for the project Wikibase. That way OpenRefine could be used to reconcile text strings against matches found through the OpenRefine reconciliation endpoints for the CONTENTdm Wikibase and could also use the similar endpoint supported by the Wikidata community to reconcile strings against Wikidata. OCLC also made use of OpenRefine endpoints developed and hosted by others to reconcile against the OCLC FAST³⁰ subject terminology system, the VIAF³¹ authority file service, and the GeoNames³² service for geographic data.

After cleaning up and reconciling the CONTENTdm metadata, OCLC staff exported the data from OpenRefine and used locally developed scripts written in the Python³³ scripting language to restructure the data to match the format specified for the Wikidata QuickStatements³⁴ application. This is a tab-separated format with a set of rules for adding data to a Wikibase, with each row representing a single component of the item's description. And OCLC utilized the Pywikibot³⁵ library to develop another application that could read the QuickStatements data and load it into the Wikibase.

The most significant barrier to quickly transforming and loading CONTENTdm metadata into the project Wikibase was the absence . . . of Wikibase entity descriptions for the people, organizations, places, concepts, and events that are represented in the CONTENTdm records.

ADDING RELATED ENTITIES TO THE CONTENTDM WIKIBASE FROM EXTERNAL SOURCES

The most significant barrier to quickly transforming and loading CONTENTdm metadata into the project Wikibase was the absence, especially in the early stages, of Wikibase entity descriptions for the people, organizations, places, concepts, and events that are represented in the CONTENTdm records. In a linked data environment, each of those related entities must have its own entity description in the system, so that relationships can be defined between the entities. For example, when transforming a CONTENTdm record for a photograph, a "photographer" property should be added to the entity describing the photograph with a link to a separate entity for the photographer.

Unless those related entities are already in the project Wikibase and can be matched through OpenRefine reconciliation, the data loading process stalls until data for the related entities can be found, transformed, and loaded.

To move the process along and create entities as quickly as possible, OCLC staff initially created entities just for the creative work and its direct string-based properties (e.g., its title, description, height, width, IIF Manifest URL, etc.). Once that step was completed, OpenRefine and the project's SPARQL Query Service were used to look up the newly created Wikibase identifier for each item and those identifiers were added into a new column in the OpenRefine project. That step was followed by the creation of one or more new OpenRefine projects focused on reconciling strings for related entities and making connections between those entities and the creative work entities in the Wikibase.

Creating entities in advance for anticipated matches

OCLC also created Wikibase entity descriptions in advance for concepts and places that were anticipated to be mentioned in the CONTENTdm source data so the OpenRefine reconciliation process would find something to match against.

Entities for concepts were based on a set of headings from OCLC's FAST subject vocabulary. Staff selected subject headings that are widely used in other databases with the expectation that these would represent headings that would also occur in CONTENTdm metadata. The subject headings were transformed and loaded into the Wikibase as concept entities. This created an initial set of about 75,000 concept entities. In a second step, the FAST data was analyzed to find "broader concept" relationships for the 75,000 concept entities, and new concept entities were created for all of the "broader concept" FAST headings. Adding broader concept entities resulted in a total of over 100,000 concept entities being added to the Wikibase to support the CONTENTdm metadata matching process.

Entities for places that were anticipated to be found in CONTENTdm metadata were created based on information from the GeoNames geographical database, beginning with data describing cities with a population larger than 15,000 along with other place descriptions from administratively higher levels (countries, states, provinces, territories, counties, etc.). The GeoNames data processing produced about 70,000 Place entities for reconciliation.

This step of prepopulating the Wikibase with descriptions of entities for anticipated CONTENTdm headings helped reduce the barriers for entity creation. But the limits that were applied to the external sources meant that there were potential matches still to be found in FAST or GeoNames that had not been included, and potentially additional or richer data available from VIAF and Wikidata. Unmatched headings were reconciled against those services in OpenRefine, and if matches were found, the external source data was retrieved, converted, and loaded into the project Wikibase, and reconciliation was attempted again. OCLC also developed a separate application called the "Retriever," described in more detail later in this report, that staff used to search for matches in Wikidata, VIAF, and FAST and create new entities with a simple web interface.

Testing an alternative openrefine reconciliation endpoint

During the second phase of the pilot, OCLC prototyped a new reconciliation endpoint for matching against headings in the project Wikibase, relying on separate indexes of entity data to speed up the reconciliation process. The performance metrics for this prototype service were very encouraging, as it does not rely on SPARQL Queries and the triplestore for matching, which can

slow the process down. The index response times were consistently much faster in OCLC tests. This efficiency gain comes at the cost of replicating and synchronizing data from the Wikibase in another index, but for this project the costs were easily managed. OCLC staff provided a detailed presentation on this prototype work as part of the OCLC DevConnect Online 2020³⁶ series. The DevConnect webinar sparked some interest from a few developers that work on OpenRefine, and the OpenRefine Reconciliation Service API and OCLC has consulted with them to determine if any of our optimizations can be incorporated into their projects.

Creating placeholder entities for things that could not be reconciled

Some entities mentioned in CONTENTdm records could not be found in the controlled vocabularies and authority control systems that were used by OCLC for reconciliation. This was an anticipated, and indeed one of the points of carrying out this pilot was to better understand how these references appear and how to account for them in a Wikibase, where the established identity of the entity is of great importance. The solution the team settled on was to create a “placeholder” entity with as much information about the referenced entity as could be extracted from the CONTENTdm description, for instance its type (person, organization, etc.), birth and death dates (if present), occupation, and a consistently applied component of the Wikibase description that would help suggest, for potential future matches during reconciliation, that the entity’s identity had not yet been established (figure 12).

A “Placeholder” Entity for a Person without an Established Identity

The screenshot shows the Wikibase entity page for 'E. Burke Wilford' (Q144548). The page layout includes a sidebar on the left with navigation links like 'Main page', 'Recent changes', and 'Tools'. The main content area has tabs for 'Item' and 'Discussion', and a search bar. The entity title is 'E. Burke Wilford (Q144548)'. Below the title is a warning: 'This is an unidentified item. Until its identity is established DO NOT CHANGE THIS DESCRIPTION.' followed by the label 'Wilford, E. Burke'. A table titled 'In more languages' has columns for Language, Label, Description, and Also known as. The English row shows the label 'E. Burke Wilford' and the same placeholder description. Below this is a 'Constraint Violation Report' and a 'Statements' section with three entries: 'sex or gender' (male), 'type' (person), and 'description' (President Pennsylvania Aircraft Syndicate, Phila (English)).

FIGURE 12. A “placeholder” entity for a person without an established identity.³⁷ View a larger image [online](#).

Representing Compound Objects

Descriptions of cultural materials that consist of multiple parts, such as a photograph album or the recto and verso views of a postcard, can be structured in CONTENTdm as “compound objects.” Compound objects maintain the sequential order of related digitized items and can include an “object description” of the whole item, along with more detailed “item descriptions” about each part.³⁸

The project team tested two ways to maintain this structure and descriptive detail in Wikibase entities that were created from CONTENTdm compound object metadata.

In the most granular and detailed approach, for a photograph album, an entity for the album was created along with separate entities for the album cover and its individual pages. Each cover or page entity has a “part of creative work” property linking back to the album entity. While this approach acknowledges the whole-part relationship of pages to the album, that relationship on its own cannot represent the sequential order of the parts. To document their sequential order, “has creative work part” statements were added to the description of the album, linking to the related parts, and each statement was qualified with a “series ordinal” property to represent the numeric sequence of the pages (figure 13).

Example “has creative work part” Statements and Sequencing for the First Four Parts of an Album









| has creative work part | |
|---|--|
|  | Front cover of "The life and Times of Nicolas Hubble, Esq.", a photograph album of Edwin Hubble's cat, Nicolas Copernicus  |
| | series ordinal 1 |
| | 0 references |
|  | Inside front cover of the photo album "The life and times of Nicolas Hubble, Esq.", Edwin Hubble's cat  |
| | series ordinal 2 |
| | 0 references |
|  | Title page 1 to photo album entitled "The life and times of Nicolas Hubble, Esq."; a photo album of Edwin Hubble's cat with photos taken by Alfred and Maxine Parish  |
| | series ordinal 3 |
| | 0 references |
|  | Title page 2 of the photo album entitled "The life and times of Nicolas Hubble Esq"  |
| | series ordinal 4 |
| | 0 references |

FIGURE 13. Example “has creative work part” statements and sequencing for the first four parts of an album.³⁹ View a larger image [online](#).

In reviewing other compound objects, a more typical pattern revealed that they included very little item-level descriptive data beyond a default caption such as “Page 1,” “Page 2,” etc. It was also noted that the IIIF Manifest that is present for compound objects maintains the structure, sequence, and captions of related images. That led to a decision to describe most compound objects as a single entity without separate entities for the items in the compound object, relying on access to the structure and sequence and caption-level metadata in the corresponding IIIF Manifest.

Syndicating Data in Standard Schemas

The data managed in the CONTENTdm Wikibase is accessible through MediaWiki APIs and the Wikibase user interface and can be transformed by Wikibase in several formats including the RDF⁴⁰ linked data formats Turtle,⁴¹ N-Triples,⁴² and JSON-LD,⁴³ along with the non-RDF formats of a proprietary Wikibase JSON⁴⁴ object and Serialized PHP.⁴⁵

The pilot study also evaluated mechanisms for transforming data from Wikibase into schemas used by other systems where this data may eventually be shared. Specifically, OCLC added equivalent class and equivalent property statements for the CONTENTdm data model’s classes and properties, which were then used by a conversion program to crosswalk the data into either the DPLA Metadata Application Profile⁴⁶ or Schema.org.⁴⁷

The Schema.org transformation was used with a CONTENTdm Custom Javascript extension, described below, to test embedding JSON-LD linked data within CONTENTdm item pages.

A separate conversion process was developed to convert the project’s Wikibase data model representation into an RDF OWL Ontology⁴⁸ description. This conversion demonstrated the portability of both the Wikibase data model and the instance data if Wikibase were to be replaced by another structured data management system. The exported ontology also provided a clear way to see the model, separate from the instance data, which helped the project team explain how the data was created and structured.

In testing how data exports could be created and used, OCLC developed a conversion process that took the Wikibase JSON data and generated JSON-LD data that conformed to the JSON-LD 1.1 specification and followed the emerging W3C best practices for JSON-LD⁴⁹ as well as IIIF JSON-LD design patterns.⁵⁰ This conversion demonstrated the versatility and portability of the Wikibase JSON data and provided “developer friendly” data for our prototype applications, such as the Explorer application described in this report, to use.

Wikibase Ecosystem Advantages

The selection of the MediaWiki environment and its Wikibase extension brings several advantages right out of the box. Without custom software development or user interface design and testing, these can be employed to produce new data management and user experience benefits.

IMPLEMENTING AUTHORITY CONTROL

CONTENTdm currently has a traditional record-oriented data model, where headings for various entities are based on a single string. Varying cataloging practices and sources for controlled vocabularies can, in that approach, create obstacles to searching for the name of a person, organization, concept, place, or event if you do not know the exact form of the heading. But in the

Wikibase environment, any number of different heading strings and in different languages can be associated with an entity, greatly increasing the effectiveness of recall while strongly supporting precision as well.

Other Names Associated with the Los Angeles Dodgers Entity

The screenshot shows the CONTENTdm linked data interface for the Los Angeles Dodgers entity (Q166325). The page includes a sidebar with navigation links, a main content area with a description and a list of alternative names, and a table of associated names in various languages.

Los Angeles Dodgers (Q166325)

baseball team and Major League Baseball franchise in Los Angeles, California, United States

Dodgers | Brooklyn Dodgers | Brooklyn Robins | Brooklyn Superbas | Brooklyn Bridegrooms | Brooklyn Grooms | Brooklyn Grays | Brooklyn Atlantics | Bums | Robins | Superbas | Bridegrooms | Grooms | Grays | Atlantics | LAD | Trolley Dodgers | LA Dodgers | Brooklyn Dodgers (Baseball team) | Los Angeles Dodgers (Baseball team)

▼ In more languages Configure

| Language | Label | Description | Also known as |
|----------|---------------------|---|--|
| English | Los Angeles Dodgers | baseball team and Major League Baseball franchise in Los Angeles, California, United States | Dodgers Brooklyn Dodgers Brooklyn Robins Brooklyn Superbas Brooklyn Bridegrooms Brooklyn Grooms Brooklyn Grays Brooklyn Atlantics Bums Robins Superbas Bridegrooms Grooms Grays Atlantics LAD Trolley Dodgers LA Dodgers Brooklyn Dodgers (Baseball team) Los Angeles Dodgers (Baseball ... |

All entered languages

Statements

| | | |
|------|----------------------------------|--------------|
| type | organization <small>edit</small> | 0 references |
|------|----------------------------------|--------------|

FIGURE 14. Other names associated with the Los Angeles Dodgers entity.⁵¹ View a larger image [online](#).

For example, in CONTENTdm a precise search to find works associated with the Los Angeles Dodgers baseball team may (depending on the cataloging practices of the institution) need to use the Library of Congress (LC) heading “Los Angeles Dodgers (Baseball team).” But in the Wikibase environment, the entity describing that organization could be found using that LC preferred form, or any of several current colloquial names or previous official names, including “LA Dodgers,” “Brooklyn Dodgers,” “Trolley Dodgers,” “Brooklyn Grays,” and others (figure 14). In the Wikibase environment each entity is registered with and retrievable with its own unique identifier, separate from any and all names with which it may be associated.

DECREASING CATALOGING INEFFICIENCIES, INCREASING DESCRIPTIVE QUALITY

In a record-oriented system like CONTENTdm, if a cataloger wants to include biographic or other descriptive information about an entity associated with a work, such as information about the photographer of an image or about a depicted person, that information needs to be added to as the value of a field in every record where it is applicable. Then, if information about the related entity needs to change, all the associated records need to be updated to keep that information current and synchronized. This data management overhead may be one reason why descriptions of related entities are not common in traditional cataloging environments.

First Parts of the Description of Jasper Wood





| | |
|-------------|---|
| description |  Cleveland free speech activist Jasper Wood was a self-taught writer and photographer. His principal subjects were the residents of the Scovill Avenue area of Cleveland, with whom he became familiar while frequenting jazz clubs in the neighborhood in the late 1930s and early 1940s. He purchased his first camera in 1946 and first exhibited his work in 1947 at the Cleveland Museum of Art's annual (English) series ordinal 1 0 references |
| |  May Show. According to the Bulletin of the Cleveland Museum of Art, he won three first place May Show awards (1949, 1951, 1953) and two honorable mentions (1947, 1952) for his photographs. His last May Show entry was in 1958. Through his photographs, many of which were taken with a Contax 35mm camera, Wood attempted to capture what he called the "felt moment seen," or the emotional essence of (English) series ordinal 2 0 references |
| |  what he, the photographer, was seeing. Jasper Wood took photographs to feel alive and connected to the world. He did not sell his photographs or create a career from them. To him, the creative act was most significant. In 1951, Wood won first place in the 31st annual competition of American Photography magazine (September 1951, page 529). In 1953, Wood made a 15 minute poetic documentary titled (English) series ordinal 3 0 references |
| |  Streetcar, which depicted life in a big American city (Cleveland) centered on the experience of riding its streetcars. The film can be viewed at the Library's YouTube account. One of Wood's images was included in Edward Steichen's 1955 exhibit The Family of Man (page 191 in the published catalog). Wood, who also took photographs in Mexico and Ohio Amish Country, took few photographs after 1960. (English) series ordinal 4 0 references |

FIGURE 15. First parts of the description of Jasper Wood.⁵² View a larger image [online](#).

In the Wikibase environment, entities for works and for things associated with the work are maintained separately. The description of the photographer, or of the depicted person, can be entered and maintained in one entity description as illustrated in figure 15, and any changes to that description can be immediately seen through the relationships the entity has to other entities. This efficiency improvement could encourage richer descriptions of related entities, including context and relations that are not typically added in existing record-oriented systems.

GENERATING DATA VISUALIZATIONS

As the system architecture diagram included in this report represents, the Wikibase ecosystem includes a component that watches for changes in the Wikibase entity descriptions, retrieves that data in the form of linked data triples, updates the data in a linked data database or “triplestore,” and provides a separate user interface for querying that data using the SPARQL Query language. The user interface has built-in tools for constructing SPARQL queries and determining how the results can be visualized. The SPARQL query language is a powerful tool for making connections between and across entities, producing results that would be difficult and, in some cases, not feasible in a traditional record-oriented system. As shown in figure 16, a simple SPARQL query can retrieve all of the entities for places that are said to be depicted by works in a collection and, using the geographic coordinates in the place entity description, locate the place in a map visualization along with information about the related work.

SPARQL Query Map Visualization of Places Depicted in Works from a Collection

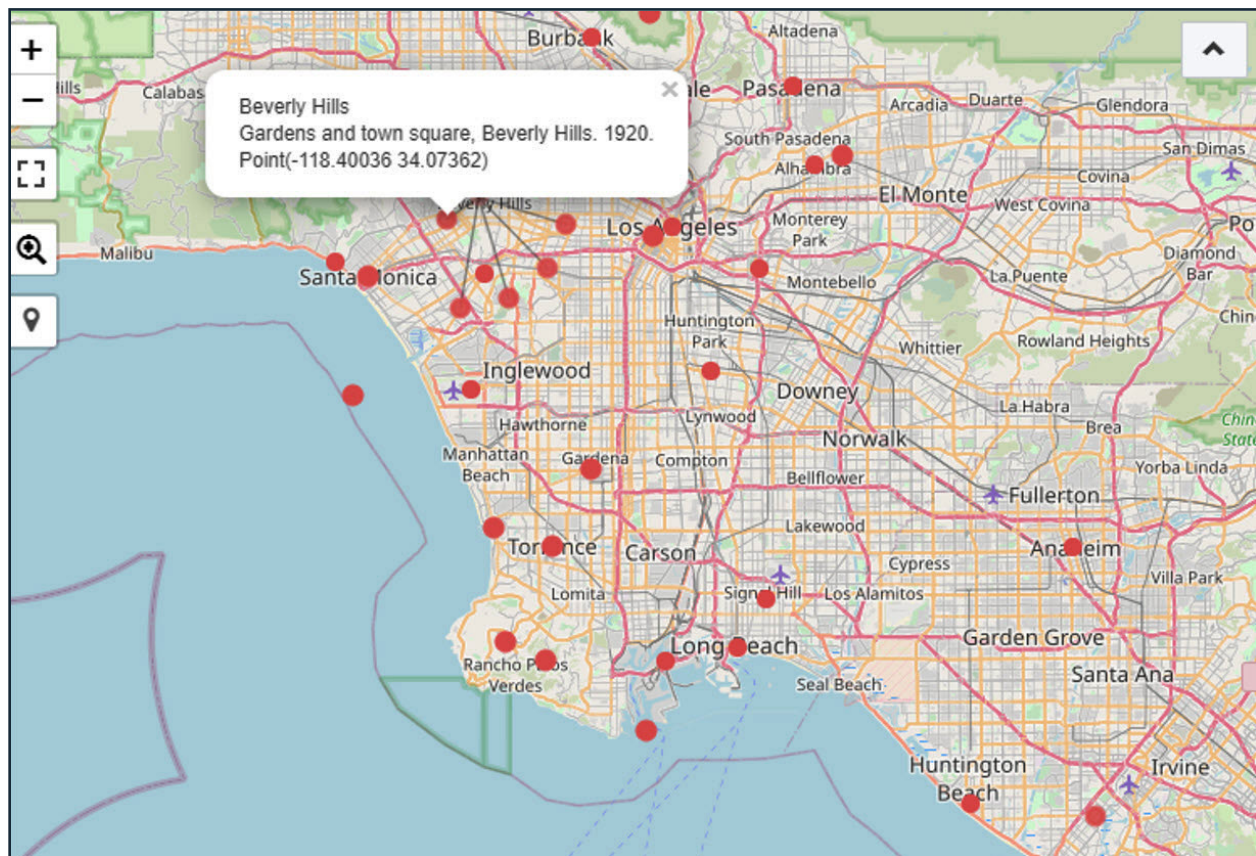


FIGURE 16. SPARQL Query map visualization of places depicted in works from a collection.⁵³ View a larger image [online](#).

User Interface Extensions

MEDIAWIKI GADGETS

The MediaWiki platform for Wikibase provides a Gadgets⁵⁴ extension that can be used to develop and add custom features to the user interface. OCLC staff took advantage of this feature to extend the interface, both to alter the user experience and to provide quality assurance tools.

Adding the Mirador viewer

Mirador⁵⁵ is a configurable, extensible, and easy-to-integrate image viewer that enables image annotation and comparison of images from repositories dispersed around the world. It can interpret the metadata and images that are included in IIIF Presentation Manifests. CONTENTdm generates IIIF manifests for all its image-based content, so Mirador was a great fit for this pilot project. Without an embedded image viewer, the Wikibase item entity displays are limited to text and are static. The Mirador viewer adds a degree of interactivity to the user experience: images can be viewed in detail, and for compound objects pages can be turned, without leaving the Wikibase user interface, as shown in figure 17.

Mirador Image Viewer Embedded in the Wikibase User Interface

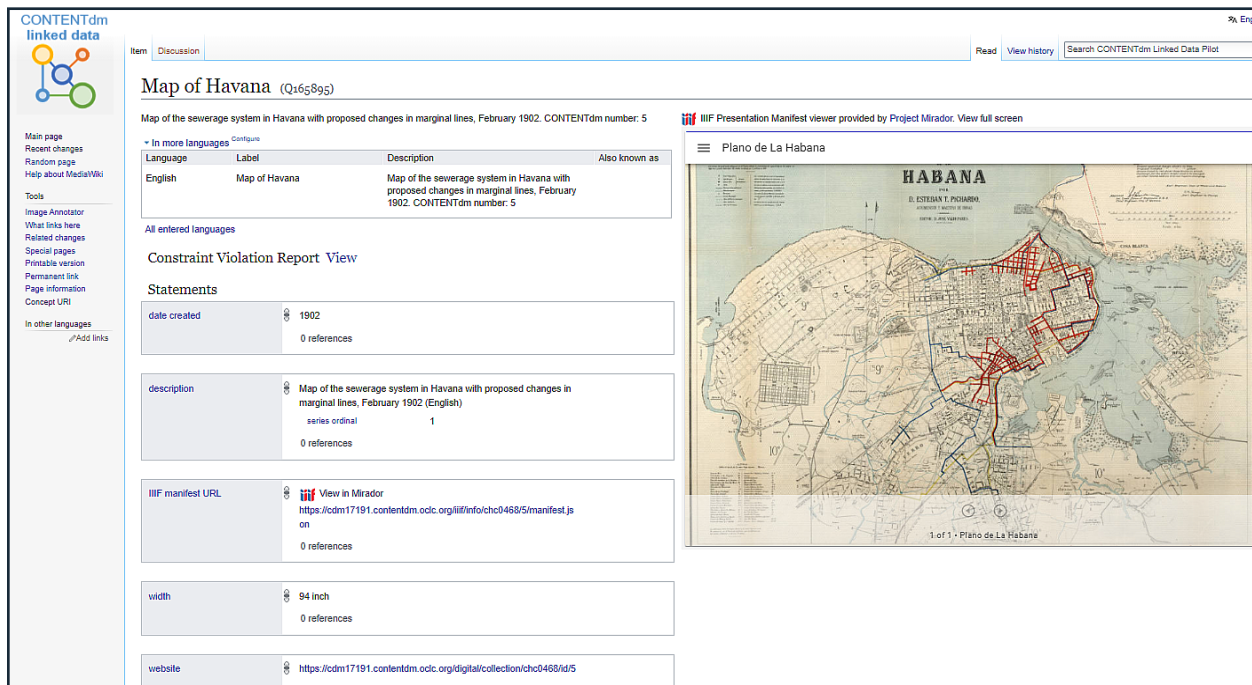


FIGURE 17. Mirador image viewer embedded in the Wikibase user interface.⁵⁶ View a larger image [online](#).

Showing contextual information from Wikidata

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. And in the linked data project Wikibase, many entities included identifiers for descriptions of the same entity in Wikidata.

OCLC developers created a Wikibase gadget that could detect the presence of the related Wikidata identifier in an entity description, make a connection to Wikidata in real-time to find an associated Wikipedia article link, and use the Wikipedia link to obtain a summary description of the entity and, in many cases, a related image from Wikimedia Commons.

OCLC developers found that this MediaWiki Gadget was simple to write. But the Gadget depended on a separate and more complex application created by OCLC developers that made all the system connections and carried out the database searches for contextual information and cached that information so as not to overburden the other shared services. The resulting contextual information included in a Wikibase entity description of San Francisco is illustrated in figure 18.

Contextual Data and Image from DBpedia and Wikimedia Commons Embedded in the Wikibase User Interface

The screenshot shows the Wikibase user interface for the entity "San Francisco" (Q71945). The interface includes a sidebar with navigation options, a main content area with a table of contextual data, and a "Context and Background" section with an image from Wikimedia Commons.

San Francisco (Q71945)
 city and county seat in California, United States
 Frisco | SF | San Francisco, CA | San Francisco, California

| Language | Label | Description | Also known as |
|----------|---------------|---|--|
| English | San Francisco | city and county seat in California, United States | Frisco SF San Francisco, CA San Francisco, California |

Constraint Violation Report View

Statements

| | | |
|--------------------------|-------------------------|--------------|
| type | locality | 0 references |
| administratively part of | San Francisco County | 1 reference |
| geographic coordinates | 37°46′30″N, 122°25′10″W | 1 reference |
| population | 864,816 | 1 reference |
| elevation | 16 metre | 1 reference |

Context and Background

San Francisco (*sæn frən sʰtskɔʊ*), officially the City and County of San Francisco, is the cultural, commercial, and financial center of Northern California and the only consolidated city-county in California. San Francisco encompasses a land area of about 46.9 square miles (121 km²) on the northern end of the San Francisco Peninsula, which makes it the smallest county in the state. It has a density of about 18,451 people per square mile (7,124 people per km²), making it the most densely settled large city (population greater than 200,000) in the state of California and the second-most densely populated major city in the United States after New York City. San Francisco is the fourth-most populous city in California, after Los Angeles, San Diego, and San Jose, and the 13th-most populous city in the United States—with a Census-estimated 2015 population of 864,816. The city and its surrounding areas are known as the San Francisco Bay Area, and are a part of the larger OMB-designated San Jose-San Francisco-Oakland combined statistical area, the fifth most populous in the nation with an estimated population of 8.7 million. San Francisco (Spanish for Saint Francis) was founded on June 29, 1776, when colonists from Spain established Presidio of San Francisco at the Golden Gate and Mission San Francisco de Asís named for St. Francis of Assisi a few miles away. The California Gold Rush of 1849 brought rapid growth, making it the largest city on the West Coast at the time. San Francisco became a consolidated city-county in 1856. After three-quarters of the city was destroyed by the 1906 earthquake and fire, San Francisco was quickly rebuilt, hosting the Panama-Pacific International Exposition nine years later. In World War II, San Francisco was the port of embarkation for service members shipping out to the Pacific Theater. After the war, the confluence of returning servicemen, massive immigration, liberalizing attitudes, along with the rise of the "hippie" counterculture, the Sexual Revolution, the Peace Movement growing from opposition to United States involvement in the Vietnam War, and other factors led to the Summer of Love and the gay rights movement, cementing San Francisco as a center of liberal activism in the United States. Politically, the city votes strongly along liberal Democratic Party lines. A popular tourist destination, San Francisco is known for its cool summers, fog, steep rolling hills, eclectic mix of architecture, and landmarks, including the Golden Gate Bridge, cable cars, the former Alcatraz Federal Penitentiary, Fisherman's Wharf, and its Chinatown district. San Francisco is also the headquarters of five major banking institutions and various other companies such as Levi Strauss & Co., Gap Inc., Salesforce.com, Dropbox, Reddit, Square, Inc., Dolby, Airbnb, Weebly, Pacific Gas and Electric Company, Yelp, Pinterest, Twitter, Uber, Lyft, Mozilla, Wikimedia Foundation, and Craigslist. It has several nicknames, including "The City by the Bay", "Fog City", "San Fran", and "Frisco", as well as older ones like "The City that Knows How", "Baghdad by the Bay", "The Paris of the West", or simply "The City". As of 2015, San Francisco was ranked high on world livability rankings.

FIGURE 18. Contextual data and image from DBpedia and Wikimedia Commons embedded in the Wikibase user interface.⁵⁷ View a larger image [online](#).

Revealing constraint violations

Constraints⁵⁸ are a Wikibase quality assurance feature that can be defined for properties and classes to describe their expected or allowed uses. For example, the property for "birthplace" might have a type constraint set indicating that the property should only be used for items that are an instance of the class "Person," that the object of the birthplace statement should be an instance of the class "Place" or one of its subclasses, and a cardinality constraint indicating that an entity should not have more than one birthplace statement.

Leveraging the project's SPARQL Query Service and its triplestore, OCLC developed a gadget that can compare the properties set for an item with any constraints set for the property and return a list of "constraint violations." In some cases, these will represent errors in the description that should be changed. In other cases, they can point to adjustments that may be needed to the data model.

As illustrated in figure 19, a constraint violation is noted for the Soviet Space Dog “Laika.” An occupation property has been set to “Astronauts,” but the type constraint for the occupation property indicates that it should only be used for instances of the type “person.” This view helps the project team see the violations generated by unexpected data and decide whether to modify these constraints, in this case based on what the community decides about occupations and whether they can be associated with other beings other than persons.

Constraint Violation Indicating the “Occupation” Property Should Only Be Used for Instances of the Type “Person”

The screenshot shows the CONTENTdm linked data interface for the item 'Laika' (Q73246). The interface includes a sidebar with navigation links and a main content area. The main content area displays the item name, a description, and a table of labels in different languages. Below this, a 'Constraint Violation Report' table shows a violation where the 'occupation' property is set to 'Astronauts', which is not an expected type for this property. The report also includes a 'Statements' section with details for 'birth date', 'death date', and 'type'.

CONTENTdm linked data

Item Discussion

Laika (Q73246)

Soviet space dog
Kudryavka | Zhuchka | Limonchik | Muttnik

▼ In more languages Configure

| Language | Label | Description | Also known as |
|----------|-------|------------------|--|
| English | Laika | Soviet space dog | Kudryavka Zhuchka Limonchik Muttnik |

All entered languages

Constraint Violation Report

| Violation | Property | Actual value | Expected Types |
|-----------------|------------|--------------|----------------|
| type constraint | occupation | Astronauts | person |

This result is cached and might be out of date by up to 10 minutes.

Statements

| | | |
|------------|-----------------|--------------|
| birth date | 1954 | 1 reference |
| death date | 3 November 1957 | 1 reference |
| type | dog | 0 references |

FIGURE 19. A constraint violation indicating that the “occupation” property should only be used for instances of the type “person.”⁵⁹ View a larger image [online](#).

CONTENTDM CUSTOM PAGES

A very useful feature of the CONTENTdm system is the ability to create Custom Pages⁶⁰ using CSS and Javascript to adjust and extend the default user interface features. You can find a wide array of examples in the CONTENTdm Customization Cookbook site.⁶¹ The CONTENTdm pilot used this customization feature to test how linked data from the pilot project Wikibase could power two enhancements to the production CONTENTdm system’s item displays.

Embedding Schema.org JSON-LD in CONTENTdm pages

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the internet, on web pages, in email messages, and beyond. By mapping the CONTENTdm Wikibase data model to Schema.org classes and properties, and by developing a conversion program to generate Schema.org-compatible descriptions of entities in the Wikibase, OCLC developed a CONTENTdm customization that embeds the Schema.org data within a CONTENTdm item page, formatted as JSON-LD, to make the content of the page easier for search engines to find and interpret (table 1).

TABLE 1. Example Schema.org JSON-LD for a CONTENTdm entity

Example Schema.org JSON-LD for a CONTENTdm entity

```
<script type="application/ld+json">
  {
    "@context": {"@vocab": "http://schema.org/"},
    "@graph": [
      {
        "@id": "https://researchworks.oclc.org/entity/Q73243",
        "@type": "CreativeWork",
        "name": "\"Buck\" of The Call of the Wild. Owned by the Bond
        Brothers. Dawson, N.W.T.",
        "alternateName": "\"Buck\" of The Call of the Wild. Owned by the
        Bond Brothers. Dawson, N.W.T.",
        "description": "Marshall Bond, Oliver H. R. La Farge, Lyman R. Cold,
        and Stanley Pearce sit with dog in front of log cabin in Dawson,
        Yukon.",
        "about": [
          {"@id": "https://researchworks.oclc.org/entity/Q73235",
          "@type": "Place", "name": "Dawson City"},
          {"@id": "https://researchworks.oclc.org/entity/Q73236",
          "@type": "Person", "name": "Marshall Latham Bond"},
          {"@id": "https://researchworks.oclc.org/entity/Q75440",
          "@type": "Intangible", "name": "Log cabins"},
          {"@id": "https://researchworks.oclc.org/entity/Q73245",
          "@type": "Intangible", "name": "Klondike Gold Rush"},
          {"@id": "https://researchworks.oclc.org/entity/Q123736",
          "@type": "Intangible", "name": "dogs"}
        ],
        "creator": {"@id": "https://researchworks.oclc.org/entity/Q73237",
        "@type": "Person", "name": "Jack London"},
        "image": "https://cdm15725.contentdm.oclc.org/digital/iiif/
        pl6003coll7/14/full/full/0/default.jpg",
        "dateCreated": "1897-01-01T00:00:00+00:00",
        "isPartOf": {"@id": "https://researchworks.oclc.org/entity/Q148324",
        "@type": "CreativeWorkSeries", "name": "Jack London Photographs
        and Negatives"}
      }
    ]
  }
</script>
```

The visibility to search engines of this embedded JSON-LD Schema.org metadata can be evaluated using applications like Google’s Structured Data Testing Tool (figure 20).⁶²

Schema.org Data Evaluated Using Google’s Structured Data Testing Tool

The screenshot displays the Google Structured Data Testing Tool interface. On the left, a JSON-LD snippet is shown, including a CreativeWork entity with properties like @context, @vocab, @graph, @id, @type, @id, name, description, creator, image, dateCreated, alternateName, and name. On the right, the tool has identified the CreativeWork entity and generated a structured data table with the following content:

| CreativeWork | |
|---------------|--|
| ID: | https://researchworks.oclc.org/entity/Q73243 |
| @type | CreativeWork |
| @id | https://researchworks.oclc.org/entity/Q73243 |
| description | Marshall Bond, Oliver H. R. La Farge, Lyman R. Cold, and Stanley Pearce sit with dog in front of log cabin in Dawson, Yukon. |
| image | https://cdm15725.contentdm.oclc.org/digital/iiif/p16003coll7/14/full/full/0/default.jpg |
| dateCreated | 1897-01-01T00:00:00+00:00 |
| alternateName | "Buck" of The Call of the Wild. Owned by the Bond Brothers. Dawson, N.W.T. |
| name | "Buck" of The Call of the Wild. Owned by the Bond Brothers. Dawson, N.W.T. |
| about | |
| @type | Thing |
| @id | https://researchworks.oclc.org/entity/Q73236 |
| name | Marshall Latham Bond |
| about | |
| @type | Thing |
| @id | https://researchworks.oclc.org/entity/Q123736 |
| name | Dogs |
| about | |
| @type | Thing |
| @id | https://researchworks.oclc.org/entity/Q73235 |
| name | Dawson City |
| about | |
| @type | Thing |


FIGURE 20. Schema.org data evaluated using Google’s Structured Data Testing Tool.⁶³ View a larger image [online](#).

Showing contextual information for headings based on Wikibase data

Similar to the Wikibase user interface gadget that adds contextual information about a single entity by connecting through Wikidata to obtain related information from Wikipedia, DBpedia, and Wikimedia Commons, an application was written that could be called by a CONTENTdm Custom Javascript component and, using the CONTENTdm item identifier as a way to find the related entity for the work in the pilot Wikibase, also find other entities related to the work in the Wikibase (the collection of which it is a part, subjects that it is about, its creator, and more), and for each of those entities look for and display more information, including an abstract and a thumbnail image.

This customization, shown in figure 21, was demonstrated to the pilot participants and there was interest in applying it to some of their collections, but the project did not see a production implementation of it, beyond OCLC’s testing, before the pilot period ended.


Additional Contextual Information Displayed in Contentdm Based on Entity Descriptions in the Pilot Wikibase



Advanced Search

Home > Jack London Photographs and Negatives > "Buck" of The Call of the Wild. Owned by the Bond Brothers. Dawson, N.W.T.

"Buck" of The Call of the Wild. Owned by the Bond Brothers. Dawson, N.W.T.



Search this record

About

Klondike Gold Rush

The Klondike Gold Rush was a migration by an estimated 100,000 prospectors to the Klondike region of the Yukon in north-western Canada between 1896 and 1899. Gold was discovered there by local miners on August 16, 1896 and, when news reached Seattle and San Francisco the following year, it triggered a stampede of would-be prospectors. Some became wealthy, but the majority went in vain. The Klondike Gold Rush ended in 1899 after gold was discovered in Nome, Alaska prompting an exodus from the Klondike. It has been immortalized by photographs, books, films, and artifacts. To reach the gold fields most took the route through the ports of Dyea and Skagway in Southeast Alaska. Here, the Klondikers could follow either the Chilkoot or the White Pass trails to the Yukon.

Sources: DBPedia Wikipedia Wikimedia Commons

Photographer

Jack London

John Griffith "Jack" London (born John Griffith Chaney, January 12, 1876 – November 22, 1916) was an American novelist, journalist, and social activist. A pioneer in the then-burgeoning world of commercial magazine fiction, he was one of the first fiction writers to obtain worldwide celebrity and a large fortune from his fiction alone. Some of his most famous works include *The Call of the Wild* and *White Fang*, both set in the Klondike Gold Rush, as well as the short stories "To Build a Fire", "An Odyssey of the North", and "Love of Life". He also wrote of the South Pacific in such stories as "The Pearls of Parlay" and "The Heathen", and of the San Francisco Bay area in "The Sea Wolf".

Sources: DBPedia Wikipedia Wikimedia Commons

Depicts

Dawson City

The Town of the City of Dawson, commonly known as Dawson City or Dawson, is a town in Yukon, Canada. It is inseparably linked to the Klondike Gold Rush. The population was 1,319 at the 2011 census.

Item Description

| | |
|----------------------------------|--|
| Title | "Buck" of The Call of the Wild. Owned by the Bond Brothers. Dawson, N.W.T. |
| Date | approximately 1897 |
| Call Number | JLP 13 |
| Physical Description | 1 photograph : print ; 9 x 15 cm |
| Location | Dawson (Yukon) |
| Description | Marshall Bond, Oliver H. R. La Farge, Lyman R. Cold, and Stanley Pearce sit with dog in front of log cabin in Dawson, Yukon. Written on back of photo: |
| Notes | Title provided by cataloger. The dog was the inspiration for the character "Buck" in <i>The Call of the Wild</i> . Though note on back of photograph says Louis Bond is pictured, scholarly sources identify him as Marshall Bond. Photograph in envelope dated 1898 with same caption as on verso of photo. Verso of photo reads: "Buck" hero of 'The Call of the Wild' with his master, Louis Bond and his cabin in Dawson." |
| Subjects | Klondike River Valley (Yukon) -- Gold discoveries. Log cabins. Dogs -- fiction. |
| Form/Genre | Black-and-white photographs. (aat) |
| Physical Collection | Jack London Collection |
| Physical Sub-Collection | Box 457, Photographs: Prints: A-H |
| Finding Aid | http://www.oac.cdlib.org/ftndaid/ark:/13030/tf8q2nb2xs |
| Rights | For information on using Huntington Library materials, please see Reproductions of Huntington Library Holdings: https://www.huntington.org/library-rights-permissions |
| Department | The Huntington Library, Art Collections, and Botanical Gardens. Manuscripts Department |
| Digital Collection | Jack London Photographs and Negatives, Huntington Digital Library |
| Unique Digital Identifier | 463141 |

FIGURE 21. Additional contextual information displayed in CONTENTdm based on entity descriptions in the pilot Wikibase.⁶⁴ View a larger image [online](#).

New Applications

The Linked Data project was well served by the “out of the box” features and functions of the MediaWiki platform, its Wikibase extension, the SPARQL Query service interface, the MediaWiki Gadgets component, and CONTENTdm Custom Pages. But for more complex investigations that were carried out during the project, the following prototype applications were developed:

- The **Image Annotator** to evaluate how subject matter experts could assist catalogers in describing images
- The **Retriever** to make the process of finding and adding new entity descriptions more efficient
- The **Describer** to investigate alternatives to the default Wikibase editing interface
- The **Explorer** and the Transportation Hub to demonstrate the value of aggregation and new discovery system features that maximize the value of linked data
- The **Field Analyzer** to assist metadata managers with analyzing their current collections

THE IMAGE ANNOTATOR

The CONTENTdm metadata transformation and reconciliation process produced descriptions of creative works that included, among other statements, relationships to other entities that the creative work either depicted or was in a more general sense “about.” The distinction between these two relationships was not always certain and a project goal was to better understand how this distinction is discerned by those managing digital collections. There was also an interest in testing whether the Wikibase platform could serve as the basis for new application development—in this case for an interface that would let domain experts and others augment the transformed CONTENTdm metadata with new annotations.

A user can review the statements that were created as part of the CONTENTdm metadata conversion process and quickly update any statements that need adjusting.

The Image Annotator application was developed and tested to investigate those questions. Given the Wikibase entity identifier for a creative work, it initially presents the work’s image along with a list of the “about” or “depicts” statements that are part of the entity description. This selective presentation of just some of the elements associated with the entity description was designed to give focus to the questions at hand: What is the image about, what does it depict, and can portions of the image be associated with depicted things?

A user can review the statements that were created as part of the CONTENTdm metadata conversion process and quickly update any statements that need adjusting, for example changing an “about” statement to a “depicts” statement if they determine that the related entity is truly depicted in the image (figure 22).

Image Annotator Initial View of an Image and Subjects

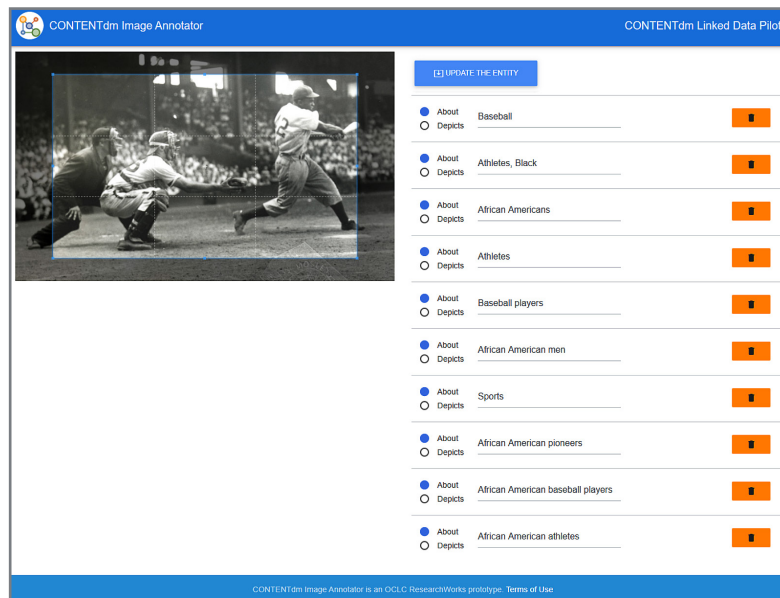


FIGURE 22. Image Annotator initial view of an image and subjects.⁶⁵ View a larger image [online](#).

And for any “depicts” statements, the user can apply the image cropping tool to associate the appropriate portion of the image with the depicted entity, providing a much finer-grained reckoning of the item and supplementing the Wikibase with new images associated with other entities (figure 23). The IIIF Image API supports the management of these selections and the persistent retrieval of the associated images.

Image Annotator Cropping an Image of a Person

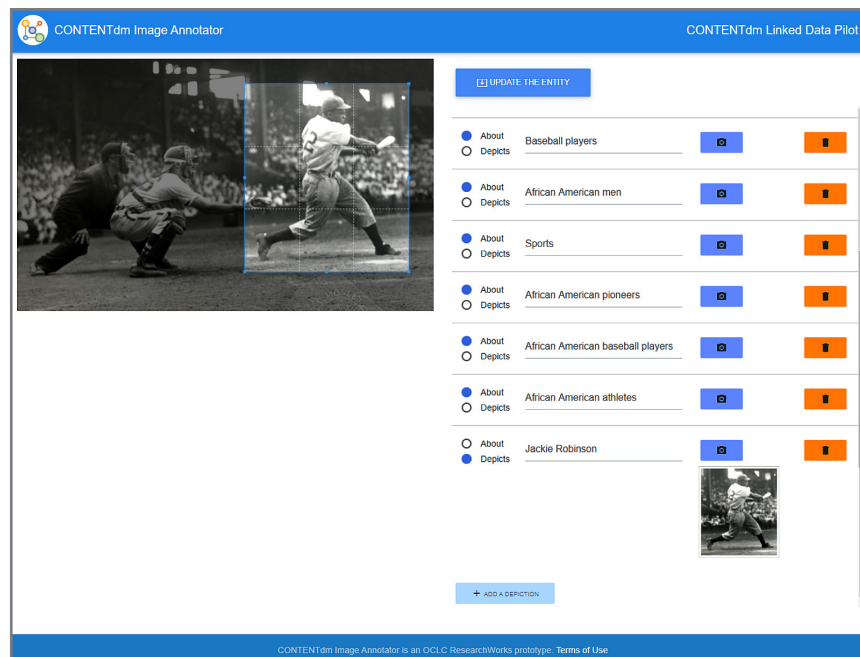


FIGURE 23. Image Annotator cropping an image of a person.⁶⁶ View a larger image [online](#).

The subject relationships shown in the Image Annotator are based on the CONTENTdm source data, but this new application gives users the opportunity to supplement the entity description with more “about” or “depicts” statements by searching for related entities and adding the new connections, with another cropped image if appropriate, as illustrated in figure 24 with the addition of the subjects “Baseball umpires” and “Catchers (Baseball)” with associated images.

This [Image Annotator] gives users the opportunity to supplement the entity description with more “about” or “depicts” statements by searching for related entities and adding the new connections.

Image Annotator After Adding More Depicted Subjects

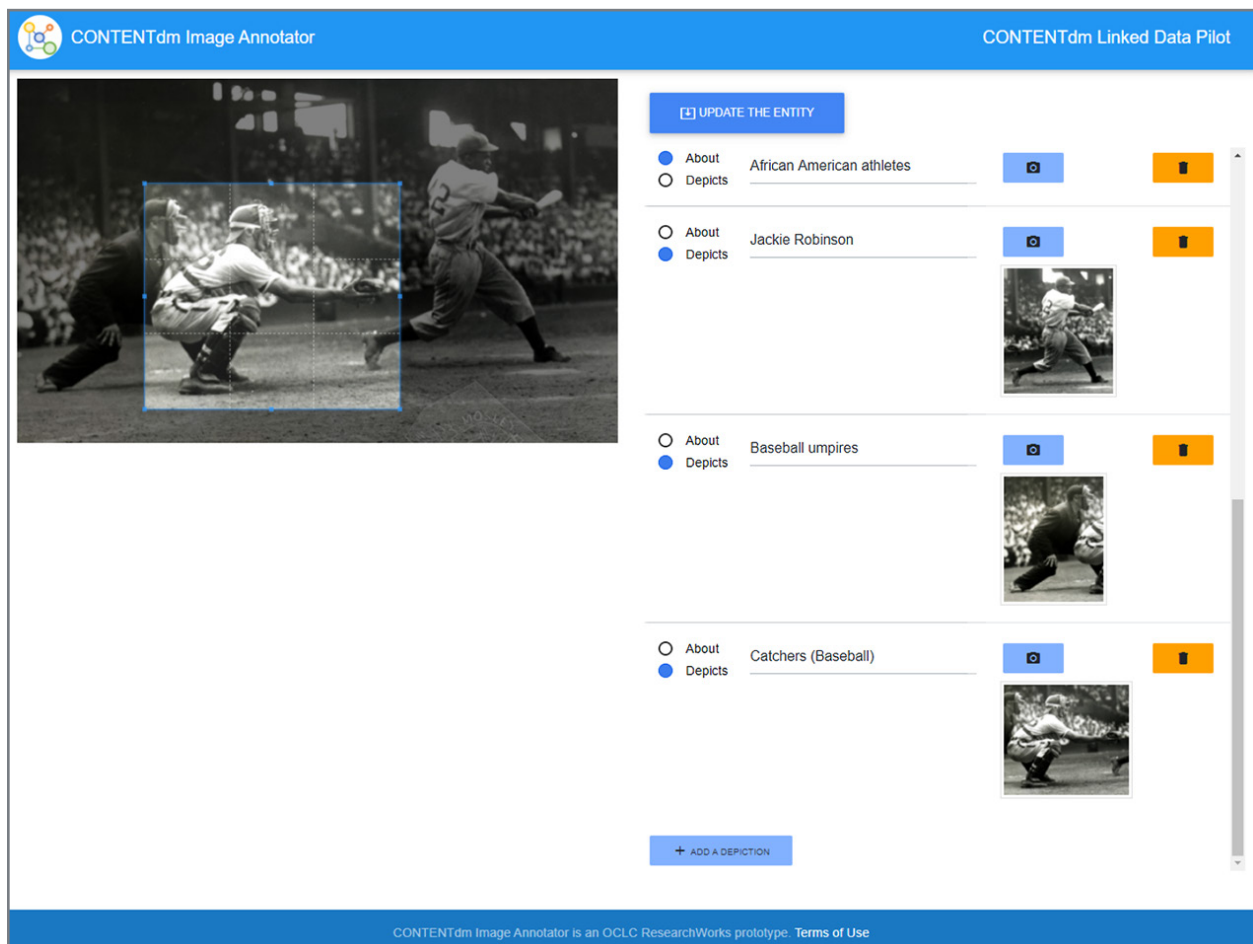


FIGURE 24. Image Annotator after adding more depicted subjects.⁶⁷ View a larger image [online](#).

Once the changes have been made in the Image Annotator, they can be saved to the Wikibase where they are immediately visible in its user interface (figure 25).

Wikibase Item Updated with Illustrated Depicts Statements

| | |
|--|--|
| depicts |  Jackie Robinson  |
| | digital representation URL https://cdm16002.contentdm.oclc.org/digital/iiif/p15037coll17/253/2354,260,1687,1938/full/0/default.jpg |
| |  0 references |
|  Baseball umpires  | |
| digital representation URL https://cdm16002.contentdm.oclc.org/digital/iiif/p15037coll17/253/58,752,1284,1610/full/0/default.jpg | |
|  0 references | |
|  Catchers (Baseball)  | |
| digital representation URL https://cdm16002.contentdm.oclc.org/digital/iiif/p15037coll17/253/915,752,1820,1610/full/0/default.jpg | |
|  0 references | |

FIGURE 25. Wikibase item updated with illustrated depicts statements.⁶⁸ View a larger image [online](#).

User study results

The usability of the Image Annotator was tested in November and December 2019 in three separate “Think Aloud” user studies.⁶⁹ In this type of study, test participants use the system while continuously thinking aloud—that is, verbalizing their thoughts as they move through the user interface. Reactions to and suggestions made about the Image Annotator were for the most part very positive, but also identified user interface and indexing improvements that would need to be implemented before it could become a truly productive tool.

The test results indicated that the Image Annotator application was usable, as everyone was able to complete the exercise steps. The results also helpfully revealed usability issues that had not been previously encountered during OCLC staff testing. The test study wrap-up discussions provided especially good feedback and suggestions for improvements. Test participants noted that the Image Annotator would be a useful tool for cleaning up metadata and would provide an easy way to bring subject matter experts from outside the library into the process of describing cultural materials.

The test results also identified several areas of needed improvement before the Image Annotator would be ready for regular use, including search and retrieval, scalability, user interface issues, and guidelines for descriptive practice.

In the area of search and retrieval, it was unclear that entering “free text” subject annotations that did not match a heading would not be retained when the entity was updated. Participants found that expected search results were not returned, for example a search for “kite” did not show a match for “kites.” And some participants hoped that vocabularies from nonlibrary domains could be included as the source for related entities: “That’s ultimately what makes digital collections meaningful.”

Scalability of the manual effort was a noted concern, in that it takes time to make annotations for individual objects, and collections can include thousands of objects. Some wondered whether crowdsourcing, under certain management and controls, could address that concern.

The Image Annotator’s mechanism for adding and removing annotations presented some usability obstacles: the absence of the “camera” button for “about” headings was confusing, a cropping icon would be preferable to a camera icon for that button, the camera and “add a depiction” buttons compete for attention, and there isn’t a way to delete a cropped image without deleting the entire depicts statement.

Some participants wondered how many subjects are “enough,” and what subjects are notable enough to deserve annotation. They wished for easier access to other descriptive metadata for the work, to help identify additional depicted subjects. The gray area between “about” and “depicts” was discussed by all without a clear consensus on when to select one or the other, though generally participants felt that “depicts” should have more of a guarantee that you will see the depicted thing in its entirety or as a significant portion in the object. For example, a photograph of Public Square in Cleveland would depict Public Square but be about Cleveland.

All the test subjects noted that the Image Annotator was enjoyable to use:

- “I like this little system.”
- “Once you get going it’s actually kind of fun.”
- “One of the things you’re offering is a way to have fun—quite literally a window into new ways of thinking about what we do.”

THE RETRIEVER

When describing an entity using the Wikibase interface, the workflow can come to a halt if you are trying to establish a relationship between the entity you are describing and some other entity when it is not yet in the Wikibase. To fill this gap, OCLC developed an application called the “Retriever” that can quickly search for an entity described in other systems and transfer those descriptions into the Wikibase as a new entity.

For example, if you are describing a photograph of Lake Vermilion in Minnesota and wanted to add a “depicts” statement linking the photograph entity to that place entity, if there isn’t already an entity in the Wikibase describing Lake Vermilion you’d need to stop editing the photograph’s description, switch to a new Wikibase editing window to create a new entity for the lake, and then return to the photograph entity description to add the statement claiming that the photograph depicts the lake.

That kind of disruption to the workflow can be reduced if there is a way to quickly add the missing entity’s description to the Wikibase. When a Wikibase is in its early stages, unless it is prepopulated with entity descriptions from another source, this situation will be commonplace. But it is also the case that in many instances the missing entity is already described in some other authority control system or vocabulary, so a tool that can find those descriptions, transform the data to align with the classes and properties in the Wikibase, provide an opportunity for human review and correction of the transformed data, and then automatically load the source data into the Wikibase as a new entity can help bridge the gap and keep the cataloging work flowing.

OCLC designed the Retriever to provide a simple keyword search interface to look for matching items in Wikidata, VIAF, and FAST (figure 26), a user interface for reviewing and editing data extracted from those sources (figure 27), and a back-end process for loading the transformed data into the Wikibase (figure 28). This application was originally developed, for the same use case, in OCLC’s Project Passage. The user interface component of the application was re-written in the Linked Data pilot to use a different Javascript framework, but the functionality was generally the same.

Retriever Search Results from Wikidata, VIAF, and FAST for “Lake Vermilion”

The screenshot shows the Retriever search interface. At the top, there is a search bar with the text "lake vermillion" and a "SEARCH" button. Below the search bar, the results are organized into three columns: Wikidata, VIAF, and FAST. Each column contains a list of search results with their respective titles, descriptions, and source links.

| Wikidata | VIAF | FAST |
|---|--|--|
| <p>Lake Vermilion large lake in St. Louis County, Minnesota, USA Wikidata</p> | <p>Minnesota Vermilion Lake Geographic name VIAF</p> | <p>Louisiana White Lake (Vermilion Parish) Place FAST</p> |
| <p>Lake Vermilion lake in Illinois, United States Wikidata</p> | <p>Louisiana White Lake (Vermilion Parish) Geographic name VIAF</p> | <p>Ohio Savannah Place FAST</p> |
| <p>Lake Vermilion State Park state park in Minnesota, United States Wikidata</p> | <p>Vermilion Lake Geographic name VIAF</p> | <p>Ontario Red Lake (Lake) Place FAST</p> |
| <p>Lake Vermilion Dam dam in Illinois, United States Wikidata</p> | <p>Lake Vermilion Singers Corporate name VIAF</p> | <p>Minnesota Vermilion Lake Place FAST</p> |
| | <p>Vermilion-Lake-Gebiet Geographic name VIAF</p> | |

Retriever is an OCLC ResearchWorks prototype. [Terms of Use](#)

FIGURE 26. Retriever search results from Wikidata, VIAF, and FAST for “Lake Vermilion.”⁷⁰ View a larger image [online](#).

Retriever Entity Editor

The screenshot shows the Retriever Entity Editor for the Wikidata item Q1801099, Lake Vermilion. The interface is in English and includes a 'CREATE ENTITY' button. It displays a table of labels and descriptions in multiple languages (English, Arabic, Dutch). Below this, it shows statements for geographic coordinates, elevation, type, and administrative part. The Wikidata Item ID is Q1801099. The footer indicates it is an OCLC ResearchWorks prototype.

| Language | Label | Description | Also Known As |
|----------|-----------------------|--|---------------|
| English | Lake Vermilion | large lake in St. Louis County, Minnesota, USA | |
| Arabic | بحيرة فيرميليون, تاور | بحيرة في الولايات المتحدة الأمريكية | |
| Dutch | Lake Vermilion | meer in de Verenigde Staten | |

Statements

| | |
|--------------------------|----------------------------------|
| geographic coordinates | 47.857908333333,-92.299611111111 |
| elevation | +420 metre |
| type | lake |
| administratively part of | Saint Louis County |

Identifiers

| | |
|------------------|----------|
| Wikidata Item ID | Q1801099 |
|------------------|----------|

Retriever is an OCLC ResearchWorks prototype. [Terms of Use](#)

FIGURE 27. Retriever entity editor.⁷¹ View a larger image [online](#).

Wikibase Entity Created by the Retriever

The screenshot shows the Wikibase entity page for Lake Vermilion (Q221424). The page includes a sidebar with navigation links, a main content area with a description and a table of labels, and a 'Context and Background' section with a detailed paragraph and an image. The page also features a 'Constraint Violation Report' and 'Statements' section.

Lake Vermilion (Q221424)
large lake in St. Louis County, Minnesota, USA

| Language | Label | Description | Also known as |
|----------|----------------|--|---------------|
| English | Lake Vermilion | large lake in St. Louis County, Minnesota, USA | |

Context and Background

Lake Vermilion is a freshwater lake in northeastern Minnesota, United States. The Ojibwe originally called the lake Nee-Man-Nee, which means "the evening sun tinting the water a reddish color". French fur traders translated this to the Latin word Vermilion, which is a red pigment. Lake Vermilion is located between the towns of Tower on the east and Cook on the west, in the heart of Minnesota's Arrowhead Region at Vermilion Iron Range. The area was mined from the late 19th century until the 1960s, and the Soudan Mine operated just south of the lake. The lake contains black crappie, bluegill, brown bullhead, largemouth bass, muskellunge, northern pike, sunfish, rock bass, smallmouth bass, tullibee (cisco), walleye, white sucker, and yellow perch. Lake Vermilion is known for its walleye and muskie fishing. In the spring of 2005, Lake Vermilion was host to the annual Minnesota Governor's Fishing Opener Weekend. Some fish consumption guideline restrictions have been placed on the lake's bluegill, cisco, crappie, northern pike, walleye, and white sucker due to mercury contamination. Many feel the increased population of muskies has had a detrimental effect on the walleye population, although walleye fishing has improved since the implementation of a walleye slot limit (18" to 26") and reduced bag limit (from 6 to 4) in recent years. The lake attracts visitors from all parts of Minnesota and the Midwestern United States, who lodge at the lake's numerous resorts and hotels. Tourists are drawn by Lake Vermilion's reputation as a fishing destination, as well as its setting in the northern Minnesota wilderness. The lake is near the Superior National Forest and the Boundary Waters Canoe Area Wilderness (BWC/WCA). The Minnesota DNR rates Lake Vermilion as the fifth largest lake by surface area within Minnesota borders. The surface area of Lake Vermilion is 39,271 acres (158.9 km²) and has a maximum depth of 76 feet (23 m). In 2007, Governor Tim Pawlenty announced the state was negotiating the purchase from U.S. Steel of a large area of land on the southeastern shore of the lake for a proposed new Minnesota state park. The sale of the land at a cost of \$18m was finalized in June 2010. Lake Vermilion State Park is being developed on the easterly southern shore of the lake, and is adjacent to and to the east of Soudan Underground Mine State Park. The claim that "in the 1940s, the National Geographic Society declared Lake Vermilion one of the top ten most scenic lakes in the United States" has been rebutted by a representative of the National Geographic Society. On May 13, 2014, it was announced that the 2015 Governor's Fisher Opener weekend would be held at Lake Vermilion again.

Sources: DBpedia, Wikipedia, Wikimedia Commons

FIGURE 28. Wikibase entity created by the Retriever.⁷² View a larger image [online](#).

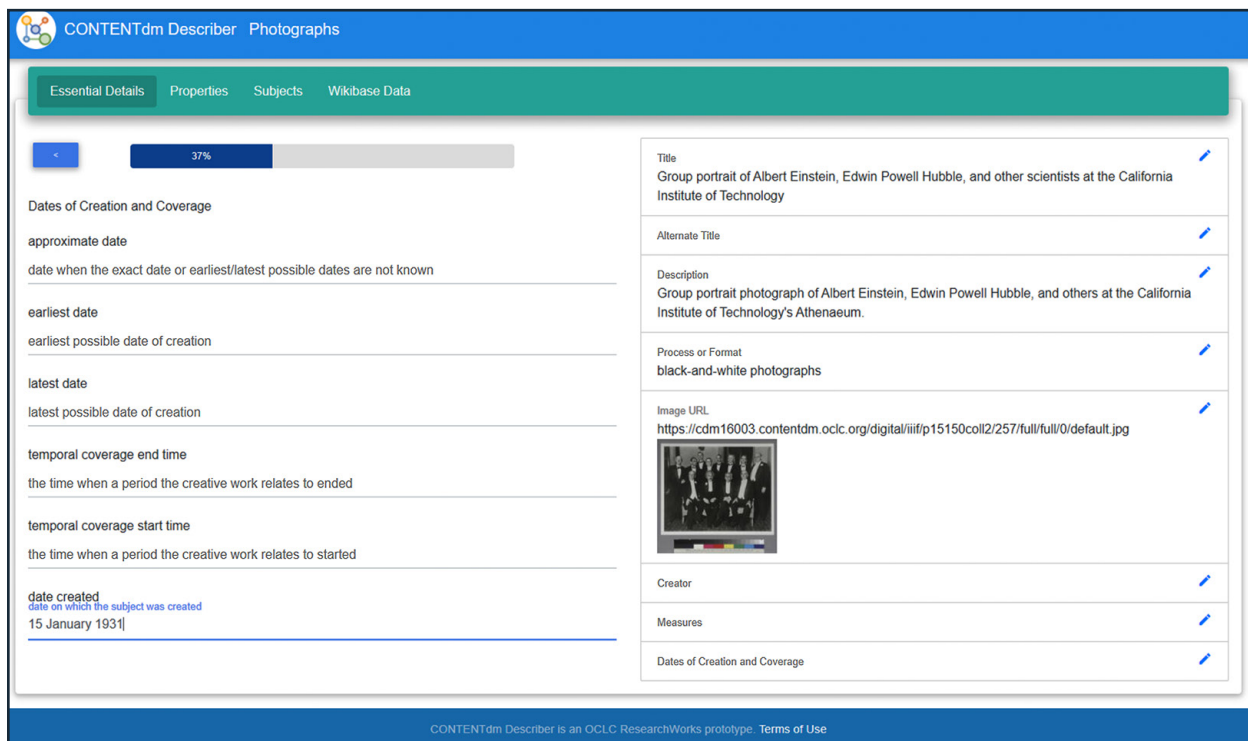
There is a server-based part of the Retriever application that takes search requests from the browser, handles mapping of external source data elements to the project Wikibase properties and classes, and utilizes the Python Pywikibot library to carry out data loading into the Wikibase.

THE DESCRIBER

The Linked Data project's goals included testing editing interface alternatives to the Wikibase default user interface. OCLC began development of a prototype web application named the "Describer" that aspired to provide a guided mode to cataloging entities for works, illustrated in figure 29. The user experience in the Describer would begin by prompting the cataloger to choose the type and classification of the material they were describing. Based on those selections, the Describer would begin prompting for additional details that would be common or expected for entities of that type and classification, factoring in property constraints and other details of the underlying data model. The Describer could also incorporate capabilities and features that had been previously tested in the Image Annotator and in the Retriever.

Work on the Describer prototype was not completed before the end of the pilot, but the initial testing suggested promise while also revealing the importance of carefully documenting the data model constraints in order to drive the user experience. Though not part of this pilot, a related investigation that could prove similarly illuminating would be to evaluate a language designed to express the shape of the data, such as SHACL⁷³ or ShEx,⁷⁴ as the mechanism for defining how the data model works and how that relates to user interface development.

Editing Essential Details for an Entity in the Describer



The screenshot displays the 'CONTENTdm Describer' interface for editing a 'Photographs' entity. The top navigation bar includes 'Essential Details', 'Properties', 'Subjects', and 'Wikibase Data'. A progress indicator shows 37% completion. The 'Essential Details' section is divided into two columns. The left column contains fields for 'Dates of Creation and Coverage', including 'approximate date', 'earliest date', 'latest date', 'temporal coverage end time', and 'temporal coverage start time'. The right column contains fields for 'Title', 'Alternate Title', 'Description', 'Process or Format', 'Image URL', 'Creator', 'Measures', and 'Dates of Creation and Coverage'. The 'Image URL' field includes a thumbnail of a group portrait photograph. The 'date created' field is populated with '15 January 1931'. The footer indicates 'CONTENTdm Describer is an OCLC ResearchWorks prototype. Terms of Use'.

FIGURE 29. Editing essential details for an entity in the Describer.⁷⁵ View a larger image [online](#).

THE EXPLORER AND THE TRANSPORTATION HUB

An important value proposition for making the transition to linked data is the ability to browse or navigate across the graph of data connections to find important related entities and reveal relationships that would be hard to see in a more traditional record-oriented search and retrieval system.

To evaluate this potential OCLC developed a prototype web application named the “Explorer” to focus on the most frequently occurring connections between entities, see relationships that were described by different institutions for different items in different collections, look for thematically-related content, and follow the graph-based connections to locate important related entities.

An important value proposition for making the transition to linked data is the ability to browse or navigate across the graph of data connections to find important related entities and reveal relationships that would be hard to see in a more traditional record-oriented search and retrieval system.

The home page of the Explorer lists entities organized across a subset of categories, sorted by frequency, to help researchers jump into the browsing experience and quickly see what the pilot project data is mostly “about” (figure 30). The Explorer also has a keyword search interface.

While the collections that were selected for evaluation in the first two phases of the pilot project were all interesting and, as a group, gave us a good idea about the range of data transformation and reconciliation challenges we’d likely encounter when working with other CONTENTdm sites and collections, they were not chosen with any special attention paid to how the materials they describe might thematically overlap.

To generate more topically related connections across the pilot participants’ data, OCLC assembled a new selection of CONTENTdm metadata records based on the topic of transportation. Using a general search for transportation-related subjects (the subject terms used were “streetcars,” “transportation,” “roads,” “highways,” “airports,” “railroads,” “automobiles,” “ferries,” “rockets,” “ships,” “boats,” “streets,” “paths”), OCLC staff applied a search across all collections for each pilot participant’s CONTENTdm site, gathered the resulting metadata records, and transformed the data for loading into Wikibase, reconciling as many headings to related entities as could be done without significant amounts of human attention. This step provided more data for us to use in assessing the scalability of this data transformation process, as we could compare this more automated and streamlined effort with the very thorough and largely manual process that had been applied to the initial set of pilot project collections.

Explorer Home Page

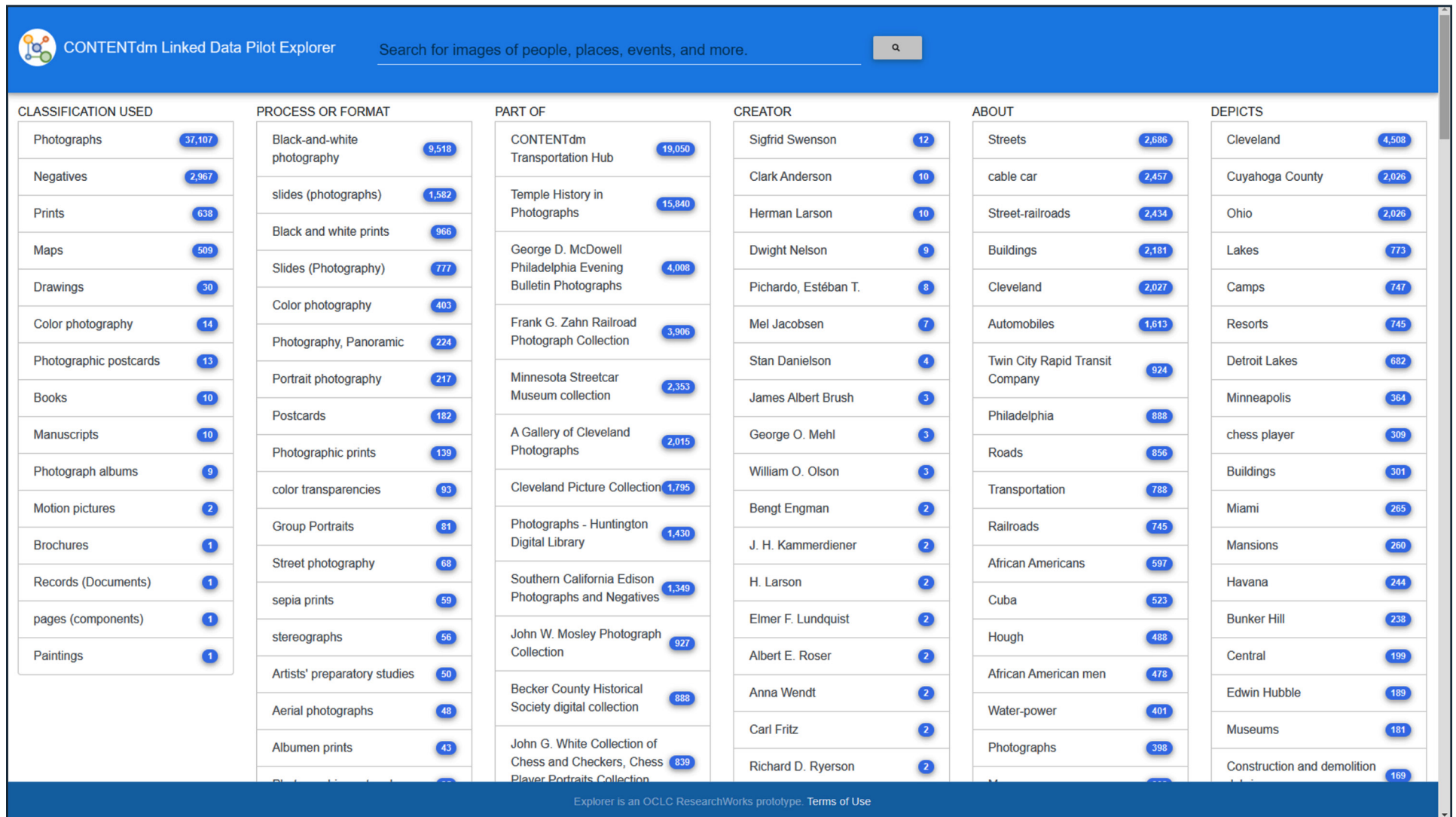


FIGURE 30. Explorer home page.⁷⁶ View a larger image [online](#).

OCLC established a new virtual collection entity in the Wikibase for a “CONTENTdm Transportation Hub” and associated all the new Wikibase items for the related works to this collection, along with their original source collection. In the Explorer, the Transportation Hub collection can be selected as the starting point for browsing and selection, with facets helping to narrow the scope to different topics, things depicted, source collections, and more (figure 31).

Explorer Transportation Hub and Related Collections

The screenshot shows the CONTENTdm Linked Data Pilot Explorer interface. At the top, there is a search bar with the text "Search for images of people, places, events, and more." and a search icon. Below the search bar, there is a navigation bar with "< PREVIOUS" and "NEXT >" buttons, and a search result range: "4061 to 4080 of 10,000 results for part_of:Q202314." On the left side, there is a list of facets with expandable arrows and item counts:

- 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 - Huntington Digital Library (967)
- Cleveland Picture Collection (650)
- University of Minnesota Duluth, Kathryn A. Martin Library, NEMHC Collections (306)
- Floyd and Marion Rinhart Photograph Collection (250)
- Jay T. Last Collection of Graphic Arts and Social History (236)

The main content area displays a grid of six image thumbnails, each with a title and a "READ MORE" link:

- East 6th Street 1930 CP06024
- Public Square 1896 CP04167
- Public Square 1915 CP04217
- Public Square 1905 CP04189
- Carnegie Avenue 1940 CPO5932
- Superior Avenue 1896, CP06853 Centennial Celebration

At the bottom of the interface, there is a footer that reads: "Explorer is an OCLC ResearchWorks prototype. Terms of Use"

FIGURE 31. Explorer Transportation Hub and related collections.⁷⁷ View a larger image [online](#).

The Transportation Hub can also be used to narrow a keyword search. For example, a keyword search for “strike” shown in figure 32 matches descriptions of items associated with labor strikes of various kinds (among other things) and narrowing the keyword search result to the Transportation Hub collection can highlight images and other works associated with transit strikes.

The Transportation Hub can also be used to narrow a keyword search.

Explorer Search Results for “Strike”

The screenshot shows the CONTENTdm Linked Data Pilot Explorer interface. At the top, there is a search bar with the text "Search for images of people, places, events, and more" and a search button. Below the search bar, the search results are displayed as a grid of six items. Each item consists of a thumbnail image, a title, and a "READ MORE" link. The items are:

- Truck brings employees home during P.T.C. walkout**: A black and white photograph showing a group of people standing around a truck.
- "Strike Pickets"**: A black and white photograph showing a line of people holding signs.
- "Trolley strike"**: A black and white photograph showing a street scene with many cars and a trolley.
- Crowded train station during the PTC strike**: A black and white photograph showing a large crowd of people at a train station.
- PTC employees photographed with strike signs**: A black and white photograph showing a group of people holding signs that say "STRIKE".
- Attempted strike break**: A black and white photograph showing a group of people standing in front of a trolley.

On the left side of the interface, there is a sidebar with various filters and categories, including "about", "classification used", "depicts", "part of", and "CONTENTdm Transportation Hub" (with 218 results). At the bottom of the interface, there is a footer that reads "Explorer is an OCLC ResearchWorks prototype. Terms of Use".

FIGURE 32. Explorer search results for “strike.”⁷⁸ View a larger image [online](#).

For a researcher interested in that topic, the Explorer can return very different perspectives on a particular transit strike; for example a Philadelphia Evening Bulletin newspaper photograph depicting the effect of the Philadelphia Transit Company strike of August 1944 on transportation options for workers (figure 33), contrasts with a John W. Mosley Photograph Collection image of a protest from the previous year in support of hiring African American trolley drivers (figure 34).

For a researcher interested in that topic, the Explorer can return very different perspectives on a particular transit strike.

Explorer View of a Truck Bringing Employees Home During a PTC Walkout

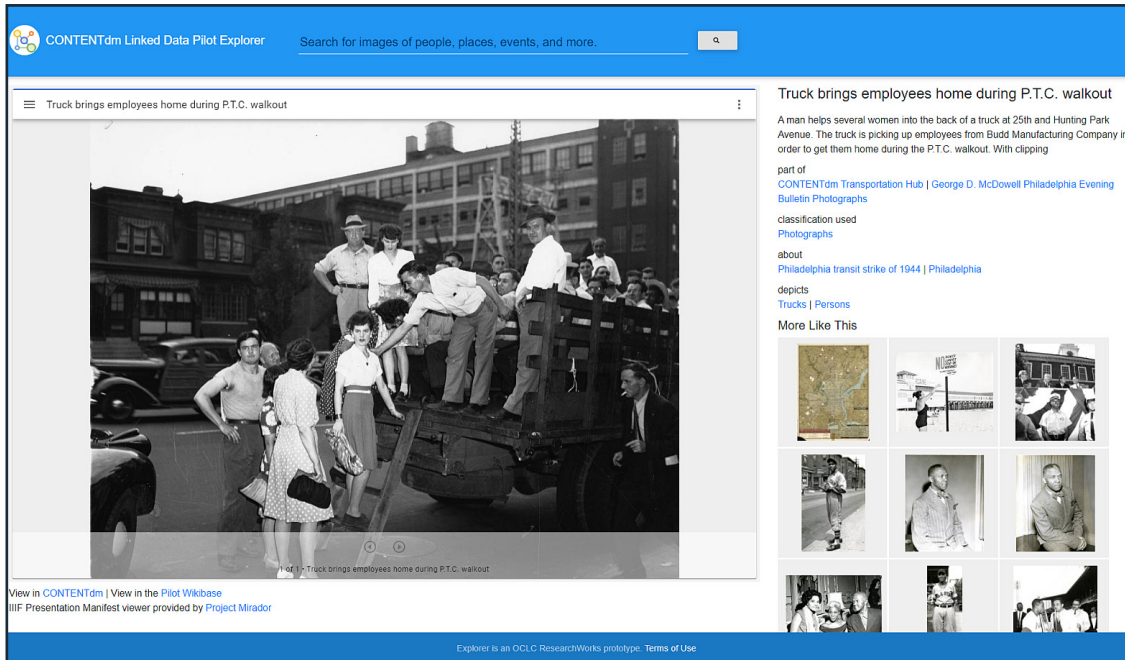


FIGURE 33. Explorer view of a truck bringing employees home during a PTC walkout.⁷⁹ View a larger image [online](#).

Explorer View of a Protest against the Philadelphia Transportation Company

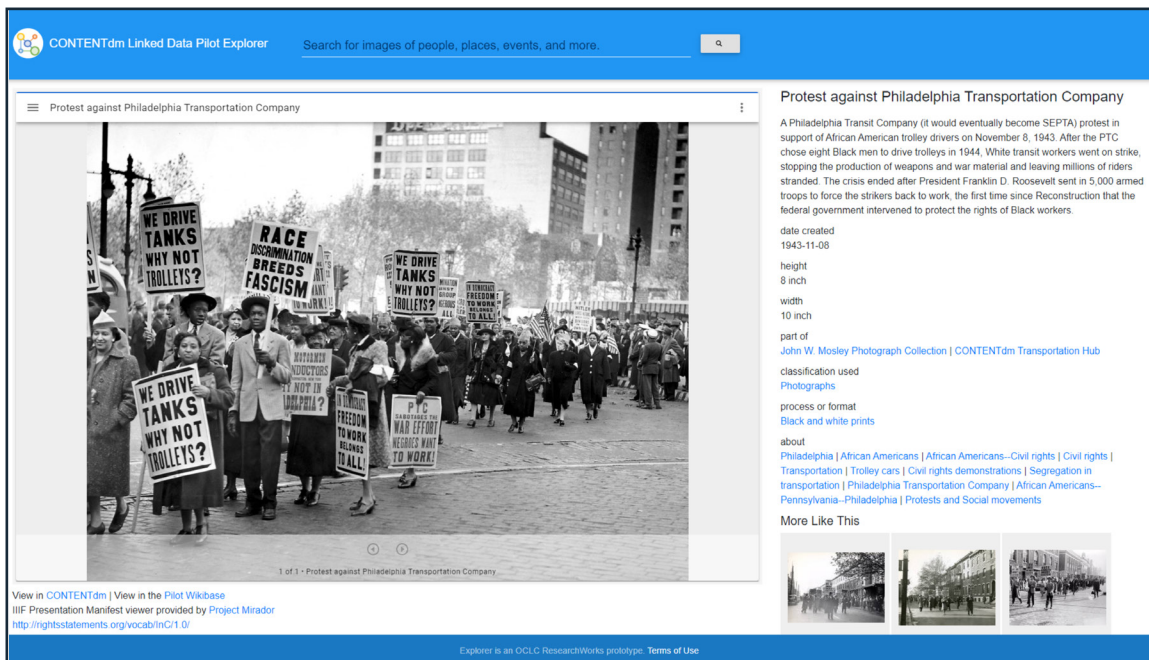


FIGURE 34. Explorer view of a protest against the Philadelphia Transportation Company.⁸⁰ View a larger image [online](#).

Explorer View of an 1899 Cleveland Transit Strike in Public Square

The screenshot displays the CONTENTdm Linked Data Pilot Explorer interface. At the top, there is a search bar with the text "Search for images of people, places, events, and more." Below the search bar, the main content area is divided into two sections. On the left, a large photograph shows a dense crowd of people gathered around a trolley car in a city square. The photograph is titled "Public Square 1899 CP04179". On the right, a metadata sidebar provides details about the image, including its date created (1899), height (6.125 inch), width (8.125 inch), and classification (Photographs). The sidebar also lists the source as "Cleveland Picture Collection | CONTENTdm Transportation Hub" and the location as "Cleveland".

Public Square 1899 CP04179

Showing the crowds surrounding the first cars to reach the downtown section during a street car strike. These cars were on the Euclid Line. The clock at the extreme right housed the store of Wm. Taylor, Son and Co.

date created
1899

height
6.125 inch

width
8.125 inch

part of
Cleveland Picture Collection | CONTENTdm Transportation Hub

classification used
Photographs

depicts
Cleveland

1 of 1 • Public Square 1899 CP04179

View in CONTENTdm | View in the Pilot Wikibase
IIIF Presentation Manifest viewer provided by Project Mirador
<http://rightsstatements.org/vocab/NoC-US/1.0/>

Explorer is an OCLC ResearchWorks prototype. [Terms of Use](#)

FIGURE 35. Explorer view of an 1899 Cleveland transit strike in Public Square.⁸¹ View a larger image [online](#).

The Transportation Hub helps to find images of transit strikes and their impacts in collections across institutions, including a Cleveland Public Library photograph of crowds surrounding a trolley car during a transit strike in during 1899 (figure 35) and a University of Miami photograph of parked trolley cars during a strike in Havana (figure 36).

The Transportation Hub helps to find images of transit strikes and their impacts in collections across institutions.

Explorer View of Streetcars Parked on the Street during a Transit Strike

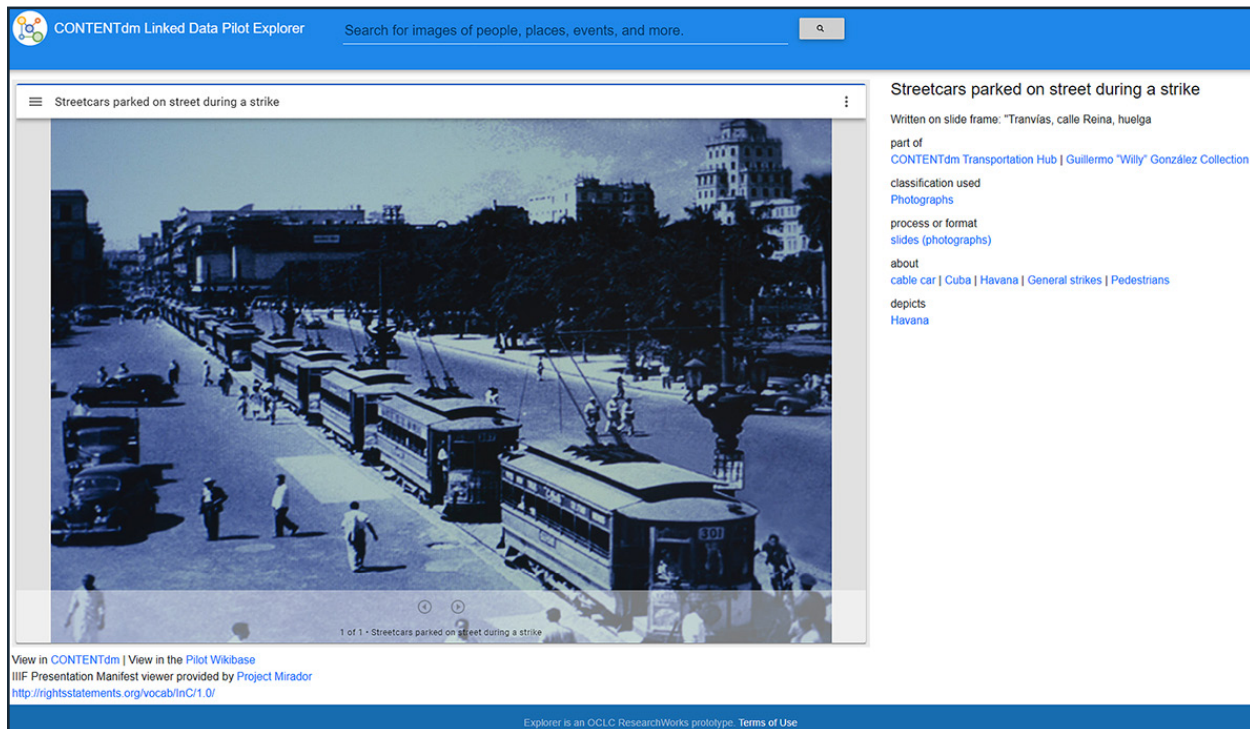


FIGURE 36. Explorer view of streetcars parked on the street during a transit strike.⁸² View a larger image [online](#).

The processes could be automated and extended to provide different views of how fields are defined and used across collections in a simple web application.

THE FIELD ANALYZER

Late in the pilot project, the OCLC developers saw a need for a new tool that could visualize how CONTENTdm fields are defined across different collections for participating institutions. This field-level analysis had been carried out in earlier phases of the project as a largely manual process, using CONTENTdm APIs and custom applications to gather data and reformat it for analysis in OpenRefine. After those manual processes had been ironed out, OCLC staff found that the processes could be automated and extended to provide different views of how fields are defined and used across collections in a simple web application.

Field Analyzer Field Usage Chart

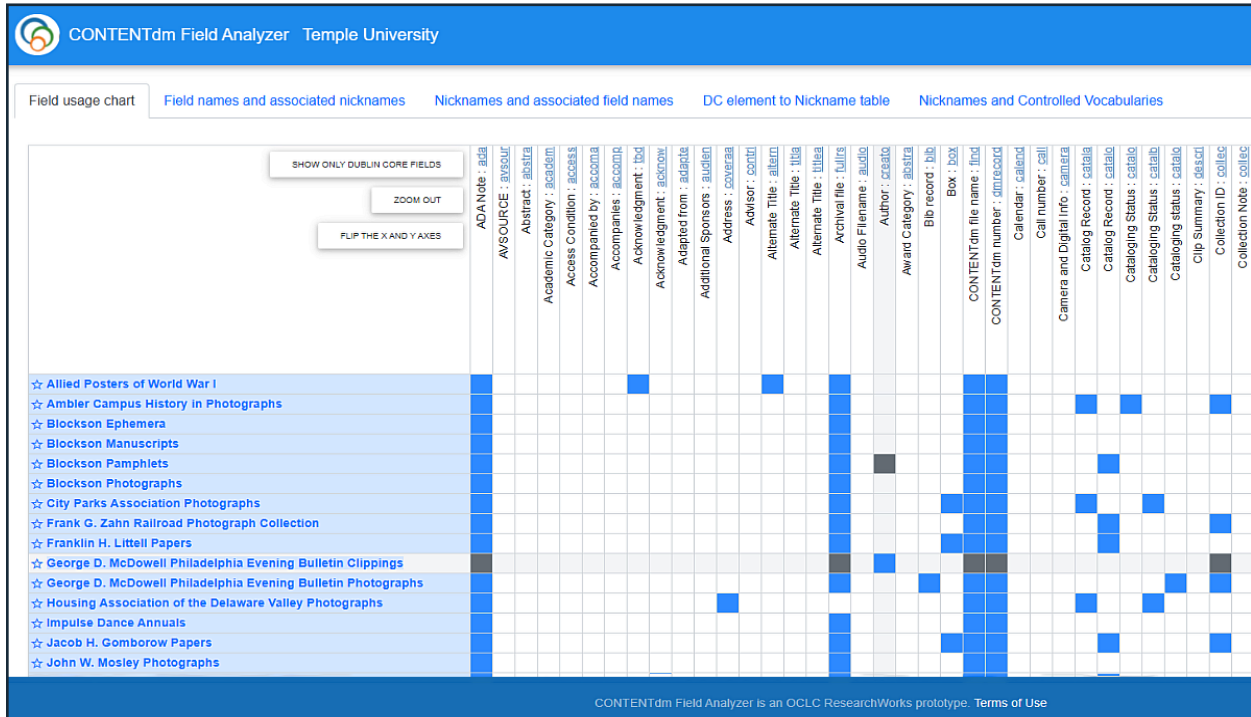


FIGURE 37. Field Analyzer field usage chart.⁸³ View a larger image [online](#).

Many participants found it to be a useful addition to their CONTENTdm toolkit, giving them a cross-collection view of how their collection vocabularies are defined.

During the pilot, the data that could be listed and visualized by the Field Analyzer was based on a “snapshot” of records copied from CONTENTdm and needed to be periodically refreshed to reflect any subsequent changes made. Pilot participants expressed interest in having the Field Analyzer maintained after the end of the pilot for ongoing use, with access to “live” or frequently synchronized data, and for all collections.

Field Analyzer List of Field Values

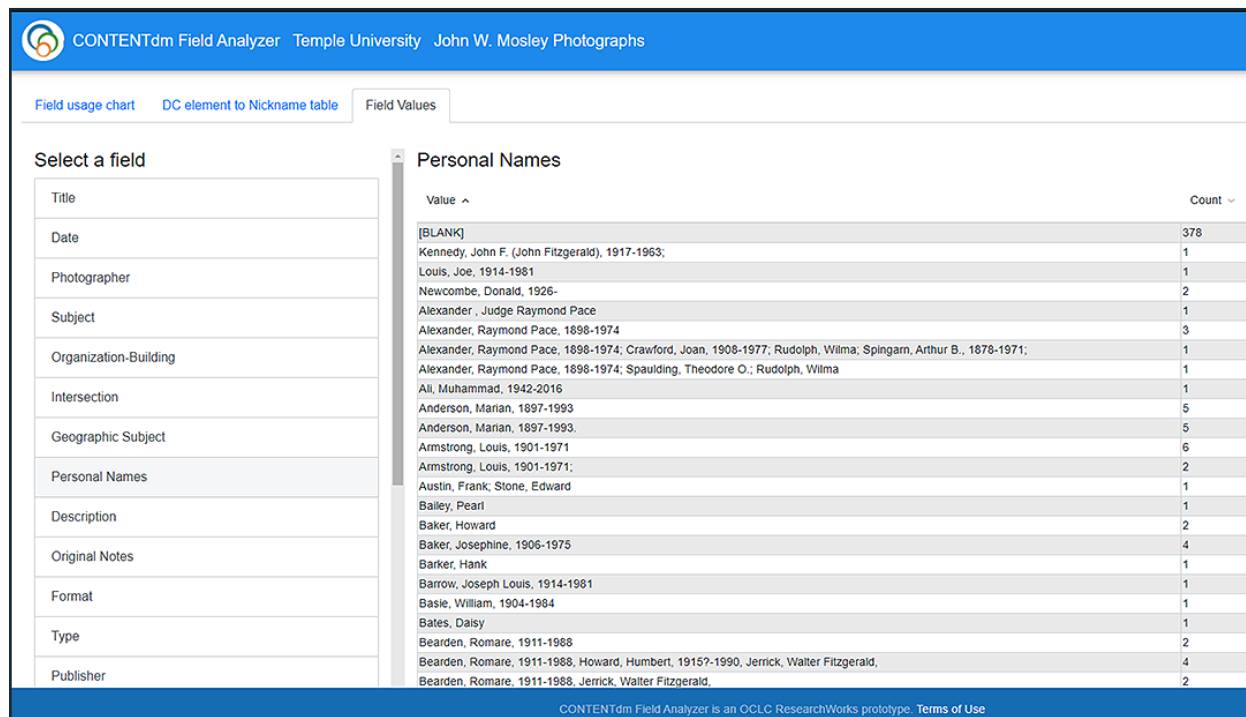


FIGURE 38. Field Analyzer list of field values.⁸⁴ View a larger image [online](#).

Cohort Communication

Communication is key to the success of any project, and was vital for effectively collaborating with the Linked Data project participants. In addition to using the CONTENTdm Community Center for addressing questions and tracking progress, the project participants and OCLC staff met every two weeks for an “office hour.” These sessions covered work in progress, planning for future stages, and demonstrations of new applications and processes. Apart from those regularly occurring topics, many sessions included a more open-ended group exploration of other questions, including:

- What local authority sources are used for reconciling headings?
- What user research practices have been applied to evaluate your systems?
- How are access, use, and reuse rights managed at your institution? Are these rights documented in CONTENTdm? Are there different rights assigned for physical vs. digital materials?
- How is CONTENTdm technical and administrative metadata managed?
- How and when should a “placeholder” entity description be created, for things that lack an established identity?
- What are current local practices for metadata cleanup, and do the work location changes made in response to the COVID-19 pandemic impact the priority of that work?
- Who is using the CONTENTdm Catcher utility, and who else might use it that isn’t yet?
- How could advancing racial equity in CONTENTdm descriptive metadata be facilitated?

The office hours served as a key point of communication and connection over the course of the project. Engaging in discussions about the challenges and day-to-day work of managing digital collection metadata and receiving real-time feedback about developed applications and tools provided OCLC staff with critical insights to inform and improve project outputs. Exploring the questions listed above as a group helped participants to share experiences and learn from one another.

The regular connection points that the sessions provided were especially valuable with the onset of the COVID-19 pandemic and ensuing facility closures. Amid many disruptions, the meetings included periodic check-ins for project participants to discuss and reflect on the effects of the pandemic on their work and their libraries in the near- to long-term; the reported impacts were varied and substantial.

Partner Reflections

At the end of the Linked Data project, the project partners provided their perspectives, representing both complementary and contrasting views of their experiences, the benefits returned, and implications for the future.

CLEVELAND PUBLIC LIBRARY (CHATHAM EWING)

Cleveland Public Library (CPL) has partnered with OCLC on metadata issues over the last several years, beginning with Project Passage in 2018-19 and then in 2019-20 with this Linked Data project. CPL believed the projects would have several potential benefits. The projects presented an opportunity to motivate our staff and institution to revisit, revise, and improve our metadata content and structure. It had the potential to lead to better and more accurate description and consequently improved discovery for our customers. The projects also provided motivation to rethink how we might enable more effective sharing through platforms such as DPLA or WorldCat.

Over the course of the projects, our digital library staff engaged with and learned from other partners and OCLC's team. OCLC's team helped us deeply consider how linked data could have an impact on our descriptive work. CPL staff presented on collections and processes using our scrapbooks project as an example, were recorded live for the purposes of a user interface usability study, submitted CPL collections for analysis through the field analyzer and got back a useful matrix that enabled analysis of our work, frequently conferred with OCLC team members as well as other partners, and more.

Our staff worked diligently to raise questions related to public library practice. Though the latter part of the Linked Data project happened during COVID-19, hampering our ability to explore some socially oriented goals with CPL digital library partners, staff were grateful for the intellectual lifeline the Linked Data project provided during the lockdown, and we eagerly anticipate working with project tools in the future to keep exploring some of the community-oriented possibilities for metadata brought up by the project. We believe that much of what we anticipated did happen, but additional insights emerged from the process. The experience and results strongly validate implementing more and more effective approaches to the use of linked data in digital library contexts in public libraries, and we strongly support the report's call for further investigation into using linked data. We also agree with the recognition that tools for reconciling data, particularly data such as name authorities and discipline specific thesauri, should be an integral part of any advance with regard to digital library tools within OCLC's suite of digital library applications.

But we feel there is another observation to make: digital collections described using linked data might be able to help explain what is “uniquely the same” about Cleveland as a place in the United States and the world. Representing what is locally unique yet making local information legible to outsiders and creating a mechanism where differences can be understood, bridged, and linked is an important part of what public libraries using newer descriptive systems can surface. Because we already involve diverse community members and community partners in the creation of digital items for our collections, it seemed natural to think about how we might include our partners in description, as well. During the project we looked at several examples of collecting information about and from CPL digital collections. We looked at our digital collection of scrapbooks, local newspapers, local theaters, and the library archives, and we tried to think about places where we had drawn description from the language of the communities we were working with rather than from internationally scoped name authority lists or cataloging thesauri. And that was promising, we thought.

Representing what is locally unique yet making local information legible to outsiders and creating a mechanism where differences can be understood, bridged, and linked **is an important part of what public libraries using newer descriptive systems can surface.**

But it was when the project took a turn and collections like these were juxtaposed against one another that things became interesting. The “Transportation Hub” was a useful example of how this concept was explored by the project. Each institution’s collecting around transportation was pulled together into a gathered collection, and the project implemented a platform that offered a glimpse of how to explore the role the digitized items played in each of the communities documented by the separate collections.

The project team at OCLC discussed the challenges of normalizing and reconciling the data in the collections for the Transportation Hub, and the process highlighted typical challenges in reconciling data and enabling searching across multiple institutional collections. And, as we mentioned before, the process also highlighted the labor-intensive nature of such work and spotlighted a long-standing need for more robust tools within the OCLC suite of applications for managing controlled vocabularies across collections in the context of digital library tools.

However, the OCLC staff’s discussion of the process also raised the question of what to do about significant local variances in uncontrolled description language for digitized items. Perhaps we should also uncover and share different communities’ understandings of more generalized concepts? It would seem that a linked data system holds up the promise of capturing some locally generated data that reflect local variances while also offering traditional authoritative descriptive data. We feel that a linked data system that includes some broader, more locally oriented mechanism for participation in description would be a powerful tool for our work doing digitization in our community. For us at CPL, we began to consider how we might describe collections that not

only made use of authorized names, subjects, and thesaurus terms to describe Cleveland's unique and local digital collections, but which also described our city and region's uniqueness and made it legible through networks of links to alternative information including local lingo, alternate lists of names, and diverse uses of descriptive language found in lists and resources that might supplement more standard descriptions.

A natural extension of this kind of thinking is that (for public libraries at least) tools need to be easy to use not only to professional catalogers, but also to community experts as well. A wiki offers a simple, public-facing user interface, but other interfaces might also be designed to be more inclusive and accessible, allowing digital projects to easily incorporate local, grounded expertise that librarians cannot often be expected to have. Linked data systems can facilitate that inclusivity by creating and making connections between related or synonymous local terminology and concepts. Perhaps using linked data for description even has the potential to decenter hierarchies and master narratives about cultural heritage that may be implicit in approved authoritative descriptive practice, allowing for alternative hierarchies and assumptions to surface and enrich descriptive practice.

And the local/global break with regard to epistemological understanding revealed through the Transportation Hub implies other breaks that could be drawn out from other collection gatherings that might also contribute to rich, differentiated hierarchies of description that will enable more diverse access through richer and more inclusive community generated description. Perhaps design a system that is usable by expert catalogers (because solid hierarchies are a backbone of effective access), comprehensible by local metadata experts (because local historians have awesome expertise), and is also open enough to capture (and sift) the kinds of description generated at the level of the general user. This might not only lead to higher quality and more comprehensive metadata, but also, if handled well, can create opportunities for deep community listening. We could generate access points to our information based upon empirical observation of how our communities create links within our information. This kind of engagement might enable libraries to engage patrons and learners, using digitization as a process for delving into what really makes their communities wonderful and unique.

THE HUNTINGTON LIBRARY, ART MUSEUM, AND BOTANICAL GARDENS (MARIO EINAUDI)

The invitation to join this project in August 2019 came as the Huntington Library was reviewing the digital collections accessible in the Huntington Digital Library (HDL), which had been launched in 2011. In 2018 it was determined that an overhaul and a full review of the metadata and structure of the 23 collections was needed. We had hopes that the Linked Data project might aid us in this endeavor. Following the initial ingest of materials selected from three of our collections, the review and initial cleanup work, along with the testing done by Bruce Washburn to feed our metadata into the Wikibase, it was quickly apparent that this pilot would not be able to help that cleanup directly. Rather the Linked Data project provided us the context to better understand our workflows, our metadata, and how we structured that metadata.

Importantly, this project did demonstrate the incredible value of linked data as a way of creating and maintaining metadata. Linked data in the Wikibase enabled the creation of a web of connections and context that is lacking in many other systems. A good example of the power of the tools developed using linked data was the Image Annotator. This tool allows the user to highlight a section of an image and then apply one of the known entities to that highlighted section. This creates links between that image and other images in the collection that would not exist—unless the cataloger remembered that x also appeared in y and z. It provided a tantalizing look at a new tool for cataloging materials.

It would have been good to test some of these tools outside of the pilot. The Image Annotator if reconfigured for use with a future CONTENTdm would be a great improvement. Subject specialists could be brought on board a project and asked to identify people or places, with the linked data providing added indexing in a controlled environment in the background. Also, the Explorer tool would enhance discovery across collections, both internally to the library, and if part of a larger linked data universe, to other libraries, large or small.

While the project was focused on the benefits of, and how to create, linked data, one tool grew out of the need to analyze the extant data in the participants systems. And that tool, 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. It has allowed us to pursue our goal of descriptive uniformity across all CONTENTdm collections. A companion tool that would replace, or build on, the Catcher interface, allowing for the cleaned-up metadata to be pushed back into our CONTENTdm site would also have been a real boon. But the complexities faced in cleaning up the data, along with the entity-based structure within Wikibase, foreclosed that option.

Throughout this Linked Data Pilot project OCLC Staff were incredible, providing guidance, soliciting input, posing questions, and seeking solutions that engaged all the participants. The tools developed and the cleanup done by Bruce Washburn and Jeff Mixter show all the power and promise of linked data, as well as some of the hurdles. Yet, this is a path that should be followed, especially as CONTENTdm shows its age. The leap forward to a new solution has been greatly helped by the solid work done by all on this project. 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.

MINNESOTA DIGITAL LIBRARY (GRETA BAHNEMANN AND JASON ROY)

Invitation

In July of 2019, the Minnesota Digital Library (MDL) was asked to join the CONTENTdm Linked Data Pilot project. Initially, we were one of three pilot partners. This invitation was an opportunity for us to see the practical application of Wikidata to MDL's collection of images. MDL is a collection of digitized cultural heritage materials comprised of images, text-based, cartographic materials, etc. with 67% of our collection represented by images. Given our high percentage of images, we were especially interested to see how our image metadata would reconcile and work with Wikidata. Would MDL's metadata withstand this kind of work?

Development of three tools by OCLC

During the Linked Data project, OCLC developed three tools to assist the project participants:

1. **Retriever**—designed to help pilot partners search for and create entity descriptions. Especially helpful for those new to the process
2. **Image Annotator**—subject analysis tool that has the potential to change how we describe cultural heritage materials
3. **Field Analyzer**—developed in response to the need of the pilot project participants but has usefulness beyond the pilot. This tool provides partners with a backend look at their data, and gives a comprehensive view of how data is mapped, field names used, etc. It quickly shows the inconsistencies in a collection's data regarding field names, mapping, etc.

4.

The Image Annotator has the most potential to change the user's understanding of digital content. With its capacity to provide both a layer of subject analysis and descriptive details to images in CONTENTdm, it is no less than groundbreaking. For example, an albumen photograph of a late 19th century home in Minneapolis can be "about" the concept/subjects of "Richardsonian Romanesque Style Architecture" and/or "Rock-face Construction;" but it can also "depict" things found in the image, such as a horse-drawn wagon, a fire hydrant, pedestrians, named individuals, etc. This added layer of meaning and contextualization can only add to the user's understanding of the image. This is a type of analysis traditionally associated with the fields of fine art, architecture, urban planning and has the potential to add more nuanced description to cultural heritage materials and change how users understand these materials. While this tool is valuable and has huge potential for changing how we describe cultural heritage materials, it can be a labor-intensive process that may not be sustainable on a large scale.

The Image Annotator has the most potential to change the user's understanding of digital content. With its capacity to provide both a layer of subject analysis and descriptive details to images in CONTENTdm, it is no less than groundbreaking.

Leveraging the power of linked data

In terms of linked data support, a lot of initial effort was spent discussing how these controlled vocabularies might best be ingested and stored within the CONTENTdm framework. The rationale, one would believe, behind this was to ensure that it would better integrate with our more hyperlocal vocabularies and taxonomies. That is, how best to blend national vocabularies alongside locally created terms to best describe the source material. Unfortunately, by bringing and storing this "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.

What may be a path forward is an opportunity for CONTENTdm to create web services that call upon these linked data sources at the point of need. This would allow catalogers and metadata creators the opportunity to align their local descriptive practices more closely with national and international initiatives. CONTENTdm would store the URI, not the term itself, thus creating linkages that would allow for more accurate and consistent sharing without "hardwiring" terms into the CONTENTdm data store.

It is, ultimately, the data store itself that is the most valuable piece of information. From this building block we can construct user interfaces and applications, share out our metadata for others to package, and scale out across multiple, shared repositories. We consume this data to create our local, default CONTENTdm view, but this same data can be packaged and shared in new ways. Applications such as that created by the Minnesota Digital Library⁸⁵ consume the same data but build it out in different ways; additionally, this same data is openly shared with and aggregated by the Digital Public Library of America⁸⁶ for use in their national initiative. Same data, different views. Ultimately, it is the data that must remain interoperable enough to work across systems and alongside other data sources. Within the limited timeframe of our project, OCLC was able to provide a proof of concept of the potential for enhancing CONTENTdm metadata through linked data integrations by way of a single new view that builds upon the existing CONTENTdm user facing discovery layer.

Ultimately, it is the data that must remain interoperable enough to work across systems and alongside other data sources.

Concluding thoughts

We believe that this work should result in the further decoupling of some of these tight integrations in order to achieve our desired results: separating out the data store from the data view layer; leveraging the URI for further linkages out toward reliable and trustworthy linked data sources within the data store itself; and allowing for the open sharing of our data (and our assets as well through the existing IIF infrastructure) with others to achieve large scales of discovery and to better network our data alongside that of our colleagues.

Included in all of this should be a discussion of the future application of the tools OCLC developed for this project. The Image Annotator and Field Analyzer could be integrated into the CONTENTdm package/workflow to help CONTENTdm users (both administrators as well as crowd-sourced end users) provide a more robust, nuanced description via the Image Annotator. The Field Analyzer can also help CONTENTdm adopters see their data, across multiple collections, in a single interface. Both tools should be developed further, thereby making CONTENTdm more user-friendly—for both administrators and end users.

The Minnesota Digital Library was excited to be a part of this pilot project. In addition to learning more about the practical application of Wikidata, it was a great opportunity to get to know staff at OCLC and speak about the potential future of CONTENTdm in a collaborative environment.

TEMPLE UNIVERSITY LIBRARIES (HOLLY TOMREN AND MICHAEL CARROLL)

In 2019 we joined the CONTENTdm Linked Data project. Focusing on how this compared to and differed from our previous experience with Project Passage, while Project Passage was more about one-by-one original description, the CONTENTdm pilot was more about batch transformation of existing metadata.

OCLC staff consulted with us about how they planned to map our metadata and to answer any of our questions, and we provided feedback about the mappings as well as any questions we had

about the data model, which was now much expanded from what we had started with in Project Passage. This gave us a sense of what a future data migration would look like, and how migration to a linked data model can be even more complex than a migration from one flat metadata model to another.

Linked data also provides different opportunities for how we can search our CONTENTdm metadata, particularly through more indirect relationships between entities in the system.

After OCLC transformed our data, we evaluated it to see how this could help us look at our metadata differently, where is there room for further data enrichment, and what are the new relationships and connections we can create with a system that is built to do so.

One thing that particularly stood out were the different ways we could browse our data using the Explorer tool that OCLC developed. At Temple, our customized library discovery layer is built on three concepts: Search, Browse, and Recommend. But so far, we have only implemented Search. As we've thought internally about Browse features, we've struggled with a way to approach this that is different from the standard Title, Author, Subject browse from the past. The CONTENTdm Explorer offers a model that provides a variety of different starting points for browsing and then allows a user to traverse a graph of relationships, which is inspiring as we continue to develop our local discovery environment.

Linked data also provides different opportunities for how we can search our CONTENTdm metadata, particularly through more indirect relationships between entities in the system. For example, we were thinking of the use case where we might have an "On This Day" feature to post on social media. We were able to develop queries in the SPARQL endpoint that could help us find images that depict people born on a certain day or images that depict people born in Philadelphia that could be used to help us select featured images for different scenarios.

Participating in the project introduced the Wikibase interface and exciting tools to enhance the discovery of and engagement with digital records. The Wikibase offered a glimpse into what a digital collections database that employs linked data might look like and how the cataloging process might change. For instance, the inclusion of clickable headings for each entity has the potential to make it even easier for student catalogers to understand the context of the terms they use to describe an image.

The Describer prototype tool was a simplified visual interface that enables cataloging based on the resource type classification. This tool felt more approachable than the Wikibase interface. The text box of the Describer tool automatically suggested verified terms like controlled vocabularies in CONTENTdm, but this tool felt more intuitive and tailored to what the user was typing.

It was very useful from a cataloging perspective to have access—supported by IIF standards and viewers—to the image and be able to zoom in to see details while describing it. We also thought the Image Annotator had a lot of potential for being able to associate a part of an image with a specific depicts or subject property, and it would be interesting to see how that could be incorporated into the end user discovery experience.

One of the potential impacts of this project would be to rethink our cataloging workflows in accordance with a linked data structure. The Temple University team described existing images as a group exercise that proved challenging without the original objects in front of us. It became clear during this exercise that there would also need to generate more nuanced descriptions when cataloging in order to develop a richer network of relationships between entities.

The Linked Data project demonstrated the amount of work involved in the transition to linked data, but also that the tools exist and that the workflows can be developed.

UNIVERSITY OF MIAMI LIBRARIES (PAUL CLOUGH AND ELLIOT WILLIAMS)

Participating in the Linked Data project was an opportunity for us to understand more concretely what it would take to transform our existing collections into linked data and what a linked data version of CONTENTdm might look like. Interacting with our metadata in the Wikibase environment raised valuable questions about how our existing metadata practices might complicate the transition to linked data, such as a lack of standardization of elements and inconsistent uses of existing vocabularies, and inspired us to focus more on data normalization and consistency. Some of the insights and tools that came out of the project, such as the Field Analyzer, will be immediately useful for our work in CONTENTdm, even outside of the transition to linked data. Participating in a cohort with other CONTENTdm users and OCLC staff was also a great opportunity to learn from and with our peers. The Linked Data project demonstrated the amount of work involved in the transition to linked data, but also that the tools exist and that the workflows can be developed. While we appreciate the promise of linked data, we believe that more work still needs to be done to show that the effort will be worth it.

KEY FINDINGS AND CONCLUSIONS

The linked data project reaffirmed some prior lessons learned and provided new insights across a range of concerns, including the expected benefits of working in a linked data environment, the potential to develop a shared data model, a reality check on the effort to transform metadata to linked data, and the essential benefits of a strong partnership.

TESTING THE LINKED DATA VALUE PROPOSITION

The project confirmed key aspects of the linked data value proposition, that cultural material discovery and data management can be significantly improved when the materials are described using a shared and extensible data model, when metadata string-based headings are transformed to linked data entities and relationships, and when those entities and relationships are brought together into a single discovery system. In this environment, the technology works in service to both the staff, who can more easily and accurately impart the expertise they have about the collections they steward, and to the researcher, who can see more robust connections between—and context about—the cultural materials that make up CONTENTdm collections.

In project prototype applications, entities can be retrieved by searches that use a persistent identifier rather than a string heading. This capability provides integrated authority control for the entities and greatly improves the precision and recall performance metrics for discovery.

As CONTENTdm string headings are reconciled and converted to entities, additional information from external data sources can automatically and efficiently enrich the entity description. This supports new discovery and data visualization capacities that would be expensive or impossible to achieve in the current CONTENTdm system. For example, place entity descriptions can be enriched with geographic coordinates, which can then be used to generate map-based visualizations of places depicted in cultural materials.

In an entity-oriented system like Wikibase, different types of entities have their own distinct representation. This design contrasts with record-oriented systems where the creative work is the primary entity and other types of things are only present as statements representing notes and headings that are associated with the work. Data management and maintenance efficiencies are gained by transforming these statements into entities. For example, a biographical statement about a person can be associated with that person's entity description, rather than repeated as a note in every record that is in some way about that person.

EVALUATING A SHARED DATA MODEL

Building an initial data model with a high-level structure informed by other standards, including Dublin Core and Schema.org, provided a solid set of initial classes and properties. The model could be effectively and responsively expanded based on new entities and relationships represented in the source metadata. The metadata and mapping discussions with pilot partners helped OCLC develop the data model, as data was encountered in the CONTENTdm sources that OCLC had not anticipated.

SELECTING AND TRANSFORMING METADATA

Data transformation tools should be shared and the workflows decentralized. This will be essential to making the conversion scalable, as the workload is too great for a central agency to carry out. Domain expertise is needed to determine how locally defined fields are used at the institution level and sometimes at the collection level.

Though it required considerable manual effort, most headings for concepts and places found in CONTENTdm source metadata could be reconciled to matching entities described in other sources, including the Wikidata knowledge base, the VIAF authority file, FAST, and GeoNames.

Not surprisingly given the relative lack of notability of some of the represented people and organizations, those headings often could not be found in one of the external sources OCLC used for reconciliation and led to manual data entry for a “placeholder” entity.

Other than the initial field mapping review, pilot participants did not get a more in-depth “behind-the-scenes” view of the data processing workflows, which could have been offered as “office hour” homework or a workshop. In retrospect that appears to be a missed opportunity.

For the transition to linked data to be comprehensive and complete, a set of new CONTENTdm tools are called for that can be applied to transformation and reconciliation workflows in a decentralized way, along with fundamental changes to the centralized CONTENTdm system. A paradigm shift of this scale will necessarily take time to carry out and calls for long-term strategies and planning.

CONTINUING THE JOURNEY TO LINKED DATA

Substantial resource commitments will be required to carry out these data transformations across all CONTENTdm institutions and collections, but the community does not need to wait for the transformation to linked data to be fully completed before they can see benefits. Data management and discovery benefits are applicable from this work in the current CONTENTdm environment, and downstream linked data transformation efficiencies accrue as metadata makes greater use of shared vocabularies and persistent identifiers. For the transition to linked data to be comprehensive and complete, a set of new CONTENTdm tools are called for that can be applied to transformation and reconciliation workflows in a decentralized way, along with fundamental changes to the centralized CONTENTdm system. A paradigm shift of this scale will necessarily take time to carry out and calls for long-term strategies and planning.

Several of the prototype applications developed during the pilot point the way to advantageous additions to the CONTENTdm toolkit. In particular, the Image Annotator encourages domain experts to enrich material descriptions, and the Field Analyzer helps CONTENTdm users make sense of the variations in field definitions and uses across their collections (a prerequisite for more holistic data rationalization and transformation). The project participants encouraged OCLC to pursue these and other improvements as part of CONTENTdm's evolution into a linked data platform.

WORKING PARTNERSHIPS REPRESENT STRENGTH IN NUMBERS

The value of library participants as partners in this project cannot be overstated. As colleagues and thought partners in the work, participants connected with project staff in regularly scheduled office hours throughout the project. Through these meetings and regular communications, project participants shared their thoughts on topics ranging from philosophical approaches and concepts to technical details and provided ongoing feedback that steered the project work toward tools and applications of greatest practical value for library staff and researchers. Recognizing the critical insights contributed by the project partners confirms the importance of involving library staff in this manner for similar technical research projects.

NOTES

1. OCLC's CONTENTdm digital content management service overview:
<https://www.oclc.org/en/contentdm.html>.
2. An overview of OCLC's history of Linked Data research projects:
<https://www.oclc.org/research/areas/data-science/linkedata/linked-data-outputs.html>.
3. W3C. "Linked Data." <https://www.w3.org/wiki/LinkedData>.
4. An overview of OCLC's linked data pilot Project Passage:
<https://www.oclc.org/research/areas/data-science/linkedata/linked-data-prototype.html>;

See also: Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Holly Tomren, and Craig Thomas. 2019. *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage*. Dublin, OH: OCLC Research. <https://doi.org/10.25333/faq3-ax08>.

5. The Wikibase environment includes several components:

The MediaWiki Platform:

<https://www.mediawiki.org/wiki/MediaWiki>;

MediaWiki. "Wikibase:Overview—MediaWiki extension for managing structured data. Updated 29 December 2020, at 19:51. <https://www.mediawiki.org/wiki/Wikibase>;

Wikipedia. "Triplestore: [...] a purpose-built database for the storage and retrieval of triples through semantic queries." Updated 12 November 2020, at 18:12 (UTC).

<https://en.wikipedia.org/wiki/Triplestore>;

Wikipedia. "SPARQL" (Query service for reading data from the triplestore). Updated 3 January 2021, at 14:42 (UTC). <https://en.wikipedia.org/wiki/SPARQL>.

6. CONTENTdm Linked Data Planned Project Phases diagram:
<https://researchworks.oclc.org/cdmlD/screenshots/phase-diagram.png>.
7. OCLC. 2020. "Guide to the CONTENTdm Catcher." Updated 7 August 2020.
https://help.oclc.org/Metadata_Services/CONTENTdm/CONTENTdm_Catcher/Guide_to_the_CONTENTdm_Catcher.
8. Wikipedia: <https://www.wikipedia.org/>.
9. Wikidata: The free knowledge base. Updated 30 December 2019, at 04:00.
https://www.wikidata.org/wiki/Wikidata:Main_Page.
10. SPARQL [Linked Data] query language for RDF. W3C Recommendation 15 January 2008.
<https://www.w3.org/TR/rdf-sparql-query/>.

11. Wikibase system architecture diagram: <https://researchworks.oclc.org/cdmld/screenshots/wikibase-system-architecture.png>.
12. CONTENTdm Class data model visualization: <https://researchworks.oclc.org/cdmld/screenshots/class-ontology.png>.
13. Background on the Dublin Core Metadata Initiative and the Dublin Core element set: <https://dublincore.org>.
14. The Dublin Core Metadata Initiative DCMI Type Vocabulary: <https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/>.
15. Linked Art project: <https://linked.art/>.
16. Example use of type, classification, and process or format properties in the description of a postcard: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q73226.png>.
17. A depicts statement for the concept of “Dogs”: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q147731.png>.
18. A type statement of “dog” for a specific dog: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q142481.png>.
19. The RDF Linked Data modeling vocabulary “RDF Schema”: <https://www.w3.org/TR/rdf-schema/>.
20. The class “dog” is defined by the concept “Dogs”: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q73829.png>.
21. Wikibase templates for proposing new properties: <https://researchworks.oclc.org/cdmld/screenshots/cdm-property-proposal.png>; <https://researchworks.oclc.org/cdmld/screenshots/cdm-property-proposal-is-defined-by.png>.
22. Unmapped CONTENTdm metadata displayed in the Wikibase user interface with a Gadget extension: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q143578.png>.
23. Collections evaluated for the pilot project:
 - Cleveland Public Library
 - Cleveland Picture Collection: <https://cplorg.contentdm.oclc.org/digital/collection/p4014coll18/search/searchterm/cleveland%20picture%20collection/field/collect/mode/exact/conn/and/order/sortda/ad/asc>;
 - Jasper Wood photos of Cleveland: <https://cdm16014.contentdm.oclc.org/digital/collection/p4014coll18/search/searchterm/jasper+wood/field/creato/mode/all/conn/and>;
 - John G. White Collection of Chess and Checkers, Chess Player Portraits Collection: [https://cdm16014.contentdm.oclc.org/digital/collection/p4014coll20/search/searchterm/Chess Portraits](https://cdm16014.contentdm.oclc.org/digital/collection/p4014coll20/search/searchterm/Chess%20Portraits).

- The Huntington Library, Art Museum, and Botanical Gardens:
 - Edwin Hubble Papers: <https://cdm16003.contentdm.oclc.org/digital/collection/p15150coll2/search/searchterm/Edwin%20Hubble%20Papers/field/physic/mode/exact/conn/and;>
 - Palmer Conner Collection of Color Slides of Los Angeles, 1950 - 1970: <https://hdl.huntington.org/digital/collection/p15150coll2/search/searchterm/Palmer+Conner+Collection+of+Color+Slides+of+Los+Angeles%2C+1950+-+1970/field/physic/mode/all/conn/and/order/nosort;>
 - Photographs of the California Missions by William Henry Jackson <https://hdl.huntington.org/digital/collection/p15150coll2/search/searchterm/Photographs%20of%20the%20California%20Missions%20by%20William%20Henry%20Jackson/field/physic/mode/exact/conn/and;>
 - Verner Collection of Panoramic Negatives [https://hdl.huntington.org/digital/collection/p15150coll2/search/searchterm/Verner+Collection+of+Panoramic+Negatives/field/physic/mode/all/conn/and/order/title.](https://hdl.huntington.org/digital/collection/p15150coll2/search/searchterm/Verner+Collection+of+Panoramic+Negatives/field/physic/mode/all/conn/and/order/title)
- Minnesota Digital Library:
 - American Swedish Institute: https://reflections.mndigital.org/?f%5Bcollection_name_ssi%5D%5B%5D=American+Swedish+Institute;
 - Becker County Historical Society: https://reflections.mndigital.org/?f%5Bcollection_name_ssi%5D%5B%5D=Becker+County+Historical+Society;
 - Kanabec County Historical Society: https://reflections.mndigital.org/?f%5Bcollection_name_ssi%5D%5B%5D=Kanabec+County+Historical+Society.
- Temple University:
 - John W. Mosley Photograph Collection: <https://digital.library.temple.edu/digital/collection/p15037coll17;>
 - Temple History in Photographs. Templana Event Album Collection: <https://cdm16002.contentdm.oclc.org/digital/collection/p245801coll0/search/searchterm/Templana%20Event%20Album%20Collection/field/reposa/mode/exact/conn/and;>
 - Temple History in Photographs. Templana Photograph Collection: <https://cdm16002.contentdm.oclc.org/digital/collection/p245801coll0/search/searchterm/Templana%20Photograph%20Collection/field/reposa/mode/exact/conn/and;>
 - Temple History in Photographs. Temple Times Photographs: <https://cdm16002.contentdm.oclc.org/digital/collection/p245801coll0/search/searchterm/Temple%20Times%20Photographs/field/reposa/mode/exact/conn/and;>
 - Temple University Libraries. YWCA Philadelphia Branches Records: <https://digital.library.temple.edu/digital/search/collection/p16002coll6!p15037coll19!p15037coll14!p16002coll2/searchterm/YWCA%20Philadelphia%20Branches%20Records/field/digitb/mode/exact/conn/and.>
- University of Miami:
 - Cuban Map Collection: [https://merrick.library.miami.edu/cubanHeritage/chc0468/;](https://merrick.library.miami.edu/cubanHeritage/chc0468/)

- Latin American and Caribbean Photograph Collection:
<https://merrick.library.miami.edu/cdm/search/collection/asm0304>;
 - Rosenstiel School of Marine & Atmospheric Science Photograph Collection:
<https://merrick.library.miami.edu/rsmas/rsmasphotos/>.
24. Wikibase Discussion page for a collection review:
<https://researchworks.oclc.org/cdmld/screenshots/cdm-item-talk-Q148309.png>.
 25. The OpenRefine software for cleaning up, analyzing, and reconciling metadata:
<https://openrefine.org/>.
 26. CONTENTdm collection metadata viewed in OpenRefine:
<https://researchworks.oclc.org/cdmld/screenshots/openrefine-project.png>.
 27. IIF International Image Interoperability Framework website: <https://iif.io/>.
 28. A triplestore is a database to manage linked data “triples”, which are a combination of a subject, predicate, and object: <https://en.wikipedia.org/wiki/Triplestore>.
 29. Wikidata OpenRefine reconciliation endpoint software. See Delpuch, Antonin. (2017) 2020. “Wetneb/Openrefine-Wikibase.” Python. <https://github.com/wetneb/openrefine-wikibase>.
 30. OCLC’s FAST (Faceted Application of Subject Terminology) system:
<https://www.oclc.org/research/areas/data-science/fast.html>.
 31. VIAF OpenRefine reconciliation endpoint service:
http://iphylo.org/~rpage/phyloinformatics/services/reconciliation_viaf.php.
 32. The GeoNames service for geographic data: <https://www.geonames.org/>.
 33. The Python scripting language. See “Manual:Pywikibot/Overview - MediaWiki.” n.d. Accessed 7 January 2021. <https://www.mediawiki.org/wiki/Manual:Pywikibot/Overview>.
<https://www.python.org/>.
 34. “Help:QuickStatements - Wikidata.” Edited on 4 January 2021, at 10:41.
<https://www.wikidata.org/wiki/Help:QuickStatements>.
 35. Pywikibot Python library overview. See “Manual:Pywikibot/Overview - MediaWiki.” n.d. Accessed 7 January 2021. <https://www.mediawiki.org/wiki/Manual:Pywikibot/Overview>.
<https://www.mediawiki.org/wiki/Manual:Pywikibot/Overview>.
 36. OCLC DevConnect Online 2020 presentation on the alternative OpenRefine reconciliation endpoint software developed during the pilot project. See Mixer, Jeff, and Bruce Washburn. 2020. “Building an OpenRefine Reconciliation Endpoint for a Wikibase project: Lessons Learned.” Produced by OCLC, 20 May 2020. MP4 video presentation, 58:01.
<https://www.oclc.org/en/events/2020/devconnect-online-2020/devconnect-2020-creating-linked-descriptive-data-for-contentdm.html>.
 37. A “placeholder” entity for a person without an established identity:
<https://researchworks.oclc.org/cdmld/screenshots/entity-Q144548.png>.

38. An example CONTENTdm compound object for a photograph album. See University of Miami Libraries. "Album Documenting a Sea Journey to Trinidad, Venezuela, and Grenada." Latin American and Caribbean Photograph Collection. Digital Collections. Accessed 7 January 2021, <https://cdm17191.contentdm.oclc.org/digital/collection/asm0304/id/1311>.
39. Example "has creative work part" statements for the parts of an album: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q73586.png>.
40. RDF Resource Description Framework standard for linked data. See W3C Semantic Web. "RDF: Resource Description Framework." Updated 15 March 2014, at 21:35. <https://www.w3.org/RDF/>.
41. RDF Turtle textual syntax. See Beckett, David, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. 2014. "RDF - Semantic Web Standards." <https://www.w3.org/TR/turtle/>.
42. RDF N-triples plain text syntax. See W3C Semantic Web. 2014. "RDR 1.1 N-Triples: A Line-based Syntax for an RDF Graph." <https://www.w3.org/TR/n-triples/>.
43. RDF JSON-LD format for linked data. See Sporny, Manu, Dave Longley, Gregg Kellogg, Markus Lanthaler, Pierre-Antoine Champin, and Niklas Lindström. 2020. "JSON-LD 1.1: A JSON-based Serialization for Linked Data." W3C Editor's draft. Edited by Gregg Kellogg, Pierre-Antoine Champin and Dave Longley. Posted 14 November 2020. <https://w3c.github.io/json-ld-syntax/>.
44. JSON (JavaScript Object Notation) data format. See Wikipedia. "JSON." Updated 31 December 2020, at 22:32 (UTC). <https://en.wikipedia.org/wiki/JSON>.
45. The PHP Group. "Object Serialization: Serializing Objects - Objects In Sessions. *PHP Manual*. Accessed 7 January 2021. <https://www.php.net/manual/en/language.oop5.serialization.php>.
46. DPLA Metadata Application Profile documentation: <https://pro.dp.la/hubs/metadata-application-profile>.
47. Schema.org metadata schema documentation. See "Organization of Schemas." 2021. <https://schema.org/docs/schemas.html>.
48. W3C Semantic Web. "Web Ontology Language (OWL)." Updated 11 December 2013, at 11:38. <https://www.w3.org/OWL/>.
49. Kellogg, Greg (ed). 2020. "JSON -LD Best Practices: W3C Editor's Draft 20 February 2020." W3C (MIT, ERCIM, Keio, Beihang). <https://w3c.github.io/json-ld-bp/>.
50. Appleby, Michael, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. 2018. "JSON-LD Design Patterns." Chap. 3 in *IIIF Design Patterns*. International Image Interoperability Framework Consortium. https://iiif.io/api/annex/notes/design_patterns/#json-ld-design-patterns.
51. Other names associated with the Los Angeles Dodgers entity: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q166325.png>.
52. First parts of the description of Jasper Wood: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q147700.png>.
53. SPARQL Query map visualization of places depicted in works from a collection: <https://researchworks.oclc.org/cdmld/screenshots/sparql-visualization.png>.

54. Wikibase Gadgets extension documentation. See MediaWiki. "Extension:Gadgets." Updated 16 October 2020, at 11:36. <https://www.mediawiki.org/wiki/Extension:Gadgets>.
55. Mirador IIF-compatible image viewer project website: <https://projectmirador.org/>.
56. Mirador image viewer embedded in the Wikibase user interface: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q165895.png>.
57. Contextual data and image from DBpedia and Wikimedia Commons embedded in the Wikibase user interface: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q71945.png>.
58. Constraints quality assurance Wikibase mechanism documentation. See MediaWiki. "Extension:Wikibase Quality Extensions." Archived 7 January 2019, at 13:45. https://www.mediawiki.org/wiki/Extension:Wikibase_Quality_Extensions.
59. A constraint violation indicating that the "occupation" property should only be used for instances of the type "person" <https://researchworks.oclc.org/cdmld/screenshots/entity-Q73246.png>.
60. OCLC CONTENTdm Custom Pages with CSS and JavaScript documentation. Updated 28 June 2018. https://help.oclc.org/Metadata_Services/CONTENTdm/Advanced_website_customization/Custom_pages/Custom_pages_with_CSS_and_JavaScript.
61. OCLC. CONTENTdm Advanced Website Customization Cookbook website: https://help.oclc.org/Metadata_Services/CONTENTdm/Advanced_website_customization/Customization_cookbook.
62. Google's Structured Data Testing Tool: <https://search.google.com/structured-data/testing-tool>. [Google has announced that this tool is being discontinued.]
63. CONTENTdm Schema.org data evaluated using the Google Structured Data Testing tool: <https://researchworks.oclc.org/cdmld/screenshots/google-structured-data-testing-tool.png>.
64. Additional contextual information displayed in CONTENTdm based on entity descriptions in the pilot Wikibase: <https://researchworks.oclc.org/cdmld/screenshots/cdm15725-p16003coll7-14.png>.
65. Image Annotator initial view with subjects: <https://researchworks.oclc.org/cdmld/screenshots/image-annotator-1.png>.
66. Image Annotator cropped image of a person: <https://researchworks.oclc.org/cdmld/screenshots/image-annotator-2.png>.
67. Image Annotator after adding more depicted subjects: <https://researchworks.oclc.org/cdmld/screenshots/image-annotator-3.png>.
68. Wikibase item updated with depicted subjects and associated cropped images: <https://researchworks.oclc.org/cdmld/screenshots/entity-Q148552.png>.
69. Nielsen, Jakob. 2012. "Thinking Aloud: The #1 Usability Tool." *Nielsen Norman Group*. Posted 15 January 2020. <https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/>.

70. Retriever search results from Wikidata, VIAF, and FAST for “lake vermilion”:
<https://researchworks.oclc.org/cdmld/screenshots/retriever-1.png>.
71. Retriever entity editor:
<https://researchworks.oclc.org/cdmld/screenshots/retriever-2.png>.
72. Wikibase entity created by the Retriever:
<https://researchworks.oclc.org/cdmld/screenshots/entity-Q221424.png>.
73. Knublauch, Holger, and Dimitris Kontokostas (eds). 2017. “Shapes Constraint Language (SHACL): W3C Recommendation 20 July 2017.” W3C. <https://www.w3.org/TR/shacl/>.
74. ShEx Shape Expressions Language W3C Recommendation. See Prud’hommeaux, Eric, Lovka Boneva, Jose Labra Gayo, and Gregg Kellogg. 2017. “Shape Expressions Language 2.0: Draft Community Group Report 27 March 2017.” W3C. <http://shex.io/shex-semantic-20170327/>.
75. Editing essential details for an entity in the Describer:
<https://researchworks.oclc.org/cdmld/screenshots/describer-1.png>.
76. Explorer home page:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-1.png>.
77. Explorer Transportation Hub and related collections:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-2.png>.
78. Explorer search results for “strike”:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-3.png>.
79. Explorer view of a truck bringing workers home during a PTC walkout:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-4.png>.
80. Explorer view of a protest against the Philadelphia Transportation Company:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-5.png>.
81. Explorer view of an 1899 Cleveland transit strike in Public Square:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-6.png>.
82. Explorer view of streetcars parked on the street during a transit strike:
<https://researchworks.oclc.org/cdmld/screenshots/explorer-7.png>.
83. Field Analyzer field usage chart:
<https://researchworks.oclc.org/cdmld/screenshots/field-analyzer-1.png>.
84. Field Analyzer list of field values:
<https://researchworks.oclc.org/cdmld/screenshots/field-analyzer-2.png>.
85. Minnesota Digital Library website. See University of Minnesota. “Minnesota Reflections.”
<https://reflections.mndigital.org/>.
86. Digital Public Library of America website: <https://dp.la/>.

For more information about our work related to digitizing library collections, please visit: oclc.org/digitizing



6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878

T: +1-614-764-6000

F: +1-614-764-6096

www.oclc.org/research

ISBN: 978-1-55653-185-9
DOI: 10.25333/fzcv-0851
RM-PR-216817-WWAE 2101