DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

OCLC's Model in WorldCat: A Focus on Relationships

Jeff Mixter
OCLC, USA
mixterj@oclc.org

Michael Phillips
OCLC, USA
phillipm@oclc.org

Kathryn Stine
(Formerly) OCLC, USA

Abstract

Conceptual models are a key component to holistically understanding data and using it in end-user applications. They provide an understandable roadmap for exploring, visualizing, and surfacing information. Library reference models serve a similar purpose by describing component parts of bibliographic materials that can help users find materials that fit their specific information needs. This work presents OCLC's thinking in how to adapt the traditional Works, Expression, Manifestation, and Item (WEMI) model in WorldCat based on experiments with WorldCat bibliographic records.

Keywords: data modeling; linked data; metadata; data clustering; OCLC; WorldCat; relationship

1. Introduction

1.1. LRM

The work of the Functional Requirements for Bibliographic Records (FRBR) and Functional Requirements for Authority Data (FRAD) working groups set the stage for the development of the International Federation of Library Associations and Institutions (IFLA) Library Reference Model (LRM). The LRM uses a Works, Expression, Manifestation, and Item (WEMI) model to describe the hierarchical relationship between bibliographic entities (Le Boeuf, et al., 2018). Resource Description and Access (RDA) has adopted the LRM conceptual model for its cataloging rules. One challenge in adopting the WEMI conceptual model is applying it to a MARC bibliographic database. MARC records are normally anchored at the WEMI Manifestation level, so Works and Expressions need to be derived based on record clustering. Another implementation problem with WEMI is that for certain material types, applying the conceptual model produces redundancies. A photograph in a digital repository is an easy-to-understand resource for an end user, but the WEMI model imposes four representations for an item that appear different only in classification and degree of specificity in the properties used to describe each (Coyle, 2022). These challenges, both from a data creation and a data reuse standpoint, have led others, such as the Library of Congress Bibliographic Framework (BIBFRAME) initiative, and the Casalini Libri Share-VDE project to rethink the high-level conceptual model for bibliographic materials.

1.2. BIBFRAME

Library of Congress's BIBFRAME initiative focuses on a slightly more compact form of the WEMI model. The BIBFRAME model contains Work, which is similar to a WEMI Expression; Instance, similar to a WEMI Manifestation; and Item. This approach allows a single MARC record to be directly mapped into each of its corresponding BIBFRAME classes, but this mapping can have problems when working with a large aggregate set of records in which duplicate records exist. Both Library of Congress and the Share-VDE project are using BIBFRAME to map MARC records into RDF, and both have adopted extensions to account for the abstract WEMI Work class. Library of

https://doi.org/10.23106/dcmi.953193572

Congress has developed a Hub class to cluster related BIBFRAME Works and similarly, Share-VDE has created an Opus class for the same purpose. This demonstrates interest in having a very abstract view of a bibliographic item but does not attempt to model the semantic relationships between the related BIBFRAME Works.

2. Defining WEMI Works in WorldCat

WorldCat bibliographic records contain WEMI Manifestation-level descriptions. Therefore, deriving abstract WEMI Work and Expression entities from them has presented a challenge for OCLC. An approach to this challenge was the development of a FRBR algorithm to analyze and cluster like records, thereby approximating the aggregation intent of the WEMI Work entity. Initial thinking at OCLC identified these clusters as the potential basis for establishing a WEMI Work entity, but this method posed two problems. First, FRBR clusters simply are not perfect. The WorldCat database joins data from thousands of sources, so differences in cataloging practices and simple human errors in records can prevent the automated processes from creating clusters with complete accuracy. Second, there is the issue of end-user discovery and display. This is an issue of scale, as a given work may have hundreds of derivative WEMI Expressions with complex relationships attached to it. Presenting a Work entity with so many related resources could present the user with an unnecessary quantity of results which they would need to traverse to find the desired resource, such as a translation or specific format (Aalberg et al., 2019). Ultimately, this is a problem of relationships; the WEMI Works do not have detailed properties necessary to fully contextualize their relationships to, and between, derivative Expressions.

Considering the challenges inherent in generating WEMI Work entities from MARC records, and their limited utility to end-user discovery, OCLC has proposed the notion of a "WorldCat Work." The WorldCat Work entity is equivalent to a WEMI Expression, and thereby provides specific attributes to aid user retrieval, such as format or language. Foregoing the notion of a WEMI Work as the aggregating entity for all derivative Expressions, the "Representative WorldCat Work" instead fulfills this purpose. The Representative WorldCat Work is based on the first Expression, or the expression which is considered the canonical resource for the WorldCat Work. It provides foundational metadata and a logical center to which all subsequent derivations can be linked.

3. Modeling

Based on experiments around FRBR clustering (Hickey et al., 2002) GLIMR clustering (Gatenby et al., 2012), and end-user testing, OCLC decided to focus its modeling efforts on WEMI Expressions, WEMI Manifestations, relationships between Expressions, relationships between Expressions and Manifestations, and relationships between Manifestations. To avoid confusion with LRM and other bibliographic data models, OCLC is referring to its two primary classes of bibliographic entities as "WorldCat Works," which are analogous to WEMI Expressions, and "WorldCat Editions," which are analogous to WEMI Manifestations. Relationships are key to understanding the structured meanings of, and connections between, bibliographic and other types of entities. They have the added effect of making the sometimes-artificial hierarchy of the WEMI model more understandable. OCLC's model, in some sense, turns

https://doi.org/10.23106/dcmi.953193572

the WEMI model on its head and instead of focusing on the hierarchical classes, focuses on the semantic relationships between the entities, which results in the classes themselves being much less important. OCLC is working on enhancing its linked data model to account for WorldCat Work-to-WorldCat Work relationships and WorldCat Work-to-WorldCat Edition relationships.

Figure 1 (below) shows a limited view of the model using *The Adventures of Tom Sawyer* by Mark Twain as an example. A noteworthy aspect of this graph is the addition of specific semantic relationships between the Representative WorldCat Work and the derivative WorldCat Works. This allows us to better understand the contextual connections between these entities, something that will be key to any users looking for specific resources. Examples of relationships shown here include "translation of," "reading of," and "based on."

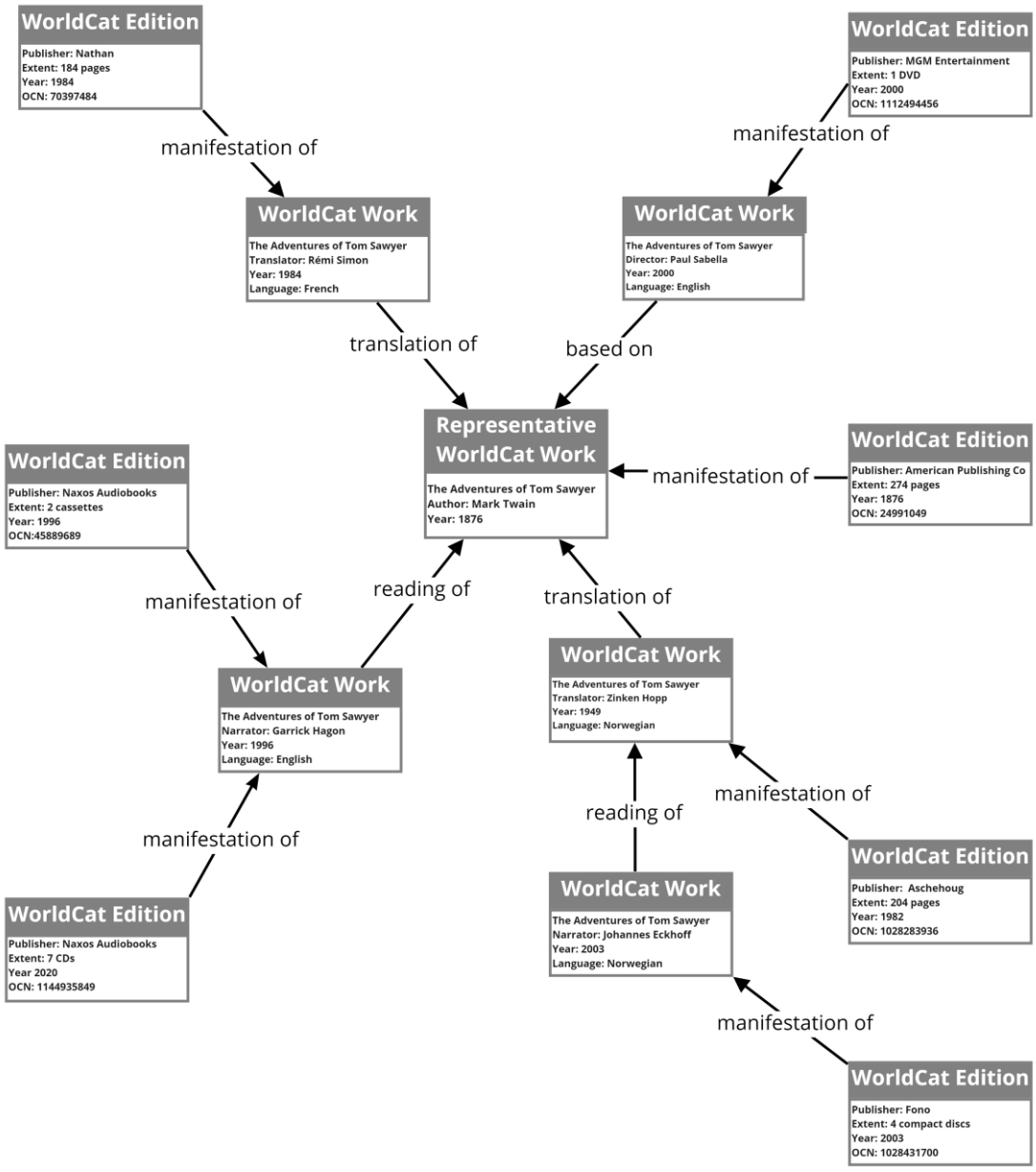

Figure 1. OCLC model for Tom Sawyer. Note the Representative WorldCat Work at the center of the graph, with WorldCat Work and WorldCat Edition entities branching from it.

https://doi.org/10.23106/dcmi.953193572

For those aggregating data, maintaining the relationships between these entities is a unique challenge. While a single institution may have only 10 different manifestations of Tom Sawyer, WorldCat has almost 4,500. Therefore, designing the model in this way provides a structure from which to base a traversable and detailed graph at scale. OCLC is able to derive WEMI Work descriptions by clustering together the WorldCat Works that link to each other via specific WorldCat Work to WorldCat Work properties. The WorldCat Works to WorldCat Works properties can also be used to identify and connect related but distinct WEMI Works. For example, figure 2 shows how the original *Tom Sawyer*, as a Representative WorldCat Work (a type of WorldCat Work), relates to the 2000 animated film *Tom Sawyer*, also a WorldCat Work, via an "based on" property.

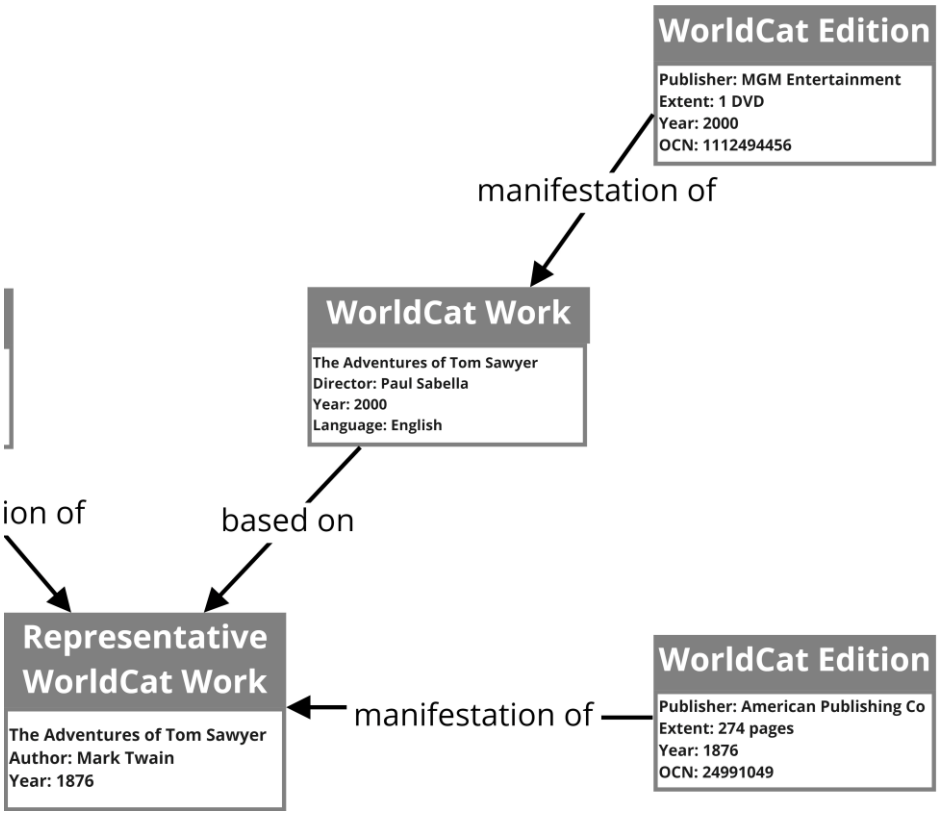

Figure 2. A focus on the "based on" relationship used for a film adaptation of the Representative WorldCat Work.

In processing this relationship, OCLC can identify two distinct WorldCat Works and semantically relate them. The emphasis on meaningful relationships allows the model to express how derivative works relate to one another, and, directly or indirectly, relate back to the Representative WorldCat Work. In this way, like entities may be clustered using relationships rather than additional classes, the way that BIBFRAME uses the "Hub" and the Share-VDE the "Opus."

The OCLC model aims to support a discovery-focused knowledge graph that can be used by library systems and services, as well as by general purpose web services, to improve the searchability and contextualization of library materials. OCLC's current work focuses on the relationships between entities rather than the hierarchical class structure of the entities. This focus has resulted in interesting WorldCat Work to

https://doi.org/10.23106/dcmi.953193572

WorldCat Work relationships that make the resulting data helpful for discovery service use. One such property is "reading of." Figure 3 shows the model for a "reading of" relationship between the Norwegian translation of *Tom Sawyer* by Hop, published in 1949, and the audio recording of the translation by Eckhoff in 2003. Without the specificity in the property, the relationship between two WorldCat Works might appear anomalous, or at the very least ambiguous. Instead, the relationships clearly describe the transformations taking place between derivative WorldCat Works.

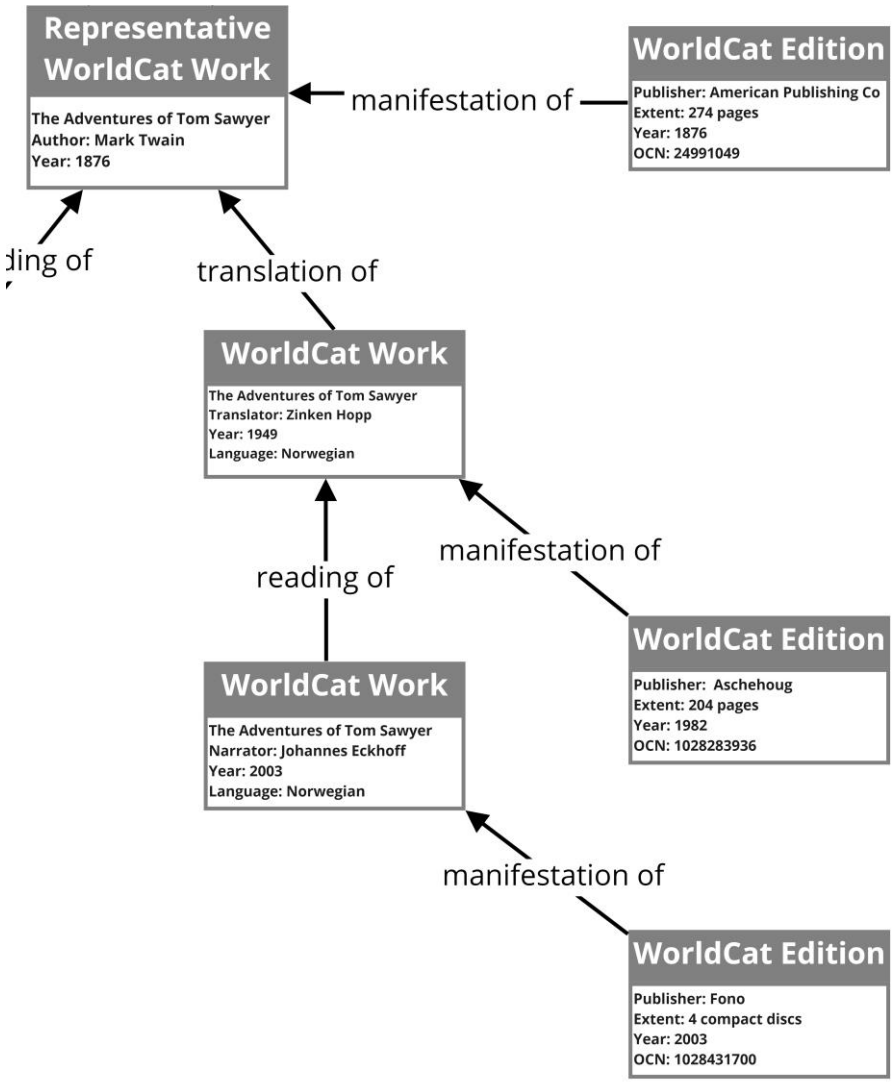

Figure 3. A focus on the reading of a translation, each a distinct WorldCat Work. The semantic relationships provide a traversable graph wherein multiple derivatives can always be traced back to the Representative WorldCat Work.

The focus on these relationships is key to the user experience and therefore a core characteristic of the graph model. In many cases, the exact relationship between expressions is fully described in bibliographic records. Where it is, it tends to be locked away in a free-text note field. This could potentially lead to further study on methods of enriching the graph with more accurate relationship data, including metrics to indicate the strength of relationships, showing where a relationship is explicitly known versus just assumed.

https://doi.org/10.23106/dcmi.953193572

4. Conclusion

OCLC is currently building out relationships among WorldCat Work and WorldCat Edition entities. In addition to modeling, OCLC plans to test the model against WorldCat data to make sure it can be instantiated at scale. Part of this work includes data mining MARC records for additional context to help improve relationships between entities and to augment the entity description. In addition to data mining WorldCat, community collaboration will be critical to improving and managing the entity data over time. The value of community involvement has been demonstrated by the success and growth of the Wikidata knowledge base. Work is also being done on modeling additional related entities that are important for bibliographic discovery and data exploration. These include People, Organizations, Events, Places, and Concepts. Other modeling challenges currently under consideration are how to model non-monograph resources, such as serials, and non-bibliographic materials, such as photographs. OCLC is planning a combined strategy of data analysis and community engagement to refine the model as this work continues.

References

Aalberg, Trond., Kim Tallerås, and David Massey. (2019). The impact of new bibliographic models on the search experience. *Proceedings of CoLIS, the Tenth International Conference on Conceptions of Library and Information Science, Ljubljana, Slovenia, June 16-19, 2019. Information Research, 24*(4), paper colis1915. Retrieved from http://InformationR.net/ir/24-4/colis/colis1915.html (Archived by the Internet Archive at https://web.archive.org/web/20191217172926/http://informationr.net/ir/24-4/colis/colis1915.html).

Coyle, Karen. (2022). Works, Expressions, Manifestations, Items: An Ontology. *Code4Lib journal*, (53). Retrieved from https://journal.code4lib.org/articles/16491.

Gatenby, Janifer, Richard O. Greene, W. Michael Oskins, and Gail Thornburg. (2012). GLIMIR: Manifestation and content clustering within WorldCat. *Code4Lib journal*, (17). Retrieved from https://journal.code4lib.org/articles/6812

Hickey, Thomas B., Edward T. O'Neill, and Jenny Toves, (2002). Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). *D-Lib magazine, 8*(9), 1-13. Retrieved from https://www.dlib.org/dlib/september02/hickey/09hickey.html

Le Boeuf, Patrick, Pat Riva, and Maja Žumer. (2018). IFLA Library Reference Model: A Conceptual Model for Bibliographic Information. Retrived from https://repository.ifla.org/handle/123456789/40

https://doi.org/10.23106/dcmi.953193572