Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC

#ldmodels

Carol Jean Godby
OCLC Research

Ray Denenberg
Library of Congress

January 2015

Library of Congress
Washington, DC 20540 USA
www.loc.gov

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

ISBN: 1-55653-489-2 (978-1-55653-489-8)
OCLC Control Number: 900616520

Please direct correspondence to:

Carol Jean Godby
Senior Research Scientist
OCLC Research
godby@oclc.org

Ray Denenberg
Senior Network and Standards Specialist
Library of Congress
rden@loc.gov

Suggested citation:
Godby, Carol Jean, and Ray Denenberg. 2015. *Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC.* Dublin, Ohio: Library of Congress and OCLC Research. http://www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015.pdf.

Acknowledgements

The authors gratefully acknowledge the feedback from their colleagues, which shaped this draft and the underlying technical analysis. From the Library of Congress, Sally McCallum, Nate Trail, and Beacher Wiggins; and from OCLC, Ted Fons, Roy Tennant, Richard Wallis, and Jeff Young.

Introduction

Since 2011, OCLC researchers have been experimenting with Schema.org as a vehicle for exposing library metadata to Web search engines in a format they seek and understand. Schema.org is sponsored by Bing, Google, Yahoo! and Yandex as a common vocabulary for creating structured data markup on Web pages. OCLC's experiments led to the 2012 publication of Schema.org metadata elements expressed as linked data on 300 million catalog records accessible from WorldCat.org.[1] In 2011, BIBFRAME was launched by the Library of Congress (LC) as an initiative to develop a linked data alternative to MARC, building on the Library's experience providing linked data access to its authority files which began in 2009.[2] Among BIBFRAME's aims were (1) to supply search engines with descriptions of library resources in a form they could use, (2) to promote the application of concepts defined in the FRBR and RDA models, and (3) to offer an extensible solution for the description of resources in the broader cultural heritage community. A BIBFRAME high-level model was developed by Zepheira LLC,[3] under contract, to provide a framework for development.[4]

During the latter part of 2012 and throughout 2013, the LC BIBFRAME modeling and development team formulated principles to guide the creation of the BIBFRAME vocabulary. A prepublication draft was evaluated by the BIBFRAME Early Experimenters, which included the British Library, the Deutsche Nationalbibliothek, George Washington University, the National Library of Medicine, OCLC, and Princeton University. One outcome was a first edition of the BIBFRAME vocabulary and the first BIBFRAME descriptions, which were algorithmically generated by LC and OCLC from millions of MARC records.

Another outcome of the Early Experimenters Group was a position paper written by OCLC describing the relationship between BIBFRAME and OCLC's models derived from Schema.org. *The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper*[5] was published in 2013 and made available from the BIBFRAME home page. The analysis highlighted lexical correspondences between the vocabularies defined by BIBFRAME and Schema.org enhanced with a small set of extensions proposed by OCLC; identified places where the underlying models were immature and could diverge; and concluded that, given the use cases motivating the two efforts, the two models should be complementary. The paper pointed out that the coverage of Schema.org is necessarily broad but shallow because library resources must compete with creative works offered by many other communities in the information landscape. Conversely, the coverage of BIBFRAME is deep because it contains the vocabulary required of the next-generation standard for describing library collections.

In the past year and a half, OCLC has focused on the tasks related to the use of Schema.org: refining the technical infrastructure and data architecture for at-scale publication of linked data for library resources in the broader Web, and investigating the promise of Schema.org as a common ground between the language of the information-seeking public and professional stewards of bibliographic description. BIBFRAME has focused on publishing additional vocabulary and facilitating implementation and testing. These new developments prompt the need to re-examine the relationship between the LC and OCLC models for library linked data. This document is an executive summary of a more detailed technical analysis that will be released later this year.

BIBFRAME since 2013

In late 2013 the Early Experimenters Group concluded its work and in early 2014 the BIBFRAME Implementation Testbed[6] was formally established. Its purpose is to encourage development of BIBFRAME test implementations; monitor implementation progress; discover errors, inconsistencies, and shortcomings in both the implementations and in the BIBFRAME model and vocabulary; and provide a forum for the development of BIBFRAME vocabulary and tools. Over the past year, 17 organizations have participated actively in this effort.

In addition, there has been lively discussion on the (public) BIBFRAME listserv.[7] Since BIBFRAME is expressed in RDF (the W3C-developed Resource Description Framework), listserv discussion has covered issues pertaining to RDF and linked data, as well as issues pertaining to the BIBFRAME model and vocabulary. To assist experimentation with the BIBFRAME model, LC has provided tools available for open download such as MARC to BIBFRAME transformers and a simple input editor,[8] and has also encouraged the community to share any software components they develop. Testing, implementation, and discussion have produced corrections and improvements to the BIBFRAME vocabulary, and LC continues to work with implementers for further enhancements.

Later in 2015, LC will publish a revised vocabulary and launch a pilot project to test whether the BIBFRAME vocabulary supports the capability for catalogers to do original cataloging, including authority work. In the pilot, LC catalogers will test the creation of cataloging data in BIBFRAME using the BIBFRAME Editor. Catalogers will create BIBFRAME descriptions for a variety of materials, in a variety of languages. LC Name/Title and Title MARC records will be converted into BIBFRAME Works and stored in an RDF triplestore. Bibliographic records will be converted and matched against the Works, with subjects and other properties merged. A search/display tool will be put on top of the triple store, as well as the BIBFRAME Editor.

Similar pilot projects are being planned at other institutions such as Stanford and Cornell. The details of these pilots are not finalized; however, there will be cooperation and sharing of information and results within the community. The pilots are expected to provide an opportunity to evaluate many of the issues that will be raised by the transition from MARC to BIBFRAME.

OCLC's experiments with Schema.org since 2013

Since 2013, the linked bibliographic data accessible from WorldCat.org has been upgraded and republished, and the linked data models of the FAST[9] and VIAF[10] authority files have been redesigned with references to classes defined in Schema.org for fundamental concepts such as "Person," "Organization," "Creative Work" and "Topic." In addition, the first draft of WorldCat Works has been published,[11] which represents Work-level descriptions produced from the latest version of OCLC's FRBR-inspired clustering and data-mining algorithms operating on library authority files and WorldCat catalog records.[12] As a result, nearly 200 million "Work" clusters are now modeled as linked data using Schema.org and associated with persistent URIs.

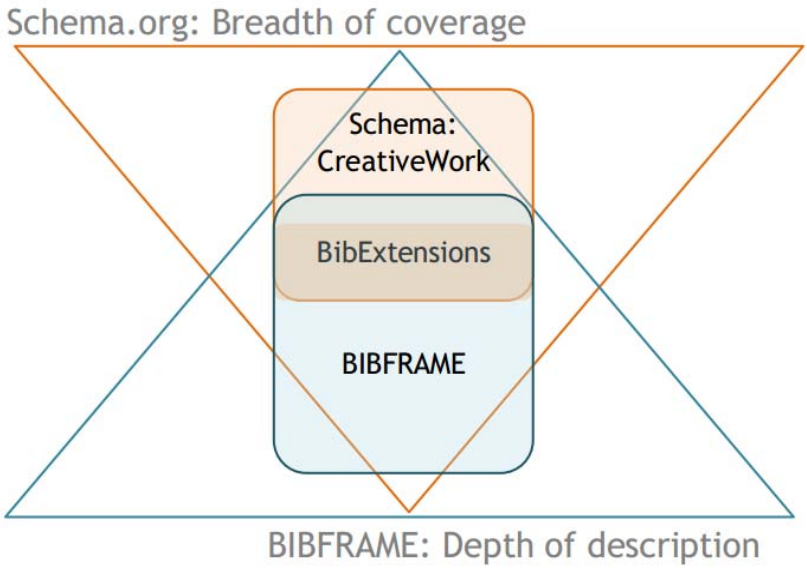
Jeff Mixter and Jean Godby, who are members of OCLC's linked data modeling team, have also been collaborating with Montana State University's Dean of the Library Kenning Arlitsch and Semantic Web Research Director Patrick OBrien to examine issues of discoverability and visibility of library resources in general-purpose search engines such as Google. One outcome is a model of some of the contents of institutional repositories, expressed primarily in Schema.org.[13] This model will be refined by OCLC in the IMLS-funded project "Measuring Up,"[14] which will also be directed by Arlitsch and OBrien. All of these projects are built on Schema.org and most focus on the generation of linked data from legacy standards with the goal of publication in a form that can be consumed by general-purpose search engines.

OCLC's linked data experts envision a need for an extension vocabulary tailored to Schema.org that fills in gaps required for the description of resources managed by libraries. In the linked data markup published on WorldCat catalog data in 2012, these extensions were described in the "library" vocabulary, a small draft ontology maintained at OCLC and developed with an awareness of Schema.org that was not explicitly formalized. These extensions are now accessible from http://BiblioGraph.net.[15] The underlying BiblioGraph vocabulary contains terms defined by those with a professional commitment to bibliographic resource description that are understandable and potentially useful outside their narrow communities of practice. Designed as a proving ground for demonstrating the potential impact of candidate extensions to Schema.org, it has the same look and feel of Schema.org and is integrated with a copy of the most recent version of the Schema.org vocabulary.

The BiblioGraph.net website is maintained by OCLC, but the ontology that populates it will be managed as a community resource. The concept was inspired by the work of the Schema Bib Extend Community Group,[16] which was convened by OCLC Technology Evangelist Richard Wallis and sponsored by the World Wide Web Consortium to evaluate the suitability of Schema.org as a standard for bibliographic description by librarians, library systems developers, and publishers, and to recommend amendments, if necessary.

Aligning BIBFRAME and the OCLC/Schema models

In 2013, the relationship between BIBFRAME and OCLC's models based on Schema.org was visualized in the diagram reproduced in the figure below.



Godby, for OCLC Research. 2013.

High-level alignment of BIBFRAME and a model derived from Schema.org[17]

At the highest levels, OCLC's linked data model is similar to BIBFRAME, particularly in the definition of entities such as Work, Instance, Organization, and Person. This redundancy reflects a convergence of two projects that have different motivations and use cases. LC is developing BIBFRAME for data exchange in the linked data environment, taking into account existing formats for resource description, as well as interactions with search engines; it must be designed as a persistent standard for library resource description. By contrast, the linked data models being developed at OCLC optimize descriptions of library resources for discovery on the Web beyond libraries, using the vocabulary designed for consumption by general-purpose search engines. If the promise of Schema.org markup is realized, the outcome should be measurable as increased click-through rates or other evidence of improved visibility for

libraries on the Web. Nevertheless, the overlap between the two projects is anticipated to be only partial. The vocabulary defined in Schema.org and BiblioGraph aims to be broadly understandable to the information-seeking public and may not include many of the details defined in BIBFRAME, which aims more to address the needs of long-term curation by libraries and other cultural heritage institutions.

The technical analysis: a summary

In the technical analysis planned for release later in 2015, Ray Denenberg and Jean Godby compare RDF descriptions conforming to the OCLC/Schema model with corresponding BIBFRAME descriptions, focusing on the two key BIBFRAME entities, Work and Instance, and the relationships between them. Other primary BIBFRAME concepts such as authorities, annotation, subjects, titles, identifiers, and agents are also discussed.

A set of dialogs

Each concept is the subject of a focused dialog that asks two questions. First, are the persistent identifiers assigned to the corresponding concepts in the two models mutually consumable? If so, it is possible to conclude that though the models have different internal details and are expressed in different vocabularies, they are describing the same objects. As a result, a BIBFRAME Work description, could, for example, contain a "same as" assertion to an identifier published in the OCLC Works Service and an OCLC/Schema description of a resource described in WorldCat catalog data could refer to a BIBFRAME Instance.

Second, the authors ask whether a BIBFRAME description can be reformulated in the OCLC/Schema model (and vice versa) without loss of information. This question is especially important to OCLC because an affirmative answer implies that it is possible to address the need of a data aggregator to import and export BIBFRAME data even if the internal linked data model is expressed in a different vocabulary. The high-level conclusion is that the alignment shown in the image in the previous section is still accurate and is perhaps even more defensible than in 2013 because the primary BIBFRAME concepts are now more consistent with the corresponding concepts defined in the OCLC/Schema model. Moreover, given BIBFRAME's terms for the description of music and maps that have no counterpart either in Schema.org or in BiblioGraph, the new analysis provides a much-needed empirical demonstration of the difference in granularity between the two models and offers technical solutions for managing it. This difference was presented merely as a theoretical possibility in 2013.

Representing the FRBR Group 1 hierarchy

BIBFRAME and OCLC's models both take a simplified view of FRBR. Both models define RDF classes for Work entities, and while a BIBFRAME and OCLC Work are not entirely the same, the analysis reveals that they are quite compatible. Both models encode FRBR Expression entities as RDF properties, or relationships. Both also recognize Manifestation entities, though in different ways: BIBFRAME defines the Instance RDF class to represent a Manifestation entity, while the OCLC model induces Manifestation and Item entities using a combination of RDF type assignments from schema:CreativeWork and schema:Product, as described in the aforementioned 2013 publication, *The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper*.

BIBFRAME has defined a set of 30 content-to-content (i.e., Work-to-Work) relationships derived from MARC and RDA, which are consistent with OCLC's modeling assumptions and can supplement a model of creative works derived from Schema.org. In addition, people, places, and organizations, which are typically described in library authority files, are represented not as curated strings or as concepts but as real-world objects in the LC and OCLC models. Thus the referents of many top-level BIBFRAME RDF classes, including Work, Instance, heldItem, and the subclasses of Authority, are ontologically similar enough that the corresponding URIs are mutually consumable between BIBFRAME and OCLC's models. This claim could not be made with confidence in 2013.

Differences

The analysis reveals at least three high-level differences in the models. The first was alluded to above: BIBFRAME defines RDF classes for Work and Instance, while OCLC defines classes for Work but not for Instance. As noted, this difference does not present an incompatibility.

Second, an Authority entity is formally defined as an RDF class in BIBFRAME, but not in OCLC's models. In OCLC's linked data models, "Authority" is simply an informal name for any resource that contains vetted information about people, places, organizations, concepts and other entities that are important for the description of the entities that populate library resource descriptions. However, the RDF data stores representing the contents of library authority files are otherwise compatible and contain descriptions of the same objects. In the BIBFRAME model, the RDF class bf:Authority is defined largely to facilitate the description of subjects. This issue will be explored more deeply in the forthcoming technical analysis, as will the treatment of subjects in general in the LC and OCLC models.

Third, the BIBFRAME RDF class defined for the Annotation entity has no counterpart in OCLC's models. Nevertheless, the BIBFRAME Annotation now contains structured data that can

describe reviews, summaries, cover art, and holdings—and most have alternative and more parsimonious formulations in the OCLC/Schema model. The BIBFRAME Annotation class is being carefully reviewed in light of the work currently being conducted by the World Wide Web Consortium on Web annotations.[18]

As expected, the analysis revealed differences in granularity. For example, if a review has an author or publisher, or if a piece of cover art has a provenance, BIBFRAME describes the object with a structured data value, defining an RDF subclass of the Annotation class with properties. The most obvious corresponding description in Schema.org typically contains only a simple data value such as a string literal or a URL and cannot represent such details.

The same issue arises in the description of several BIBFRAME concepts, such as titles and identifiers. In BIBFRAME, a title can be expressed as a string literal or as a structured resource (including main title, subtitle, part number, and several other information elements), while an OCLC title is always expressed as a string literal (via the property schema:name). But since both models allow titles to be expressed as literals, there is sufficient compatibility. Identifiers are more complex and will get comprehensive treatment in the forthcoming technical analysis. OCLC's linked data experts are exploring generic solutions for expressing BIBFRAME's additional granularity in Schema.org, while also engaging in debate about whether it is always necessary.

The vocabulary of discovery and curation

Of course, BIBFRAME descriptions can also be more detailed because they include the specialized vocabulary required for professional curation. For example, the analysis compares a hand-crafted BIBFRAME description of a celestial map held at the Library of Congress with an algorithmically generated description of the same object in the OCLC/Schema model. The BIBFRAME description contains the technical terms bf:cartographicScale, bf:cartographicEquinox, and bf:cartographicAscensionAndDeclination. The OCLC description does not contain these terms because the OCLC source record does not represent this information and these concepts are not defined in Schema.org or BiblioGraph. They illustrate BIBFRAME's focus on vocabulary development to support upgraded machine-understandable descriptions of the resources uniquely held by libraries, such as maps, sheet music, audiovisual materials, and archives. The OCLC/Schema model can refer to this description and enhance its own simply by adding a "same as" assertion containing the BIBFRAME URI. But to generate comparable descriptions or to pass them through OCLC's data processing stream without loss of information, the OCLC/Schema model must use the BIBFRAME vocabulary directly. This is the "depth of description" mentioned in the figure that is supplied by BIBFRAME and will perhaps always be missing from a data model optimized for discovery.

BiblioGraph is mentioned throughout the technical analysis as a vehicle for promoting the vocabulary of expert description to the vocabulary of discovery, and it may have a role in the description of the celestial map. For example, "map" has been defined as a resource in Schema.org, but the list of defined properties is too sketchy to meet the stewardship needs of librarianship. But the BIBFRAME terms are defined as RDF properties that can be theoretically positioned in the schema:Map class using BiblioGraph as a testing ground. A representation in BiblioGraph can be interpreted as a claim that other communities of practice might have a need for these terms, which makes them candidates for eventual absorption into Schema.org. Among library standards experts, much analysis is required to determine which terms have commonly understood semantics and which are specialized, and perhaps it could be concluded that bf:cartographicScale is a candidate for broader use, while the others may not be. Nevertheless, BiblioGraph is designed as a place to consolidate the results of this analysis.

Some recommendations for closer alignment

Much of the common ground between the LC and OCLC linked data models has not yet been exploited because of solvable technical and conceptual barriers. These are prescriptions for future collaboration, but many are addressed in the technical analysis. They include:

OCLC

- Develop and test the technical solutions for capturing the granularity expressible in BIBFRAME but not the OCLC/Schema model and demonstrate that OCLC can import and export BIBFRAME without loss of information.
- Publish acceptance criteria that defines the scope of BiblioGraph and propose terms defined in BIBFRAME that satisfy them.

LC

- Produce BIBFRAME descriptions that refer to OCLC's Work identifiers.

OCLC and LC in partnership

- Develop and test an implementation of a common model of one or more resource types held by libraries that are not easily describable in BIBFRAME or in Schema.org, such as maps or audiovisual materials.
- For any given vocabulary term (defined as either an RDF class or property) required for library data and not in Schema, analyze and compare its usage within BIBFRAME and BiblioGraph. Is it in both vocabularies and are the definitions similar? Can the BIBFRAME term be used in conjunction with Schema (in the same manner as a BiblioGraph term)?

Notes

1. OCLC. 2012. "OCLC adds Linked Data to WorldCat.org." 20 June. http://www.oclc.org/en-US/news/releases/2012/201238.html.
2. Library of Congress Linked Data Service: Authorities and Vocabularies. http://id.loc.gov.
3. Zepheira LLC: http://zepheira.com/.
4. Library of Congress. 2012. Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Washington, DC: Library of Congress. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.
5. Godby, Carol Jean. 2013. *The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-05.pdf.
6. BIBFRAME (BF) Implementation Testbed: http://www.loc.gov/bibframe/implementation/testbed.html.
7. To subscribe to the BIBFRAME Listserv (version 14.5) see http://listserv.loc.gov/cgi-bin/wa?SUBED1=bibframe&A=1.
8. BIBFRAME Editor: http://bibframe.org/tools/editor/.
9. FAST (Faceted Application of Subject Terminology) Linked Data. Last updated 9 January 2015. http://experimental.worldcat.org/fast/.
10. VIAF: The Virtual International Authority File: http://viaf.org/.
11. OCLC. 2014. OCLC Releases WorldCat Works as Linked Data. 28 April. https://oclc.org/news/releases/2014/201414dublin.en.html.
12. OCLC Research. 2015. "OCLC Research Activities and IFLA's Functional Requirements for Bibliographic Records." Accessed 22 January. http://www.oclc.org/research/activities/frbr.html?urlm=159763.
13. Mixter, Jeff, Patrick OBrien and Kenning Arlitsch. 2014. "Describing Theses And Dissertations Using Schema.org." In *2014 Proceedings of the International Conference on Dublin Core and Metadata Applications*. 138-146. http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/269/238.
14. OCLC Research. 2014. "'Measuring Up: Assessing Use of Digital Repositories and the Resulting Impact' Project Receives IMLS Grant." 21 October. http://www.oclc.org/research/news/2014/10-21.html.
15. BiblioGraph.net: http://bibliograph.net.
16. See http://www.w3.org/community/schemabibex/wiki/Main_Page.
17. This image originally appeared in the report referenced above in note 5.
18. W3C Web Annotation Working Group: http://www.w3.org/annotation/.