DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2015

Exposing Library Holdings Metadata in RDF Using Schema.org Semantics

Myung-Ja K. Han
University of Illinois at Urbana-Champaign, USA
mhan3@illinois.edu

Timothy W. Cole
University of Illinois at Urbana-Champaign, USA
t-cole3@illinois.edu

Patricia Lampron
University of Illinois at Urbana-Champaign, USA
lampron2@illinois.edu

M. Janina Sarol
University of Illinois at Urbana-Champaign, USA
mjsarol@illinois.edu

Abstract

Libraries have been busy transforming and publishing their data as linked open data by testing already existing semantics and developing new sets of semantics. So far, most of the efforts have focused on the bibliographic data, not the holdings and item related data that are unique to individual libraries and that help users access the information resources they need. The University of Illinois at Urbana-Champaign Library experimented with a subset of its bibliographic records (5.4 million) describing print resources and associated holdings data to examine options and best practices so far identified for expressing library holdings data using schema.org semantics. The experimentation suggests that the mappings for holdings data recommended by the BibExtend Community Group are in some ways incomplete and that some proposed uses of schema.org types and properties to describe library holdings go beyond current schema.org definitions. Existing schema.org enumerations should be extended (e.g., regarding availability) to better describe library use cases, and some extensions to schema.org are needed to fully describe library holdings data and to maximize their utility. This paper highlights issues, suggests potential extensions identified during the transformation to schema.org semantics, and discusses options to make essential library holdings data fully visible as linked open data.

Keywords: Linked Open Data; library catalog; holding data; schema,org; Semantic Web

1. Introduction

Libraries today are both producers and consumers of linked open data (LOD). In describing library resources, libraries need to identify what unique information they can and want to contribute to the growing Web of Data (aka the Semantic Web) and to assess which semantics will be most effective for sharing resource descriptions. Their role as consumers of LOD can help inform these decisions. To date, libraries have tried and tested a variety of data models and semantics to publish catalog records as linked data (Cole, Han, Weathers, & Joyner, 2013). Two initiatives, the Library of Congress (LC)' BIBFRAME (Library of Congress, 2015) and schema.org (schema.org, 2015) as used, for example, by the Online Computer Library Center (OCLC) (OCLC, 2014) have garnered the majority of interest. The graphs produced by transforming library catalog records to BIBFRAME or schema.org are useful, but incomplete; less attention has been given so far to holdings data, which is essential to help users know where to locate information resources and how to access them. This is because libraries maintain holdings data separate from their bibliographic descriptions, e.g., in acquisitions and circulation modules in Integrated Library System (ILS) or Electronic Resource Management (ERM) systems.

Using a snapshot of the University of Illinois at Urbana-Champaign (UIUC) Library's print bibliographic and holdings data, this paper examines options and best practices so far identified for expressing library holdings data using schema.org semantics. Web search engine vendors

https://doi.org/10.23106/dcmi.952137119

collaborated to create schema.org, and the perspective is decidedly commercial. Preliminary mappings of holdings data to schema.org have been proposed, notably by the W3C BibExtend Community Group (Schema BibExtend Community Group, 2015), but our examination suggests that these mappings are incomplete and some proposed uses of schema.org types and properties (i.e., Resource Description Framework (RDF) classes and predicates) go beyond current definitions. Enumerations are insufficient in a few cases (e.g., regarding availability and borrowing terms), and extensions to schema.org are needed to fully describe library holdings data. We also found that the holdings data contained in our ILS acquisitions and circulation modules, while adequate to generate RDF descriptions of print holdings, were not adequate to generate RDF descriptions of electronic holdings. In this paper, we highlight issues identified from the experimentation, suggest potential extensions to schema.org semantics, and discuss options libraries may want to pursue to make their holdings data visible as LOD. This work remains incomplete and further research is ongoing. For example, conflation of work, expression, manifestation, and item data makes matching and de-duping across collections difficult, but OCLC's Work Identifiers (OCLC, 2015) may provide a solution.

2. Library Holdings Data

2.1. Holdings Data in MARC

Libraries have been using the MAchine Readable Cataloging (MARC) format as a cataloging tool since the early 1960s. Although both bibliographic and holdings data can be encoded in MARC (Library of Congress, 2015), most ILS manages bibliographic and holdings data in different modules. Because of this, the term 'library MARC record,' usually refers only to the bibliographic data without the holdings data.

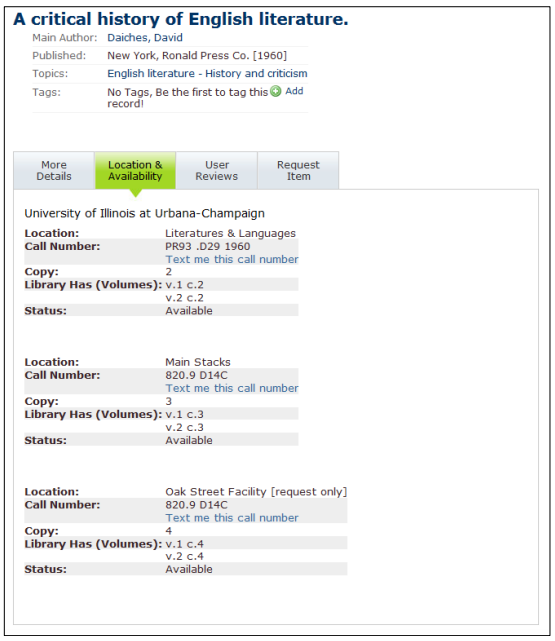

FIG. 1-1.

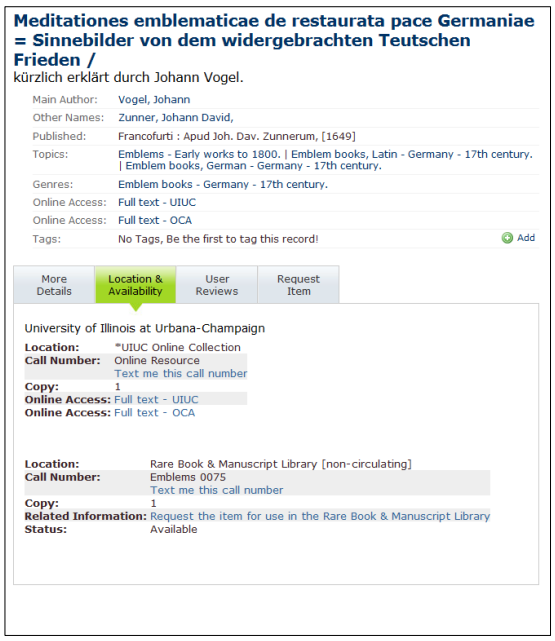

FIG. 1-2.

FIG. 1. Holdings and item specific data displayed in the UIUC Library's OPAC. Figure 1-1 shows the multipart items and Figure 1-2 shows the print and online holdings shown together.

"Copy-specific information for an item; information that is peculiar to the holding organization; information that is needed for local processing, maintenance, or preservation of the item; and version information" are collectively referred to as holdings data. Holdings are sub-classed into three different types "single-part, multipart, or serial item (Library of Congress, 2006)," and each copy will have one holdings record, i.e., there are three serial holdings if there are three copies of

https://doi.org/10.23106/dcmi.952137119

the title. Not encoded in MARC, the ILS also has circulation or status of the item data in the system. In addition to these three traditional types of holdings, we also consider items with both print and electronic holdings as a new class. Figure 1 illustrates the types of holdings (and item status) data as displayed in the UIUC Library's Online Public Access Catalog. As shown in Figure 1-1 for the title holdings information of David Daiches' *A Critical History of English Literature*, users can see information about the item's location, the call number, copy number, and available volume information specific to the copy of the title (as well as availability and the status of the item which are not normally encoded in MARC). As illustrated in Figure 1-2, a link to an online copy of the item is also provided if it is available. A barcode of each item is also available in the ILS system but is not displayed to users, because it is not used for the search. According to our analysis of types of holdings data, while a majority (72%) of the titles in the sample set of 5.4 million records are associated only with a single copy of a single-part holding, multiple copies and other holding classes are also represented, and 6% of the titles have both print and online holdings as shown in Figure 2. (Note, we treated links to the full contents included in the data field 856 (Electronic Location and Access) as additional online holdings.)

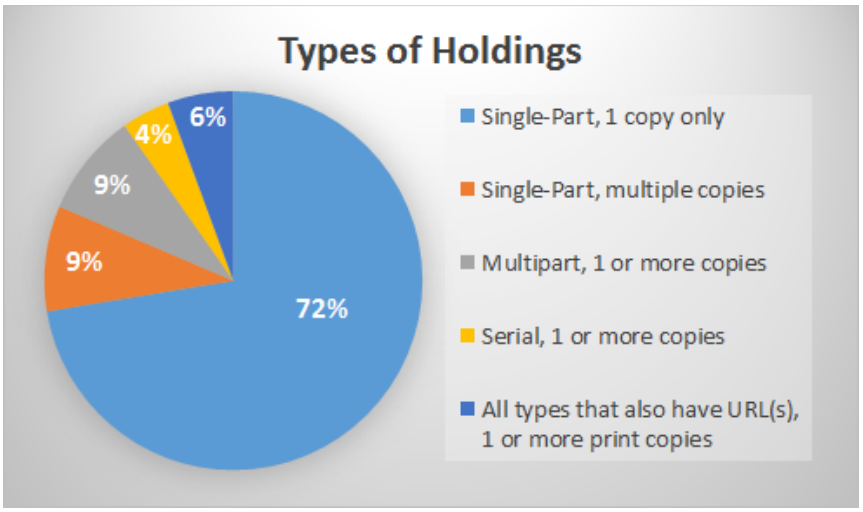

FIG. 2. Types of Holdings associated with UIUC Print Catalog Records

2.2. Relationship Between Bibliographic and Holdings Records

While bibliographic data describes manifestation and higher level information according to the Functional Requirements for Bibliographic Records (FRBR)'s group 1 entities (Tillet, 2004), the holdings data includes both manifestation and item level information. Because one manifestation can link to more than one item, the relationships between bibliographic data, holdings data, and item data may vary depending on the number of copies each library has, or the type of resource, e.g., whether the resource is a monograph, multipart item or serial. For example, one bibliographic record can have more than one holdings record, and each holdings record can have more than one related item with item specific information, such as a barcode and coverage (or volume) information if the item is part of a serial or multipart item. So the relationship between bibliographic record and holdings record(s) is one-to-one or one-to-many, and the relationship between holdings record and item data is also one-to-one or one-to-many. (If an item is a part of series, then the item data may have two different bibliographic records, one that describes a series and the other that describes the specific item.)

Traditionally, the library manages holdings data at the copy level, i.e., the unit of the holdings data is based on the copy number. For example, if the copy is for a multipart item or serial, volume specific information is organized and added under the copy level. So if a user wants to know the availability of a specific issue or volume of the serial or multipart item, the user has to

https://doi.org/10.23106/dcmi.952137119

check the item at copy level first, and then check the volume information from the next layer of the information structure.

3. Holdings Data in schema.org

OCLC decided to use schema.org semantics for expressing bibliographic metadata in RDF for the simple reason that the majority of search engines support schema.org, which means these resources are more easily searchable and discoverable on the web. More than 97% of UIUC's sample set of 5.4 million MARC bibliographic records include OCLC numbers. This allowed us to focus on how best to describe holdings data, i.e., data unique to UIUC that can be easily integrated with the bibliographic LOD graphs already published by OCLC. We acknowledge that the Document Availability Information (Voß and Reh, 2015) group has been working on describing holdings as LOD and developed its own ontology. However, we decided to use schema.org in line with bibliographic data already available in OCLC in order to improve discoverability of holdings data on the web. To create schema.org RDF, we used two transformation stylesheets; a modified version of LC's transformation stylesheet (Library of Congress, 2014) to transform MARCXML records with holdings data to Metadata Object Description Schema (MODS), and a locally created transformation stylesheet to transform MODS to schema.org RDF. We decided to use MODS as a transit metadata format because the MODS top element <location> can properly contain library holdings data, and the schema allows us to include linked data source URIs as values.

Our starting point for this work was the recommendation offered by the BibExtend Community Group (2014). For cases where a bibliographic record is associated with a single copy of a single-part print item (holding), the BibExtend Community Group recommendation works well (see Table 1). With modest modification the recommended "Holdings via Offer" approach allowed us to express most of the critical information contained in this type of holdings data. For instance, we can express the holding organization (schema predicate seller) and branch (availableAtOrFrom) using the Offer class. The IndividualProduct class allows us to express the barcode (serialNumber) and call number (sku). We do, however, deviate from the BibExtend Community Group's recommendation in some particulars; these are further explained in Section 5 below.

TABLE 1: Similarities and differences between the BibExtend Community Group's recommendations and UIUC approaches in mapping holdings and item specific information to schema.org semantics

Holdings and item specific information	**Schema.org/BibExtend Community Group Recommendation**	**UIUC Approaches**
Library	Seller	seller
Shelving Location	availableAtOrFrom	availableAtOrFrom/Place/name
Call number	Sku	itemOffered/IndividualProduct/sku
Item barcode	serialNumber	itemOffered/IndividualProduct/serialNumber
Copy number	Not mapped	itemOffered/IndividualProduct/name
Borrowing terms	businessFunction	Not mapped
Item status	availability	Not mapped
Online Holdings	Not mapped	itemOffered/IndividualProduct /url

4. More Complex Holdings Scenarios

The purpose of exposing the library's holdings data as LOD is to allow users to find, identify, select, and obtain (IFLA, 1998) the exact information resource they need even when it is only available in print. Holdings information can also be a way to filter search results. As illustrated in Figure 2, bibliographic data can be linked to single-part, serial, or multipart holdings as well as to online holdings. A library may hold multiple copies in one or more locations. An item may be available both in print and online as a scanned set of page images. A monographic resource may

https://doi.org/10.23106/dcmi.952137119

have been published in multipart volumes (e.g., a 19th-century triple-decker novel). Serial bibliographic entities are published over time, one issue or volume at a time. Users need access to both the bibliographic information and to exactly which issues and volumes are available from an institution. It is not enough to know that ten libraries in the country own at least some volumes of an obscure journal; a user wants to know which of the ten libraries own volume 4 (the volume that the user needs).

```
schema:hasPart [
  a schema:PublicationVolume ;
  schema:volumeNumber "1972" ;
  schema:offers [
   a schema:AggregateOffer ;
   schema:seller <http://id.loc.gov/authorities/names/n79066210> ;
   schema:itemOffered [
    a schema:SomeProducts ;
    schema:offers [
     a schema:Offer ;
     schema:availableAtOrFrom [
      a schema:Place ;
      schema:name "Oak Street Facility [request only]"
     ] ;
     schema:itemOffered [
      a schema:IndividualProduct ;
      schema:sku "324.23 Un3m" ;
      schema:serialNumber "30112071980053" ;
      schema:name "Copy Number: 1"
     ]
    ], [
     a schema:Offer ;
     schema:availableAtOrFrom [
      a schema:Place ;
      schema:name "Oak Street Facility [request only]"
     ] ;
     schema:itemOffered [
      a schema:IndividualProduct ;
      schema:sku "324.23 Un3m" ;
      schema:serialNumber "30112063348632" ;
      schema:name "Copy Number: 2"
     ]
    ]
   ] ;
   schema:offerCount "2"
  ]
] ;
schema:offers [
            a schema:Offer ;
            schema:seller <http://id.loc.gov/authorities/names/no2015002503> ;
            schema:itemOffered [
             a schema:IndividualProduct ;
             schema:url <http://purl.access.gpo.gov/GPO/LPS6982> ;
             schema:description "electronic resource"
]
```

FIG. 3. Multipart library print and electronic holdings information serialized with schema.org semantics.

To accommodate complex holding scenarios involving multiple copies (i.e., multiple holdings), we employed schema.org's AggregateOffer and SomeProducts types. These were suggested in the BibExtend Community Group recommendation as possibly being of use when describing consortial holdings, but we found them useful for our single institution to deal with multiple copies in different department libraries. This approach also anticipates aggregation of

https://doi.org/10.23106/dcmi.952137119

holdings LOD from multiple sources. Multiple holdings from the same institution– print and local digital copies – are grouped within the same AggregateOffer. Using AggregateOffer allows us to provide the number of copies from each institution through the offerCount property. Local online holdings are also included within an Offer rather than using the url property under the CreativeWork class in order for descriptive information about the electronic copy to be included. For monographic resources published in multiple volumes (multipart) and serial publications, we used the hasPart property under the CreativeWork class. Each volume has a type of PublicationVolume, which allows the enumeration and chronology to be specified using the volumeNumber property. Multiple copies of a single volume appear within an AggregateOffer description. This usage of AggregateOffer and SomeProducts is illustrated in Figure 3.

5. Discussion and Recommendations

5.1. Variations from BibExtend Community Group Recommendations

Our explicit use of IndividualProduct deviates from the BibExtend Community Group recommendation to use sku and serialNumber predicates under Offer class, which is a shortcut way to express the serial number and barcode of the product implied in the Offer. Having Offer as the domain for these properties created problems when we looked at more complex holdings examples. Our bibliographic records with multiple holdings consist of several products, so we decided not to add Product as an additionalType property to the bibliographic record. Instead, we decided to define each item or digital instance as an IndividualProduct.

Another way that we deviated from the BibExtend Community Group's approach was by not including the borrowing terms (businessFunction) or the item status (availability). In both cases we felt the enumeration of possible values for these predicates was insufficient (see below). In addition, we did not include the borrowing terms, which the BibExtend Community Group recommends changing to LeaseOut (http://purl.org/goodrelations/v1#LeaseOut), because some of our holdings are not eligible to be loaned (e.g. non-circulating items) and the definition of LeaseOut does not adequately describe library loan policies.

Because it provides additional identifying information about print volumes, we include the copy number in the IndividualProduct description, something not anticipated by the BibExtend Community Group recommendations.

5.2. Challenges of Working with Holdings Data

Gathering Holdings Data from Various Sources: Information about availability (schema property) is difficult or impossible to acquire through static holdings data, and requires cooperation with a more dynamic and live data source, such as a circulation database. For online access to multipart items and serials, it has proven difficult to express which resources are available through which services and the coverage of the serials. For instance, the Offer information for a journal with electronic issues divided by provider would require harvesting that information outside of the traditional bibliographic and holdings data in an ILS, possibly through close work with vendors or through a separate ERM system maintained elsewhere in the library, in order to correctly display this availability information to users.

Irregular Formatting of Volume Information: Some multipart/serial enumeration and chronology fields in holdings data are irregularly formatted. The value may be a volume, year, or another pattern. For example, one holding may contain volumes 1 and 2 of a serial publication, while another holding for the same record only contains volume 1. In some cases, the information can be completely different from others based on historic binding decisions. This makes it difficult to share serial holdings information across institutions, sometimes even across branches within the same institution. However, we think that this kind of string-based practice can be corrected easily by assigning a permanent identifier for each volume when the item is published.

https://doi.org/10.23106/dcmi.952137119

Data Created Using Different Practices: The way libraries create catalog records has changed over the years, and because of this, catalogs often contain various bibliographic records that follow different rules and standards. For example, the UIUC catalog records include bibliographic records that describe manifestation and expression. The UIUC library has a single record approach when a print book's digital copy is available in open access and the Library has not locally digitized it, i.e., the print record has a url of the digital copy. However, the Library creates a separate record for all purchased electronic and in-house digitized books (separate record approach). The separate record approach can result in disconnected CreativeWork descriptions (one linked to a print item Offer, and one linked to an online Offer) for essentially the same intellectual content. On the other hand, the single record approach results in some CreativeWork descriptions linking simultaneously to print item and online Offers, although the bibliographic data only describes the print copy.

5.3. Representing Holdings Data into RDF

The library manages holdings data at the copy-level and provides available volumes in the next layer. For our experimentation, when we transformed our records into RDF, we changed this relationship. We mapped the holdings data at the volume-level (using hasPart for serial and multipart items), and provided the copy information in the next layer. We think this approach benefits users by allowing them to find the item related data directly from the bibliographic data without searching from the copy level data. The volumes available to a user can be easily expressed with this approach. However, we recognize that this doesn't mean that libraries have to expose their entire holdings and item related data as LOD on the web since not all data that libraries use to manage and organize their resources are beneficial for discovery and access. In addition, some holdings data is not easy to integrate with the data in the ILS and in MARC format.

5.4. Limitations of schema.org For Expressing Holdings

Because schema.org is designed to accommodate structured commercial data, there are instances where schema.org semantics do not align conveniently with library data.

Immediate Availability: While the BibExtend Community Group recommends expressing item availability using: InStock, OutOfStock, PreOrder, or InStoreOnly, this does not capture all possible information about the item status and availability used in the library. It would also be beneficial to our users to further provide availability and accessibility data by adding information describing loan periods or class reserves, but providing this data requires new properties. Additionally terms like OutOfStock do not really describe that an item is currently loaned out and expected back at the end of the loan period. One practitioner has suggested using availabilityStarts to indicate when an item is expected back from loan (Scott, 2014). We recommend better enumeration values for item availability, for example, AvailableToLoan, OnLoan, and InLibraryUseOnly (or RoomUseOnly). We also recommend adding more information such as how long an item can be loaned (eligibleDuration), and for electronic holdings, until when an item can be accessed (validThrough).

Borrowing Terms: The BibExtend Community Group recommends adding the businessFunction property with a value of LeaseOut (http://purl.org/goodrelations/v1#LeaseOut) to describe that a library item is available to be borrowed. While better than the default value, which is for sale, this does not adequately describe library loans, nor does it account for items that cannot be loaned (e.g., in-library use only). We recommend adding more enumerations to borrowing terms (Loan, NonCirculating, Request).

Eligible Customers: Print loan requires a current and valid library ID, and may also require additional conditions be met, e.g., enrollment in class. Customer type may also dictate the loan period or other access constraints. Further complicating the issue of online access is the conditions of use prescribed by various vendors, e.g., requiring a campus IP address, VPN connection, or number of concurrent users. The eligibleCustomerType property in schema.org

https://doi.org/10.23106/dcmi.952137119

expects it to be of type BusinessEntityType, which can have values of Business, Enduser, PublicInstitution, or Reseller from the GoodRelations (http://purl.org/goodrelations/v1) data model. These enumerations are inadequate to describe such information, and the other requirements such as having a library ID or a login account cannot be described. We propose adding an Offer property (requires) to describe the eligibility requirements.

6. Conclusions

Complete holdings data is essential for the user to locate and obtain/access the precise representation or component of the item that the bibliographic data describes, and this unique holdings and item related data can be provided only by an institution that holds the particular information resources. The UIUC Library's research on exposing library holdings data as LOD revealed that holdings data has unique challenges that require a community-wide discussion and collaborative efforts to solve them. Several key elements of holdings data are not encoded in MARC or stored in the same ILS module with bibliographic data. This requires a coordinated effort with ILS vendors as well as publishers. In addition, the way that some item related data is organized and managed must be adjusted based on characteristics of each item, e.g., there is no consistency in representing enumeration/volume information for multipart item or serials. In case of items in special collections (and online resources), the eligibility and availability information is hard to capture and represent as LOD without proper semantics. Additionally, gathering this data requires working with systems where the information is stored and updated dynamically.

While the schema.org and BibExtend Community Group's recommendations provide libraries a good guideline on how to express holdings data as LOD, it is apparent that further discussion and research are also needed to understand how the library has been creating, using, and managing holdings data, in both data structure and systems. Our analysis and experimentation suggests that libraries should change the traditional way of structuring holdings and item data in the library catalog – from inventory focused to discovery and access focused. Differences between types of data that libraries have and the semantics that schema.org has established and the BibExtend Community Group recommends should also be reconciled, possibly in conjunction with the vocabularies for holdings data available in the BibFrame model led by the Library of Congress.

Finally, it would be helpful for libraries to better understand what today's users use and need for holdings and item related data on the web, to locate and obtain/access the information resources they want. Since not all holdings and item related data are useful for discovery and access services on the web, more research is required on which types of information are beneficial for libraries to expose to the web and to establish the holdings data model in LOD.

References

BibExtend Community Group. (2014). Holdings via Offer. Retrieved, April 1, 2015, from https://www.w3.org/community/schemabibex/wiki/Holdings_via_Offer.

Cole, T.W., M.J. Han, W.F. Weathers, and E. Joyner. (2013). Library MARC Records into Linked Open Data: Challenges and Opportunities. Journal of Library Metadata, v.13/Issue 2-3: pp. 163-196.

Library of Congress. (2006). MARC 21 Holdings. Retrieved, April 1, 2015, from http://www.loc.gov/marc/holdings/hdintro.html.

Library of Congress. (2014). MARCXMML to MODS 3.5. Retrieved July 9, 2015, from http://www.loc.gov/standards/mods/v3/MARC21slim2MODS3-5.xsl.

Library of Congress. (2015). Bibliographic Framework Initiative: Bibframe. Retrieved, April 8, 2015, from http://www.loc.gov/bibframe/.

Library of Congress. (2015). MARC Standards. Retrieved, April 1, 2015, from http://www.loc.gov/marc/.

International Federation of Library Associations and Institutions. (1998). Functional Requirements for Bibliographic Records: Final Report. Retrieved, April 1, 2015, from http://archive.ifla.org/VII/s13/frbr/frbr3.htm.

OCLC. (2014). OCLC Releases WorldCat Works as Linked Data. Retrieved, April 1, 2015, from https://www.oclc.org/news/releases/2014/201414dublin.en.html.

https://doi.org/10.23106/dcmi.952137119

OCLC. (2015). WorldCat Work Descriptions. Retrieved, April 1, 2015, from https://www.oclc.org/developer/develop/linked-data/worldcat-entities/worldcat-work-entity.en.html.

schema.org. (2015). What is schema.org? Retrieved, April 8, 2015, from http://schema.org.

Schema Bib Extend Community Group. (2015). Schema Bib Extend Community Group. Retrieved, April 8, 2015, from https://www.w3.org/community/schemabibex/.

Scott, D. (2014). RDFa with schema.org codelab: Library holdings. Retrieved, April 1, 2015, from http://stuff.coffeecode.net/2014/lld_preconference/rdfa_exercises/2_holdings/.

Tillet, B. (2004). What is FRBR? A Conceptual Model for the Bibliographic Universe. Library of Congress. Retrieved, April 1, 2015, from http://www.loc.gov/cds/downloads/FRBR.PDF.

Voß, J. and U. Reh. (2015). Document Availability Information API (DAIA). Retrieved July 9, 2015, from http://gbv.github.io/daiaspec/daia.html.

https://doi.org/10.23106/dcmi.952137119