ARTICLE EXCERPTED FROM:

INFORMATION STANDARDS QUARTERLY WINTER 2013 | VOL 25 | ISSUE 4 | ISSN 1041-0031



TOPIC

EVOLUTION OF BIBLIOGRAPHIC DATA EXCHANGE

ARE CURRENT BIBLIOGRAPHIC MODELS SUITABLE FOR INTEGRATION WITH THE WEB?

A TRANSFORMATIVE OPPORTUNITY: BIBFRAME AT GEORGE WASHINGTON UNIVERSITY, AN EARLY EXPERIMENTER

INTERVIEW WITH GILDAS ILLIEN, BnF

SCHEMA BIB EXTEND





ARE CURRENT BIBLIOGRAPHIC MODELS SUITABLE FOR INTEGRATION WITH THE



Libraries are traditionally seen as the gatekeepers to information. A defined process guides the selection of which information enters the library and the cataloging process creates the metadata necessary for the discovery of (non-digital) resources. The advent of the World Wide Web and full-text search has been a game changer in that online publications and resources are better incorporated into the major general-purpose search engines than is (non-electronic) library material.

To enable a better integration into modern, web-based workflows be it the identification of a book for private reading or the construction of a bibliography for a PhD thesis—it is important that library (meta) data is not only available on the web, but really an integral part of it,¹ thus helping to build what Tim Berners-Lee calls the Giant Global Graph.² Given the structure and rich interlinking of this information, an obvious option to realize this is to publish it as linked data.



The publication of library data as linked data not only helps search engines to improve the findability of library resources, it also makes library data (authorities and bibliographic information) more accessible to organizations outside of the library sphere.

The publication of library data as linked data not only helps search engines to improve the findability of library resources, it also makes library data (authorities and bibliographic information) more accessible to organizations outside of the library sphere. Following the lead of Kungliga biblioteket (the Swedish National Library),³ several libraries—e.g., Országos Széchényi Könyvtár (National Széchényi Library, Hungary),4 the Bibliothèque nationale de France (French National Library, BnF),⁵ the British Library,⁶ the Biblioteca Nacional de España (Spanish National Library),⁷ and the Deutsche Nationalbibliothek (German National Library),⁸ —and library service centers (such as OCLC⁹ or the German library networks) have sparked projects and initiatives to include bibliographic information in the linked data cloud. The transformation of traditional, records-based bibliographic data to RDF¹⁰ made it necessary to deal with the actual semantics of the elements of a bibliographic description. Whereas the translation of some elements was fairly straightforward, other elements posed a major difficulty and revealed that we often are at odds with what bibliographic information actually is and that the bibliographic universe lacks an agreed-upon model. Such a model would have large advantages when it comes to explaining the structure and the value of this information to non-librarians and would also simplify interoperability with data adhering to other models. Currently, however, the main discussion in the library community seems to focus more on the formats (e.g., MARC 21¹¹) than on an underlying model that can be expressed/ serialized in different ways. This focus on the format is insofar counter-productive in that it tends to encourage the use of literals (strings) without analyzing what the information is about and how it relates to other pieces of information (things)—within or outside of a specific bibliographic description. Further, the preoccupation with data in the context of a particular format tends to prevent real innovation, since it is more focused on carrying the existing data forward than on analyzing which data would be necessary for what operation. A shift to a more model-driven view on bibliographic information would increase the possibilities to interlink the individual parts of a bibliographic record to other entities outside of the library domain, particularly within the cultural heritage sector, but also in settings like academia and e-commerce.

The bibliographic data itself

The bibliographic world still very much mirrors the card catalogs. The problem is that the card display was not built around the concept of pivot points (e.g., authorities) but for sequential display organized according to certain criteria (title, headings). ISBD,¹² the format for sharing bibliographic information in a standard, human readable form, has an inner structure and groups the description elements into eight distinct areas composed of multiple elements. But it still focuses very much on the bibliographic record and does not build on an explicit model based on entities and their relations. Many linked data representations of bibliographic data—e.g., the recently published DINI-KIM recommendation for the RDF representation of bibliographic information¹³—still mimic the traditional record-based structure and are more an application profile aiming to provide an easy-to-implement bridge from the library world into the linked data domain than an actual bibliographic model.

There are currently several initiatives working on creating a recognized model for bibliographic information. The most well-known is probably IFLA's Functional Requirements for Bibliographic Records (FRBR)¹⁴ where the entities in the bibliographic universe are first separated into three groups (bibliographic, authority, and topic) and then within the first group into work, expression, manifestation, and item. FRBR is a well-recognized model that was developed from the user tasks of find, identify, select, and obtain. The model is not without problems and there is work underway in IFLA to improve it and also to harmonize it with the other members of the IFLA FR* family: Functional Requirements for Authority Data (FRAD)¹⁵ and Functional Requirements for Subject Authority Data (FRSAD).¹⁶ Nonetheless, the FRBR approach to group elements and properties common to different versions of the same publication obviously struck a chord with the semantic web community as shown by the transformation of FRBR into RDF by Ian Davies and Richard Newman in 2005.17 The FRBR model was later adopted by the upcoming cataloging code RDA,18 and the European Commission's CESAR service19 uses FRBR concepts to model the publication of semantic assets in different revisions and formats. Further, research has shown that users intuitively relate specific abstractions of a bibliographic description to the appropriate FRBR group 1 entity.^{20,21} RDA is currently in the process of defining relations between the entities that go beyond what FRBR specifies and given that the archives' community is interested in adopting RDA standards, RDA has the potential to serve as a common foundation for data models in the cultural heritage communities.

Another major initiative for modeling cultural heritage data is CIDOC-CRM²² which is an event-based model originally designed for museum materials. There has been work undertaken to harmonize FRBR and CIDOC-CRM through FRBRoo, "a formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information".23 Even if some institutions use CIDOC-CRM (e.g., WissKI^{24,25}) and FRBR (e.g., BnF) as models for their electronic services, it is important to bear in mind that both are conceptual models, and that it might not be intended to implement them verbatim. Instead we should look at them as what they are-modelsand discuss what elements and relations are useful in which context, as in the Europeana Data Model (EDM)²⁶ used by europeana²⁷ and serving as the basis for the data model of the German Digital Library (DDB),²⁸ and how we can encode the instances of our models in an interoperable fashion using widely agreed-upon exchange formats.

The conflation of model and exchange format becomes very visible in the work of the BIBFRAME initiative.²⁹ The primer declares that a "major focus of the initiative will be to determine a transition path for the MARC 21 exchange format to more Web based, Linked Data standards" and talks about the initiative as "Bibliographic Framework as a Linked Data Model".³⁰ In the introduction it is stated that the "goal of this initial draft is to provide a pattern for modeling both future resources and bibliographic assets traditionally encoded in MARC 21." Indeed the intention seems to be to create a complete replacement to MARC 21 as a format, both as an exchange format, as a cataloging format, and as the internal format of integrated library systems.³¹ Further, BIBFRAME is



In order to transport instance data adhering to other models, the plan is to create community profiles that map the community model to the BIBFRAME model.

intended to be both "rule agnostic" (i.e., not tied to a particular cataloging code) and "model agnostic" (i.e., flexible enough to accommodate both "flat" record-based as well as highly interlinked FRBRized data). But whereas the first of those two is relatively easily achieved, the de-coupling from any specific model is questionable. The BIBFRAME architects have chosen not to adopt FRBR (the reasons are not completely clear, but it seems that they consider FRBR too complex for the Semantic Web) but instead they have created their own model based on the entity types Creative Work, Instance, Authority, and Annotation. In order to transport instance data adhering to other models, the plan is to create community profiles that map the community model to the BIBFRAME model.³² A complete round-trip transformation of data between two models, however, is only possible if both models are equally granular and their entity types and relations have (approximately) the same semantics (in which case it is questionable why there are two different models in use). If not, there will be a loss of specificity when transforming in either direction. Within BIBFRAME, the focus seems to be equally on the format and on the model. This is not explicitly stated in the BIBFRAME documents, but the use of concrete XML syntax to illustrate core concepts and relations gives the format (syntax) an unnecessary emphasis that occasionally puts the actual model in the background. The use of XML instead of RDF serializations (e.g., RDF/XML³³ or Turtle³⁴), since "support for RDF is not yet as widespread as support for XML,"35 is a valid argument when looking at actual implementations and data transfer. If the intention is to focus on the model, however, it would be preferable to have a graphic notation showing the entities and how they are connected and give examples for how this construct can be expressed in several serializations, including at least one RDF syntax.

CONTINUED »

FE



In the LoC-NAF, John Lewis Burckhardt is both a name (*madsrdf:PersonalName*) and a *skos:Concept*. This is in line with cataloging tradition, but a non-librarian would be surprised that the *rdf:type* is not, for example, *foaf:Person*.

CONTINUED »

Another discussion of entity types and their relations is currently taking place within the scope of Schema.org bibliographic extension group.³⁶ In contrast to most library initiatives that model top-down, Schema.org takes a bottom-up approach when incorporating new resource types into their vocabulary. The discussion within the group focuses on what constitutes a specific entity type (e.g., eBook), what are its specific properties, and what properties does it have in common with other entity types so that they can be generalized to a higher level in the hierarchy. It has been argued that the schema.org ontology "is deep enough to create rich and subtle descriptions of many library resources and the events that impact them,"³⁷ which might be true or not, depending on whom the data is intended for: e.g., the bibliographic description necessary for a national bibliography is different from the one needed for a freshman course reading list.

Authorities

A case where library models sometimes differ from what customers might expect are the non-bibliographic items such as people, places, and things that bibliographic descriptions often rely on-the authorities. In the Anglo-American cataloging tradition, the role of the authority is to provide a unique name or a unique heading for an entity that can then be used consistently throughout the catalog. Since the advent of electronic cataloging, libraries, library networks, and (national) bibliographic agencies have collected authorities into authority files that were first distributed directly to interested parties (e.g., other libraries) on magnetic tapes and now increasingly are published in RDF in order to make the data reusable for parties outside of the library sphere.

Those RDF-based authority services are often a core part of a library's linked data service since the authority data acts as a hub for all information relating to a specific entity (e.g.,

a person or a particular topic). One example is the Library of Congress's service id.loc.gov,³⁸ where the LoC publishes a rich set of the commonly-used authority data and value vocabularies that it maintains. An inspection of the site and of some of the descriptions reveals several points where the model used differs from what the non-library community might expect. As an example we can look at the following piece of RDF about the publication Travels in Nubia by John Lewis Burckhardt:

dnb:956706967 a bibo:Book ; dc:title "Travels in Nubia"@en ; dct:creator lc-naf:n50045595 ; dct:subject lc-naf:n81103291 . lc-naf:n50045595 rdfs:label "John Lewis Burckhardt"@en . lc-naf:n81103291 rdfs:label "Nubia"@en .

Without deeper knowledge of the library domain, a developer would intuitively assume that lc-naf:n50045595 identifies a person (books are written by people) and that lc-naf:n81103291 identifies a place (in this case Nubia). The actual data, however, reveals another world-view:

In the LoC-NAF, John Lewis Burckhardt is both a name (madsrdf:PersonalName) and a skos:Concept. This is in line with cataloging tradition, but a non-librarian would be surprised that the rdf:type is not, for example, foaf:Person. There is a hint in the description that the entity described identifies a RWO (real world object) of type foaf:Person, but in order to find the description of that person you need to follow the link to the external authority in VIAF.³⁹

<http://viaf.org/viaf/sourceID/

LC%7Cn++50045595#skos:Concept> a skos :Concept ; skos:prefLabel "Burckhardt, John Lewis, 1784-1817" ; foaf:focus <http://viaf.org/viaf/59176329> . <http://viaf.org/viaf/59176329> a foaf:Person, rdaGrlEntities:Person .

The description of Nubia in the LC NAF might be even more confusing to a non-librarian since the main rdf:type given is madsrdf:Geographic which suggests that the URI lc-naf:n81103291 identifies a geographic area. Again, however, lc-naf:n81103291 is a madsrdf:Authority and a skos:Concept and it is only through the link to VIAF that we can find out that it is linked to a dbpedia:Place.

Another approach was taken by the German National Library (Deutsche Nationalbibliothek, DNB) and the German library networks when in a cooperative project they revamped the authority files used in the German-speaking countries. Until April 2012, descriptions about persons, topics, geographic areas, corporate bodies, and work titles were kept in four separate authority files. When designing the new, common authority format for the Integrated Authority File (Gemeinsame Normdatei, GND),⁴⁰ one of the requirements was that the data model should be directly reusable in the DNB's linked data services in order to expose the information in the authority file better on the web and allow third parties to more easily reuse that information. The result of the design process was an entity-based model featuring seven different types: Corporate Body, Conference or Event, Topic, Work, Place or Geographic Name, Personal Name, and Person. A core feature of that model is that the URIs for the entities in the GND identify, as far as possible, the world objects (e.g., persons, places, or corporate bodies) and only for subject headings, for example, the authority is a concept instead of the RWO. Further, the model reduces redundancy in that there is only one record (and thus one URI) for each entity regardless of the number of roles it can occur in. As an example, we refer to the German author Hermann Hesse using the same URI both for Hesse as an author and for Hesse as the topic of a PhD thesis or a biography.

Using the GND data model, the representation of John Lewis Burckhardt makes it obvious that the authority record is about the real person and does not only model his name, which is more in line with the general expectation that if you dereference a URI pointing to a book's author, you will retrieve a representation of the actual person (or agent), not only a representation of its name:

BIBFRAME introduces a so-called lightweight abstraction layer for representing authorities in their model. Those authorities are local to a specific library's data but can be linked to other commonly used authority providers such as id.loc.gov, GND, or RAMEAU.⁴¹ The authorities in BIBFRAME's lightweight abstraction layer are identified by URIs but again the authority's URI does not denote the real thing; it only denotes the authority record, thus introducing an extra, non-intuitive level of indirection. Since BIBFRAME is still very much a work-in-progress, it cannot be anticipated if there are plans to switch to a different view on authorities. On the other hand, the model used by VIAF very nicely bridges the two views of how to represent real-world objects in library data. For people from outside the library domain, however, a common model would simplify the understanding.

Discussion

The library community needs to enter into a deeper discussion on the actual semantics of bibliographic descriptions. In order to create descriptions where the various parts can be reused outside of the library domain, we need an entity-centric model based on real-world objects (as in the GND) and not on traditional library authorities. In order to find a suitable model for the bibliographic information, more research is necessary. We can expect that different serialization formats with various levels of granularity will be necessary depending on the target application, but in order to retain interoperability a common



It is noticeable that the most problematic entity in the FRBR model the expression—also is one that is core to the user task "find": users often search for a specific text in a certain language and then in the next step pick the edition (manifestation) of their choice, be it hardcover, paperback, or e-book.

CONTINUED »

conceptual model will be necessary. It has been argued⁴² that we can make models interoperable by using vocabulary alignment and rdfs/owl reasoning but this requires that the semantics of the aligned elements are very similar in order to achieve true interoperability and then the question is if it would not be better to reuse the other vocabulary anyway. Further, the applicability of this approach has not been tested in a large-scale setting where data adhering to several different models is brought together.

The BIBFRAME approach is so far very promising in that it thoroughly analyzes the existing data and builds its model from that. The discussion, however, seems too much focused on replacing MARC 21 as both a cataloging and an exchange format instead of analyzing the elements of bibliographic descriptions in the light of its constituting parts and entities. It is noticeable that the most problematic entity in the FRBR model—the expression—also is one that is core to the user task "find": users often search for a specific text in a certain language and then in the next step pick the edition (manifestation) of their choice, be it hardcover, paperback, or e-book. Further, the FRBR Work level overlaps with work descriptions in authority data and can enhance the value and the reusability of those descriptions.

On the way to the future model, we will have to deal with some elements of bibliographic descriptions where the semantics are extremely fuzzy. The best example is the publication statement (e.g., London: Topographical Society, 1898), which merely is a transcription of information found on a publication's title page and where the exact meaning of the parts is not clear. It can be argued that in the example above the string "London" refers to the real place London (the capital of the United Kingdom) and that this denotes the place of business of the Topographical Society. When confronted with a publication statement like "Berlin, Heidelberg, New York, NY, London, Paris, Tokyo, Hong Kong: Springer, 1990"⁴³ the question arises if the strings "Berlin," "Heidelberg," etc. really denote place names and what is their relation to the publisher "Springer." A future bibliographic model needs to clarify what entities are involved in the publication statement and if this can be modeled using corporate bodies from a library authority file.

Conclusion

The publication of bibliographic information as linked data has left the laboratory and is increasingly entering a stage where it is part of everyday library operations. The work done so far clearly shows that there is no one-size-fits-all model for bibliographic information. In order to replace the current records-based model with one that allows library information to be reused in other settings and also allows libraries to make better use of data originating outside of the library domain, it is necessary to agree on a common model that reduces the complexity of that data integration. To build such a model, librarians—as the domain experts—need to cooperate with potential data consumers from industry and from other cultural heritage institutions.



LARS G. SVENSSON (L.Svensson@dnb.de) is Advisor for Knowledge Networking at the German National Library (Deutsche Nationalbibliothek, DNB).

ORCID: http://orcid.org/0000-0002-8714-9718

Acknowledgements

The author is very much indebted to Reinhard Altenhöner, Christine Frodl, Julia Hauser, Reinhold Heuvelmann and Brigitte Wiechmann for their comments on this article.

REFERENCES

- Kett, Jürgen, Sarah Beyer, Mathias Manecke, Yvonne Jahns, and Lars G. Svensson. The Deutsche Nationalbibliografie as linked open data: applications and opportunities. Presented at 78th IFLA General Conference and Assembly, Helsinki, Finland, August 11-17, 2012. http://nbn-resolving.de/urn:nbn:de:101-2012052306
- 2. Berners-Lee, Tim. Giant Global Graph. http://dig.csail.mit.edu/breadcrumbs/node/215
- Malmsten, Martin. Making a Library Catalogue Part of the Semantic Web. DC-2008 - Berlin Proceedings.
- http://dcpapers.dublincore.org/pubs/article/view/927
 4. National Széchényi Library (National Library of Hungary) on the semantic web http://nektar.oszk.hu/wiki/Semantic web
- 5. Bibliothèque nationale de France (French National Library) semantic data http://data.bnf.fr/
- 6. British Library linked open data http://www.bl.uk/bibliographic/datafree.html#lod
- Biblioteca Nacional de España (Spanish National Library) linked data http://datos.bne.es/
- Deutsche Nationalbibliothek (German National Library) linked data service http://www.dnb.de/EN/lds
- Ted Fons, Jeff Penka, and Richard Wallis. "OCLC's Linked Data Initiative: Using Schema.org to Make Library Data Relevant on the Web." Information Standards Quarterly, 2012 Spring/Summer, 24 (2/3): 29-33. http://dx.doi.org/10.3789/isqv24n2-3.2012.05
- Resource Description Framework (RDF) http://www.w3.org/standards/techs/rdf#w3c_all
- 11. MARC 21 http://www.loc.gov/marc/
- International Standard Bibliographic Description (ISBD) http://www.ifla.org/publications/international-standard-bibliographicdescription
- Empfehlungen zur RDF-Repräsentation bibliografischer Daten [Recommendation for the RDF representation of bibliographic data], v.1.0, September 2013. http://edoc.hu-berlin.de/series/dini-schriften/2013-14/ PDF/14.pdf
- Functional Requirements for Bibliographic Records (FRBR) http://www.ifla.org/publications/functional-requirements-forbibliographic-records
- Functional Requirements for Authority Data (FRAD) http://www.ifla.org/publications/functional-requirements-for-authority-data
- 16. Functional Requirements for Subject Authority Data (FRSAD) http://www.ifla.org/news/functional-requirements-for-subject-authoritydata-frsad-final-report-published
- 17. Davis, Ian, and Richard Newman. Expression of Core FRBR Concepts in RDF. http://vocab.org/frbr/
- 18. **RDA: Resource Description and Access** http://www.rda-jsc.org/rda.html
- Community of European Semantic Asset Repositories (CESAR) https://joinup.ec.europa.eu/community/cesar/description
- Pisanski, Jan, and Maja Žumer. "Mental models of the bibliographic universe. Part 1: mental models of descriptions." *Journal of Documentation*, 2010, 66 (5): 643 - 667.
 - http://dx.doi.org/10.1108/00220411011066772
- Pisanski, Jan, and Maja Žumer. "User verification of the FRBR conceptual model." Journal of Documentation, 2012, 68 (4): 582 - 592. http://dx.doi.org/10.1108/00220411211239129
- 22. CIDOC Conceptual Reference Model (CRM) http://www.cidoc-crm.org/

- 23. FRBRoo [FRBR object-oriented] http://www.cidoc-crm.org/frbr_inro.html
- 24. WissKI http://wiss-ki.eu
- 25. Hohmann, Georg. "Die Anwendung des CIDOC CRM für die semantische Wissensrepräsentation in den Kulturwissenschaften [The use of CIDOC-CRM for semantic knowledge representation in the humanities]." In: Sieglerschmidt, Jörn and Peter H. Ohly, eds. Wissensspeicher in digitalen Räumen: Nachhaltigkeit, Verfügbarkeit, semantische Interoperabilität. Proceedings der 11. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation Konstanz [Knowledge storage in digital space: persistence, availability, semantic interoperability]. Würzburg: Ergon, 2010, pp. 210-222.
- 26. Europeana Data Model (EDM) Documentation http://pro.europeana.eu/edm-documentation
- 27. Europeana http://europeana.eu/
- 28. Duetsche Digitale Bibliothek http://www.ddb.de/
- 29. BIBFRAME: Bibliographic Framework Initiative. Library of Congress. http://bibframe.org/
- Bibliographic Framework as a Web of Data [BIBFRAME model primer]. Library of Congress, November 21, 2012. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf
- 31. See for example, the e-mail from Eric Miller on the BIBFRAME list where he agrees to the notions that the BIBFRAME format must have the capacity to function as an internal format, and that it must be a replacement for both carrier and content formats.
- http://listserv.loc.gov/cgi-bin/wa?A2=indt310&L=bibframe&T=0&P=2403 32. Miller, Eric. BIBFRAME Community Profiles [slides]. Presented at LC
- Bibliographic Framework Initiative Update Forum, June30, 2013. http://de.slideshare.net/zepheiraorg/alaBIBFRAME-Ic20130630
- RDF/XML Syntax Specification (Revised). W3C Recommendation, February 10, 2004. http://www.w3.org/TR/2004/REC-rdf-syntaxgrammar-20040210/
- 34. Turtle: Terse RDF Triple Language. W3C Candidate Recommendation, February 19, 2013. http://www.w3.org/TR/turtle/
- 35. BIBFRAME model primer, op. cit. [no. 30], p. 25.
- 36. Schema Bib Extend Community Group http://www.w3.org/community/schemabibex/
- Godby, Carol Jean. The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper. OCLC, June 2013. http://www.oclc.org/content/dam/research/publications/ library/2013/2013-05.pdf
- 38. LC Linked Data Service: Authorities and Vocabularies http://id.loc.gov
- 39. VIAF: The Virtual International Authority File http://viaf.org/
- 40. Integrated Authority File (Gemeinsame Normdatei, GND) http://www.dnb.de/EN/gnd
- RAMEAU: Répertoire d'autorité-matière encyclopédique et alphabétique unifié http://rameau.bnf.fr/
- Dunsire, Gordon, Diane Hillmann, Jon Phipps, and Karen Coyle. A Reconsideration of Mapping in a Semantic World. DC-2011, The Hague proceedings. http://dcevents.dublincore.org/index.php/IntConf/dc-2011/ paper/view/52/6
- 43. DNB catalog record for cited example: http://d-nb.info/891684808