Serials, FRBR, and Library Linked Data: A Way Forward

Laura Krier
Principal Metadata Analyst
California Digital Library

Short title: Serials, FRBR, and Library Linked Data

Abstract: This article proposes a new way of cataloging serials using linked data and Resource Description Framework (RDF), as well as how the concepts of Functional Requirements for Bibliographic Records (FRBR) can be expanded to apply to journal content at both the journal level and the article level, all with an eye toward ease of access and understanding for users.

Statement: This article has not been published elsewhere, nor submitted simultaneously for publication elsewhere.

Serials, FRBR, and Library Linked Data: A Way Forward

The practice of cataloging has not changed significantly over the last century. Catalogers still create records that describe objects in the library's collection. Libraries maintain local databases that attempt to detail the circumference of the knowledge the library possesses. But as information resources become ever more complicated, these practices begin to feel constrained. The traditional functions of the cataloger might not be sufficient to bring libraries into a world where resources are networked and scholarly communication is a changing practice. Library ownership of objects is giving way to licensing of digital resources. Scholarly communication is slowly shifting to an open access model, wherein the library no longer needs to be a financial and digital gateway for research. More and more, libraries are moving into a world where traditional descriptive cataloging will not meet users' needs. The functions of cataloging will likely change significantly: Instead of providing descriptions of static resources, catalogs will provide links to a dynamic world of digital information, helping patrons make connections and explore. Serials management is one area that could specifically benefit from a change in the function of bibliographic control.

In the early twentieth century librarians made the decision not to catalog the articles published in scholarly journals. The business of creating article-level metadata was given over to abstracting and indexing services like H. W. Wilson, and library metadata was thereafter split between different library systems: the catalog and published abstracts and indexes. This made some degree of sense when librarians were the parties responsible for locating resources for patrons, but with the advent of the Web, and the increasing ease of access to networked resources, patrons have begun to take their information searching into their own hands. The split between catalogs and journal databases is not intuitive to patrons, and librarians have been trying for the past few decades to figure out how to unify these resources in ways that are easier to understand and navigate. Federated search engines have been implemented, but they present problems around the quality of searches (Baer, 2004). Discovery layers like VuFind and AquaBrowser were built to provide access to library resources across multiple systems, but their effectiveness depends on the existence of rich metadata in a variety of environments.

In recent years, Web-scale discovery systems, including EBSCO's Discovery Service and Serials Solution's Summon, have appeared in the library environment, promising to solve data silo problems by providing a single point of access to all library materials. But these systems rely on unstable means of bringing together diverse resources. Data is aggregated into a discovery system based on the vendor's agreements with publishers, and the vendor's ownership of abstracting and indexing services. A library's ability to access the entirety of the resources owned or licensed through the discovery system is dependent upon the whims, deep pockets, and negotiating abilities of the vendor. Current discovery systems are not built using a standard for aggregating article-level and traditional bibliographic metadata, and our existing metadata doesn't adequately serve this purpose.

In order to effectively develop library discovery systems that incorporate article-level, journal-level, and monograph resources, developers and catalogers need to start thinking about how we represent these resources and the relationships between them in new ways. Librarians have long struggled with how to present these relationships to users and represent them in bibliographic and authority records. Many articles have been written that explore the relationship between a print journal and its digital counterpart, the relationship of a particular journal title to volumes published under a previous title, and the relationship of an article to the journal in which it was published. For decades catalogers have searched for answers to the problem of multiple versions or format variation. Graham (1989) presented a paper at the Multiple Versions Forum in which she identified the developments that have led to the "crisis" of multiple versions. The proliferation of new formats and the need to create new records for each format have resulted in cataloging bottlenecks and this has confused patrons trying to access resources and libraries involved in resource sharing. Graham argues that existing cataloging practices "have hopelessly intertwined the terminology for the physical pieces and the ideas they contain" (p.11). Although Graham gives some recommendations for differentiating between works and versions, she doesn't suggest any significant cataloging changes that might effectively solve the problems.

Oliver (2004) explores the impact of the Functional Requirements for Bibliographic Records (FRBR) on the format variation issue by detailing the single record and the separate record approaches: in the single record approach, a cataloger would create a single MARC record in which all manifestations of the work are described; the separate record approach requires each format to be described by its own record. Oliver discusses how these different approaches for serials might work with the FRBR model and concludes that "FRBR offers another way of looking at the same problem and indicates a way to resolve it," and proposes that notes, linking fields, and uniform titles might be sufficient to pull together different manifestations of a title (p. 35)

Riva (2004a) also details the ways in which FRBR can help solve the multiple versions dilemma by suggesting that linking entry fields can be put to use to collocate manifestations. She points out, however, that there are significant differences in the scope of the "three distinct categorizations of bibliographic relationships" that FRBR details (p. 138). Riva suggests that "understanding how precisely MARC 21 coding maps to theoretical taxonomies of bibliographic relationships can be a consideration in future format development" (p. 138).

The solutions and ideas suggested by these authors all assume libraries will continue to operate in a record-based data environment. Each presents solutions for a MARC-based library system, but this is a system that will always struggle to adequately deal with networked library resources. In their examinations of the multiple version problem, they all acknowledge the difficulty, if not the impossibility, of adequately resolving the problem in the current MARC-based cataloging environment. However, there is an open and reliable way to pull together disparate sources of bibliographic metadata through the use of Resource Description Framework (RDF) and linked data. RDF is a Web language used to represent information about resources "that can be *identified* on the Web, even when they cannot be directly *retrieved* on the Web" (Manola & Miller, 2004a, para. 1).

Linked data practices may offer a way to answer some of these questions, and to change cataloging practices by defining relationships between entities, instead of creating discrete records for static displays of data. It can be used to build links between many different types of metadata, and to manipulate that metadata in ways that work best for users at any given point in time. However, shifting to this model entails a radically different concept of library resources and library data.

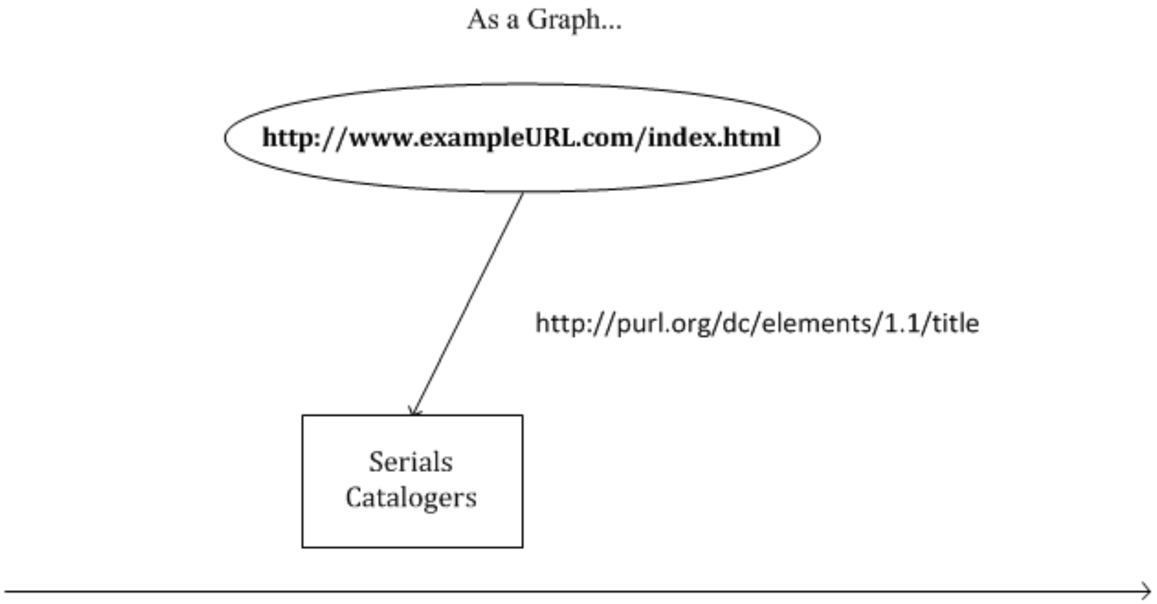
This paper looks at the use of linked data in libraries and how it can be used to catalog serials in order to bring together all library resources in Web-scale discovery systems. It looks at some of the ways people have tried to incorporate serials into the Functional Requirements for Bibliographic Records (FRBR) model. FRBR can be expanded to apply to journal content not just at the journal title level, but also at the article level. These problems are examined with an eye toward ease of access for patrons, and openness of metadata for widespread use and re-use. Finally, the paper explores the ways that linked data can be used to show the relationships between various library resources and allow patrons easier access to the information they need, and proposes some ideas about the kinds of discovery systems that might be built to best meet both librarian and patron needs.

Linked data refers to structured data and metadata that is published in a way that allows links to be created between various data sets, element sets, and value vocabularies. It allows data to be read by computers, and to be re-used and extended in a variety of ways. Linked data challenges traditional models of bibliographic metadata, because the data model differs markedly. Traditional library data lives in discrete records: each record contains a set of elements (e.g. MARC tags) and their associated values, and each record describes a unique resource. Linked data is not record-based; it is based on a graph data model. The graph data model centers around statements, not records. Baker, et al. (2011) writes that "in a graph-based ecosystem an organization can supply individual statements about a resource, and all statements provided about a particular uniquely identified resource can be aggregated into a global graph" (para. 9).

In a linked data ecosystem, each resource, element, and vocabulary term is assigned a unique identifier, a Uniform Resource Identifier (URI) that will allow a resource to be accessed using the protocol of the Web: Hypertext Transfer Protocol (HTTP). These identifiers allow a resource to be used unambiguously in a variety of different places on the open Web.

In RDF "statements may be either URIrefs, or constant values (called literals) represented by character strings, in order to represent certain kinds of property values" (Manola & Miller, 2004b, para. 7). Statements are in the form of subject, predicate, and object. These statements can be in the form of a graph or, when not convenient, "each statement in the graph is written as a simple triple of subject, predicate, and object, in that order" (para. 8). For example, Figure 1 below shows both a graph example and a typical 'triples' for a fictional journal named 'Serials Catalogers' constructed in the same manner as the examples found in the RDF Primer.

Figure 1 Example of linked data with RDF statements


As a Graph...
http://www.exampleURL.com/index.html
http://purl.org/dc/elements/1.1/title
Serials
Catalogers

As a Triples….

<http://www.exampleURL.com/index.html>< http://purl.org/dc/elements/1.1/title><Serials Catalogers>

What does this mean for bibliographic metadata? An item to be cataloged as a resource is assigned a URI that is available on the open Web. A cataloger would then use element sets such as the Dublin Core Metadata Initiative terms, the International Standard Bibliographic Description (ISDB) terms, or FRBR concepts in RDF to describe that resource by making statements about it. Value vocabularies like the Library of Congress Subject Headings (LCSH) and Getty Union List of Artist Names are used, as they are in traditional cataloging, to ensure uniformity in bibliographic description and to help collocate items. Many bibliographic resources, elements sets, and vocabularies have already been published in RDF on the Web, and are ready for use in bibliographic descriptions.

The significant difference between traditional record-based cataloging and cataloging in a linked data model revolves around the existence of discrete records in local databases: In a linked data model, records largely disappear. Rather than downloading and editing, or creating, a new record for each item added to the collection, a cataloger would find data already available about an item, and make statements that link the item to the library, indicating that it is held in the library's collection. The cataloger could also publish any locally-specific notes as additional RDF statements. Library systems would in turn pull data from many places on the Web to dynamically assemble a display for a user. Different data elements could be pulled together depending on the need of the user; no single, consistent record exists, but a "record" is created at the point of need. In this model, Integrated Library Systems would cease to consist of databases storing records locally and would instead become editors and servers of RDF statements. Rather than maintain redundant sets of records at every library, librarians could store bibliographic data in centralized databases, maintained collectively and used as needed by individual libraries and users.

The linked data model works especially well for serials, where the relationships are some of the more complicated in the bibliographic universe. Even before the advent of electronic journals and the multiple versions (or MulVer) problem, serials presented issues for catalogers through frequent title changes, publisher changes, and shifts in scope (Everett Allgood, 2006). When serial titles began to be published in multiple different formats, the problem became more intractable, and existing cataloging practices were 'contorted' to try to accommodate users' needs. Catalogers went back and forth between creating single records for each serial title, attaching holdings records for each unique manifestation, and creating a new bibliographic record for each format available. The changes in cataloging practice over time resulted in catalogs containing inconsistent metadata, where user confusion was practically guaranteed.

A shift away from the record data model for bibliographic metadata could more readily accommodate the serials environment that exists today. As resources become more and more available in an electronic format, description becomes less important than identification, and the relationships between resources become crucial (Antelman, 2004). Users need links to resources more than they need descriptions of those resources.

The International Federation of Library Associations and Institutions (IFLA) recognized the importance of bibliographic relationships when it drafted the FRBR model. FRBR is centered around identifying relationships between bibliographic resources, and the metadata that help people to discover those resources. The Group 1 entities of work, expression, manifestation, and item are intended to allow different versions of a particular bibliographic resource to be more easily collocated. However, it was clear from the start that the FRBR model works well for monograph resources but not necessarily for modeling continuing resources. Over the past decade, several people have proposed changes to the FRBR model to incorporate serials. Adams, Santamauro, & Blythe (2008) draw on the work of Frieda Rosenberg and Diane Hillman, and suggest that in the serials world, there might be three levels, rather than four, in the Group 1 entities. They suggest that work and expression might be combined into a *superworkspression*, as “an umbrella record that collects the bibliographic information relating to a serial’s content...serving all its manifestations in different formats” (p. 195). This record would change infrequently; the manifestation record would reflect the data specific to different formats. Riva (2004b) suggests relationship-based clusters created at the work level to show the relationships between serials as they shift over time. Curran (2009) suggests that perhaps each serial change is itself a new manifestation.

The problem with the FRBR model for serials is that serials publications contain two related entities that could both be identified on the work level: the journal itself, and each individual article within a particular issue. The separation between the journal and the articles it contains has been deeply embedded in cataloging practices for over century, and current ideas about bibliographic relationships haven’t adequately addressed the issue of trying to bring these two work-level resources together.

However, if FRBR is viewed as a conceptual model for defining relationships, rather than a strict template, the model can expand in several directions. If the relationships between resources are re-conceptualized as the central component of cataloging, then the links between resources created using the RDF model become the engine that drives collocation and discovery. The journal and its articles can be brought back together in discovery systems, making it easier for users to find what they need no matter how they begin their search.

If the journal and article can both be considered as the “work,” users can begin searches from any entry point, and displays of “records” can shift depending on the ways a user navigates through bibliographic metadata. Beginning at the article level, a particular article, in an abstract sense, is defined as a work. The expression of the work might be the article as written by the author, and in a particular language. The first manifestation of the article could be the print version as published in a particular journal and then a second if it published as a digital version. If the article goes on to be re-published in a monograph anthology, another manifestation (and possibly another expression) is created, and can be linked to the article as published in the journal.

The work can also be defined from the journal title level; a user can navigate from a journal title search to see all articles contained in individual issues of the journal. This introduces a new question: What is a serial work? How the boundaries of a particular journal have been defined has changed over time: the debate between latest-entry and successive-entry cataloging illustrates the shifts in thinking around this issue. Latest-entry cataloging, in which title and publisher changes for a particular journal are made in the main entry fields of a MARC record, with the previous title moved to a former title field and indexed in the library system, suggests that the serial as a whole constitutes the work, each part of its history contained within the work and subsumed to its current incarnation. The shift to successive-entry cataloging, in which a new record is created with each change in title, signified a different way of thinking about the serial as a work. The new title becomes a discrete resource, separate from its earlier incarnations. The gap between these two models could be bridged with the concept of the serial work in FRBR: if one considers the journal family, encompassing its whole history, as a work, each title change or change in scope becomes a new expression, and an expansion of the journal family. Each manifestation of the journal, with a new title or publisher, can be cataloged as a separate entity, and linked to its previous incarnations through a unique identifier assigned to the journal family.

The concept of linking journals together as a family already exists in the MARC record, through the use of 76X-78X linking fields. However, attempting to create these links in an isolated library system is flawed because the single library might not hold the entirety of the journal family: the links are easily broken (Riva, 2004b). If bibliographic data is shifted away from the isolated library system and into a network of bibliographic data, these links can be maintained, and resources can be more effectively grouped together for easier discovery. Data can be pulled in when needed, and disregarded when it is not, based on point-of-use needs.

The current record-based bibliographic model requires that the cataloger make a decision about what constitutes the work: the journal or the article. A linked data model will allow either article or journal to be positioned as the work, depending on the needs of the user at a given point in time. Rather than treating resources as static entities, the linked data model allows for a fluidity that reflects the real existence of these resources in a networked environment. The following example illustrates how this might work in a discovery system.

A user has a citation in hand for a particular journal article: “Sweet Revolution” by Adam Gopnik, published in *The New Yorker* in 2011. If she searches the article title, the search will treat the article as the work. The article has a unique identifier that has been linked to the identifiers for the publication *The New Yorker,* both in its electronic and print incarnations. This article was also re-published in a book called *The Table Comes First.* This book also has an identifier, and the article was linked to this book, as well, by way of that identifier. When the user searches for the article by the article’s title, an on-the-fly display will be created that shows all of these versions. The electronic and print versions of the article as published in *The New Yorker* constitute one expression of the work, in two different manifestations. The library might own the print copy (identified by an item identifier: a barcode), as well as the electronic copy (which will have its own identifier based on the database in which it's published and on how the library has access to it). The article as re-printed in the book is another expression and manifestation. If there is an e-book available, then the book expression has two manifestations.

The user may not care which is the manifestation and which is the expression. She may simply want to know how she can read this article. The on-the-fly display will show that she can read it online, either as originally published in *The New Yorker* or as published later in the ebook version of the anthology, and that she can obtain a print copy of either of these items.

If she determines that she's interested in the print copy of *The New Yorker*, she may decide to see what else has been published in journal. Clicking on the journal title will change the display, and the journal-as-work, rather than as expression, would become the organizing principle behind the display. The user could then see the history of the publication, and click into any volume or issue of the journal to see which articles it contains. The user could also take the author name or subject heading as an entry point, and subsequent changes in the display could reflect information about the author or subject, as well as displaying what related resources are available. These displays could go beyond what is available through traditional authority records by pulling data about authors, places, and terms from outside of traditional library sources.

Once bibliographic data is shifted away from the record-based model, it no longer matters whether the article or the journal is considered the work, or whether a serial title is cataloged as a single item or as a group of successive, related items. Relationships are defined, and related resources are linked together by unique identifiers. The descriptive and access data that is required in a particular search can be pulled in and used at the point of need. If links are created between different format manifestations and different title variations of a publication, then they are easy to pull together in order to show the entire journal history, or just as easily pulled apart to find only a distinct issue or volume or article. Merely creating the links allows a user to see the data in whatever permutation or at whatever level is needed. If searching at the article level, then the user can see the different physical and digital formats that exist and the different modes of retrieval for that particular issue or article. If the user searches at the journal title level, a user can be shown the "formerly published as" or "currently published as" links, as well as the journal title level metadata, and the options available for viewing content at an issue/volume level.

This model also allows bibliographic metadata to be used in new ways, beyond allowing a user to find a particular resource. Researchers may be interested in understanding publication history for a number of reasons: to understand a field of study in a new way, or to see how scholarly communication has shifted over time. Publishing this information in an open, linked data form allows it to be used for any number of new purposes, and can open up new avenues of scholarly understanding. As research in the digital humanities expands, this kind of data could prove very useful for new inquiries.

The shift to a linked data model would not only help users better understand the bibliographic universe; it will save immense amounts of time for catalogers, too. Rather than maintaining a whole catalog's worth of data, updating records with each publication change or deleting and re-loading records every time a package of electronic resources is re-negotiated, catalogers can work collaboratively to maintain bibliographic metadata, and take advantage of metadata released by publishers and vendors. The vision of collaboration that was begun with cooperative cataloging ventures like OCLC and RLIN can be truly realized.

This notion of a linked data-based bibliographic ecosystem seems far away from a technical services perspective still mired in traditional MARC records. But there are signs of forward movement. The announcement in October 2011 by the Library of Congress that they are exploring the future of bibliographic metadata through the Bibliographic Framework Transition Initiative (Marcum, 2011) and the work toward the adoption of Resource Description and Access are signs that catalogers are starting to think differently about bibliographic metadata. In the meantime, there are concrete projects librarians can undertake to make local data identifiable and usable in a linked data environment. Institutions that maintain institutional repositories can explore the release of that metadata on the open Web as RDF triples. Librarians who work with faculty to create data management plans and archive their research can ensure that identification systems like EZID are used to give data sets unique identifiers and allow them to be found on the Web. Most importantly, cataloging librarians can begin to learn about RDF and graph-based data models, and how they might be useful in their own context.

Current bibliographic systems were created to reflect a certain information landscape, one in which resources were discrete items, located and accessed in physical spaces. But today's librarians and patrons are not working in that landscape anymore, and library systems struggle to keep up, and to reflect the new world of information resources. In order to create systems and provide access patrons can use and understand, we need to start thinking about our work very differently, and to imagine new possibilities unhindered by current practices. The fluidity of the linked data model offers a way out of the binds that we too often find ourselves in now.

References

Adams, K., Santamauro, B., & Blythe, K. (2008). Successive entry, latest entry, or none of the above? How the MARC21 format, FRBR and the concept of a work could revitalize serials management. *The Serials Librarian*, *54*(3/4), 193-197. doi:10.1080/03615260801974099

Antelman, K. (2004). Identifying the serial work as a bibliographic entity. *Library Resources & Technical Services*, *48*(4), 238-255.

Baker, T., Bermès, E., Coyle, K., Dunsire, G., Isaac, A., Murray, A….Zeng, M. (2011, October 25). Library linked data incubator group final report. Retrieved from http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Baer, W. (2004). Federated searching. *College & Research Libraries News*, *65*(9), 518–519.

Curran, M. (2009). Serials in RDA: A starter's tour and kit. *The Serials Librarian*, *57*(4), 306-323. doi:10.1080/03615260903218825

Everett Allgood, J. (2006). Serials and multiple versions, or the inexorable trend toward work-level displays. *Library Resources & Technical Services*, *51*(3), 160-178.

Graham, C. (1990). Definition and scope of multiple versions. *Cataloging & Classification Quarterly*, *11*(2), 5-32. doi:10.1300/J104v11n01_02

Manola, F. and Miller, E. (2004a). Introduction. In B. McBride (Series Ed.), RDF Primer W3C Recommendations 10 February 2004. Retrieved from http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#intro

Manola, F. and Miller, E. (2004b). 2.2 RDF Model. In B. McBride (Series Ed.), RDF Primer W3C Recommendations 10 February 2004. Retrieved from http://www.w3.org/TR/rdf-primer/#rdfmodel

Marcum, D. (2011, October 31). A bibliographic framework for the digital age. Library of Congress. Retrieved from http://www.loc.gov/marc/transition/news/framework-103111.html

Oliver, C. (2004). FRBR is everywhere, but what happened to the format variation issue? *The Serials Librarian*, *45*(4), 27-36. doi:10.1300/J123v45n04_02

Riva, P. (2004a). Mapping MARC21 linking entry fields to FRBR and Tillett's taxonomy of bibliographic relationships. *Library Resources & Technical Services*, *48*(2), 130-143.

Riva, P. (2004b). Defining the boundaries. *The Serials Librarian*, *45*(3), 15-21. doi:10.1300/J123v45n03_02