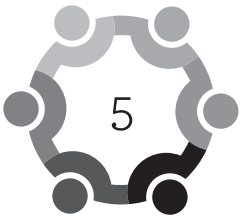


A DIVISION OF LABOR

The Role of Schema.org in a Semantic Web Model of Library Resources

Carol Jean Godby

LIBRARY METADATA AS LINKED DATA

This chapter describes some of OCLC's experiments with Schema.org as the foundation for a linked data model of library resources. The most important reason for giving Schema.org serious consideration is that this is the vocabulary endorsed by the world's major search engines at a time when a user's quest for information is more likely to begin on the broader Web than in a library or even a library website.[1] But there is a simpler and more pragmatic reason for taking a serious look at Schema.org. At a time of unprecedented uncertainty, Schema.org offers the promise that at least some of the effort involved in designing the next-generation standards for library data can be undertaken by a trusted third party. As a general-purpose vocabulary that is compliant with linked data principles, Schema.org addresses the requirements expressed in the Library of Congress publication *On the Record* for a solution that recognizes the Web as a technology platform and as a means for the discovery and delivery of resources that fulfill a user's information request.[2]

Yet many standards experts view the Schema.org vocabulary as too superficial and narrowly focused on the commercial sector to serve the needs of libraries. Some researchers at OCLC were among the initial skeptics. We noted, for example, that the "schema:CreativeWork," represented as an RDF class, would form the core of a Semantic Web model of the contents of a library collection. But a reasonable first impression is that this term offers little to a sophisticated discussion. The definition appears simplistic— "The most generic kind of creative work, including books, movies, photographs, software programs, etc." —and it is accompanied by a flat list of attributes, or RDF properties, such as "schema:author," "schema:isbn," "schema:publisher," and "schema:about," which could be used to assemble a set of statements resembling a primitive Dublin Core record.[3] Yet an accumulating body of model fragments and other experimental results is starting to show that Schema.org is actually rich enough to keep pace with mature thinking about the replacement for MARC and other standards that aim to modernize the data architecture for the library community. This chapter recounts the highlights of the experiments conducted at OCLC by focusing on three works: a published memoir, a children's fairy tale, and a recorded multimedia performance. It summarizes some of the arguments presented in the book *Library Linked Data in the Cloud.*[4] But the perspective given here is more personal, is focused more directly on the arguments in favor of data models featuring Schema.org, and is closer to the leading edge of our analysis.

THE FIRST MATURE MODEL OF BIBLIOGRAPHIC RESOURCES AND A REFORMULATION IN SCHEMA.ORG

In the months before Schema.org was published in the summer of 2011, linked data research at OCLC was focused on the task of expressing the contents of MARC records as a network of RDF statements. In the preceding years, OCLC researchers had contributed to the development of Semantic Web standards, worked in multiple venues with the international community of library standards experts to shape the argument that the linked data paradigm was consistent with the values of librarianship, and published the first linked data versions of VIAF, FAST, and the Dewey Decimal Classification.[5] Attention had been initially directed to the task of converting library authority files to RDF, not only at OCLC but also in the broader library community in the United States and Europe. This focus made sense. Authority files were about people,

places, organizations, and concepts, and were consistent with the linked data directive to create descriptions of persistent objects in the real world.[6] From a technical perspective, the contents of library authority files were already normalized, and the underlying semantics of the MARC authority record could map relatively easily into first-generation Semantic Web standards such as the Simple Knowledge Organization Scheme, or SKOS.[7] In 2008, researchers affiliated with the Library of Congress published a seminal study that produced a linked data representation of the Library of Congress subject and name authority files, paving the way for the development of http://id.loc.gov, one of the most widely accessed RDF datasets published in the library community.[8]

However, the description of creative works, or bibliographic resources, was identified as a more challenging task and a top priority for future work. This was the conclusion of the Library Linked Data Incubator Group, an international committee of library standards experts convened by the World Wide Web Consortium.[9] Bibliographic description is problematic because library standards are large, semantically complex, and designed primarily for human readers, not machine processes. But just as the *Library Linked Data Incubator Group Final Report* was being finalized, the British Library published a data model and the first large-scale proof of concept that bibliographic records that support a sophisticated use case for a major national library could be decomposed into networks of RDF statements. The result was a dataset representing approximately three million holdings in the British National Bibliography. A high-level view of the model would become an iconic diagram. Details of the model were later described in a series of tutorials by Dodds.[10]

The design of the British Library Data Model is conceptually simple and has been widely replicated, despite the deceptively cluttered appearance of the model diagram. At the center of the model is a "bibliographic resource" defined in the Bibliographic Ontology, which may either be realized as a self-contained work or a member of a series.[11] The resource came into being through a "publication event," which defines relationships between the bibliographic resource and a set of people and organizations acting in a particular time and place. The creator or "author," can be a person, who has a name and a birth date; or an organization, which has a name and an address. Finally, the bibliographic resource may have a "subject" defined from the various kinds of things that populate traditional library authority files—that is, people, places, and concepts.

The model is expressed as RDF instance data containing persistent identifiers for people, places, organizations, and concepts, supplemented by publisher-maintained resource identifiers such as ISBNs or ISSNs. Relationships

are expressed primarily by concepts defined by Dublin Core Terms, such as "dct:subject," "dct:contributor," and "dct:isPartOf." Other features of the model are described by a vocabulary maintained by the British Library and supplemented with twelve other sources of terminology: standards endorsed by the W3C, RDF vocabularies published by the library community such as Dublin Core and VIAF, and ontologies developed by other researchers interested in the intersection of the Semantic Web and language, linguistics, and references to written works, such as de Melo.[12]

When it was first published, the British Library Data Model represented a major advance over previous Semantic Web models of resources managed by libraries, both conceptually and technically. Conceptually, the model succeeds at reducing the complexity of library standards for bibliographic description to a few simple building blocks that represent the most important real-world objects and their relationships to resources managed by libraries—people who are creators or contributors; organizations that publish or distribute; people, places, or topics that the resources are about; and so on. In addition, the model is described in accordance with conventions established in the linked data paradigm. Thus it uses published vocabularies wherever possible; it refers to as many entities as possible with URIs from published RDF datasets; and it minimizes literal text strings while maximizing cross-linking. Technically, the model exhibits attention to the details of the linked data architecture, such as the specification of persistent URI patterns for every entity defined in the model. To encourage consumption, the outputs of the British Library experiment are available from a mature website with institutional branding, where data consumers can download the entire dataset or view descriptions in RDF/XML, JSON, RDF/Turtle, and a human-readable format.[13]

A FIRST DRAFT OF THE OCLC MODEL OF BIBLIOGRAPHIC DESCRIPTION

Also in 2011, as researchers were evaluating the British Library Data Model as a possible foundation for OCLC's model of linked bibliographic data, Schema.org was published. Like many other modeling experts, they initially wondered whether Schema.org supported rather than undermined the linked data paradigm because the Schema.org vocabulary appeared to be shallow and permitted the use of text strings where URIs would have been more appropriate. Though

a commitment to linked data appeared to be optional, the first announcement assured Semantic Web advocates that Schema.org indeed supported RDFa encodings.[14] Additional investigation revealed that the chief data architect for Schema.org was R. V. Guha, who has a long history of involvement in knowledge engineering research and the development of Semantic Web standards, including RDF Schema.[15]

OCLC researchers were also skeptical of Schema.org because the vocabulary seemed too focused on commercial products, which overlap only partially with the curatorial needs of libraries. But a quick test revealed that Schema.org offered nearly the same coverage as the fourteen vocabularies used in the British Library Data Model. These equivalences are illustrated in figure 5.1, which represent comparable descriptions of a memoir described in WorldCat at www.worldcat.org/title/memoir-my-life-and-themes/oclc/43477327 and in the British Library dataset at http://bnb.data.bl.uk/doc/resource/008140605. To highlight the commonalities, the descriptions contain text strings, not URIs. In the real instance data, of course, all RDF predicates except string literals such as titles and dates are expressed as URIs that conform to the patterns suggested by figure 5.1 in the British Library data and derived primarily from FAST, VIAF, and LCSH in the OCLC data. The small number of gaps, such as the British Library model of the statement of responsibility as an event, would be addressed by an extension vocabulary for Schema.org proposed by OCLC, a topic discussed in more detail below.

BRITISH LIBRARY DATA MODEL	OCLC MODEL
a bibo:BibliographicResource;	a schema:Book;
dc:title “Memoir : my life and themes”;	schema:name “Memoir : my life and themes”;
dcterms:creator “Conor Cruise O’Brien”;	schema:author “Conor Cruise O’Brien ;
blterms:datePublished “1999”;	schema:datePublished “1999”;
blterms:publication <event>;	schema:publisher “Profile”;
dcterms:language “English”;	schema:inLanguage “English”;
dcterms:subject “Statesmen”;	schema:about “Diplomats”;
bibo:isbn10 “1861971516”;	schema:isbn “1861971516”;
isbd:01953 “470 p.”;	schema:numberOfPages “470 p.”;
dcterms:description :”includes index”;	schema:description :”includes index”;
dcterms:spatial “Ireland”.	schema:place “Ireland”.

Figure 5.1 | **Two Descriptions of a Memoir in the British National Bibliography**

In short, a reformulation using Schema.org vocabulary appeared to be simpler without significant loss of expressiveness. Because the description is formulated from terms defined in a single namespace, the concept space is clearly specified. But the British Library solution raises questions. Is "dcterms:accrualPeriodicity" in or out of scope? Does the presence of the BIBO definition of 'ISBN' preclude the use of the DC-Terms solution for describing ISBNs? Is an "foaf:Document" the same thing as a "bibo:BibliographicResource," and if so, is the property "foaf:topic" equivalent to "dcterms:subject" because both assert an "aboutness" relationship between a resource and a real-world object? Are these equivalences formally expressed in the model, or are they informally assumed? Technically, a small set of namespaces also means that instance data is easier to produce and maintain. In addition, the simpler solution has a greater chance of widespread adoption because prospective advocates do not have to engage in the task of vetting each vocabulary and substituting other choices that are more appropriate or timely. But in one important respect, the vocabulary choices for encoding instance data are a side issue because the most important details of the underlying high-level model can be preserved in the transformation. Thus a description encoded with Schema.org vocabulary is still centered on a creative work, with properties that identify the people, places, organizations, and concepts that brought it into being and define a context for interpreting it.

Arguments about conceptual simplicity and scope of coverage can be stated without considering the stature of Schema.org as a de facto standard, but the case is stronger when this fact is acknowledged. Right after Schema.org was announced, bloggers pointed out that the new ontology might curtail some of the experiments with vocabulary development made visible by projects such as the British Library Data Model. Nevertheless, the declaration of a shareable semantics from a group of influential organizations with a commercial incentive could only be interpreted as a positive development for the Semantic Web.[16] In a recent presentation, Guha argued that the major Internet search engines have long been interested in structured data, but have been unable to solicit enough high-quality input from data providers.[17] Schema.org is the latest incentive, and as in earlier solutions such as the HTML <meta> tag or microdata, the promised return is a Rich Snippet, a Knowledge Card, and generally greater visibility in the marketplace where most users now begin their quest for information.[18]

The Schema.org vocabulary is designed to strike a balance between simplicity and sophistication, making it easy for webmasters to say simple things, while

providing a platform for data managers representing specialized communities of practice to say complex things, to paraphrase a point made by Guha in his 2014 presentation. Proposals for enhancement are managed by community groups sponsored by the W3C, following a pattern established by other web standards initiatives such as Dublin Core, SKOS, and RDF. As Schema.org evolves, the vocabulary is enriched with ontologies developed by third parties, some of whose interests align with those of libraries. For example, GoodRelations defines an ontology for e-commerce and has recently been integrated with Schema.org.[19] Some of the vocabulary defined in GoodRelations being evaluated by librarians as a model for library holdings is more easily consumed by machine processes and more readily understood by the major search engines than the text-heavy standards used in the library community.[20]

PUBLISHING AND EXTENDING SCHEMA.ORG

In 2012 OCLC published the first draft of a linked data model for bibliographic description expressed in Schema.org as RDFa markup on approximately 300 million MARC catalog records accessible from WorldCat.org.[21] These descriptions were built upon previous successes with linked data models of library authority files, featuring URIs from RDF datasets representing the Dewey Decimal Classification, the Library of Congress Subject Headings, the Library of Congress Name Authority File, VIAF, and FAST. The outcome was the largest set of linked bibliographic data on the Web by many orders of magnitude, and a proof-of-concept demonstration of linked data as a viable next-generation data architecture for library resource description. The WorldCat linked data was updated in 2014 with URIs from the recently published WorldCat Works dataset, which uses the latest generation of FRBR-inspired clustering algorithms to group bibliographic records with similar content.[22]

The demonstration has a technical as well as an ontological component. Technically, the result demonstrates that the technology stack, which featured map-reduce jobs implemented on Hadoop clusters and Semantic Web development tools such as those described in the compilation maintained by the World Wide Web Consortium, could handle the sheer magnitude of a data-processing task that produces tens of billions of RDF triples and renders them searchable in real time.[23] Once mature, the processes could refresh the markup on WorldCat records in a matter of hours, not days, producing BIBFRAME instead of Schema.org encodings in a late processing step when

this result was requested in a joint experiment with the Library of Congress in early 2013. Conceptually, even the first experiment demonstrated that a simple entity-relationship model expressed in Schema.org could capture assertions about every resource type described in the WorldCat catalog data. In fact, the Schema.org definitions of "Person," "Organization," "Creative Work," "Place," and "Topic" appeared to be robust enough to serve as the real-world referents for the URIs defined in the linked data versions of VIAF and FAST, which were republished in 2014. In these revisions, a VIAF or FAST "Person" such as Peter Tchaikovsky is defined as a "schema:Person," matching the RDF assertions about the Russian composer published in the bibliographic descriptions accessible from WorldCat.org.

EXTENSION VOCABULARIES

A more detailed examination of the descriptions generated from the OCLC experiments is deferred to the next section, but even a cursory look at Schema.org reveals gaps that need to be filled to keep pace with linked data experiments being conducted elsewhere in the library community. For example, figure 5.1 shows that the Schema.org description that could be generated in 2011 was slightly less expressive than the corresponding British Library Data Model description because of missing or imprecise terms that should be synonymous with "dcterms:isPartOf" and "dcterms:spatial." But such problems are now being addressed. In 2012 OCLC founded the W3C-sponsored Schema Bib Extend Community Group, which includes librarians, publishers, and integrated library system (or ILS) vendors who share OCLC's perspective regarding the importance of Schema.org in the conversion of legacy bibliographic descriptions to linked data.[24]

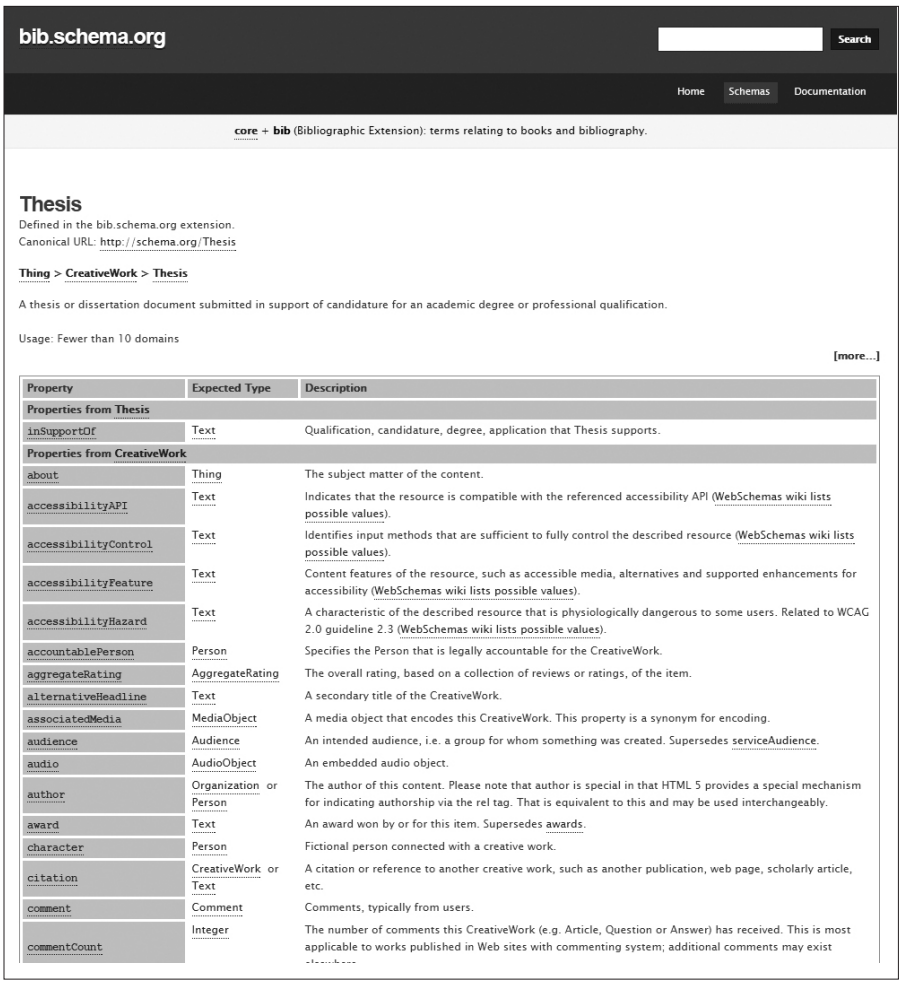
As at OCLC, the Schema Bib Extend group begins a modeling exercise by selecting a problem in library resource description and hand-crafting a set of statements using terms defined in Schema.org. The analysis demonstrates what Schema.org successfully covers, but it also reveals gaps, inconsistencies, or terms that are incorrectly placed in the ontology. Possible amendments to Schema.org are discussed on W3C-managed mailing lists such as "public-schemabibex" or "public-vocabs," some of which advance to the status of a formal recommendation to the managers of Schema.org.[25] In October 2014 Schema.org adopted the recommendations from the Schema Bib Extend group for "schema:hasPart" and schema "isPartOf." Schema.org has also adopted the

refinements of the structured parts of a journal citation recommended by the Bib Extend group for the property "schema:issueNumber" on "schema:PublicationIssue" and for "schema:volumeNumber" on "schema:PublicationVolume." In addition, Schema.org now contains the "schema:CreativeWork" properties "schema:workExample" and "schema:exampleOfWork," which emerged from OCLC's experiments with the implementation of Work-to-Work relationships concepts defined in FRBR and RDA. The use of these relationships is discussed later in this chapter.

In the absence of a comprehensive standard, however, model development requires a proving ground for experimenting with candidate vocabulary even before it has been submitted for public review. To meet this need, OCLC introduced the BiblioGraph.net extension vocabulary in 2014, which has the look and feel of Schema.org because it has been constructed from a common code base. One outcome is a visualization of the impact of proposed extensions on the Schema.org vocabulary. For example, "Agent" can be defined as an extension of "schema:Thing" and has "schema:Person" and "schema:Organization" as subclasses, borrowing the definition created in the Friend of a Friend (or FOAF) ontology. According to the FOAF documentation, "Agent" is "useful in a few places . . . where 'Person' would have been overly specific."[26] In effect, the BiblioGraph definition acts as a technical "pass-through" from FOAF to Schema.org that shows how terms defined in an external vocabulary could be positioned in an ontology that can be directly consumed by general-purpose search engines. In BiblioGraph, the "Agent" class has been proposed as a useful place to define properties such as "bgn:publishedBy" and "bgn:translator," both because people and organizations can have these relationships to creative works and because it is not always possible to distinguish between them in MARC records or other bibliographic descriptions. BiblioGraph and the informally named "cherry-picking" approach observable in the British Library Data Model may have superficial similarities, but there is one important difference. While both efforts aim to expand the scope of Semantic Web vocabularies to advance the development of resource descriptions that can support services offered by libraries, a solution built on Schema.org has the equally important goal of extending the boundary of a shared semantics. And the outcome is a single ontology that is explicitly specified, comprehensible to human readers, and actionable by machine processes.

Thus the models developed at OCLC emerge from the assumption that many concepts already defined in Schema.org are essentially the same as those defined by library standards experts, including top-level classes such as

"schema:Thing," "schema:Person," "schema:Organization," "schema:Topic," and "schema:Place." In fact, most properties defined for "schema:CreativeWork" have such a shareable semantics, such as "author," "director," "publisher," "ISBN," "genre," "copyrightYear," and "audience." With an extension vocabulary such as BiblioGraph, OCLC provides a forum for representing other concepts defined by the library community that are also understandable and useful outside the narrow community of professional catalogers. For example, Schema.org is especially deficient in the inventory of relationships among creative works, as well as formats and resource types, such as "microform," which is defined in



bib.schema.org

Search

Home Schemas Documentation

core + bib (Bibliographic Extension): terms relating to books and bibliography.

Thesis

Defined in the bib.schema.org extension.
Canonical URL: http://schema.org/Thesis

Thing > CreativeWork > Thesis

A thesis or dissertation document submitted in support of candidature for an academic degree or professional qualification.

Usage: Fewer than 10 domains

[more...]

Property	Expected Type	Description
Properties from Thesis		
inSupportOf	Text	Qualification, candidature, degree, application that Thesis supports.
Properties from CreativeWork		
about	Thing	The subject matter of the content.
accessibilityAPI	Text	Indicates that the resource is compatible with the referenced accessibility API (WebSchemas wiki lists possible values).
accessibilityControl	Text	Identifies input methods that are sufficient to fully control the described resource (WebSchemas wiki lists possible values).
accessibilityFeature	Text	Content features of the resource, such as accessible media, alternatives and supported enhancements for accessibility (WebSchemas wiki lists possible values).
accessibilityHazard	Text	A characteristic of the described resource that is physiologically dangerous to some users. Related to WCAG 2.0 guideline 2.3 (WebSchemas wiki lists possible values).
accountablePerson	Person	Specifies the Person that is legally accountable for the CreativeWork.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alternativeHeadline	Text	A secondary title of the CreativeWork.
associatedMedia	MediaObject	A media object that encodes this CreativeWork. This property is a synonym for encoding.
audience	Audience	An intended audience, i.e. a group for whom something was created. Supersedes serviceAudience.
audio	AudioObject	An embedded audio object.
author	Organization or Person	The author of this content. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.
award	Text	An award won by or for this item. Supersedes awards.
character	Person	Fictional person connected with a creative work.
citation	CreativeWork or Text	A citation or reference to another creative work, such as another publication, web page, scholarly article, etc.
comment	Comment	Comments, typically from users.
commentCount	Integer	The number of comments this CreativeWork (e.g. Article, Question or Answer) has received. This is most applicable to works published in Web sites with commenting system; additional comments may exist

Figure 5.2 | **"Thesis" defined in bib.schema.org**

BiblioGraph as a subclass of "schema:Product" because it is a physical object that can be bought and sold. The expertise of library standards experts is especially strong on these topics, of course, and OCLC's modeling experts believe that definitions of translations, adaptations, product forms, and other derivatives that could result from their engagement with Schema.org are likely to be useful to other professional managers of bibliographic resources and many information seekers.[27]

Much of BiblioGraph.net has recently been absorbed experimentally into Schema.org as a more closely integrated "reviewed/hosted Extension."[28] Reviewed extensions are also interpreted as an overlay on the core Schema vocabulary, to which subclasses and properties would typically be added. An example is shown in figure 5.2. Here, "bib:Thesis" is defined as a subclass of "schema:Creative" and inherits all of its properties, such as "schema:name," "schema:about," "schema:author," and so on. This core description is enhanced with the additional property "bib:inSupportOf," which applies only to a thesis and describes the academic purpose for which it was produced. Extension vocabularies such as bib.schema.org are formally recognized only if they are maintained by a recognized community of practice such as the Schema Bib Extend community group. Other extensions are being discussed for the domain of library resource management, and for communities of practice elsewhere on the Web.

TOWARD A MODELING DIVISION OF LABOR

The "reviewed extension" proposed by the Schema.org editors presents the opportunity for librarians and other communities of practice to engage more closely with an important target audience, eliminating some of the guesswork about whether domain-specific vocabularies are visible to general-purpose search engines. Other communities are confronting the same issue. For example, a person suffering from hives might issue a Google search for advice on treatment, which returns a Knowledge Card, and associates the condition with the medical term "urticaria."

The user's query is more likely to yield rich results if authoritative health care sites are published with structured descriptions that bridge the gap between the frames of reference of the general population and the professional. For both the medical and library communities, it is possible to imagine a model in which a broadly understandable vocabulary is defined in Schema.org, while

a more complete description is modeled with terminology defined with the greater precision required by researchers, practitioners, and other specialists.

By designing such a model, data architects acknowledge the utility of the "linguistic division of labor" made famous by the philosopher of language Hilary Putnam.[29] His work addresses a classic question: does the meaning of a word reside only inside a person's head, or is it in the public sphere? Difficulties ensue if the answer is the first choice. It would be impossible, for example, to verify that the tasteless, odorless clear liquid that sustains life called "water" by one community is the same substance assigned the chemical structure H_20 by chemists, and is different from the embalming fluid called "water" in some drug subcultures. But as Putnam argues, if the meaning of "water" is instead a social construct, it can be built up by communities that have different interests, levels of expertise, and requirements for scientific truth who cooperate through a network of trust. Thus most of us in the lay public have no practical need for knowing the chemical structure of water, but we sometimes have to rely on experts to vouch for the fact that the clear liquid in our drinking glasses is life-giving and not poisonous. The conception of word meaning as a social construct is still relevant today, and is at the heart of the theory of reference defined by the Semantic Web. Linked data principles identify the primacy of real-world objects, and conventions such as "Cool URIs" interpret the Web as a corpus of facts about these objects, which require collaborative effort and a range of expertise to arrive at the truth.[30]

But it remains to be determined how the division of labor should be implemented in the design of library linked data. In the work underway at OCLC, community input would be solicited to distinguish between the terms that would be formally proposed as candidates for inclusion in Schema.org, and the vocabulary that would be maintained in BiblioGraph or bib.schema.org, which defines the high-level concepts in the domain of library resources and the transactions that involve them. This distinction is a proxy for the differences in the language of the public and that of experts, and it serves two use cases. The first supports discovery through general-purpose search engines, and the second is required for long-term curation and other functions of libraries that enable the fulfillment of the user's information request.

The same distinction emerges from our joint analysis with the BIBFRAME team at the Library of Congress.[31] But many details remain to be specified. For example, the vocabulary of experts will likely be much larger and more complex than the stub defined in BiblioGraph. Perhaps most of it will never be directly

consumable by search engines, except for high-level concepts that comprise a metalanguage of sorts that could be exposed through the pathway we have defined. For example, the properties defined for "schema:MedicalCondition" include "treatment," "risk factor," "cause," "pathophysiology," and "prevention," whose expected values are text strings or URIs defined in specialized vocabularies, some of which are maintained by the library community. Conversely, the vocabulary that can be exposed through a general-purpose ontology such as Schema.org and the BiblioGraph extension may be larger than is acknowledged in MARC or other library standards. This is the working hypothesis of much of the linked data work now being conducted at OCLC, as we argued in Godby, Wang, and Mixter and in the examples in the next section of this chapter.[32] For example, BiblioGraph defines "translator" and "translation of work," but so does RDA.[33] By defining these terms in BiblioGraph, we are claiming they exist in the language of the general public, which makes them candidates for eventual absorption into Schema.org. The same observation motivates the work to produce "unconstraine" definitions of all RDA relationships, which reduce to commonsense definitions with no dependency on the ontological claims of FRBR.[34] In other words, "translation" is like "water." Many users probably consider its definition to be clear enough, so they don't need to understand its chemical structure, as long as a recognized community of trusted experts does.

MODELING BEYOND PUBLISHED MONOGRAPHS

A focus on the relatively well-understood descriptions of published monographs such as the excerpt shown in figure 5.1 makes it easy to underestimate the true scope of the job required to transform the data architecture for library resource description. Unfortunately, the change will not happen completely through a simple record-by-record mapping from MARC to the new format. The transformation also requires the development of an entity-relationship model with ever-increasing granularity and algorithms that can discover evidence for the model in existing data. But when the algorithms reach their inevitable upper limits, it will be necessary to design a more aspirational model that is populated with human guidance by enacting recommendations for future descriptive practice. Though the larger set of tasks is challenging and multidimensional, OCLC's experiments make us confident that the Schema.org ontology can evolve with the demands that will have to be put on it.

A FAIRY TALE

An informal progress report on OCLC's approach to the modeling of linked bibliographic data can be assembled by examining two MARC records accessible from WorldCat.org. The first is a description of *The Nutcracker and the Mouse King,* a fairy tale written by the nineteenth-century German author E. T. A. Hoffmann and translated into English by Joachim Neugroschel. The second record describes a movie that captures a live performance of Tchaikovsky's ballet *The Nutcracker* at the Bolshoi Theatre in Moscow in 1989. The ballet is based on Hoffmann's tale.

Figure 5.3 shows critical details of the MARC record describing *The Tale of the Nutcracker*. The 041, 240, 245, 500, 700, and 740 fields reveal the relationships between the English derivative and a German original. The 041 field specifies the language of the cataloged work as English and the language of the original as German. The title of the cataloged work is listed in the 245 field; and the title of the original German work is listed in the 240, or "Uniform Title" field, which is prescribed by cataloging rules "when a work has appeared under varying titles, necessitating that a particular title be chosen to represent the work."[35] E. T. A. Hoffman is listed as the author of the English translation, perhaps because the cataloged work is identified as a variant of the original. The 500 field identifies the translator as Joachim Neugroschel, and the 245 $c lists a relationship between Hoffmann's work and *The Tale of the Nutcracker*, which has an unspecified link to Alexander Dumas. A list of 700 fields associate Hoffmann, Neugroschel, and Dumas with entries in the Library of Congress Name Authority file and the Dumas work is listed in the 740 field as an uncontrolled related title.

In sum, the MARC record excerpted in figure 5.3 describes a complex network of Agent-to-Work and Work-to-Work relationships. But not every detail is algorithmically recoverable. The details of the model that can be populated automatically by mapping from the semantics of the MARC record are captured in the RDFa markup, which is accessible from a tab labeled "Linked Data" at the bottom of the page in the associated WorldCat.org display. Figure 5.4 is a graphical representation of the most important details. The entities, or classes, are shown as white boxes; relationships, or properties, are shown as labeled arrows; and literals, such as "2007," are shown as floating strings. The names of the entities—"Person," "Creative Work," "Organization," and "Topic"—as well as the names of the relationships, displayed in italics, are imported from the corresponding RDF markup.

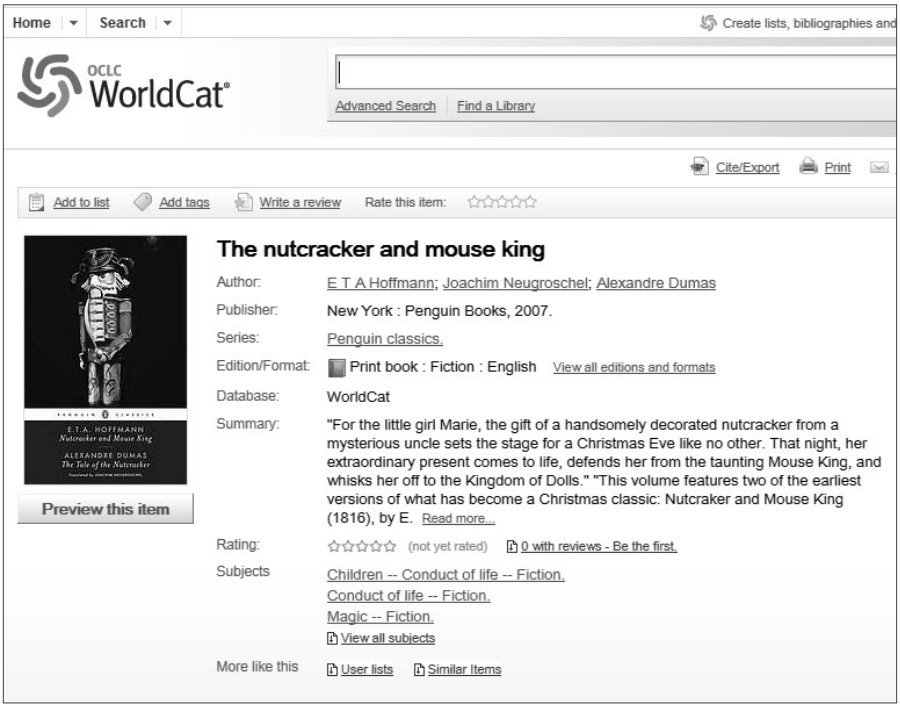


Figure 5.3 | **A MARC record for a 2007 edition of *The Nutcracker and the Mouse King*, WorldCat ID # '76967162'**

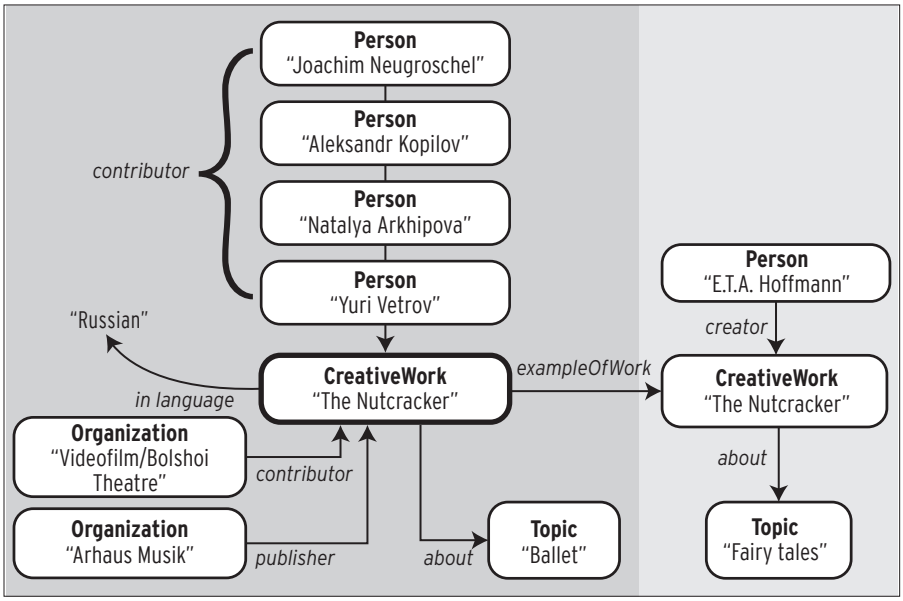


Figure 5.4 | **Entities and relationships revealed in the RDF statements for *The Nutcracker and the Mouse King***

As figure 5.4 implies, the RDF representation of the MARC record identifies multiple creative works connected through a translation relationship. The properties associating the Creative Work with other entities form machine-understandable RDF statements that correctly assert that "E. T. A. Hoffmann is the author of the creative work with the title *Nussknacker and Mausekőnig*," "E. T. A. Hoffmann is the author of the creative work with the title *The Nutcracker and the Mouse King*," "The creative work with the title *The Nutcracker and the Mouse King* is published by an Organization with the name Penguin Books," which has a publication date of 2007 and is about Magic. But the RDF representation identifies Joachim Neugroschel only as a "contributor," not as a "translator," because the MARC-to-RDF conversion algorithms do not parse the free text in the MARC 500 field or other notes. For a similar reason, the RDF does not encode precise details about the Alexander Dumas version of *The Nutcracker* because the MARC record does not contain machine-understandable encodings of the relationship, perhaps because it is not clear even from the human-readable text.

The MARC and corresponding RDF descriptions of *The Nutcracker and the Mouse King* point to two areas of research in OCLC's next-generation models of bibliographic description. First, the Multilingual WorldCat project recognizes that nearly two thirds of WorldCat catalog records now represent non-English works—works that have the most complex publication histories, are the most widely held by libraries, represent arguably the most significant contributions to the world's literary canon, and are the most commonly translated.[36] The Multilingual WorldCat project aims to improve the quality of the bibliographic descriptions for these materials, with the long-term goal of delivering the description as well as a copy of the work in the searcher's chosen language. Though linked data is a means to this end, incremental improvements can made to the source MARC records as a useful side effect, a point that is developed in more detail in Smith-Yoshimura and Godby.[37]

For example, the important entities and relationships can be discovered more easily if the 041 field names the source and target languages and a "Uniform Title" field identifies the translation source. In addition, the Multilingual WorldCat team recommends that the description contain the name of the translator and the translation relationship in a machine-processable form. Thus if the 700 field for "Joachim Neugroschel" had contained $4 subfield populated with "trn," the MARC Relator code for 'translator,' the "schema: contributor" property could have been promoted to the more specific value, "bgn:translator."[38] This is a minor shortcoming in an otherwise high-quality

record—which, unsurprisingly, is easily transformed into a set of expressive RDF statements.

The second area of research can be inferred from implicit references to the concepts defined in the FRBR Group I model.[39] A bibliographic record accessible from a library catalog or an aggregation such as WorldCat.org typically describes a Manifestation; a translation is an Expression of a Work; and an Item must be delivered to fulfill the information request that originated in the catalog. The linked data descriptions now available as RDFa markup on WorldCat catalog records build on nearly fifteen years of research conducted at OCLC on the problem of algorithmically discovering FRBR concepts in collections of MARC records.[40] Three conclusions emerge from this research. First, it is still necessary to distinguish among Work, Expression, Manifestation, and Item because these concepts are motivated by the user's need to find, identify, select, and obtain resources that satisfy his or her request for information, regardless of whether the search begins in a library catalog or on the Web. Second, the essential distinctions defined by FRBR and RDA elaborations can be captured using the properties defined for "schema:CreativeWork" and some commonsense extensions defined in BiblioGraph or other vocabularies maintained by library standards experts. The main problem is that Schema.org needs to be enhanced with richer content-to-content and agent-to-content relations. Finally, and more narrowly, the model of translations developed by the Multilingual WorldCat project can be viewed as a prototype for defining other content-to-content relations.

A rough draft of the solution is already visible in the high-level view shown in figure 5.4. On the left side, against the orange background, is a description based on the definition of the FRBR Manifestation, which describes a volume that was published by Penguin Books in 2007. On the right side, against the green background, is a more abstract description based on the definition of the FRBR Work. The Manifestation description is derived from a single MARC record, but the Work description is automatically generated from a set of similar MARC records and published as an entry with a persistent URI in the WorldCat Works dataset.

The clustering process is described in more detail in Godby, Wang, and Mixter, but it is conceptually simple.[41] The first step produces a group of records whose MARC 1xx "author" and 245 "title" fields match in a string comparison. A subsequent step considers information extracted from the corresponding VIAF descriptions for the authors, which contains lists of their published monographs in multiple languages and has the effect of pulling translations into

the cluster. In the final step, a WorldCat Works description is constructed from the clustered records by extracting properties that describe the content, such as "schema:description," "schema:about," and "schema:genre"; as well as agents responsible for the content, such as "schema:contributor" and "schema:editor."

Important details of the Work and the Manifestation descriptions and the property that associates them is shown more explicitly in the RDF statements excerpted in figure 5.5. In literal terms, the object described when the URI <http://www.worldcat.org/title/nutcracker-and-mouse-king/oclc/76967162> is de-referenced—that is, the Penguin edition of the book *The Nutcracker and Mouse King*—is a product with a model number, or an ISBN. As a result, the Manifestation has class assignments of "schema:CreativeWork" and "schema: ProductModel." But since the "Work" described in the document accessible from the URI <http://worldcat.org/entity/work/id/839867> has no properties

```
<http:// http://www.worldcat.org/title/nutcracker-and-mouse-king/oclc/76967162>
    a schema:CreativeWork, schema:ProductModel
        schema:name "The nutcracker and mouse king";
        schema:exampleOfWork <http://worldcat.org/entity/work/id/839867>;
        schema:contributor "Neugroschel";
        schema:creator "ETA Hoffmann";
        schema:isbn "xxx";
        schema:publisher "Penguin Books";
        schema:publicationDate "2007";
        schema:about "Magic".

<http://worldcat.org/entity/work/id/839867>
    a schema:CreativeWork
        schema:name "Der Nussnacker…
        schema: creator "E.T.A Hoffmann;
        schema: description "After hearing how her toy nutcracker got his ugly face,
        a little girl helps break
        the spell and watches him change into a handsome prince."
        schema:about "Mice";
```

Figure 5.5 | **RDF/Turtle excerpts of Manifestation and Work descriptions for the 2007 Penguin edition of *The Nutcracker and Mouse King***

that allude to a physical presence, the single class assignment of "schema: CreativeWork" is sufficient. The property "schema:exampleOfWork" establishes that a semantic relationship exists between the Manifestation and the Work. But the above discussion implies that the 041 and 240 fields of the Manifestation description contain some of the detail required to assign the more specific property recently absorbed from Bibliograph.net, "schema:translationOfWork." Algorithms that exploit this information and operate on the entire corpus of translated resources accessible from WorldCat,org and VIAF are now being tested, and the newly published terms for translators and translations are beginning to appear in published RDF statements.

Though simple, this example has unexpected theoretical and practical implications for the implementation of FRBR as linked data. At a high level, this data illustrates the configuration shown in the right-hand panel of figure 5.6, which contrasts with the familiar diagram depicting hierarchical relationships among FRBR Works, Expressions, Manifestations, and Items shown on the left.[42] In other words, OCLC's linked data model of bibliographic description is simply a network of objects typed as "schema:CreativeWork," all accessible through persistent URIs, with no presumptive hierarchical relationships. If they form a cluster of related creative works, they are at least connected via the semantically null "exampleOfWork" relationship, which may be upgraded to the more meaningful "translationOfWork" if certain details are present.

One consequence of this design is that Expression as a class is remodeled as a set of relationships among creative works, such as "translation," "adaptation," and others defined in RDA Work-to-Work and Work-to-Expression ontologies.[43] Another consequence is that the need to partition a set of properties across a hierarchy is eliminated. In most implementations of FRBR and RDA, creators, titles, or subjects are assigned only to Works, while contributors, series titles, or publication details can be assigned only to Manifestations, and

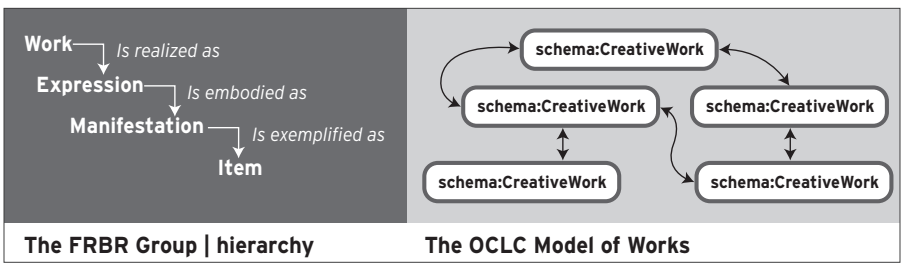


Figure 5.6 | **Hierarchical and graph models of creative works**

so on. Instead, the RDF statements reproduced in figure 5.5 show that Works and Manifestations both have subjects, titles, and authors, but only the Manifestation description has properties describing a publisher, page count, and other evidence of a physical presence. Likewise, an Item description would be assigned the Schema.org types "CreativeWork" and "IndividualProduct" if it contained a bar code. Thus, the hierarchical data models of FRBR and RDA have been replaced with a "trigger" model, which recognizes that most properties defined for the description of a creative work do not induce further ontological distinctions; only those that indicate physicality, uniqueness, or membership in a set of identical objects do. If present, these properties trigger the assignment of an additional Schema.org type.

A BALLET

With this background, it is now possible to examine a bibliographic description of a resource that is not a published monograph. This work is far from mature, but it reveals that OCLC's model of creative works encoded in Schema.org is potentially far richer than the evidence that can currently be discovered in legacy MARC records. The analysis is anchored in the data underlying the WorldCat.org record mentioned earlier describing a video of a live performance of Tchaikovsky's ballet *The Nutcracker* performed in 1987 at the Bolshoi Theater in Moscow. Similar examples are described in Smith-Yoshimura and Godby.[44]

The essential fields in the MARC source for this display are excerpted in figure 5.7. The 245 field contains title and creator of the performed work and the 260 field contains the name of the publisher of the DVD. The names of the primary stage performers are listed in the two 511 fields, from which a human reader can infer that the first 511 field has a list of dancers, while the second contains the name of the orchestra and conductor. Production credits for the performance, such as music, choreography, set design, and producer, are listed in the 508 field. The 650 fields indicate that the work is about ballet and mice. The 700 and 710 fields contain authority-controlled forms of some of the names listed in the 245, 508, and 511 fields.

Figure 5.8 is a graphical depiction of the most important RDF statements published on the WorldCat record. As in the previous example, the entity-relationship model expressed in RDF reveals the presence of two objects typed as "schema:CreativeWork"—a Manifestation-like description on the left and a Work-like description on the right, connected by the "schema:exampleOfWork"

245 04 $a The Nutcracker $h [videorecording]/ $c by Peter Tchaikovsky

260 __ $a [S.I.] : $b Arthaus Musiik, $c 1989

300 __ $a 1 videodisc (DVD) (101 nub,) : $b sd., col.,; $c 4 ¾ in.

511 1_ $a Natalya Arkhipova, Irek Mukhamedov, Yuri Vetrov, Bolshoi Ballet

511 0_ $a Bolshoi Theatre Orchestra conducted by Aleksandr Kopilov

508__ $a Producer, Takeshi Hara ; artistic director of the Bolshoi Ballet, Yuri Grigorovich ; director for NHK, Motoko Sakaguchi ; music by Pyotor Illyich Tchaikovsky ; scenario by Ivan Vsevolozhsky and Marius Petipa ; original choreography by Lev Ivanov ; revised choreography by Yuri Grigorovich.

650 $a Ballets.

700 $a Vetrov, Yuri

700 $a Arkhipova, Natalya.

700 $a Mukhamedov, Irek.

700 $a Kopilov, Aleksandr.

710 $a Videofilm/Bolshoi Theatre.

Figure 5.7 | **A MARC record for a video recording of *The Nutcracker,* WorldCat ID # 81750960**

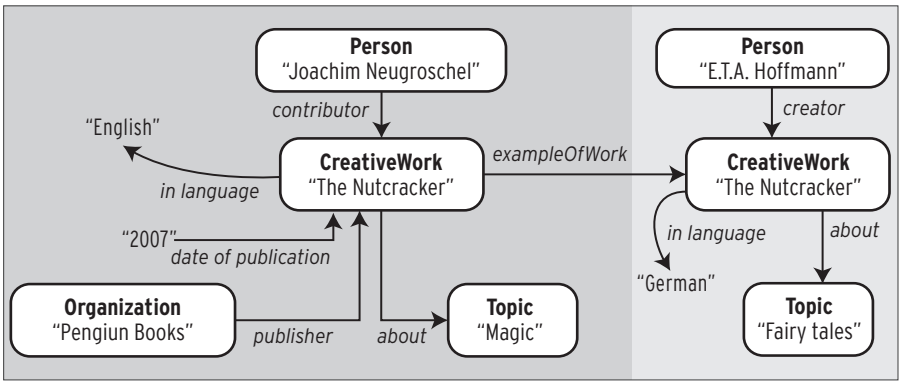


Figure 5.8 | **Relationships revealed in the RDF statements for the video recording of *The Nutcracker***

property. As in the first example, all of the RDF statements except literal strings such as titles and dates contain URIs in the subjects and predicates. But this RDF output is disappointing. The Manifestation description consists primarily of a long list of "schema:Person" and "schema:Organization"

agents with a "schema:contributor" relationship to the creative work, while the Work description is a cluster of recorded performances of *The Nutcracker* at the Bolshoi Theater on multiple dates with an overlapping set of performers. The algorithm correctly associates E. T. A. Hoffmann with the ballet, but only accidentally because one of the records in the Work cluster mentions the relationship between the ballet and the E. T. A. Hoffmann fairy tale. The Manifestation description states that the language of the work is Russian, but real-world knowledge is required to infer that the performance of the ballet captured on the videodisc has some association, possibly erroneous, with a Russian-language libretto.

The model shown in figure 5.8 is best interpreted as a first-draft view of the entities and relationships that can be extracted from MARC records describing filmed performances. But the less-than-impressive result is a symptom of two unresolved problems.

First, the network of relationships is encoded primarily as free text in the 245, 260, 508, and 511 fields, not the controlled fields that are more recoverable algorithmically. Second, the conceptual model that underlies the MARC record works best for monographs and is not suited to the description of multimedia works with complex interdependencies among agents responsible for variations in content and execution. At OCLC, the free-text problem is being addressed in a research project that uses text-mining algorithms to extract roles from the free-text fields and match them with the names listed in the 700 and 710 fields, which would have the effect of upgrading the relationships from "contributor" to a more specific value using evidence supplied by human catalogers who contributed the source records to WorldCat. Though free text is always unpredictable, partial successes are possible because much of the content in the 245, 508, and 511 fields is stylized and can be parsed with a named-entity recognizer and a simple grammar. The data in the 511 field is easiest to process because it almost always contains a list of names delimited by commas. Once the names are extracted, the MARC semantics of the field enables a machine process to promote "contributors" to "performers." More subtly, a machine process can also infer the existence of an event in the underlying model because "schema:performer" is defined as a relationship between a "schema:Person" and a "schema:Event."[45]

Of course, a data-driven approach would be more effective if complemented with a top-down analysis that produces a richer model of the entities and relationships required for describing a filmed performance. An important start has been made in the study of audiovisual materials commissioned by the

Library of Congress.[46] After surveying the treatment of films by MARC, RDA, BIBFRAME, and more specialized standards used in audiovisual cataloging communities, the authors conclude that these materials are not adequately served by any current standard. An improved model would recognize that a film captures live action in elapsed time and can be interpreted as an event, though it is not always a product of the human imagination because it could be a singing bird or an unfolding natural disaster. Moreover, the production of a film requires agents acting in multiple roles, many of whom make primary contributions that are understated in data models such as MARC, FRBR, RDA, and BIBFRAME that draw a bright line between creators and contributors. Finally, the Library of Congress study recognizes that the physical artifacts of filmmaking exist in multiple versions and are often collected into aggregations.

Schema.org is also mentioned in the literature survey, but the report is sketchy on details about how audiovisual materials can be described with this standard. Nevertheless, OCLC's linked data researchers are confident that classes and properties defined in Schema.org, with minor extensions defined in BiblioGraph, can address many of the requirements specified in the Library of Congress study.

Figure 5.9 shows a high-level model for the filmed version of *The Nutcracker* inspired by the requirements specified in the study. Some of the features were first proposed by Mixter.[47] The model shows relationships among four creative works: the original Hoffmann fairy tale, the Tchaikovsky ballet, a performance of the ballet at the Bolshoi Theater, and a film of the performance recorded on a video object. All are assigned the type "schema:CreativeWork," subclasses such as "schema:Movie," or plausible BiblioGraph extensions such as "bgn:Ballet." As in the previous example, the creative work may have more than one type assignment that can be interpreted as an ontologically distinct facet. For example, the ballet performance is a creative work as well as a "schema:Event" because it is anchored in a time period during which a group of agents realize an artistic creation. Similarly, the video object is a creative work and a "schema:ProductModel" because it is a manufactured object with a model number that can be bought, sold, borrowed, and tracked. The creative works are linked through two properties defined in Schema.org, "schema:encodesCreativeWork" and "schema:workPerformed"—as well as the property "isBasedOn," which could be defined in BiblioGraph with a meaning equivalent to the corresponding RDA term that has been "unconstrained," or redefined to exclude a literal reference to the definitions of FRBR Work, Expression, Manifestation, and Item, as mentioned earlier.[48]

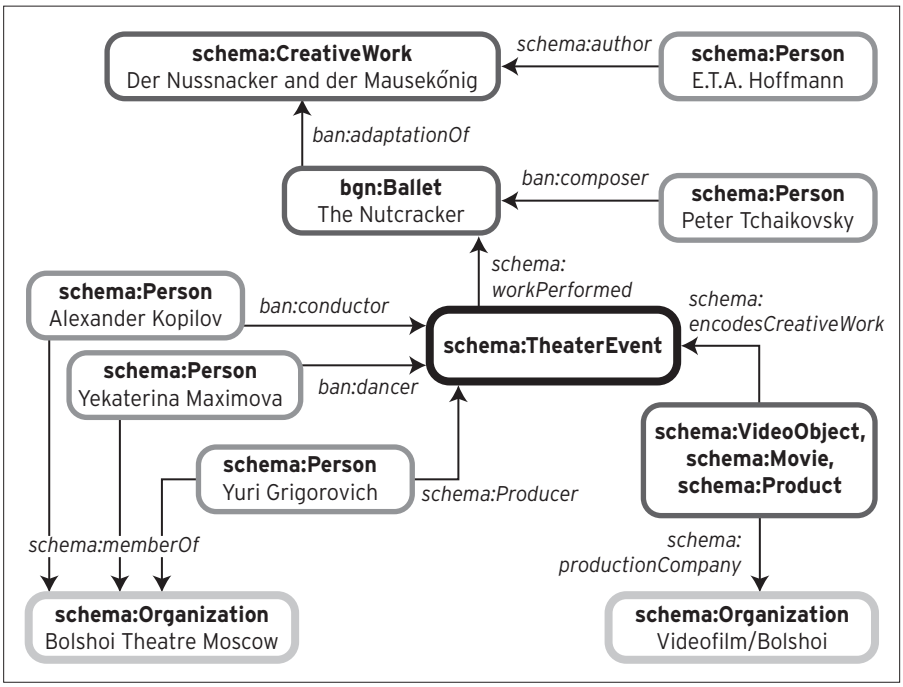


Figure 5.9 | **An aspirational model of a recorded creative work**

Once the four creative works have been identified, the agents can be distributed appropriately, clearing up a major source of confusion in the original MARC record and the RDF statements derived from it. Thus E. T. A. Hoffmann is the author of the fairy tale, Tchaikovsky is the composer of the ballet; Kopolov, Maximova, and Grigorovich are members of the Bolshoi Theater and performers in the theater event with identifiable roles, and Video-Film/Bolshoi is a distributor of the video object. Roles that are not defined in Schema.org, such as "composer," "conductor," or "dancer" could be defined in BiblioGraph. Or they could be entered as string values in a generic "Role" class, which as Wallis shows, permits a refinement or a narrower meaning for a defined role such as "contributor" or "performer."[49] Though the technical result is the same, the BiblioGraph representation would serve as a tool for negotiating with Schema.org to adopt a controlled list to which URIs might be assigned, thus anticipating that users on the broader Web might also need normalized terms to describe dancers, composers, and others involved in the production of a multimedia work.

The model shown in figure 5.9 is still a sketch because details that would be important to some users have been suppressed. But what is clear is that this example is dramatically more complex than the description of the memoir discussed earlier in this chapter, whose referent was a single creative work and whose properties could be specified through a set of lexical mappings directly from the MARC source. Though the model shown in figure 5.9 is not discoverable with current algorithms, it can at least be defined with relative ease in Schema.org and BiblioGraph by starting with the model developed for published monographs and enhancing it with a more complex network of relationships that results when identifiable acts of creation are realized as variant genres, content types, and physical realizations. As in the earlier examples, the core of the model is a set of creative works connected through relationships to one another and to the people and organizations that brought them into being.

The model also addresses two problems identified in the survey of options in the Library of Congress audiovisual study. The first is the assignment of "Agent" role relationships to the respective creative works. Without the constraints of a data model derived from FRBR, the model derived from Schema.org can simply name the "producer" relationship between Yuri Grigorovich and the performance, as well as all other Agent-Work relationships shown in figure 5.13, without having to consider whether the performance is a Work, Expression, or Manifestation, and whether Grigorovich is a primary creator or relegated to a contributor. This problem results only if properties are partitioned among Works, Expressions, Manifestations, and Items. But as the discussion of the RDF statements in figure 5.5 shows, most properties defined in OCLC's Semantic Web models of bibliographic description can float freely among these concepts. The result is the ontologically more natural set of statements asserting that a dancer is a primary creator of a performance event, a composer is the primary creator of a musical score, and an author is the primary creator of a fairy tale.

At a more abstract level, the authors of the audiovisual study argue that filmed works imply the existence of an event that unfolds in elapsed time, which is problematic for MARC and nearly every subsequently defined model of resource description that does not define "Event" as a primary concept. In Schema.org, "Event" is defined as a subclass of "schema:Thing," which implies an existence in multiple contexts that may or may not have a relationship with a creative work. The event implied in the *Nutcracker* description obviously does, but the same model could just as easily describe a video object that records a tsunami or a hurricane. The video object would still be a creative work with a producer and other human agents, but the event itself would not be doubly

typed as "schema:CreativeWork" and would not be associated with identifiable human agents or have a traceable connection to another creative work. The network of relationships would be much simpler than that shown in figure 5.9.

CONCLUSION AND NEXT STEPS

The model for multimedia resources presented above is only a thumbnail sketch that shows the feasibility of expanding the use of Schema.org to describe a set of resources that are especially challenging and underserved by current library standards. It must be expanded to cover a broader range of examples and refined with input from the audiovisual cataloging community. Not coincidentally, these resources also show that OCLC's current strategy of populating linked data models through retrospective conversion are reaching an upper limit and will have to be upgraded—first, with more sophisticated text-mining algorithms and improved procedures for distinguishing high-quality data from errors, and ultimately with user interfaces that accompany the transformation of the data architecture for library metadata and support the yet-to-be-defined workflows of next-generation cataloging.

But the audiovisual model and the other examples discussed in this chapter also prescribe priorities for future development. In the modeling division of labor, we are still awaiting the specialist's view, which will define the vocabulary of curation for resources managed in cultural heritage institutions by experts with graduate degrees. As a result, current standards efforts in the library community are focused on the development of Semantic Web models of the ordinary-user view. Thus it is reasonable to point out that a cluster of creative works involving a fairy tale, a performance event, a ballet, a movie, and multiple agents can be described in Schema.org with a relatively small set of extensions. Given the intense public interest in recordings of multimedia performances, popular and authoritative resources such as the Internet Movie Database and MusicBrainz can be consulted to obtain additional clues for defining a model of such resources that is widely understood and may even turn out to be richer than current MARC models of audiovisual materials.[50] The resulting descriptions promise to facilitate discovery on the open Web because they would be expressed in the language that search engines can consume and would describe concepts that are important to the information-seeking public. And once all of the pieces are in place, the outcome will be a compelling demonstration of the

benefits that accrue when libraries are more closely integrated into the Web, a goal to which the linked data paradigm is well-suited, and to which all of the library community's next-generation modeling effort aspires.

ACKNOWLEDGMENTS

This chapter describes work being done by a team of linked data experts at OCLC—in particular, Jon Fausey, Tod Matola, Jeff Mixter, Karen Smith-Yoshimura, Stephan Schindehette, Bruce Washburn, Richard Wallis, and Jeff Young. I have contributed to this effort. But I am solely responsible for the errors of analysis and perspective presented here.

Notes

1. Cathy de Rosa, Joanne Cantrell, Matthew Carlson, Peggy Gallagher, Janet Hawk, and Charlotte Sturtz, *Perceptions of Libraries: Context and Community,* a report to the OCLC membership (Dublin, OH: OCLC, 2010). www.oclc.org/content/dam/oclc/reports/2010perceptions/2010perceptions_all.pdf.
2. Library of Congress, *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control Library of Congress,* January 9, 2008, www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf.
3. Schema, "Thing–>CreativeWork," 2015, http://schema.org/CreativeWork.
4. Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter, *Library Linked Data in the Cloud: OCLC's Experiments with New Models of Library Resource Description: Synthesis Lectures in Linked Data and the Semantic Web,* 2015. A publication in the Morgan & Claypool Publishers series Synthesis Lectures on the Semantic Web: Theory and Technology. doi: 10.2200/S00620ED1V01Y201412WBE012.
5. Eric Miller and Bob Schloss, eds., "Resource Description Framework (RDF) Model and Syntax," Version 1, October 2, 1997, World Wide Web Consortium, www.w3.org/TR/WD-rdf-syntax-971002/; VIAF, "VIAF: Virtual International Authority File," 2014, http://viaf.org; FAST (Faceted Application of Subject Terminology), 2014, "FAST Linked Data: FAST Authority File," OCLC Experimental, http://experimental.worldcat.org/fast/; DDC. "Dewey Decimal Classification / Linked Data." 2014. OCLC. http://dewey.info.
6. Tim Berners-Lee, "Linked Data," in *Design Issues: Architectural and Philosophical Points,* July 27, 2006, World Wide Web Consortium, www.w3.org/DesignIssues/LinkedData.html.

7. Alistair Miles and Sean Bechhofer, "SKOS Simple Knowledge Organization System: Reference," W3C Recommendation, August 18, 2009, World Wide Web Consortium, www.w3.org/TR/2009/RECskos-reference-20090818/.
8. Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech. "LCSH, SKOS, and Linked Data," in *DC-2008: Proceedings of the International Conference on Dublin Core and Metadata Applications*, 25–33, Berlin, Ger.: Dublin Core Metadata Initiative, 2008, http://edoc.huberlin.de/conferences/dc-2008/summers-ed-25/PDF/summers.pdf.
9. Thomas Baker, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, et al., "Library Linked Data Incubator Final Report," W3C Incubator Group Report, October 25, 2011, World Wide Web Consortium, www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.
10. Tim Hodson, Corine Deliot, Alan Danskin, Heather Rosie, and Jan Ashton, "British Library Data Model – Book, V.1," British Library, August 4, 2012, www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf; Leigh Dodds, "An Introduction to the British National Bibliography, Part I," in *Lost Boy* (blog), October 28, 2014, http://blog.ldodds.com/2014/10/08/an-introduction-to-the-british-national-bibliography/.
11. BIBO, "The Bibliographic Ontology: Bibliographic Ontology Specification," 2009, http://bibliontology.com.
12. Gerard de Melo, "Lexvo.org Main Page," 2014, Lexvo.org. www.lexvo.org.
13. British Library, "Collection Metadata: Data Services," 2014, West Yorkshire, United Kingdom: British Library.
14. Google, "Introducing Schema.org: Search Engines Come Together for a Richer Web," in *Webmaster Central Blog: Official News on Crawling and Indexing Sites for the Google Index*, June 2, 2011, http://googlewebmastercentral.blogspot.com/2011/06/introducing-schemaorg-search-engines.html.
15. Dan Brickley, R.V. Guha, and Andrew Layman, "Resource Description Framework (RDF Schemas)," W3C working draft, April 9, 1998, World Wide Web Consortium, www.w3.org/TR/1998/WD-rdf-schema-19980409/.
16. Alan Morrison, Gabriel Kniesley, Marie Wallace, and Matt Everson, "Is Schema.org Good or Bad for the Semantic Web?" in Quora (online reference service), June 14–July 2, 2011, www.quora.com/Is-Schema-org-good-or-bad-for-the-Semantic-Web.
17. R. V. Guha, "What a Long, Strange Trip It's Been," presentation at the 2014 Semantic Technology and Business (SemTech) Conference, San Jose, CA, August 20, 2014, +www.slideshare.net/rvguha/sem-tech2014c.
18. Ian Hickson, "HTML Microdata," W3C working group note, October 29, 2013, World Wide Web Consortium, www.w3.org/TR/microdata/; Google, "Promote Your Content with Structured Data Markup," in *Google Developers*, February 12, 2015, https://developers.google.com/structured-data/?hl=ta&rd=1; Google, "Introducing the

Knowledge Graph: Things, Not Strings," in *Google Official Blog*, May 16, 2012, http://googleblog.blogspot.com/2012/05/introducing- knowledgegraph-things-not.html.
19. Martin Hepp, "GoodRelations: The Web Vocabulary for E-Commerce," 2014, www.heppnetz.de/projects/goodrelations/.
20. SBX, "Holdings via Offer," Schema Bib Extend Community Group, W3C Community and Business Groups, World Wide Web Consortium, 2014, www.w3.org/community/schemabibex/wiki/Holdings_via_Offer.
21. OCLC, "OCLC Adds Linked Data to WorldCat.org," OCLC news release, June 20, 2012, www.oclc.org/news/releases/2012/201238.en.html.
22. OCLC, "OCLC Research Activities and IFLA's Functional Requirements for Bibliographic Records," 2015, OCLC Research, www.oclc.org/research/activities/frbr.html; Barbara Tillett, "What Is FRBR? A Conceptual Model for the Bibliographic Universe," Library of Congress Cataloging Distribution Service, 2004, www.loc.gov/cds/downloads/FRBR.PDF.
23. W3C, "Semantic Web Development Tools," World Wide Web Consortium, 2014, www.w3.org/2001/sw/wiki/Tools.
24. SBX, "Schema Bib Extend Community Group," W3C Community and Business Groups, 2014, World Wide Web Consortium, www.w3.org/community/schemabibex/.
25. W3C, "public-schemabibex@w3.org: Mail Archives," last modified August 11, 2015, World Wide Web Consortium, https://lists.w3.org/Archives/Public/public-schemabibex/; W3C, "public-vocabs@w3.org: Mail Archives," last modified August 11, 2015, World Wide Web Consortium, https://lists.w3.org/Archives/Public/public-vocabs/2015Aug/author.html.
26. Dan Brickley and Libby Miller, "FOAF Vocabulary Specification 0.99," Namespace Document–Paddington Edition, January 14, 2014, http://xmlns.com/foaf/spec/.
27. Carol Jean Godby, "The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper," 2013, http://oclc.org/content/dam/research/publications/library/2013/2013–05.pdf.
28. Schema, "Extension Mechanism," 2015, https://schema.org/docs/extension.html.
29. Hilary Putnam, "Meaning and Reference," in *Naming, Necessity, and Natural Kinds*, ed. Stephan P. Schwartz (Ithaca, NY: Cornell University Press, 1977), 119–34.
30. Leo Sauermann and Richard Cyganiac, "Cool URIs for the Semantic Web," W3C Interest Group note, December 3, 2005, World Wide Web Consortium, www.w3.org/TR/cooluris/.
31. Library of Congress, "BIBFRAME AV Modeling Study: Defining a Flexible Model for Description of Audiovisual Resources," last modified May 15, 2014, www.loc.gov/bibframe/pdf/bibframe-avmodelingstudy-may15–2014.pdf.

32. Godby, Wang, and Mixter, *Library Linked Data in the Cloud*. doi: 10.2200/S00620ED1V01Y201412WBE012.
33. RDA (Resource Description and Access), "RDA Toolkit: Resource Description and Access," 2010, www.rdatoolkit.org.
34. RDA, "RDA Element Sets: Unconstrained Properties," RDA Registry, last modified April 7, 2015, www.rdaregistry.info/Elements/u/.
35. Library of Congress, "240—Uniform Title (NR)," in *MARC 21 Format for Bibliographic Data*, *1999 Edition*, Network Development and MARC Standards Office, Library of Congress, 2014, www.loc.gov/marc/bibliographic/bd130.html.
36. Karen Smith-Yoshimura, "Multilingual Bibliographic Structure," OCLC research update at the Annual Conference of the American Library Association, Las Vegas, NV, June 30, 2014, OCLC Research, YouTube video, https://www.youtube.com/watch?v=NG1tkE03WJo.
37. Karen Smith-Yoshimura and Carol Jean Godby, "An OCLC Perspective on What It Takes to Make Linked Data Work," presentation at the ALCTS ALA Preconference, "Beyond the Looking Glass: Real World Data: What Does It Take to Make It Work?" San Francisco, CA, June 26, 2015.
38. Library of Congress, "Code List for Relators," Network Development and MARC Standards Office, last modified May 13, 2010, www.loc.gov/marc/relators/.
39. Tillett, "What Is FRBR?" www.loc.gov/cds/downloads/FRBR.PDF.
40. OCLC, "OCLC Releases WorldCat Works as Linked Data," OCLC news release, April 28, 2014, https://www.oclc.org/news/releases/2014/201414dublin.en.html.
41. Godby, Wang, and Mixter, *Library Linked Data in the Cloud*. doi: 10.2200/S00620ED1V01Y201412WBE012.
42. Tillett, "What Is FRBR?" www.loc.gov/cds/downloads/FRBR.PDF.
43. RDA, "RDA Toolkit: Resource Description and Access," 2010, www.rdatoolkit.org.
44. Smith-Yoshimura and Godby, "An OCLC Perspective on What It Takes to Make Linked Data Work."
45. Schema, "Thing–>Property>Performer," 2015, http://schema.org/performer.
46. Library of Congress, "BIBFRAME AV Modeling Study: Defining a Flexible Model for Description of Audiovisual Resources," last modified May 15, 2014, www.loc.gov/bibframe/pdf/bibframe-avmodelingstudy-may15–2014.pdf.
47. Jeffrey K. Mixter, "Linked Data in VRA Core 4.0: Converting VRA XML Records into RDF/XML," thesis submitted to the College of Communication and Information in partial fulfillment of the M.S. and M.L.I.S. degrees, Kent State University, 2013, http://jmixter.s3-website-us-east-1.amazonaws.com/thesis/LinkedDataInVRACore4.pdf.

48. RDA (Resource Description and Access), "RDA Element Sets: Unconstrained Properties," RDA Registry, last modified April 7, 2015, www.rdaregistry.info/Elements/u/.
49. Richard Wallis, "The Role of Role in Schema.org," in *Data Liberate* (blog), April 15, 2014, http://dataliberate.com/2014/09/a-step-for-schema-orga-leap-for-bib-data-on-the-web/.
50. IMDB, "IMDb" (Internet Movie Database), 2015, www.imdb.com; MB, "MusicBrainz," 2015, https://musicbrainz.org/.