
*PCC SCS/LDAC Task Group on the
Work Entity*

Preliminary White Paper

1 October 2017

Karen Coyle
Nancy Fallgren
Steven Folsom
Jean Godby
Stephen Hearn
Ed Jones, *chair*

Executive summary

The PCC SCS/LDAC Task Group on the Work Entity was charged with producing a white paper to give a high-level outline of the issues surrounding the identification of work entities, considering the contrasting conceptions of the work entity emerging in various communities and the implications of working with multiple data models. Additionally, the white paper was to explain the role of work identifiers in a Linked Data environment, including the modeling and metadata management issues they raise, and propose feasible options to advance the provision of work-level metadata.

The white paper provides a history of the concept of the work in the library community, and its description is explored from the nineteenth century to the present (FRBR, RDA, and MARC 21). We discuss the nature and use of identifiers for works along with the role of graphs and classes in RDF to manage and correlate metadata about works.

Works have been and are being modeled in multiple relevant standards. The paper discusses the work as modeled in the library community by the FRBR entity-relationship model, RDA, BIBFRAME, BIBFRAME Lite, the FRBR object-oriented model (CIDOC CRM, FRBR_{OO}, and PRESS_{OO}), Schema.org, and Dublin Core, as well as the way it is modeled in the publisher and intellectual property communities, observing the varying ways in which the concept has been realized in different systems to meet different user needs.

While the Task Group has been able to assemble a considerable amount of information on topics outlined in the charge to the group, there are many important questions that cannot be answered, either by this group or by anyone at this moment in time. Many of these questions have to do with how the concept of the work and its encoding will be implemented in systems in the future. Other questions are primarily bibliographic in nature.

In keeping with its charge, the task group does not offer a set of recommendations. Instead we provide a framework for a clearer discussion around PCC policies relating to the work entity.

The white paper addresses some preconceived ideas that are common in discussions relating to FRBR-inspired bibliographic models. In particular, the Task Group rejects the assumption that there will be a single, canonical work identifier and work “record” that prevails in future bibliographic practice. Instead, the library community will need to embrace bibliographic descriptions that vary as needed to serve diverse user groups while using available technology to promote sharing among these varying practices. The white paper also promotes a “post-record” view of metadata that is introduced by RDF, questioning the common assumption that the metadata describing the bibliographic work is stored in a work record. Unless we challenge

these assumptions, we risk carrying over to a future model technical requirements that are not suited to either entity-relation or RDF capabilities.

Some of the questions that are raised in this paper are:

1. What is the relationship between works in the FRBR sense and work authorities (title and name/title authorities) as defined in today's cataloging?
2. What functionality is desired that would require a work *entity* to be created? Does it need to be created for every cataloged resource?
3. What are the cataloging workflow concerns that relate to the work ...
 1. as a bibliographic description?
 2. as a defined bibliographic entity?
4. How can the scope of the work description be defined? Does it include the properties drawn from the creator and subject entities? How much of the overall bibliographic graph is needed for different functions such as cataloging a new expression, user displays, etc.?

In nearly every area where the white paper discusses the nature of the work—from usage in various communities to the role of identifiers and the algorithmic conversion of legacy data—there is no single solution. We encourage PCC in particular and the cataloging community in general to be active participants in seeking answers to these questions. Although technology can be developed to implement works and other emerging bibliographic concepts, technology itself should not drive solutions to what are essentially bibliographic concerns.

These questions and others are discussed in greater length in the final section of the report. They should, however, be seen as an incomplete list of areas for further exploration by the cataloging and library technology communities. We should expect that other questions will arise as library systems and practices evolve.

The appendices provide additional background on the concept and utility of the work in the publishing and intellectual property communities and about the challenges posed by legacy data.

1: Introduction

Background

Action 3.3 in the PCC 2015-2017 Strategic Plan called for the Standing Committee on Standards to charge and establish a task group that would

Produce a white paper to give a high-level outline of the issues surrounding the identification of work entities. This document should consider the contrasting conceptions of work entities emerging in different communities (e.g. BIBFRAME, JSC, PRESS₀₀, Zepheira serials group and others) and the implications of working with multiple data models. The white paper will explain the role work identifiers can play in a linked data environment, outline the modelling and metadata management issues they raise, and propose feasible options for the PCC to advance the provision and use of work-level metadata.¹

This task group was duly charged and established in February 2016 and made interim reports at the PCC Policy Committee meeting in November 2016 and the PCC Operations Committee meeting in May 2017.

Organization of the white paper

The White Paper comprises four main sections, questions for the PCC, and appendices:

- An introduction providing the background, organization, and the assumptions underlying the paper
- A discussion of the work as it has been used in the cataloging community over time
- A discussion of the role of work identifiers and descriptions
- A discussion of the work as it is used in various models in the cataloging community and related communities
- Questions for the PCC
- Appendices 1 and 2 elaborating respectively the use of the work by the publishing and IP communities
- Appendices 3 and 4 elaborating respectively the representation of the work in legacy data and OCLC's efforts to discover the work in that data

Assumptions

To carry out its work the task group necessarily made certain assumptions about the future cataloging environment based on past PCC commitments, in particular that

- Resource Description and Access (RDA) would continue to be the cataloging standard

¹ Program for Cooperative Cataloging. Vision, Mission, and Strategic Directions, January 2015-December 2017, revised: November 20, 2015 <https://www.loc.gov/aba/pcc/about/PCC-Strategic-Plan-2015-2017.pdf>

- PCC cataloging would continue to take place as it does today, using a shared standard record for serials cataloging and authority control, but with more tolerance for varying catalog records for monographs

The task group also recognized that the standards employed in PCC cataloging were in a state of flux, with plans underway to

- Supersede the FR family of entity-relationship models with a consolidated model (IFLA LRM) that includes many changes, some of which affect the work entity (principally as it relates to aggregates and serials)
- Supersede the current version of RDA with a restructured and redesigned version (3R Project) that will bring RDA into conformity with IFLA LRM
- Supersede the current record syntax for PCC catalog records (MARC 21) with a syntax that is grounded in Linked Data techniques (BIBFRAME)

As of this writing, none of these transitions has been completed, though IFLA LRM, on which both RDA and BIBFRAME are to some extent dependent, is complete except for its endorsement by the IFLA Professional Committee. The RDA 3R Project has been undertaken in the expectation that there will be no further significant changes to IFLA LRM.

BIBFRAME, after completion of a phase one pilot (September 8, 2015-March 31, 2016), was revised in March-April 2017 (BIBFRAME 2.0), and a phase two pilot has now begun. Unlike IFLA LRM and the RDA 3R Project, BIBFRAME is not an updating of an existing standard but rather a new standard that remains in the testing phase. It is unclear at this point how the simpler entity structure and looser domain and range specifications of BIBFRAME will be reconciled in practice with the disjoint work and expression entity definitions of RDA

Work description

This paper will use the phrase *Work description* to refer to any graph of RDF triples that describes a Work entity. This differs from a Work record both in that it employs the open-world assumption (OWA)² and that the set of triples is not predefined: any triple with a property that can take a Work as its subject is a potential member of the graph.

The FRBR Work and Expression entities

This paper will necessarily examine both the FRBR Work and Expression entities inasmuch as these two abstract entities share an indistinct boundary, and several of the models we will be discussing ignore, downplay, or redefine that boundary.

² Open-world assumption (Wikipedia) https://en.wikipedia.org/wiki/Open-world_assumption

2: The Work

2.1: A brief historical review

The work as a bibliographic concept has long been part of Anglo-American library practice, if not terminology. As early as 1674, Thomas Hyde tried to bring together books published under a variety of titles for his Bodleian Library catalog.³ In 1847, Sir Anthony Panizzi explicitly invoked the *work* when he set out before a royal commission the rationale for his elaborate cataloging rules for the British Museum library: so that “a reader may know the *work* he requires; he cannot be expected to know all the peculiarities of the different *editions*; and this information he has a right to expect from the catalogues.” In fact, so central is the idea of the *work* that in 1979 Seymour Lubetzky warned that a new cataloging code about to be introduced at the time—AACR2—was in danger of abandoning it.⁴ With the FRBR conceptual model in 1997, the *work* roared back, at least in cataloging theory.

Although there was not a specific definition of the term until the FRBR final report (“a distinct intellectual or artistic creation”), the *work* has long had a central role in Anglo-American cataloging, where it is bound up with the related concept of *author*, a concept that has changed over time. Until the second edition of the Anglo-American Cataloguing Rules (AACR2, 1978), “authors” included editors (responsible for bringing together the writings of others in a single book) and corporate bodies (responsible for the publications issued in their names). At the title level, works published under various titles originally simply referenced one another from their separate locations in the catalog (1908 Anglo-American code), though by 1941 they were being brought together—at least in American practice—via the uniform title, defined as “The distinctive title by which a work which has appeared under varying titles and in various versions is most generally known.”⁵

The combination of author and title (if necessary, a uniform title), was the technique used in alphabetical catalogs (book catalogs, card catalogs, and early online catalogs) to bring together the editions and translations of a work, a sequence of elements known as the main entry. Authority records were created as necessary to direct the catalog user to the main entry from variants under which they might look for the work (for example, from the author and title proper of a translation to the author and uniform title of the work, followed by the name of the language into which it had been translated, or from a co-author and title proper to the author and title used as the main entry).

³ Richard P. Smiraglia, *The Nature of “A Work”* (Lanham, Md.: Scarecrow Press, 2001), 16-17.

⁴ Seymour Lubetzky, The Fundamentals of Bibliographic Cataloging and AACR2. In *The Making of a Code: the Issues Underlying AACR2*. Ed. By Doris Hargrett Clack. (Chicago: American Library Association, 1980), 16-25

⁵ A.L.A. *Catalog Rules: Author and Title Entries*, preliminary American 2nd ed. (Chicago: American Library Association, 1941), xxxi [https://hdl.handle.net/2027/uc1.\\$b354381](https://hdl.handle.net/2027/uc1.$b354381)

Flaubert, Gustave

Sentimental education

see **Flaubert, Gustave**

Éducation sentimentale. English

[Example under AACR2 rule 26.4B1 (2002 revision, 2005 update)]

Child, Lincoln. Ice limit

see

Preston, Douglas J. Ice limit

[Example with presentation according to RDA E.1.3.2]

All this is embodied in the MARC21 record syntax, with its dedicated fields for the author (100-111) and uniform title (130 and 240).⁶ In cases where an explicit uniform title is not needed, it is present implicitly in the title proper (field 245). In MARC records, therefore, the work is made explicit only “as needed” and is created using a combination of data from various fields. It is presented together in one field only when it is used as an access point on the bibliographic record for a different work.

In this context, authority records for works existed only to control access points for those works in an alphabetical catalog. However, most authorized access points for works are established only indirectly, in the process of creating authority records for expressions of those works, such as translations. Only incidentally might they *describe* the work or the expression as defined in FRBR. Unlike authority records for names and subjects, if there was no need to control access points (e.g., a work had not been published under more than one title or had not been translated), no authority record would be created for the work or any of its expressions. This principle still governs the creation of authority records for works and expressions, and consequently the works present in the vast majority of cataloged resources (with some exceptions, such as musical works) today are not represented by authority records. Note also that the need for a work or work/expression access point is relative to the context of a specific catalog. The inclusion of uniform titles differs across collections, as well as library types and sizes.

The FRBR final report (1998) defines a work as “a distinct intellectual or artistic creation,” and in the FRBR conceptual model the work entity has a small (non-exhaustive) set of attributes. Although this was the most precise exposition of the bibliographic concept of work to date, the working group that developed FRBR admitted that the boundaries of the work entity would vary

⁶ On printed cards, uniform titles in field 130 were printed in boldface with a hanging indent, while those in field 240 were printed between square brackets with a regular indent.

across communities of practice, and for that reason the boundary between work and expression represented by the various work-work, work-expression, and expression-expression relationships in the final report was valid not universally for all bibliographic communities but only for purposes of the study.⁷

The FRBR conceptual model underlies the current cataloging standard, Resource Description and Access (RDA), which includes the FRBR definition of the work entity, slightly clarified by the Statement of International Cataloguing Principles (ICP) (2009) as “[a] distinct intellectual or artistic creation, that is, the intellectual or artistic content.”⁸ The instructions for describing that entity (as is the case for other entities in RDA) are not so precise as to guarantee uniformity of definition nor a uniform description of the work entity. The cataloger must also take into account the various transformations represented by relationships between works and expressions in RDA as well as a great deal of heuristics in distinguishing tolerable from intolerable variation within a particular cataloging context .

Data model designers favor precise models and clear definitions, but a data model should not impose restrictions that limit the expressivity of the data itself. Any model that can accommodate works as defined in RDA must be able to accommodate all the interpretations of work that RDA allows. It is the specific application of the cataloging standard that will determine the boundaries for works and will enforce consistency of bibliographic control of works where that is desired.

The definition of the work in FRBR, the development of RDA, and the experimentation with linked data models are converging, although they are not yet fully coordinated. It should be noted that the encoding of works as entities does not require linked data. Such encoding could be accomplished using other data models, such as the markup language XML. Coordination with linked data, however, means that the various bibliographic relationships between works and works, works and expressions, and expressions and expressions—as well as the other bibliographic relationships that the FRBR model has introduced—can be easily accommodated.

A more detailed treatment of the Work in knowledge organization (through 1999) can be found in Richard Smiraglia’s *The Nature of “A Work”* (2001).⁹

⁷ IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report* (München: K.G. Saur, 1998), 16-17
<https://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf>

⁸ RDA Toolkit (through April 2017 update), Glossary, s.v. “work”

⁹ Richard P. Smiraglia, *The Nature of “a work”* (Lanham, Md.: Scarecrow Press, 2001). See especially chapter 2: The Concept of the Work in Anglo-American Cataloging and chapter 3: Bibliographic Relationships Give Parameters to the Concept of a Work.

2.2: Three kinds of Work

In general, the term *work* is understood in at least three senses:

- (1) As the categorization arrived at by an intellectual decision-making process employed by the cataloger. This is the process by which the cataloger identifies the "distinct intellectual or artistic creation" embodied in the manifestation being cataloged, distinguishing it from other similar but distinct works within the cataloging environment.
- (2) As the description that results from applying this process within the RDA context. This work description represents the work using a collection of appropriate RDA elements, sometimes including a discrete authorized access point for the work. Catalogers use these elements in descriptive cataloging even when the work is not defined as a standalone entity, as is the case when encoding RDA cataloging using the MARC 21 Bibliographic record syntax.
- (3) As the entity represented by organizing this description in a particular data structure. This data structure could be realized in any number of different encodings, such as ISO 2709 (the underlying record structure for MARC21), XML, or JSON. The latter two can also accommodate linked data in the form of serializations such as RDF/XML and JSON-LD.

The answers to many of the questions posed in this paper hinge on the extent to which the work description and work entity can capture the often-tacit steps comprising the intellectual process of identifying the work.

As we see with the implementation of RDA cataloging in the MARC21 environment, not every work description must be stored at the data structure level as a work entity. Whether work decisions are required for every cataloged manifestation is a cataloging policy decision. Whether work entities are required for every work is a data design decision. The latter may or may not require some explicit action on the part of the cataloger, depending on the interface in which cataloging takes place.

This decision will be influenced to some extent by how works are identified and described in a given system, a topic we address in the next section.

3: Identifiers

An identifier is a token--usually an alphanumeric string--agreed upon by a particular community to represent an object of interest within a defined environment. The identifier must be unique within that environment.

A general rule for identifiers is that each identifier identifies one thing and always that one thing. The consequences of this rule are:

- If two different things are given the same identifier, they become the same thing for the purposes of identification.
- If an identifier for a thing changes, that has the effect of creating a new “thing” in the identified universe.
- Identifiers must not be re-used, even if the thing they identify is no longer in existence.

However, it is not the case that every identified thing will have only one identifier. Different identifiers may be assigned for different purposes or by different systems. This is commonly the case for us in our daily lives where a person may have a passport number that identifies her, a state driver’s license number, a number on a doctor’s office chart, a student ID number, and a Social Security Number. Each of these is unique within its context.

In the past, the library community has often relied on display strings as identifiers. While a string of any type can be used to identify something, the problem with display strings is that they may need to change when the desired display form changes. A person identified as:

Doe, Jane, 1937-

may become:

Doe, Jane, 1937-2016

In this situation, the display form has become obsolete as an identifier because the facts it reflects have changed (the second bullet above). Properly managed identifier systems allow us to separate identification from metadata intended for indexing and display, and should make our data more stable over time.

Much of the effort required by authority work is spent on formulating and ensuring the uniqueness of an authorized access point. The unique AAP was a necessity in the context of browse indexes. In the context of a search-driven environment and descriptions with designated data elements and unique identifiers, the formulation of representations to assist users with identifying and selecting a resource may become more flexible and less onerous. The larger challenge will be moving beyond the flat files of resources often found in library catalogs now to hierarchical, networked files enabling searchers to navigate from generalized representations of entities of varying types to more granular representations of particular entities and resources. If

that can be managed, the traditional authority file of AAPs could be replaced by shared entity representations integrated with the local catalog.

3.1: Linked Data Identifiers (IRIs)

The identifiers used in linked data are called IRIs (Internationalized Resource Identifier). They have the same format as URLs, beginning with `http://` or `https://`, but IRIs can accept any script, not just Latin-based scripts.

IRIs are what the technology community calls “opaque”: they are just strings and they don’t have any meaning in the human understanding sense. What is special about them is that each IRI is unique in the context of the web because the first segment of the identifier uses a domain name owned by the party minting the identifier. This means that it should not be possible for two parties to accidentally create the same IRI. IRIs are generally created by and managed by systems, much in the same way that OCLC numbers are assigned by the OCLC system when a user saves a new record.

The linked data identifier makes use of the technical platform of the Web. Because it begins “`http://`” it has all the functionality of a web address: it is managed by the domain name system of the Internet; it can also be used as a locator, primarily for additional information about the thing identified; it can be used anywhere on the web. IRIs are unique on the web and use the same naming conventions and controls that are governed by the Domain Name System. It is this functionality that will support linking using the Web as the technical platform.

IRIs in RDF provide an agreed on identifier for a thing, and a thing is anything, whether physical or conceptual. Because RDF is based on graphs, not records, the IRI provides a stable identifier for the thing but does not define a record or a set of descriptors that would carry the information about that thing. That information must be defined elsewhere. (See section “Descriptions” below.)

Because IRIs are designed for use by machines, they are not intended to be human-friendly or to have human-understandable meaning. In the majority of cases where identifiers are shown to people, users and data creators will see display labels, not identifiers. This means that a cataloger will see “Le Carré, John, 1931-” and not “`http://id.loc.gov/authorities/names/n79083252`” or “`http://viaf.org/viaf/109254932`”.

Where identifiers themselves are meaningful in the workflow, they can have a human-readable label that functions very much like the identifiers that are familiar to library cataloging, such as LCCNs and ISBNs. In the abbreviated example below, the VIAF information about a person has an IRI that is the identifier for the person in the VIAF system. There are display forms for both the person (“Julia Pettee”) and for the VIAF identifier (“76683403”).

```
<rdf:Description rdf:about="http://viaf.org/viaf/76683403">
  <skos:prefLabel>Julia Pettee</skos:prefLabel>
  <dcterms:identifier>76683403</dcterms:identifier>
</rdf:Description>
```

Ideally, humans will not see raw IRIs, but will instead see friendlier identifiers, like:

VIAF ID: 76683403

Where needed, identifiers can be searchable, and the search will use the friendly string. For example, with the LCCN, a person will search on 8002589, not <https://lccn.loc.gov/8002589>, and preferably it is the former that will be displayed to the person, not the latter, possibly with a label such as:

LCCN: 8002589

Note that there is no requirement that the displayed form of the identifier be a part of the eye-readable IRI. However, if the display forms will be used for searching and linking, it should be possible to render them as unique IRIs to avoid ambiguity.

As described so far, linked data identifiers should look to the cataloger very much like standard bibliographic identifiers look today. The difference, however, is that because they are actionable and may be used in the open environment of the web for linking, IRIs must be assigned automatically by systems during data creation processes, and not by humans. They also should not be modified by human data creators, but always managed by machine processes that enforce the requirements for IRI uniqueness.

3.2: Identifiers for Works

It is generally assumed that work descriptions will have identifiers. While this is true in certain well-defined domains such as western classical music (thematic catalog numbers), it is not true for schemes such as the International Standard Text Code (ISTC) and the International Standard Musical Work Code (ISWC), which have seen limited uptake in their domains.

Where work entities are created in RDF models, those will necessarily have identifiers. Most likely, the concept in people's heads is that one work = one unique identifier. We need to begin with the recognition that the idea that a work will be universally assigned a single identifier that is recognized by everyone is unrealistic. Following the rule above, any given identifier must always identify a single work, but we will undoubtedly have multiple identifiers for the same, or for almost the same works, especially works identified in different contexts, such as RDA and BIBFRAME. This is a situation that can be managed by creating machine-actionable statements of equivalence or similarity. A mechanism included in the linked data technology allows one to declare that two identifiers represent either the exact same thing, or two things that are similar, and this function can be either human-mediated or may be an algorithm, such as those used by OCLC and VIAF to bring together data from different sources.

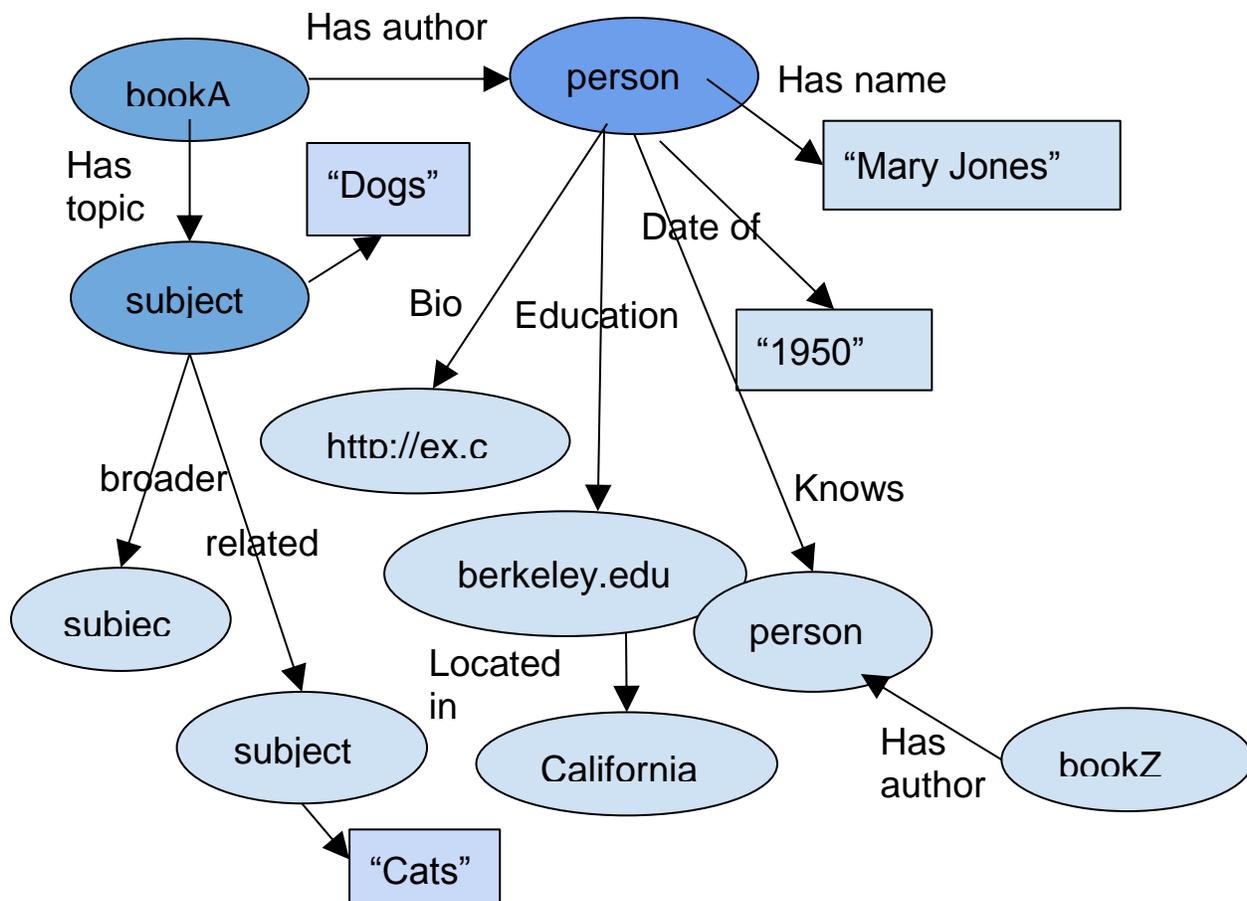
This begs the question, however, of how the *identity* of a work is determined. The options appear to be two: either the cataloger makes that decision by coding a set of properties as belonging to the work, or the identity of the work is inherent in the properties that describe it. The latter is described by Elaine Svenonius as a “set-theoretic” method,¹⁰ where a certain set of properties (such as the same author and same title) define two bibliographic items as being the same work. The advantage of this is that it instills consistency in the members of a work set because it is algorithmically determined. However, this method does not allow the cataloger to exercise judgment about “workness and about “sameness”--whether two things are two works or two versions of a single work. The disadvantage of the former is that identified works may not be described with a consistent set of properties, and thus may be more difficult to manage from a systems point of view.

3.3: Descriptions

As stated above, the identifier identifies the entity, such as the work, but does not define the scope of the properties (or attributes) that make up the description of the work. Think of the work identifier as being like an ISNI or an ORCID. These identify a person, but it is the metadata about the person that helps us understand who that person is. The work identifier has the same relationship to the descriptive attributes of the work. The identifier itself does not define attributes. Rather, this is left up to standards such as RDA and BIBFRAME.

In the record-based environment with which we are familiar, a record identifier always represents a specific set of fields. An RDF identifier represents the object that is identified, not a specific set of metadata. It is a “thing” identifier, not a record identifier. The triples with that IRI as a subject are statements (subject - verb - object) about the subject, such as “IRIx bf:creator ‘IRly’.” However, any statements built on IRly are about IRly, not IRIx. This means that seemingly essential elements of a bibliographic description, such as the display form of a creator’s name, is not included in that first level of relationships to the bibliographic resource. Also, because any RDF graph can lead to any number of relationships, following the paths in the graph can often lead to statements that are not suitable for all functions. In the graph below, the IRI for the author links both to information about the author as well to multiple publications. If IRIx is declared to be a work, it can logically include the relationship to IRly, but properties with relationships to IRly have an undefined relationship with IRIx. The extent of the work graph is not inherent in the identification of the work, and must be defined by other means.

¹⁰ Svenonius, E. (2009). *The intellectual foundation of information organization*. Cambridge, Mass: MIT., p. 33



Because of the graph structure of RDF, a single identifier can be used as a starting point for any number of different graphs, depending on the needs. For example, when accessing information about a work during the cataloging function it may be necessary to include cataloger notes and provenance information relating to the preferred label (“work title”) for the work. A work display in the online catalog will not include the information that is only of use to catalogers (and would likely confuse most users). Each of these functions defines a selection of elements from the overall graph. Using the same work identifier as a beginning point, different profiles of the data can be applied.

Clearly there needs to be some certainty for applications and users in what metadata will be associated with an identifier. At the moment this information is included in program documentation and code because there is no existing standard for the definition of profiles of RDF graphs. Work on graph validation and on dataset exchange in W3C, the standards body over the RDF technology, appears to be setting the groundwork for machine-actionable profiles of metadata to be used for the description of identified entities. There will undoubtedly be the

option to have more than one profile that is anchored by a particular identifier, as described in the paragraph above. Some profiles will include metadata describing more than one defined entity, as in the case with works and their creators: both a work and a creator are identified, described entities, and some functions will make use of metadata about both.

3.4: The Role of Classes in RDF

The IRIs in RDF create a stable identifier for a thing, but do not in themselves reveal anything about the nature of the thing. As stated above, identifiers are opaque and have no meaning in themselves, and RDF identifiers identify a thing, not the attributes that describe that thing. The creation of groups of attributes that make up a coherent description can be facilitated through the use of RDF classes.

Classes in RDF provide conceptual groups of data elements, not unlike taxonomic classes. In RDF, properties (data elements) can be associated, through their defined domains, with particular classes. In many of the RDF-based vocabularies we are looking at, a work is described using a set of properties that belong to the class that defines a “work”. A class is a quality of the property as it is defined in the vocabulary. In FRBRer, the property “form” is a member of the class “frbrer:work”; in RDA, “subject relationship” is a member of the RDA class work. BIBFRAME has work and instance classes, but also uses classes for some logical types of properties; for example, all title properties are members of the bf:Title class.

Classes have a number of functions in RDF, and are often used in searching. One could search on the keyword “remembrance” in a database of BIBFRAME data, qualifying that as a search only in members of the title class. A search within strings of class bf:Title would search bf:title, bf:mainTitle, bf:subTitle, bf:partNumber, bf:partName, and bf:variantType, and any subclasses of these titles. Searches can also include the subclasses of a class, which in the BIBFRAME case would include all titles and variant titles, which are all subclasses of title. Once a search is performed, the search engine can return whatever parts of the bibliographic description that the application needs: individual headings, the entire description, or a short form for display. This is very similar to the way that a search on a keyword retrieves MARC-based bibliographic data.

Classes in RDF are conceptually complex, but in fact can at times be more useful in a workflow than identifiers, depending on the specific design. For this reason it would be a mistake to focus solely on identifiers as creating useful sets of descriptive properties. Classes are more flexible than identifiers because any one property can be a member of more than one class. For example, the title of the work can be a member of the work class as well as the class of title. The latter class will include all titles in the record, which can be useful to support a broad title search. Any search that is limited to the properties of a work will almost certainly use a defined “work” class because identifiers are less useful for the search function.

Any descriptive element can be within the domain of more than one class. This is analogous to the fact that a person can be, for one purpose, a parent, for another an employee, and for yet another a voter. Because a property can be a member of more than one class, classes can be

used to provide alternate views of data. It would be possible to assign FRBR-defined bibliographic classes to BIBFRAME data elements along with the BIBFRAME classes. In this way an application can view either BIBFRAME work and instance, or FRBR work, expression, and manifestation. In a sense this could create a “cross-walk” between BIBFRAME and FRBR models.

Classes and their subclasses are applied to the vocabulary, whereas identifiers are assigned to individual sets of triples. Vocabulary documentation will indicate which classes each data element is assigned to, but there is no way to know a priori the identifiers that have been assigned to particular bibliographic descriptions. Identifiers are very important, but they are used primarily in the background of machine operations that are generally not visible to data creators or users.

Bearing all this in mind, we look next at how the work is modeled in various contemporary bibliographic standards.

4: Modeling the work

This section presents two aspects of the work as defined in the related standards:

- 1) How the concept of work is presented in its conceptual model;
- 2) How the work is implemented in the related vocabulary (code).

Not all models have been implemented in code. The models are presented below in the following order [section numbers in brackets]:

- The FRBR entity-relationship (ER) model [3.1], followed by its implementation in RDA [3.2], and the vocabulary (BIBFRAME [3.3]) proposed for carrying RDA work descriptions, as well as an experimental variant, BIBFRAME Lite [3.4]
- The FRBR model expressed in an object-oriented (OO) formalism, FRBRoo [3.5], as an extension of the CIDOC Conceptual Reference Model (CRM) used in the museum community,
- The Schema.org model for embedding work metadata in web pages [3.6]
- The general Dublin Core (DC) model [3.7] which does not differentiate works
- Models from neighboring (publisher and intellectual property) communities [3.8]
- Other Linked Data models being used in the library community [3.9]
- Implementing the work algorithmically [3.10], which deals with issues that have arisen in the implementation of the work concept during conversion of pre-existing (MARC) data

The work concept is defined fuzzily in all these models. To the extent that the concept can be defined, it must be extracted from the set of relations that are valid between instances. For example, if translation is not a valid relation between works in RDA, then the translation of a work does not result in a new RDA work, but since translation *is* a valid relation between works in BIBFRAME, the translation of a work can result in the creation of a new BIBFRAME work (though a metadata application profile may specify that use of this relation in BIBFRAME [or its parent hasDerivative relation] simultaneously creates a hasExpression relation to clarify that this instance of BIBFRAME work represents an RDA/FRBR expression).

At present these vocabularies do not reference one another. For example, classes and properties in the RDA vocabularies are not explicitly related (e.g., via OWL or SKOS properties) to similar or identical classes and properties in the BIBFRAME vocabularies. Such linking will presumably be added in the future, if only to facilitate interoperability.

4.1: FRBR

The Work entity in the FRBR model

Functional Requirements for Bibliographic Records (FRBR) defines a work as “a distinct intellectual or artistic creation.” It further states that “[b]ecause the notion of a *work* is abstract, it is difficult to define precise boundaries for the entity. The concept of what constitutes a *work* and where the line of demarcation lies between one *work* and another may in fact be viewed differently from one culture to another.”¹¹ This means that there is no agreed-upon definition of FRBR work that is suitable for all materials and all contexts. In fact, this is demonstrated by other emerging bibliographic standards, in particular Resource Description and Access (RDA), BIBFRAME, FRBR_{OO}, and Schema.org as employed by OCLC. The FRBR final report recognized this and assigned a fixed scope to “work” only “for the purposes of this study.”¹²

The FRBR work is characterized as a bibliographic entity that has its own attributes as well as relationships with other entities. The attributes describe the work, while the relationships are with other FRBR entities. Works have relationships with their expressions—a so-called primary relationship—as well as more general relationships with FRBR Group 2 entities (persons, corporate bodies, and families), Group 3 entities (subjects), and other Group 1 entities (works, expressions, manifestations, items). For example, a work can be an adaptation of another work.

In a given implementation, the set of relationships that are valid between works define the scope of the entity for that implementation.

Caveat: While FRBR is the conceptual model underlying current PCC cataloging practice, it is expected to be superseded—along with the related Functional Requirements for Authority Data (FRAD) and Functional Requirements for Subject Authority Data (FRSAD)—by a revised consolidated entity-relationship model called the IFLA Library Reference Model (IFLA LRM).¹³

FRBR as Code: FRBR_{ER} in RDF/OWL

The FRBR entity-relationship model is represented in code as FRBR_{ER}. FRBR_{ER} describes itself as “an element set of native RDF classes and properties described in the current text (February 2009) of the Functional Requirements for Bibliographic Records (FRBR) entity-relationship model.”¹⁴ It attempts to faithfully reflect the intention of the FRBR Study Group using the entities, attributes, and relationships as defined in the FRBR final report. It comprises ten classes

¹¹ IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report* (München: K.G. Saur, 1998), 17

¹² Ibid.

¹³ Pat Riva, Patrick Le Boeuf, and Maja Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information* (Den Haag, Netherlands: IFLA, ©2017)

<https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf>

¹⁴ <http://iflstandards.info/ns/fr/frbr/frbrer.rdf>

(representing the FRBR entities) and 206 properties (representing the FRBR attributes and relationships). Similar vocabularies exist for FRAD and FRSAD.

FRBR_{ER} is a strict interpretation of the description of bibliographic entities, attributes, and relationships that is found in the FRBR final report. In FRBR_{ER}, each entity is defined in terms of its attributes, and no attribute is valid for more than one entity. FRBR_{ER} likewise realizes the valid relationships between entities of the three FRBR groups. Each of the FRBR Group 1 entities is disjoint from the others and the attributes and relationships are specified as pertaining only to a specific Group 1 entity, with no overlap between them. This means that a FRBR_{ER} work can only have a work-work relationship with another FRBR_{ER} work, and that an attribute of a work, like “has subject”, can be used only with FRBR_{ER} work and not with any other FRBR_{ER} Group 1 entity nor with an entity from another vocabulary.

The FRBR_{ER} element set was last updated in July 2015, and this may be related to development of the IFLA Library Reference Model (IFLA LRM). IFLA LRM is a high-level conceptual reference model that encompasses all three of the FR entity-relationship models: FRBR_{ER}, FRAD, and FRSAD. At the time of writing no announcement had been made concerning a planned instantiation of FRBR-LRM in RDF or RDF/OWL.

Note that there are RDF vocabularies outside of the library community that make use of FRBR in RDF using vocabularies known as “FRBR core” and “FRBR extended.”¹⁵ These were developed by Ian Davis and Richard Newman, members of the Talis engineering group, in 2005. It describes itself as a “work in progress” and does not include FRBR attributes. It has not been updated.¹⁶¹⁷¹⁸

4.2: RDA

The Work entity in RDA

Resource Description and Access (RDA) has as its underlying conceptual models FRBR_{ER}, FRAD, and FRSAD. “The RDA data elements for describing a resource generally reflect the attributes and relationships associated with the entities work, expression, manifestation, and item, as defined in FRBR.”¹⁹

¹⁵ <http://vocab.org/frbr/core>

¹⁶ Expression of Core FRBR Concepts in RDF, ©2005. <http://vocab.org/frbr/core>

¹⁷ Expression of Extended FRBR Concepts in RDF, ©2005. <http://vocab.org/frbr/extended>

¹⁸ Gordon Dunsire, Declaring FRBR Entities and Relationships in RDF, 25 July 2008 (5 p.) <https://www.ifla.org/files/assets/cataloguing/frbrng/namespace-report.pdf>

¹⁹ RDA Toolkit, 0.2.2 Alignment with FRBR [revised 2015/04]

RDA defines a work as “[a] distinct intellectual or artistic creation, that is, the intellectual or artistic content.”²⁰ The first part of this definition is identical with the definition of work in Functional Requirements for Bibliographic Records (FRBR) 3.2.1, while the clarification (following “that is”) represents an addition made to the definition in the revised Statement of International Cataloguing Principles (ICP).²¹ RDA also states that “[t]he RDA data elements for describing a resource generally reflect the attributes and relationships associated with the entities work, expression, manifestation, and item, as defined in FRBR,” though “[a]ttributes and relationships associated with these four entities [work, expression, manifestation, and item] whose primary function is to support user tasks related to resource management (e.g., acquisition, preservation) are currently out of scope.”²² Specific instructions in RDA reflect the cataloging decisions that isolate and define the work in the context of a library catalog, however narrowly or broadly this may be defined. However, the instructions include alternatives (e.g., RDA 6.2.2.9.2 for two or more parts of a work and RDA 6.2.2.10.3 for compilations of two or more works) and are subject to interpretation, which will vary from one cataloging community to another, so the functional definition of the work in RDA, while narrower than in FRBR, remains fuzzy at the edges.

RDA in Code

The RDF implementation of RDA²³ defines RDA properties in two discrete vocabularies: one that is constrained by the FRBR entities, and one that is unconstrained, the latter called “unconstrained RDA.” The constrained RDA vocabulary represents RDA as an implementation of FRBR_{ER}, where attributes and relationships are associated with specific entities (or, in the context of RDF, where each property has one and only one class as its domain). For example, there cannot be a property for “title” that could be used for work titles, expression titles, and manifestation titles. Instead, a separate “title” property must be defined in relation to each entity.

Because constrained RDA is rigid in this way, it cannot easily be used in conjunction with other data models.²⁴ To provide an alternative approach, most properties are therefore also represented in the unconstrained RDA vocabulary, in which properties are not related to specific FRBR entities, although in all other ways the bibliographic description limitations are the same. For example, the RDA unconstrained property “has title” <http://rdaregistry.info/Elements/u/P60369> is defined as a property that “[r]elates a resource to a word, character, or group of words or characters that names a resource or a resource embodied

²⁰ Ibid.

²¹ Statement of International Cataloguing Principles (ICP), 2016 edition with minor revisions, 2017, p17. https://www.ifla.org/files/assets/cataloguing/icp/icp_2016-en.pdf

²² RDA Toolkit, 0.2.2 Alignment with FRBR [revised 2015/04]

²³ <http://www.rdaregistry.info/>

²⁴ Thomas Baker, Karen Coyle, Sean Petiya. Multi-Entity Models of Resource Description in the Semantic Web: A comparison of FRBR, RDA, and BIBFRAME. *Library Hi Tech*, v. 32, n. 4, 2014 pp 562-582
DOI:10.1108/LHT-08-2014-0081

in it” but has no defined domain or range. The unconstrained RDA vocabulary can be used to align RDA with bibliographic models that do not use the bibliographic entities that are defined in FRBR_{ER}, such as, BIBFRAME.

Although there is a defined RDF vocabulary for RDA, it is not currently used in library cataloging systems, though it has been used by the RIMMF²⁵ software that has been demonstrated at “Jane-a-thons”²⁶ and other events for experimenting with bibliographic data as Linked Data.

Note that using MARC21 as an RDA record syntax does not result in discrete work entities, as modeled in FRBR, even in those cases where an authorized access point for the work is encoded in the catalog record. As discussed below in section 3.10 on implementing the work algorithmically from legacy bibliographic data, even with cataloging data created using RDA instructions it has not proved possible to reliably extract work entity descriptions with reasonable accuracy.

Note: The English text of RDA was frozen in April 2017 so that, inter alia, it could be restructured to implement the IFLA LRM.²⁷

4.3: BIBFRAME

The Work entity in the BIBFRAME model

Note that BIBFRAME is essentially a vocabulary, and provides only the most general description of the conceptual model that underlies it. Like the MARC 21 record syntax that it is expected to supersede, it is designed to be independent of any particular cataloging standard or practice. However, the fact that BIBFRAME is being developed at the Library of Congress and tested largely by PCC member institutions means that the ability to successfully convert legacy MARC 21 data (including RDA data) to RDF will be a prerequisite for BIBFRAME implementation.

Within the BIBFRAME model, a work “reflects the conceptual essence of the cataloged resource: authors, languages, and what it is about (subjects).”²⁸ This model relates the Work to subjects, agents, and—an innovation—events (i.e. “an occurrence, the recording of which, may be the content of a Work”). In the BIBFRAME vocabulary itself (see below under BIBFRAME as Code), the class Work is defined as a “Resource reflecting a conceptual essence of a cataloging

²⁵ RIMMF3 Home, <http://www.marcofquality.com/wiki/rimmf3/doku.php?id=rimmf>. RIMMF stands for RDA in Many Metadata Formats

²⁶ Jane-a-thons are hackathons for metadata about Jane Austen and her works

²⁷ Implementation of the LRM in RDA, posted 3 February 2017. <http://www.rda-rsc.org/ImplementationLRMinRDA>

²⁸ Overview of the BIBFRAME 2.0 Model, April 21, 2016 <http://www.loc.gov/bibframe/docs/bibframe2-model.html>

resource.”²⁹ Although this is similar to the functional definition provided by FRBR and RDA, BIBFRAME does not use the FRBR entities, but instead defines its own.

BIBFRAME as Code

The BIBFRAME model has three primary bibliographic entities—Work, Instance, and Item—to the four in FRBR_{ER} and RDA. Instance and item are conceptually similar to the manifestation and item in FRBR/RDA. In the initial design of BIBFRAME, the BIBFRAME work entity was defined as covering both the FRBR work and expression. The BIBFRAME 2.0 vocabulary moves even further from the FRBR entity definitions because it removes many of the vocabulary constraints that would define works and instances as having distinct sets of properties. The vocabulary definitions in BIBFRAME 2.0 avoid a strong definition of the entities, making the language potentially usable for bibliographic data that does not adhere to FRBR-like definitions. In BIBFRAME 2.0 many, if not most, properties are defined as being suitable to describe either works or instances or items.

This does not mean that the FRBR entities cannot be expressed. There is a difference between the vocabulary definition, which is deliberately loose so that it can accommodate a broad set of practices, and the metadata application profile that may be adopted by a particular community. The rules that would govern adherence to specific bibliographic entities are not embedded in the base vocabulary, but could be provided by application profiles or software.

In this sense, BIBFRAME is not a FRBR vocabulary, although it may nevertheless encode FRBR-based bibliographic data. Software now being used experimentally to convert legacy MARC 21 bibliographic data to BIBFRAME uses a set of rules that results in a division of properties between BIBFRAME Works and Instances, but not Works and Expressions as defined in FRBR. The vocabulary, however, does contain properties (*expressionOf* / *hasExpression*) that can be used to map the FRBR work-expression primary relationship as a BIBFRAME Work-Work relationship while implicitly preserving the FRBR distinction between works and expressions.

Note: At the time of writing the Library of Congress had just begun the BIBFRAME 2.0 Pilot. In the pilot, Work entities for legacy data are created by merging expression- and work-level data from MARC bibliographic records and data (including administrative metadata) from MARC authority records for titles and name/titles. Work entities for new works and expressions are created in the BIBFRAME Editor by minting blanknode Work identifiers as appropriate.³⁰

²⁹ BIBFRAME 2.0 Vocabulary List View, s.v. “Work” <http://id.loc.gov/ontologies/bibframe.html#c.Work>

³⁰ MARC 21 to BIBFRAME 2.0 Conversion Specifications <https://www.loc.gov/bibframe/mtbf/> and email from Les Hawkins, Library of Congress

4.4: BIBFRAME Lite

Work entity in BIBFRAME Lite

BIBFRAME Lite is based on the deprecated BIBFRAME 1.0 (now superseded by BIBFRAME 2.0 above) that was developed by Zepheira, the company that originally contracted with the Library of Congress to develop BIBFRAME, along with the National Library of Medicine, the George Washington University, and the University of California, Davis.³¹ However, while BIBFRAME Lite is not based on BIBFRAME 2.0, the work entity in BIBFRAME 1.0 is the same as in BIBFRAME 2.0, using the same expressionOf/hasExpression properties to relate FRBR works and expressions as BIBFRAME works.

The BIBFRAME Lite Work, like the FRBR work, is defined as “a distinct intellectual or artistic creation” (<http://bibfra.me/view/lite/Work/>).

BIBFRAME Lite as Code

BIBFRAME Lite Work has very few properties of its own (contributor, creator, genre, subject, and title) for distinguishing one work from another; however, it inherits additional properties from the superclass Resource.

In BIBFRAME Lite, as in BIBFRAME, bibliographic attributes can be related to either Work or Instance. For example, as a property of the superclass Resource, language can be related to either of the subclasses Work and Instance in BIBFRAME Lite, while format and medium are properties of Instance. The main bibliographic relationship available is “is version of” and this can be used between any bibliographic entities.

BF Lite originated as an offshoot of BIBFRAME, but with a different modeling and development approach. BF Lite is developed mainly from the data and data needs of legacy conversion rather than through creation of new data in RDF. Zepheira works with communities of practice to determine what data should be converted from legacy standards and where that data belongs in the model. Adhering to the philosophy that a flexible and extensible standard is never ‘finished’, therefore BF Lite development is designed to be agile, with updates pushed to production as often as needed following testing and community feedback.

The Zepheira modeling approach is modular, i.e., there is a core BIBFRAME Lite vocabulary with additional vocabulary modules built to meet the more specific needs of different communities of practice. As those community specific modules are built, if classes and properties are found in common across communities they may be moved up to the BIBFRAME Lite core. Current modules include Library (based on MARC), Relation (based on explicit

³¹ MacKenzie Smith, Carl G. Stahmer, Xiaoli Li, and Gloria Gonzalez, BIBFLOW: A Roadmap for Library Linked Data Transition, prepared 14 March 2017. XI: Survey of Current Library Linked Data Implementation. <https://bibflow.library.ucdavis.edu/xi-survey-of-current-library-linked-data-implementation/>

relationships in MARC, e.g., relator codes), Archive (based on DACS), and Rare Materials (based on the BF Lite core and Library module).

Zepheira seeks to align BF Lite with BIBFRAME and other RDF vocabularies (e.g., schema.org) to the extent possible. Toward that end, certain classes and properties will include ‘same as’ relationships which are visible in the human readable web pages.

4.5: CIDOC CRM, FRBR_{OO}, and PRESS_{OO}

The Work entity in the FRBR_{OO} and PRESS_{OO} models

The FRBR family of conceptual models—FRBR_{ER}, FRAD, and FRSAD—was developed using an entity-relationship formalism. FRBR_{OO} defines these models using an object-oriented formalism that introduces temporal entities, events, and time processes, refines the Group 1 entities, and analyzes the creation and production processes.³²

Intended to help provide a common view of cultural heritage information, FRBR_{OO} is also a harmonization of the FRBR models with the Conceptual Reference Model of the Comité international pour la documentation (CIDOC) of the International Council of Museums, usually referred to as the CIDOC CRM.³³ The “oo” refers to the object-oriented formalism that is used in the CIDOC CRM, FRBR_{OO}, and PRESS_{OO}. The CIDOC CRM has many aspects that do not appear in library cataloging. In particular, it views creation, including publication, as a process with multiple steps. It defines the work as a class that: “... comprises distinct concepts or combinations of concepts identified in artistic and intellectual expressions, such as poems, stories or musical compositions.” It allows that a work may be individual or complex, and further defines it as “...the product of an intellectual process of one or more persons, yet only indirect evidence about it is at our hands.” FRBR_{OO} subclasses the work into Individual Work, Complex Work, Container Work, and Recording Work. These in turn have more specific subclasses. One of the distinctions of FRBR_{OO} is that it includes the processes that connect works to expressions and manifestations, such as recording or publication or production.

As with FRBR_{ER}, FRBR_{OO} acknowledges an imprecise border between works and expressions in the model:

³² Definition of FRBR_{OO}: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism, version 2.4 (November 2015), 16-22

https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf

³³ CIDOC CRM Conceptual Reference Model, trial version, version 6.2 (May 2015) <http://www.cidoc-crm.org/Version/version-6.2>

On a practical level, the degree to which distinctions are made between variant expressions of a work will depend to some extent on the nature of the **F1 Work** itself, and on the anticipated needs of users.³⁴

PRESS₀₀ is an extension of the FRBR₀₀ model specifically for continuing resources (serials and integrating resources). It assumes the FRBR₀₀ definition of work, and adds classes and properties specific to the creation of continuing resources. It is unique among the FRBR conceptual models in that its documentation includes an element-by-element mapping from an existing (MARC) format--the one used in the ISSN Manual--to the related properties in PRESS₀₀, FRBR₀₀, and the CIDOC CRM.

Continuing resources are modeled using classes and properties from PRESS₀₀, FRBR₀₀, and the CIDOC CRM. Both PRESS₀₀ and FRBR₀₀ are IFLA standards and have been approved by the CIDOC CRM Special Interest Group as indirect extensions of the CIDOC CRM.³⁵ FRBR₀₀ introduces the class F18 Serial Work, which is at once a subclass of FRBR₀₀'s F15 Complex Work and F19 Publication Work, defined as comprising "works that are, or have been, planned to result in sequences of Expressions or Manifestations with common features."³⁶[2] This definition ties the Serial Work to a plan relating to a particular Expression and Manifestation, and in this respect the FRBR₀₀ model differs from the FRBR_{ER} model.

Information elements that, in the FRBR_{ER} conceptualisation, were directly attached to the Expression and Manifestation entities, are in FRBR₀₀ seen as being in reality part of the issuing rule for the serial work (represented as an instance of **E29 Design or Procedure**). It is at the very core of the definition of **F18 Serial Work** that it plans that issues are published by a particular publisher and contain texts in a particular form.³⁷

What this means in practice is that each Manifestation of a Serial Work is itself a distinct Serial Work. Consequently, any translation (for example) is itself a distinct Serial Work, and any version in a different format (for example, an online version) is a distinct Serial Work. This is quite different from the FRBR_{ER} model—and the current (2017) version of RDA, which is aligned with that model (cf. RDA 0.2)—in which both translations and format versions were considered expressions of a common Work.

This exceptional treatment of serials--a single work realized in a single expression embodied in a single manifestation--is carried over into IFLA LRM, with which the redesigned (2018) RDA will be compliant.³⁸

³⁴ Definition of FRBR₀₀, p55

³⁵ Definition of PRESS₀₀: A Conceptual Model for Bibliographic Information Pertaining to Serials and Other Continuing Resources, version 1.3 (August 2016) p6
https://www.ifla.org/files/assets/cataloguing/PRESSoo/pressoo_v1-3.pdf

³⁶ Definition of FRBR₀₀, p66

³⁷ Definition of FRBR₀₀, p20-21.

³⁸ "Implementation of the LRM in RDA" <http://www.rda-rsc.org/ImplementationLRMinRDA>

Work has now begun on a third edition of FRBR_{OO} (tentatively named IFLA LRM_{OO}) that will conform with IFLA LRM (4.6 below). This will be more simplified than the existing model. For example, it is expected that **F3 Manifestation Product Type** and **F24 Publication Expression** will be conflated, that **F14 Individual Work** will be dropped, and that **F4 Manifestation Singleton** will be defined as a set of one.³⁹

FRBR_{OO} and PRESS_{OO} as Code

FRBR_{OO} extends the FRBR_{ER} work entity using super- and sub-classes. In FRBR_{OO} **F1 Work** has an abstract super-class called **E89 Propositional Object** as well as sub-classes for types of **F1 Work**, specifically **F14 Individual Work** (with sub-class **F17 Aggregate Work** and its sub-classes **F19 Publication Work** and **F20 Performance Work**), **F15 Complex Work** (with its sub-class **F18 Serial Work**), and **F16 Container Work**, and **F21 Recording Work**. The FRBR_{OO} RDF vocabulary has been published in the IFLA namespace, and links have been provided to related CIDOC CRM classes.⁴⁰

The **F2 Expression** class in FRBR_{OO} is a sub-class of CIDOC CRM **E73 Information Object**, and instances of **F2 Expression** may also be instances of other CIDOC CRM classes when these expressions have a particular form. For example, a textual expression may be both an instance of **F2 Expression** and of **E33 Linguistic Object**, and so it may also use properties related to this latter class.⁴¹ **F2 Expression** has five sub-classes: **F22 Self-Contained Expression**, **F23 Expression Fragment**, **F34 KOS**, **F35 Nomen Use Statement**, and **F43 Identifier Rule**.

PRESS_{OO} has not yet been published as Linked Data.

4.6: IFLA LRM

The Work entity in IFLA LRM

IFLA LRM (Library Reference Model) consolidates and supersedes the three entity-relationship models (FRBR [4.1 above], FRAD, and FRSAD) currently in use in the library community.⁴² The LRM also takes into account the object-oriented FRBR_{OO} model (4.5 above), and many differences with the original entity-relationship models reflect the incorporation of concepts from the object-oriented model. The differences with the earlier models are numerous, but most are extraneous to the discussion of the work entity. The disjoint Work, Expression, Manifestation,

³⁹ Pat Riva and Maja Žumer, *The IFLA Library Reference Model, a Step toward the Semantic Web* (IFLA WLIC 2017 Wrocław), ©2017, p. 7 <http://library.ifla.org/1763/1/078-riva-en.pdf>

⁴⁰ FRBR_{OO} Model <http://iflstandards.info/ns/fr/frbr/frbroo/>

⁴¹ Definition of FRBR_{OO}, p55.

⁴² Pat Riva, Patrick Le Bœuf, and Maja Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information* (Den Haag, Netherlands: IFLA, ©2017) <https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf>

and Item entities are retained. However, two new work-related differences with the older model are germane to our discussion and involve the work entity as it applies to aggregates and serials.

IFLA LRM models aggregates not as works but as manifestations embodying multiple expressions, one of which is an expression of what is termed the aggregating work, that is, the selection and arrangement criteria applied to the manifestation. This differs significantly from current practice, which views aggregate manifestations as potentially members with other manifestations of a common expression, which in turn may be a member with other expressions of a common work.

Similarly, IFLA LRM explicitly defines a serial work--a type of aggregate work--as being realized in a single expression that is embodied in a single manifestation. In contrast, current practice is to treat serials as generic works that can be realized in multiple expressions embodied in multiple manifestations.

Despite their distinctive features and behaviors, both aggregating works and serial works are represented in the LRM model, along with all other works, by the **LRM-E2 Work** entity.

The model permits the definition of additional entities “that comprise, say, the paper edition of a journal and its edition on the web; all linguistic editions of a journal that is published in more than one language as separate editions; all local editions of a journal, etc., according to the needs that need to be met in a given implementation of the model.”⁴³ It is not clear how these additional entities might relate to works in a given implementation.

At the time of writing we do not know how IFLA LRM will be expressed in RDA.

IFLA LRM as Code

IFLA LRM has not yet been expressed as code, though such an expression is planned within the FRBR Vocabularies at <http://iflstandards.info/ns/fr/>.⁴⁴

⁴³ Ibid., p. 96.

⁴⁴ Pat Riva and Maja Žumer, *The IFLA Library Reference Model, a Step toward the Semantic Web* (IFLA WLIC 2017 Wrocław), ©2017, p. 6 <http://library.ifla.org/1763/1/078-riva-en.pdf>

4.7: Schema.org⁴⁵

[See a more detailed discussion in appendix A4]

The Work entity in Schema.org

Schema.org is a metadata standard developed for use within Web pages. Schema.org provides only minimal definitions for its terms, deliberately allowing for broad interpretation of meanings. It has a high-level term for intellectual resources (directly under the generic term “thing”) called CreativeWork, which is defined as “[t]he most generic kind of creative work, including books, movies, photographs, software programs, etc.”⁴⁶ Schema.org does not include input rules, so this is the only information given to define how the concept might be used, and it has been applied variously for what FRBR would call manifestations, as well as for works and expressions. There are sub-classes for about two dozen types of works, such as books, maps, TV series.

An advantage of schema.org is that it is a function of the online display service of a Web site, and as such does not require any change in the data stored by the site that is used in the generation of the display. OCLC makes use of schema.org, adding coded metadata to the site display for individual WorldCat resources, while still storing data in a local version of the MARC format.

Schema.org provides usage ranges for its elements in terms of Web domains. The most heavily used types of creative work reported are Blog and Article (both more than one million domains). CreativeWork is reported as used by between 250,000 and 500,000 domains, and Book by between 10,000 and 50,000 domains, presumably including www.worldcat.org.

Schema.org as Code

Schema.org uses basic RDF concepts but is less strict than true RDF. The reason for this is that many creators of schema.org metadata are not expected to have standardized data.

Schema.org development was initiated by key Internet indexing services: Google, Yahoo, Bing, and Yandex. The metadata encoded in schema.org should facilitate smarter displays of Internet data, such as product descriptions, prices, store locations and hours.

Schema.org has an extensive vocabulary of products and services. While the set of data elements available for documents is considerably less detailed than any library standard, it is not intended to replace the bibliographic metadata from which it is derived, but only to highlight selected characteristics for online indexing and display. On the other hand, for many types of

⁴⁵ R.V. Guha, Dan Brickley, and Steve Macbeth, Schema.org: Evolution of Structured Data on the Web, *acmqueue* 13 (November-December 2015) <http://queue.acm.org/detail.cfm?id=2857276>

⁴⁶ CreativeWork (schema.org) <http://schema.org/CreativeWork>

resources, the properties available in Schema.org are more plentiful than in traditional bibliographic metadata.

Schema.org does not distinguish between works, expressions, and manifestations. It has a main class, CreativeWork, that has more specific types (e.g., Article, Book, Map, Movie, MusicComposition, MusicRecording). CreativeWork includes many properties that describe what would be treated as works, expressions, and manifestations in the FRBR conceptual model. It also has a set of properties that express relationships between bibliographic descriptions, such as exampleOfWork, isBasedOn, isPartOf, and the bib extension property translationOfWork. In effect, schema.org's CreativeWork can be used to describe works as defined by FRBR, but there will not be a coded separation between works and other FRBR-defined entities, nor are there separate properties to describe elements of the FRBR entities. A CreativeWork title can be used for the title of a FRBR work, the title of a FRBR expression, or the title of a manifestation. The entity described is defined in part by the context provided by the Web page. More specific types of CreativeWork have type-specific identifier properties such as isrcCode (MusicRecording) and iswcCode (MusicComposition), though these two do not report any domain usage. (See below under Publisher and Intellectual Property Standards.)

4.8: Dublin Core

The Work in Dublin Core

Dublin Core is a general metadata set for the description of resources of all types. It does not define levels of abstraction that would be analogous to FRBR Group 1 entities. There is a single concept referred to as “the resource” which is described by the Dublin Core terms.

Elements cover bibliographic resources, their physical media, agents (such as creator and publisher), provenance, and rights. There are both descriptive properties (creator, date, identifier) and a property for bibliographic relationships (relation) that can be further refined (hasFormat, hasPart, hasVersion, isReferencedBy, isReplacedBy, isRequiredBy, and their inverses). These can be used between any bibliographic descriptions and are not specific to a bibliographic level such as work. As an example of a significant difference between Dublin Core and much more controlled vocabularies for bibliographic resources, note that the 2005 Guidelines for creation of content for the property hasVersion give as an acceptable use “Romeo and Juliet” hasVersion “West Side Story”⁴⁷ Although this is given as an example, Dublin Core does not provide any rules for use of its terms and practice varies among communities using this vocabulary. Also note that Dublin Core is rarely used alone, but is most often combined with other terms to complete a vocabulary. Libraries rarely use Dublin Core for

⁴⁷ Using Dublin Core – Dublin Core Qualifiers (2005-11-07)
<http://dublincore.org/documents/usageguide/qualifiers.shtml>

cataloging, but DC terms are often present in non-library vocabularies. For this reason it is useful as a switching language between vocabularies.

Dublin Core as Code

Note that the original Dublin Core vocabulary of 15 elements has been superseded by a more extensive vocabulary called “Dublin Core Terms”. Both the original set and the extended set are currently defined in RDF.⁴⁸

4.9: Publisher and Intellectual Property Standards

[See a more detailed discussion in appendices A1 and A2]

There are several abstract models defined by the publishing and intellectual property (IP) communities. These standards are used in the book trade, by publishers and others, including libraries and those that provide services to libraries. This section provides only an overview of these models. More detail is provided in Appendix 1 (publisher community) and 2 (intellectual property [IP] community).

<indecs> was a project partly funded by the Info 2000 initiative of the then European Community (precursor of the European Union) and several organizations representing the music, rights, text publishing, authors, library and other sectors in 1998-2000. It has since been used in a number of metadata activities, including the digital object identifier (DOI).⁴⁹ <indecs> is a metadata framework to accommodate and facilitate the commercial, transactional aspect of created content. Consequently, the general approach is that of content creators, their exchange partners, and the e-commerce life-cycle of the products involved in that exchange. Works are involved only to the extent that they facilitate that exchange.

<indecs> does not use the term “work” but instead uses “abstraction”, which it defines in FRBR terms as “[a]n abstract creation whose existence and nature are inferred from one or more expressions or manifestations.” Abstractions are expressed through an expression event. However, in practice the abstraction in <indecs> corresponds to the FRBR expression, as variants like translations and editions are treated as different abstractions. The <indecs> Framework is mainly employed by EDItEUR, the trade standards body for the international book and serials trade.

⁴⁸ The original 15 elements: Dublin Core Metadata Element Set, version 1.1 (2012-06-14) <http://dublincore.org/documents/dces/> and the current extended Dublin Core Terms <http://dublincore.org/documents/dcmi-terms/> (2012-06-14)

⁴⁹ The <indecs> Metadata Framework: Principles, Model, and Data Dictionary, June 2000 http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

EDItEUR is also the standards body for the ONIX EDI (electronic data interchange) standards (ONIX for Books, ONIX for Serials, etc.), which are XML formats used by publishers to describe their products.⁵⁰ Analogous to a MARC21 record, each ONIX record contains the data about a single product. EDItEUR also maintains ONIX registration formats for International Standard Book Numbers (ISBNs), International Standard Text Codes (ISTCs), and Digital Object Identifiers (DOIs).

Although some of these standards use the term “work” none of them describe or identify the abstraction that is the FRBR work. The ISTC was developed to identify texts for IP purposes as part of the publisher workflow. Its definition of work, similar to the FRBR expression, is defined as:

a distinct, abstract intellectual or artistic creation predominantly comprising a combination of words, whose existence is revealed (i.e. “published”) or intended to be revealed, through one or more textual manifestations.⁵¹

We have been told that the ISTC has not seen much uptake in the publishing community since its introduction in 2009, and the standard may be revised in an attempt to increase its use.

To receive an International Standard Musical Work Code (ISWC), a musical work may be published or unpublished. Arrangements, adaptations, and translations are considered to be separate works, so the ISWC work, like the ISTC work, is similar to the FRBR expression.

For the other product identifiers the level of description is that of the product, determined by the packaging of the content. For example, a DOI is can be applied to any entity, such as a textual publication, a data set, or a broadcast program, and at any level of granularity desired by the applicant. As used by CrossRef, the DOI can be used to identify an entire journal, a single issue, an article, or an article stored in a particular format. Similarly, an ISBN can be applied to an individual volume or for a multivolume set. The ISMN is the music identification number that serves the same function for manifestations of musical notation as the ISBN does for books. In these cases, the sales model determines the level of granularity. None of these identifiers correspond to the FRBR work entity.

By its nature, the FRBR work entity assembles expressions that may have different rights attached. The publishing and intellectual property communities do not produce metadata at this level as their focus is on the resource that has specific rights attached.

While these publisher and IP standards do not directly impact library practice, metadata supplied for libraries’ resource discovery services may reflect them. For example, CrossRef

⁵⁰ ONIX <http://www.editeur.org/8/ONIX/>

⁵¹ International ISTC Agency. *International Standard Text Code (ISTC) User Manual*, version 1.2, April 2010 http://www.istc-international.org/multimedia/pdfs/ISTC_User_Manual_2010v1.2.pdf

supplies DOI metadata to services such as Ex Libris' Primo used to create citation trails, inter alia.

4.10: Other Linked Data models

There are other Linked Data models in use by the library community that do not make use of the work entity. Among these are the British Library Data Model (used with the British National Bibliography)⁵² and the two data models in use for a couple of collaborative Web-scale cultural heritage: the Europeana Data Model (EDM)⁵³ and the Digital Public Library of America (DPLA) Metadata Application Profile⁵⁴ (which builds on the EDM). EDM defines its Information Resource class as “the union of IFLA FRBR Work, Expression and Manifestation, E89_Propositional_Object (CIDOC CRM).” Because these models do not make use of a work entity, they are not examined here.

4.11: Implementing the Work algorithmically

The automated discovery of works in legacy data such as MARC21 and Dublin Core has a long history, dating back to FRBR's inception. The key problem is that these formats do not recognize the work as an entity. As a result, the work must be inferred from evidence or clues in the data, which was intended for other purposes and is often incomplete, incorrect, or not expressed in an easily parsed format.

In the early 2000s, OCLC published the 'FRBR Work-Set algorithm'.⁵⁵ The primary goal of this research was to realize the promise of the FRBR Group I entities for simplifying the display of bibliographic records in the library OPAC. Research projects with the same goals were conducted at Kent State University and elsewhere.⁵⁶

The OCLC production stream later incorporated the Work-Set algorithm to generate a hierarchical display in WorldCat, which had the effect of simplifying the results for the complex

⁵² British Library Data Model – Book, v. 1.5 (British Library/Talis, March 2017)

<https://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf>

⁵³ Definition of the Europeana Data Model v5.2.7 (25/04/2016)

http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.7_042016.pdf

⁵⁴ Digital Public Library of America. Metadata Application Profile, version 4.0 (release date: 3/4/2015)

<https://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf>

⁵⁵ Thomas B. Hickey, Edward T. O'Neill, and Jenny Toves, Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR), D-Lib Magazine 8, no. 9 (September 2002)

<http://www.dlib.org/dlib/september02/hickey/09hickey.html>

⁵⁶ FRBR-Based Systems to Effectively Support User Tasks and Facilitate Information Seeking: Project Publications (2007-2012) <http://frbr.slis.kent.edu/publications.htm>

publication histories of the most important works ever published, such as Shakespeare's *Hamlet*. Experimental applications of the algorithm to genre-constrained collections such as works of fiction⁵⁷ and cookbooks⁵⁸ also produced simplified displays that facilitate subject navigation. Another tool, which underlies the feature named 'Find it in a library' that was once accessible on Amazon, allowed searchers to identify books held by libraries whose ISBNs appeared in the same FRBR work cluster. The result was a kind of fuzzy search for those who were relatively unconcerned about the format or edition of the content they were seeking.

FRBR-clustered data has also proven to be useful in back-office operations. For example, in the 'Metadata Services for Publishers' project conducted at OCLC from 2007-2011, metadata records ingested from ONIX sources originating in the publisher supply chain were automatically assigned to a work cluster generated from WorldCat records, to which the library community's controlled names and subject headings could be applied.⁵⁹ The OCLC research prototype Classify^{60, 61} recommends subject classifications based on data mined from work clusters to which a bibliographic reference most likely belongs. Classify is now one of OCLC's most heavily used experimental services.

Research focused on applications of FRBR preceded experimentation with linked data, and at OCLC, it has continued in parallel with the Linked Data research program. In 2014, these efforts produced the WorldCat Works dataset,⁶² accompanied by the publication of work URIs in the 'Linked Data' tab for over 300,000 records accessible from WorldCat.

Though the original use cases for works remain compelling, the encounter with Linked Data introduces a few more, such as:

- *Work as a success measure*. The conversion of library authority files to linked data⁶³ counts as the first at-scale success by the library community. But the description of any creative work or bibliographic resource is more challenging, regardless of where it falls in the FRBR Group I hierarchy, perhaps because the bibliographic standard is much larger than the authority standard and contains much more uncontrolled text.

⁵⁷ OCLC WorldCat Fiction Finder <http://experimental.worldcat.org/xfinder/fictionfinder.html>

⁵⁸ OCLC WorldCat Cookbook Finder <http://experimental.worldcat.org/xfinder/cookbookfinder.html>

⁵⁹ Carol Jean Godby, From Records to Streams: Merging Library and Publisher Metadata, Proceedings of the International Conference on Dublin Core and Metadata Applications 2010
<http://dcpapers.dublincore.org/pubs/article/view/1033/989>

⁶⁰ Diane Vizine-Goetz, Classify: a FRBR-Based Research Prototype for Applying Classification Numbers, NextSpace 14: 14-15 [2009]
http://www.oclc.org/content/dam/oclc/publications/newsletters/nextspace/nextspace_014.pdf#page=16

⁶¹ Classify: An Experimental Classification Web Service <http://classify.oclc.org/classify2/>

⁶² OCLC Releases WorldCat Works as Linked Data [press release], 28 April 2014
<http://www.oclc.org/en/news/releases/2014/201414dublin.html>

⁶³ Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech, LCSH, SKOS, and Inked Data, Proceedings of the International Conference on Dublin Core and Metadata Applications 2008
<http://dcpapers.dublincore.org/pubs/article/viewFile/916/912>

How far have we come? In 2011, the British Library Data Model was hailed as a breakthrough. It was accompanied by RDF descriptions of 2.5 million resources that comprised the British National Bibliography. It set a new standard for machine-understandability of bibliographic descriptions because nearly every statement in the source MARC record except string literals such as titles and summaries was represented in a machine-understandable format, via URIs that were either public or locally coined. The British Library RDF dataset also establishes a baseline for future work. It describes what can be interpreted as manifestations, but not works, expressions or items; and the dataset describes mostly books, not archival materials or digital objects.

- *Work as a psychologically important abstraction.* Outside the library community, descriptions of works are obscured by irrelevant details about individual editions or formats. This is true of Wikipedia's infoboxes, Google's knowledge cards, and Wikidata's resource identifiers, which sometimes list WorldCat record IDs but never the results from the application of the FRBR work-set algorithms. What is missing is at least a two-part model, which distinguishes between what users are searching for and the details of the object that might fulfill the request, such as the language of the text and the file format. These concepts are modeled in the 'Issue 53' standard for content negotiation published in 2006 by the World Wide Web Consortium.⁶⁴
- *Work as a collection point for links.* Proponents of library linked data have long argued that work identifiers implanted in third-party locations favored by the information-seeking public, such as Google knowledge cards, would boost the visibility of libraries by a form of page ranking. For example, if libraries and the reading public pointed to the same work, or even to a small number of work descriptions of Shakespeare's *Hamlet*, the relative importance of *Hamlet* would be raised, and so would the visibility of every library collection that has a copy. This result is theoretically possible, but is still only a conjecture. It has not been demonstrated, in part because the work entity itself is not yet mature enough to conduct a formal test, and a critical mass of instance data describing works is not yet widely available outside the context of research projects.

Of course, every use case mentioned so far would be easier to address if work descriptions were published to the more rigorous web-friendly standards such as those described in this document. Conversion from legacy is simply an indirect but more accessible means to the same end. But as of Sept 2017, most of the library community's data is still expressed in MARC 21

⁶⁴ On Linking Alternative Representations to Enable Discovery and Publishing, TAG Finding 1 November 2006 <https://www.w3.org/2001/tag/doc/alternatives-discovery.html>

and other legacy formats. Thus, experiments that address the use cases for the work like those described here will most likely continue.

As we wait for larger collections of work descriptions that are written to improved standards, algorithmic processing can be applied as a test to assess progress toward the goal of greater machine understandability. What can be discovered? What are the barriers? For example, work and expression are conceptually difficult to distinguish, so it should not be surprising that automated discovery is also difficult and error-prone. But sometimes we introduce unnecessary problems. For example, some MARC encodings that identify the details of translated works are potentially robust but are not widely used. To facilitate the transition to linked data, Smith-Yoshimura recommends⁶⁵ that the translator should be recorded as a code in the MARC \$4 subfield, not as uncontrolled text in \$.e. But this recommendation runs counter to current cataloging practice. This observation is embedded in a larger research study of works and their translations, which seeks to embrace the fact that MARC records in more than 460 languages are now accessible from WorldCat.^{66, 67}

The task of algorithmic discovery may even produce requests to those whose main task is the definition of modeling detail. For example, the Library of Congress and OCLC have largely similar practices for discovering works in MARC data,⁶⁸ starting with the same two inputs: MARC uniform title authority records and bibliographic records. The first is preferred because it is the output of an editorial process and is more easily interpreted. If the source is a set of bibliographic records, a work must be produced from a two-step process: isolating the work-level properties such as author, subject, and description, and clustering the results. But the results are not precise enough to identify many of the modeling details described in this document, prompting continual refinement of the algorithms that produce them. If the starting point is a uniform-title record instead, the task is to associate it with bibliographic descriptions, either through an automated process or a human-guided workflow. This is a more manageable task. But uniform title records are exceedingly rare and sparse, and it is not clear what their theoretical status is in the landscape for works as they are now being defined. Could a specification of the uniform-title replacement be expedited? If so, what properties would it have, and who would do the work of populating it?

⁶⁵ Jean Godby and Karen Smith-Yoshimura, How You Can Make the Transition from MARC to Linked Data Easier, Technical Advances for Innovation in Cultural Heritage Institutions (TAI CHI) Webinar Series, 5 November 2015 https://www.youtube.com/watch?time_continue=2&v=3erkTu3oW4c

⁶⁶ Karen Smith-Yoshimura, Linked Data: Bringing the World Closer Together, ALIA Information Online, 15 February 2017 <https://informationonline.alia.org.au/sites/default/files/1130%20Smith-Yoshimura.pptx>

⁶⁷ Karen Smith-Yoshimura, Challenges of Multilingualism, Visualizing Digital Humanities, 21 June 2017 <http://www.oclc.org/content/dam/research/presentations/smith-yoshimura/oclcresearch-challenges-of-multilingualism-lorentz-workshop-2017.pptx>

⁶⁸ Carol Jean Godby and Diane Vizine-Goetz, BIBFRAME and OCLC Works: Defining Models and Discovering Evidence, Library of Congress BIBFRAME Update Session, 26 June 2017 <http://www.loc.gov/bibframe/news/bibframe-update-an2017.html>

The issues raised in this section imply that the relationship between modeling and algorithmic discovery of the work is complex. On the one hand, algorithmic discovery is fundamentally agnostic about the subtle differences described in this document. The goal is utility, and the usual outcome--on legacy data, at least--is that some details cannot be easily discovered. But emerging from this baseline is a reality check that can inform model development and raise questions for future discussion. For example:

- If algorithmic discovery falls short of the goal of capturing all details in the model specification, what would a minimum viable result look like?
- Echoing an observation stated at the beginning of this document, it is unlikely that the library community will agree on a single model of work. If so, this creates an additional task for algorithmic discovery: how will the models be harmonized--and in linked-data terms, what sorts of 'same as' relationships should be established?
- What are the institutional barriers to making the results of algorithmic discovery of the work more widely available, and how can they be overcome?

5: Open Questions for PCC Resolution

The library catalog is no longer strictly a silo. The outside world has been making inroads for more than a decade, principally in the form of websites like WorldCat that provide a portal to local catalogs, and various browser add-ons that make use of ISBNs to link catalogs to book-oriented websites like Amazon.com. Likewise, other products and services have vastly expanded the local search environment beyond the catalog, first via federated search, where the same search can be replicated across multiple content sources, and now via resource discovery services, where metadata from many content sources is brought together into a single local search environment, returning a consolidated result set. This process of tearing down walls can be expected to accelerate in the future.

For the bibliographic work as a concept this de-siloing creates something of an existential crisis. Always a fuzzy concept, and often applied in previous cataloging practice only to the works most frequently encountered in library catalogs, today the work and its companion, the expression, are both defined more granularly than ever. At the same time these concepts are applied only to cataloged library resources which represent an ever-shrinking percentage of the resources in the user's environment.

The question surrounding the application of the work concept is first and foremost a cataloging practice question. As the integration of the FRBR model is relatively new, and since library systems that support the FRBR entities still do not exist, there are many more questions than answers. In nearly every area where the nature of the work is discussed in this report, there is not an obvious single solution. Some of the questions that arise in this report are discussed in this section. All of them should be viewed by PCC as areas for further investigation and decision-making.

5.1: Works and authorities

*What is the relationship between works in the FRBR sense and work **authorities** as defined in today's cataloging?*

As we note in section 2 on the history of the work, the “uniform title” as defined by the previous cataloging rules, AACR2, and the current rules, RDA, is a derived heading that should be represented in an authority file. Uniform titles serve the function of providing a title heading which, alone or combined with an authoritative creator heading, brings together a set of works or expressions that would not otherwise have the same title heading. Currently most works are implicit rather than explicit, represented by a combination of elements in the description of a resource. In a MARC 21 bibliographic record, these elements may be in fields 1XX, 240, or 245—used alone or in some combination to identify the work—and in some other content-related field, such as the language, subject headings, and summary and/or table of contents. In

the future environment, whether something is considered a work will depend on the model in which it is expressed.

Unlike name headings, uniform title and name/title authorities are created on an “as needed” basis, and practice has varied among libraries. What we can say is that only a small number of uniform title entries exist in the authority file. Existence in the authority file implies that prescribed due diligence has been done in establishing the heading, and that certain traces of that effort have been recorded along with the authoritative title heading.

Given the effort that is required to develop an authority file entry, it would be hard to argue that all works--even those with no variant titles--would merit an entry in the authority file. Yet all resources now have an aspect of “work-ness,” something that was not explicitly part of the thinking in cataloging prior to the FRBR concept of works.

Much of the effort currently required by authority work is spent on formulating and ensuring the uniqueness of an authorized access point. The unique AAP was a necessity in the context of alphabetical catalogs accessed via browsing through indexes. In the context of a search-driven environment and descriptions with designated data elements and unique identifiers, the formulation of representations to assist users with identifying and selecting a resource may become more flexible and less onerous. The larger challenge will be moving beyond the flat files of resources now found in library catalogs to hierarchical, networked files enabling searchers to navigate from generalized representations of entities of varying types to more granular representations of particular entities and resources. If that can be managed, the traditional authority file of AAPs could be replaced by shared entity representations integrated with the local catalog.

Another aspect of this question has to do with file management and sharing. Separate from the question of whether the authority record models work characteristics sufficiently to serve as a work description is the question of whether using a shared, uniform, distributed file of authoritative data is an appropriate or necessary model for sharing work descriptions in a given library community. Authority files have long differed from bibliographic files in this regard. While some bibliographic records have been considered more authoritative than others in some libraries, there has been no consensus (with the exception of CONSER) around a single authoritative bibliographic record for representing a resource. But this is the claim made for each authority record--that it solely and uniquely represents an entity within its authority file domain.

What are the expectations of work descriptions as a body of authoritative data? Will there be many work descriptions of the same work in a shared pool for libraries to select from, or will there be consensus around the creation and maintenance of one shared description? In a linked data environment, it may be possible to develop a model allowing for many diverse elements related to a work to be linked in a single graph and harvested in various ways for local use. In such an environment, the strictures on participating in the creation and augmentation of such a graph may be looser, enabling a wider community engagement in the task of work description. In any case, PCC needs to be clear about what policy it will pursue for managing work descriptions in the aggregate as well as individually.

It is not clear how authoritative entries will be expressed in a future bibliographic data format, although the current separation of bibliographic and authority records may be subsumed under a single approach to entities, some of which will be considered authoritative. This question is less problematic when dealing with persons or corporate bodies because those have always been routinely represented in the authority file.

5.2 What is included in “a work”?

How can the extent of the work be defined? Does it include the contents of the creator and subject entities? How much of the graph is needed for different functions such as cataloging a new expression, user displays, etc.?

Because we think of our current bibliographic description as being contained in records it is second nature to assume that the information about the work will be contained within a work record. The linked data model, however, does not include the concept of a record, but instead uses open graphs. As it is defined in FRBR and in RDA, the work is an entity with relationships to agents (persons, collective bodies), subjects, and possibly to other works. Agents and subjects are described as entities in their own right, and can have additional relationships to other entities. A simple example is that a person can have an agent relationship with more than one work, and can have different types of relationships with different works. The person entity may include information relating to the person but not necessarily relevant to any of the agency the person has with a particular work.

In processing graphs one follows the relationships that link information about entities, but there may be many graphs and relationships in any given set of data. Because these graphs are not bounded in records, it is necessary to define what relationships and properties are suitable for any given function, such as cataloging, discovery, and user display. This may seem like a

disadvantage, but the great advantage is that one can define any number of “views” of one’s data by including or excluding specific relationships and properties. These views may make use of entities that have been defined with an entity identifier, or they can cut across defined entities without selecting all properties of any single entity.

Thus, when the question “What is included in a work?” is reframed as “What is included in a public-facing work description?”, the answer can vary. Libraries may have different local configurations for what data to harvest from work and other entity graphs to represent “a work” to their users. To support such options, the approach to creating and managing work and other entity descriptions will likely favor an inclusive approach to data elements directly related to each entity, but not necessarily to data elements with only an indirect relationship. Policy decisions about what to include in the cataloger’s work description will depend on what expectations PCC has regarding the ways systems will be able to navigate across and harvest from descriptions for diverse entity types, not just from work descriptions.

5.2a Works characterized by conventional collective titles (compilations and aggregates)

RDA provides for characterizing certain compilations and aggregates of works as works in their own right via the mechanism of conventional collective titles (e.g., 6.2.2.10 [Recording the Preferred Title for a Compilation of Works by One Agent] and the alternatives for employing the term “Selections” under 6.2.2.9.2 [... for Two or More Parts of a Work] and 6.2.2.10.3 [... for Other Compilations of Two or More Works]), and LC-PCC applies these instructions and alternatives. The use of the “Selections” alternative aligns with much usage in legacy data and past practice for the LC-PCC community, and has demonstrable utility for users of the catalog, grouping similar collection of works in support of the Find and Select user tasks. However, it aligns poorly with RDA’s focus on distinct works. The authorized access points which result from adopting these “Selections” alternatives define criteria for sets of distinct works rather than the distinct works themselves. Yet RDA does not indicate that anything other than an RDA work is being named with the use of “Selections.”

This raises questions about the nature and scope of the work entity. RDA defines a “work” as “A distinct intellectual or artistic creation, that is, the intellectual or artistic content.” This implies a clear particularity for works. Two short stories are not the same work simply for having the same literary form. They must also have a commonality of text or, in the case of translations, of narrative to be considered two expressions of the same work. By the same argument, two collections of selected short stories are not instances of a single work unless some or all of the

selected texts are the same. When the contents differ to a significant degree, they become two separate works. This is the approach adopted in IFLA LRM.

The use of “Selections” in the RDA alternatives preserves a practice from older cataloging which was intended to collocate certain kinds of aggregations--not to identify particular works. Yet RDA and PCC do not differentiate the intent of the alternative practice from RDA’s instructions for naming works, and thus conflates the work groups named by the “Selections” access points with authorized access points for individual works.

The use of these authorized access points becomes more convoluted when catalogers seek to establish AAPs for particular works, given that they have a conventional collective title included in that process. This has resulted in awkward and lengthy formulations such as

Hemingway, Ernest, 1899-1961. Short stories. Selections (Men without women) –
[no2016130264]

Hughes, Langston, 1902-1967. Correspondence. Selections. (Crawford and Patterson)
[no2015103764]

Guzmán, Martín Luis, 1887-1976. Works. Selections. 1987. Fondo de Cultura
Económica
[n 89623522]

In each case, users would have been better served by following the basic instruction in RDA and accepting the found title as the preferred title rather than by inserting a generalizing characterization of the work in place of the preferred title and then qualifying the resulting AAP string to restore its particularity as a work.

The “Selections” alternatives provided in RDA effectively depart from the concept of particular works. The utility of being able to express more general categorizations of works is not in question; but the wisdom of using the work AAP for this purpose is.

We recommend that PCC explore alternative ways to conceptualize and express its adoption of the “Selections” alternatives that would recognize these access points as expressing categories of works rather than work preferred titles, and not as the basis for true AAPs for particular

works, which should be formulated when needed following the basic instructions in RDA Chapter 6.

We also recommend that PCC explore ways to ensure that the terms used in these alternatives--terms for conventional collective titles and “Selections”--are recorded in work descriptions as designated “form of work” or other data and are available for use in faceting (perhaps as a vocabulary encoding scheme [VES]). A faceted approach to these characterizations of works would offer users more flexible metadata for accessing compilations than the current AAP strings. A faceted approach would also make it possible to exclude partial compilations from a search for a creator’s works, which might also be desirable.

PCC should also be prepared to craft policies to address issues likely to emerge around the treatment of aggregate works. The RDA Steering Committee is still in the process of formulating instructions to account for aggregate works, both conceptually and in terms of authorized access points. It is premature to formulate a response to these developments, but the task force notes that this development is on the horizon and will potentially have a significant impact on the nature and treatment of the work entity in RDA cataloging.

5.3: The work entity

What functionality is desired that could require a work entity to be created? Does it need to be created for every cataloged resource?

As described in section 3 on modeling the work, not all schemas that define a work conceptually also define a machine-readable metadata instance for the work. Many descriptions of the FRBR work assume that there will be a separately identified machine-readable work entity for each described resource. This is an assumption that needs to be tested based on use cases.

Decisions about the work entity must be made in terms of the desired functions for which works are useful or necessary, in the catalog and in other library system modules. Some work has been done on “FRBR-ized” catalogs, most notably for music catalogs where work titles are provided for the majority of cataloged resources. However, little is known of how works will serve other libraries. Use cases need to be developed for a wide range of libraries and bibliographic functions; these should include libraries of various sizes and constituencies, and some careful thinking about how works could enhance discovery and selection of resources.

In considering questions about functions that make use of the work entity there is a tendency to focus on certain sets of resources with long and complex publishing histories, the works of

Shakespeare being a prime example. Studies of functional uses of bibliographic works must, however, also fully explore the role of the work entity for that majority of resources that exist as a work with a single expression. This is not only true for the lifetime of some resources, but is also the case for the first instance of any creative resource. There may be questions of efficiency that arise when contemplating whether every bibliographic description requires an identified work entity. This must be looked at as an issue for systems design, cataloging workflows, and user services.

5.4: Workflow questions

What are the cataloging workflow concerns that relate to the work as a description? as an entity?

One of the functionalities of incorporating specific work description and possible work entities into library practices is the effect these may have on the cataloger's workflow. This question is directly related to the previous question about the work entity but is of particular importance to PCC and the cataloging community.

Careful consideration of the role of work description and the potential for sharing of descriptions or entities needs to be explored. Above all, this needs to be explored without prior assumptions about the existence of work "records" in future bibliographic systems. Both FRBR and BIBFRAME have been interpreted as necessarily resulting in a separate machine-readable structure that holds the work-related description, but this decision may be premature due to the absence of a thorough analysis of workflow needs.

One aspect of traditional cataloging workflow is the concept of the "file" against which cataloging is being done. For Works, this may be a collection of descriptions originally derived from Library of Congress MARC bibliographic and authority records, or the broader VIAF Works, currently derived from a large number of national authority files, or any number of other sources. To what extent, and how, will works created in the course of cataloging be linked to works and expressions beyond the immediate cataloging environment?

Another workflow-related issue is the question of authorized access points for works. Would the creation of a work description necessarily entail the creation of a unique authorized access point as currently defined in RDA for the work? If so, then the creation of work descriptions especially for works entered under title could be an onerous addition to current practice. If not--if the identification of works by a unique aggregation of descriptive data elements rather than by a single, uniform authorized access point was regarded as practicable--then the difficulty of

differentiating work descriptions from one another would be significantly reduced, though not entirely eliminated.

5.5: Correspondences with works in the Intellectual Property community

How will we relate to works created in the IP community?

In a Linked Data environment, entities will link to entities across domains (and communities). If the scope of the linked entities corresponds, then the place of those entities in their respective conceptual models may not be so important, i.e. it may not matter whether X is characterized as an expression in one domain (with all the attendant relationships) but characterized as a derivative work in another (with its own attendant relationships). Even if the correspondence of scope is somewhat fuzzy, what may matter more is the expanded universe of resources opened up by the linking.

The distinction between works and expressions in the library community reflects the history of the printed library catalog, where the text string that identified and collocated expressions was an extension of the string that identified the overarching work. Hence, in our cataloging instructions the authorized access point for an expression takes the authorized access point for the overarching work as its basis. In the past, this formed a shorthand for the relationship between the work and its various expressions, and supported the limited facilities for collocation in book and card catalogs. In a Linked Data environment, relationships and clusters of relationships can serve the same function.

The question then becomes how well classes of works and expressions in the cultural heritage domain map to classes of original and derivative works in the IP domain. Taking the 10 derivation types specified in ONIX for ISTC (see appendix A2), will we need something analogous to the RDA/ONIX Framework for Resource Categorization to support such a mapping? If so, there will still be cases where the mapping simply fails (e.g., the RDA distinction between a free translation [work] and a translation that adheres more closely to the original [expression] and the rather more challenging distinction between the abridgement of a work and of one of its expressions).

A larger question may be whether, in practice, there will be anything to link to, even with a robust mapping mechanism. While the International Standard Work Code (ISWC) and, to a lesser extent, the International Standard Textual Work Code (ISTC) are being assigned to works in their respective domains, the metadata associated with them is not expressed in a vocabulary that conforms to RDF. While the Metadata Committee of BISG has set up a Schema.org

Working Group, its task is to map ONIX elements—presumably, but not necessarily, including ONIX for ISTC—to Schema.org types and properties rather than to recast ONIX itself in a Linked Data structure.

Appendices

A1: The concept of the work in the publishing community

A2: The concept of the Work in the intellectual property (IP) community

A3: The Work in legacy data

A4: Discovering the Work in Legacy Data: OCLC's Experience

A1: The concept of the work in the publishing community

Introduction

Unlike other communities under discussion, the publisher community is principally concerned with the manifestation (product) rather than the work, and with communicating metadata relating to commercial transactions involving the manifestation. Metadata relating to the work is included only to the extent that it facilitates these transactions. Consequently, there is no distinct work entity in the publishing community, though transactions may include identifiers assigned to the work in other contexts. In particular, the digital object identifier (DOI) may be used to facilitate transactions involving electronic content where a selection of electronic formats is involved (see below under Digital Object Identifier).

The <indecs> Framework

The <indecs> (interoperability of data in e-commerce systems) Framework is a generic ontology-based approach to identification, supporting interoperability across media, functions, levels of metadata, and semantic and linguistic barriers, and embodying four principles:

- Unique identification: Every entity is uniquely identified within an identified namespace
- Functional granularity: It should be possible to identify an entity whenever it needs to be distinguished
- Designated authority: The author of an item of metadata should be securely identified
- Appropriate access: Everyone requires access to the metadata on which they depend, and privacy and confidentiality for their own metadata from those who are not dependent on it

Relationships lie at the heart of the <indecs> analysis and underline the importance of the unique identification of entities on which relationships depend. Beyond this, <indecs> emphasizes authority: identifying the person making the claim in a given case.

Among the applications using the <indecs> approach are ONIX, the DOI system, DDEX (the music industry's messaging and data dictionary applications), and the Linked Content Coalition (an organization aimed at facilitating transactions between companies and individuals who want to trade in rights).

The <indecs> Framework acknowledges other models—specifically the CIDOC CRM and FRBR—and asserts their mutual compatibility:

Different models of the life cycle of content may have important differences, not least in the specific meaning attached to the names of terms they employ. The <indecs> approach also has much in common with the CIDOC Conceptual Reference Model (CRM), an ontology for cultural heritage information, and the Functional Requirements for Bibliographic Records report (FRBR) in the library world. CRM, FRBR, and <indecs> were each informed by different functional requirements, and so evolved different mechanisms for dealing with the issues that seemed most important to them. Broadly, they are compatible, and effective integration of metadata from schemes based on them should be achievable, but they must be handled with care. As an example: the terms abstraction, manifestation, item and expression are often used in considering content life cycles (e.g. a sound recording is the expression of a musical work during

a recording session at a particular place and time, and is distinct from, say, the master tape made, which is a manifestation). These were dealt with in <indecs>, but may have slightly different meanings in other schemes. Such an analysis of meaning of a term from a scheme is possible in <indecs> by mapping the precise definitions into further terms with precise definitions within the <indecs> Framework. <indecs> and other frameworks continue to be developed and refined through the process of implementation.

ONIX

ONIX is a family of standards in an XML format for exchanging messages between parties. It includes ONIX for Books, ONIX for Subscription Products, ONIX for Publication Licenses, Rights Information Services, and Reproduction Rights Organizations, and ONIX for Identifier Registration (ISBN, ISTC, and DOI). While there are no current plans to make ONIX data available as Linked Data, work is under way to express ONIX for Books data in JSON and to map it to schema.org.

In 2012 Jean Godby produced a crosswalk from ONIX for Books 3.0 to MARC 21. Although ONIX for Books describes manifestations rather than works, like MARC 21 it also carries elements that relate to the work or expression, such as contributor roles, edition types, BISAC/Thema subject terms and codes, audience codes, and certain parts of the description. ONIX also supports relationships between manifestations that may correspond to broader relationships between expressions and works, some more granular than those supported by MARC 21.

Digital Object Identifier (DOI)

Because English syntax in this case is ambiguous, it should be pointed out that a Digital Object Identifier is a *digital identifier of an object* rather than an *identifier of a digital object* (though it may also be the latter). In this way it is analogous to the URI, which can identify real-world objects as well as digital objects.

DOIs may be assigned at the work or manifestation level, with the work including resources that would be considered expressions under RDA, and the manifestation comprising resources that would be considered partial expressions (i.e., just those print manifestations or online manifestations of a given expression that originate from a given publisher).

The relationships between works are stated broadly:

- Includes
- Is part of
- Is a new version of
- Has a new version
- Is a different language version of
- Is a resource about
- Is continued by
- Is a continuation of

Related works must include a work identifier, which may be proprietary, an ISTC, or a DOI.

Works may also be related to products (manifestations):

- Is manifested in

The ONIX DOI Registration Formats include formats for whole monographs, chapters of parts of monographs, serials, serial issues, and serial contributions, with other types of material to be added in future.

The Publishers International Linking Association (PILA) operates Crossref, a collaborative linking service which, inter alia, promotes the use of DOIs in electronic scholarly information.

A2: The concept of the Work in the intellectual property (IP) community

Berne Convention

The Berne Convention defines “literary and artistic works” (the works covered by the convention) as

every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, such as books, pamphlets and other writings; lectures, addresses, sermons and other works of the same nature; dramatic or dramatico-musical works; choreographic works and entertainments in dumb show; musical compositions with or without words; cinematographic works to which are assimilated works expressed by a process analogous to cinematography; works of drawing, painting, architecture, sculpture, engraving and lithography; photographic works to which are assimilated works expressed by a process analogous to photography; works of applied art; illustrations, maps, plans, sketches and three-dimensional works relative to geography, topography, architecture or science.

It also states that “[t]ranslations, adaptations, arrangements of music and other alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work.” The convention characterizes these as “derivative works.”

Finally, “[c]ollections of literary or artistic works such as encyclopaedias and anthologies which, by reason of the selection and arrangement of their contents, constitute intellectual creations shall be protected as such, without prejudice to the copyright in each of the works forming part of such collections.”

The World Intellectual Property Organization provides the following definition of “derivative work”:

In copyright law, the term “derivative works” refers to the translations, adaptations, arrangements and similar alterations of preexisting works which are protected under Article 2(3) of the *Berne Convention for the Protection of Literary and Artistic Works* (1971) as such without prejudice to the copyright in the preexisting works. Sometimes, the term is used with a broader meaning, extending to the compilations/collections of works protected under Article 2(5) of the Convention, (as well as under Article 10.2 of the World Trade Organization (WTO) *Agreement on Trade Related Aspects of Intellectual Property Rights*, 1994 (the TRIPS Agreement), and Article 5 of the *WIPO Copyright Treaty*, 1996 (WCT)). (WIPO Guide to the Copyright and Related Right Treaties Administered by WIPO and Glossary of Copyright and Related Rights Terms, WIPO.)

In this sense, a “derivative work” includes compilations of data or other material, whether in machine-readable or other form, which, by reason of the selection or arrangement of their contents, constitute intellectual creations. (Art. 2(5) Berne Convention, Art. 10(2) TRIPS Agreement, Art. 6 World Copyright Treaty.)

Some jurisdictions have adapted the definition of derivative works in the field of traditional cultural expressions. According to the Pacific *Regional Framework for the Protection of Traditional Knowledge and Expressions of Culture* (2002), the term refers to any intellectual creation or innovation based upon or derived from traditional knowledge or expressions of

culture. (Pacific Regional Framework for the Protection of Traditional Knowledge and Expressions of Culture, 2002, Part I. 4.)

The United States Copyright Office provides a further elaboration:

To be copyrightable, a derivative work must incorporate some or all of a preexisting “work” and add new original copyrightable authorship to that work. The derivative work right is often referred to as the adaptation right. The following are examples of the many different types of derivative works:

- A motion picture based on a play or novel
- A translation of an [sic] novel written in English into another language
- A revision of a previously published book
- A sculpture based on a drawing
- A drawing based on a photograph
- A lithograph based on a painting
- A drama about John Doe based on the letters and journal entries of John Doe
- A musical arrangement of a pre-existing musical work
- A new version of an existing computer program
- An adaptation of a dramatic work
- A revision of a website

From this we can see that “derivative work” in copyright law includes some entities (e.g., motion pictures based on novels, a sculpture based on a drawing) that would be considered works under RDA, others (e.g., revisions of websites or previously published books) that would be considered expressions, and yet others (e.g., abridgements, translations) that might be considered either, depending on the circumstances. For copyright law, the key is that the content of a derived work derives from the original work. This will present challenges in any database that uses a work model based on copyright law, especially in cases where there is not a one-to-one correspondence between a type of derived work and an RDA FRBR Group 1 entity.

International Standard Musical Work Code (ISWC)

The ISWC (ISO 15707:2001) is a relatively successful code for identifying musical works. It is part of the CIS (Common Information System) plan with CISAC (International Confederation of Societies of Authors and Composers) and is administered by 47 agencies in 68 countries (with the notable exception of Russia). Operating as they do within the context of intellectual property rights, ISWC agencies assign codes to works only after they have also uniquely identified all the associated creators. As at 3 February 2016 there were 18 million ISWC musical work records accessible at ISWC-Net.

[T]o obtain an ISWC, a publisher must provide the following minimum: at least one original tile for the work; all [composers, authors, composer/authors, arrangers, publishers, administrators, and sub-publishers] of the work identified by their Interested Parties Information (“IPI”) code; and whether the work is derived from an existing work. One significant issue with ISWCs, then, is that they cannot be assigned until all the songwriters on a musical work are identified. This has the benefit of assuring that data are complete before an identifier is attached. But it also leads to a substantial lag time before the ISWC for a particular musical work can be assigned—unfortunately, this can occur well after a record is released, so that digital files embodying the

individual tracks often will not include ISWCs identifying the underlying musical works. ASCAP and BMI—which also use proprietary numbering systems to track works internally—add ISWCs to their databases as those codes are assigned.

The descriptive metadata associated with an ISWC includes:

- The title of the work
- All composers, authors, and arrangers of the work, identified by their IPI numbers or ISNI and role codes
- The work classification code (from the CIS standards list)
- In the case of “versions”, for example arrangements, identification of the work from which the version was made

A musical work, in the terms of the ISWC, is a result of an intangible creation of one or more people (creators); it is composed of a combination of sounds with or without accompanying text.

Even with 18 million records, the metadata in the ISWC Network seems of uneven quality. An arrangement for two guitars of Debussy’s *Children’s Corner Suite* by Jan Žáček and Richard Jackman shows a relationship to the identifier Debussy as a creator but not with the original work. This exemplifies a shortcoming in the ISWC system—and in IP systems generally—from the point of view of cultural heritage institutions. Because the primary purpose of such systems is to ensure that authors’ rights are respected, they are more concerned with the living than the dead and likewise more concerned with works that enjoy copyright protection than those that do not. Consequently, older works—which also tend to be those that are more likely to have derivative works associated with them—are much less likely to be represented (as are other works in the public domain or under Creative Commons licenses).

The IPI (Interested Parties Information) system defines two entities of relevance to our discussion: a creation class (a class of products of human imagination and/or endeavor) and a creation subclass (combinations of a creation class and their creation subclasses), each represented by a two-character code. The IPI system is implemented as a kernel server object providing services to connected clients using EDI (Electronic Data Interchange). Unfortunately, no further information is publicly available.

International Standard Text Code (ISTC)

The ISTC (ISO 21047:2009) is a more recent code used as a numbering system supporting the unique identification of textual works. According to the International Federation of Reproduction Rights Organisations, the ISTC “is particularly useful in administering copyright, licensing, collocation, royalty/fee payment, improved discovery services and sales analysis.” If the numbers at the ISTC beta search facility are to be believed—185,518 ISTCs assigned since the standard was approved by ISO in 2009—this identifier is off to a slow start. For example, just two codes have been assigned to the *Adventures of Huckleberry Finn*: one for the 1884 original (ISTC A03-2012-0000B469-0) and one (ISTC A02-2009-00000A87-C) for an abridged version, and these with limited metadata. In an article published in 2012 Margaret Hepp Harrison questioned the viability of the ISTC, as Michael Holdsworth had two years earlier. Harrison noted that “In a survey conducted for this paper of 10 major publishing trading partners, including Amazon, Barnes & Noble, Ingram, and Baker & Taylor, 100% of partners reported

that they were not using the ISTC at all.” Following a five-year systematic review of the standard in 2014, the Secretariat expressed “serious concerns about the viability of the standard.” A working group recommended revising the standard to address problems of granularity and the fact that “responsibility for registration [is] placed with publishers, who [are] not always motivated to do this.”

Beyond questions of uptake, there are questions about the sufficiency of the metadata. According to the ISTC User Manual,

The ISTC database used by the STRS system is designed merely to enable different works to be distinguished from one another. It is not intended to provide a comprehensive repository of information about each work. Therefore, while it is always desirable to have comprehensive information on each record because it makes them as distinctive as possible, it is not strictly necessary in every case; if a reference work (such as an encyclopaedia) can be distinguished from previous versions using its edition number, there will be no benefit in specifying all the numerous contributors’ names.

While the ISTC beta search facility presumably does not provide access to all the metadata associated with an ISTC, it currently returns very little to the searcher, and is often missing such seemingly fundamental elements as a year of creation.

Having said this, BISG *Best Practices for Product Metadata* (2015) states that

It is a best practice to begin to rely upon standard identifiers such as ISTC (International Standard Text Code) to identify the work underlying a product and ISNI (International Standard Name Identifier) to identify the party underlying a name. Variants in the spelling or presentation of the titles of works or the names of parties will thus not result in mismatches between products and their underlying works or mismatches between names and their underlying parties. The use of proprietary work and contributor IDs can be an effective interim step in the identification of works and parties until the international standard identifiers are adopted more widely.

So while the ISTC may not yet have taken off, passengers have not yet been refunded their ticket price.

For purposes of this discussion, the ISTC distinguishes between original and derivative works as in international copyright law. The ONIX format for ISTC specifies 11 derivation type codes as follows (corresponding RDA FRBR Group 1 entity in square brackets):

1. Unspecified
2. Abridged edition [E/W]
3. Annotated edition
4. Compilation
5. Critical edition
6. Excerpts
7. Expurgated/edited edition
8. Non-text material added (enhanced ebook)
9. Revised edition [E]

10. Translation [E/W (if free translation)]

11. Adaptation [W]

One important point to bear in mind regarding ISTCs is that they apply solely to text. That is, they are mute regarding other types of content. Consequently, the presence or absence of illustrations does not affect the validity of the ISTC. So in one important sense the work represented by the ISTC will always be less—at least potentially—than the corresponding work represented in a library catalog. This primacy of text, however, can still result in different ISTCs for illustrated editions *if the text references the illustrations*. Likewise, by virtue of their containing text in their labels and legends, cartographic materials are eligible for ISTCs.

A3: The Work in legacy data

The work entity in legacy bibliographic and authority data is implicit rather than explicit. According to section Z1 of the Library of Congress Descriptive Cataloging Manual (DCM), authority records are typically created for works and expressions only when one of the following is true:

- A variant access point is needed for the work
- Research performed in the course of constructing the authorized access point for the work needs to be recorded
- An authorized access point for the work is needed as a related work or subject access point on the bibliographic record for a different work
- Certain information about the work needs to be recorded, such as the citation title for a law

Consequently, for the vast majority of works, no authority record exists. This practice did not change with the international implementation of Resource Description and Access (RDA) on 31 March 2013.

Extracting work data from existing authority records

In theory, work data can be extracted from existing authority records for translations, where the work authorized access point will constitute the authorized access point for the expression, less any additions relating to the expression (e.g., language, translator, date). This will produce rudimentary work records for this subset of works. It may be possible to extract work data from other classes of authority record as well.

Extracting work data from existing bibliographic records

While it is acknowledged that the overwhelming majority of bibliographic records represent the sole expression of a single work, there is no foolproof way to identify these works. Cataloging rules prior to RDA did not require that works be distinguished from one another in the catalog. Such distinguishing was typically left to filing rules. For example, the ALA filing rules in force at the time of the implementation of the second edition of the Anglo-American Cataloguing Rules (AACR2) directed that

... [T]itle main entries for separate works and serials other than periodicals and newspapers [be] subarranged in groups in the following order:

- a. Those with nothing following the title, subarranged by place of publication
- b. Those with subtitles or other phrases following the title, subarranged alphabetically by the subtitles or phrases

This represents a case where it may not always be possible to reliably identify or distinguish works when two or more manifestations share the same title proper. This may also be true when manifestations of works by the same corporate body share the same title proper.

Changes in the work entity over time

Legacy bibliographic data represents works cataloged under various cataloging codes. Older codes distinguished works by different criteria than RDA. For example,

- Prior to AACR2, a work produced under editorial direction was distinguished by its editor (i.e. the authorized access point for the work comprised the a.a.p. for the editor followed by the preferred title of the work).
- Prior to AACR2—and to a lesser extent AACR1—works issued by corporate bodies were distinguished by corporate body (i.e. the authorized access point for the work comprised the a.a.p. for the corporate body followed by the preferred title of the work).
- Prior to AACR1—prior to 1971 in the US—serial works were distinguished by changes in the numbering scheme of the parts rather than by changes in title proper. In these cases, the authorized access point for the serial work was the latest title proper borne by the serial.

Although bibliographic records for revised editions that involve either a change of author or editor—personal or corporate—or a change of title would often include an authorized access point for the immediately preceding edition, the a.a.p. for that earlier edition will not identify itself as such (though the bibliographic record may include a note to that effect).

In some cases, it is hard to see how sufficient data might be automatically collected for a given work. For example, when a work is the product of joint authorship but only the first-named author is identified as such (1XX) in the bibliographic data. In the early days of MARC such relationships were indicated by an appropriate indicator value in the 7XX field, but unfortunately this was subsequently removed when it was felt not to serve a useful purpose.

This is not an exhaustive list of differences in the definition of the work over time, but it serves to highlight the challenges of retroactively identifying works in legacy data. Any system that postulates a work entity as an essential constituent will face these challenges, and others.

A4: Discovering the Work in Legacy Data: OCLC’s Experience

OCLC’s FRBR algorithms. Since the early 2000s, OCLC has been experimenting with algorithms that discover evidence for FRBR Group 1 entities and relationships, or the so-called ‘WEMI hierarchy’. Since a Work is an abstract entity or a conceptual object, according to the IFLA Library Reference Model, its presence must be inferred.

OCLC’s algorithms discover evidence for entities and relationships described in the classic FRBR documents in collections of bibliographic and authority records expressed primarily in MARC. The outcome is not a new set of definitions for FRBR concepts, but a set of business rules for discovering evidence for established definitions. Since the algorithms are undergoing continuous revision as new data is analyzed, the business rules represent an open-ended set representing what has been successful so far. They are listed in Table 1 below.

FRBR	OCLC’s business rules
Work “A distinct intellectual or artistic creation”	<ul style="list-style-type: none"> • Author + title + format • Content properties: subject, description, author, contributor, title, language, resource type, genre • A Uniform Title Authority description
Expression “The specific intellectual or artistic form that a work takes each time it is ‘realized’”	<ul style="list-style-type: none"> • A bibliographic or uniform title authority description with a “language of content” tag • A bibliographic or uniform title description showing evidence of a translator, or a “translation” relationship
Manifestation “The physical embodiment of an Expression of a Work”	<ul style="list-style-type: none"> • Evidence of an ISBN, ISSN, or other product identifier • The name of a publisher and/or a physical description showing pagination, extent, or physical dimensions.
Item “One exemplar of a Manifestation”	<ul style="list-style-type: none"> • Evidence of uniqueness, such as a barcode or unique physical location, such as a set of geospatial coordinates

Table 1. OCLC’s business rules for discovering FRBR WEMI concepts in MARC data

The most complex process starts with MARC bibliographic records. In the simplest terms, source records are assembled into Work clusters, which are identified by considering data that describes creators or contributors, titles, and formats, and genres. Fuzzy matching considers additional evidence in fields that describe publishers and extents, with details differing according to format because discriminating information may be stored in various places. The resulting Work clusters are distributed as follows:

- 50% of WorldCat records (77% of clusters) are singletons. The cluster count is smaller than the raw record count because clusters for resources with complex publication histories may contain many contain dozens of records.
- 25% of WorldCat records (9% of clusters) do not require fuzzy matching.
- 25% of WorldCat records (15% of clusters) need fuzzy matching.

Outputs from the FRBR algorithms operating on bibliographic data are used throughout OCLC to organize the displays on WorldCat.org, mine data for cataloging productivity tools such as

[Classify](#), and develop more responsive experimental user interfaces such as [FictionFinder](#) and [Cookbook Finder](#).

Works can also be identified from MARC Authority records in a process that is conceptually much simpler. If the input is a MARC Uniform Title authority record, the output is labeled as a Work record; and if the input contains descriptors related to the language of the content, the output is labeled as an Expression record. These decisions create the Work and Expression records accessible from VIAF.

OCLC's model of creative works. OCLC's RDF datasets are another product of the Work algorithms applied to bibliographic and authority records. The inputs are mapped to a Semantic Web vocabulary and decomposed into RDF triples, to which the following are added:

- RDF Type assignments
- Persistent URIs
- Correspondences to FRBR WEMI categories
- Descriptions in Schema.org; and in some limited circumstances, to the extension vocabulary bib.schema.org. This vocabulary is one outcome of the [Schema Bib Extend](#) community group, hosted by the World Wide Web Consortium, which represents linked data experts from the library, library services, and publishing communities.

Figure 1 shows a high-level model of creative works, which is consistent with OCLC's public datasets such as WorldCat's catalog data, VIAF, and FAST. The model resembles to the [British Library Data Model](#), or BLDM, which served as its inspiration. Like its predecessor, the OCLC model has a set of properties corresponding to Author, Subject, Publisher, and membership in a series, which are highlighted with multicolored backgrounds in Figure 1. Likewise, all RDF objects are expressed as URIs, except for string literals such as descriptions, dates, ISBNs, and language tags. However, OCLC's model of creative works is technically simpler than the BLDM because entities and relationships are expressed in just two namespaces, Schema.org or bib.schema.org, not the BLDM's fourteen. Some internal details are also simpler because the primary properties in the OCLC model that describe authorship, subjects, and publication details associate real-world-objects to the creative work, differing from the BLDM's reference to SKOS Concepts that refer only indirectly to RWOs.

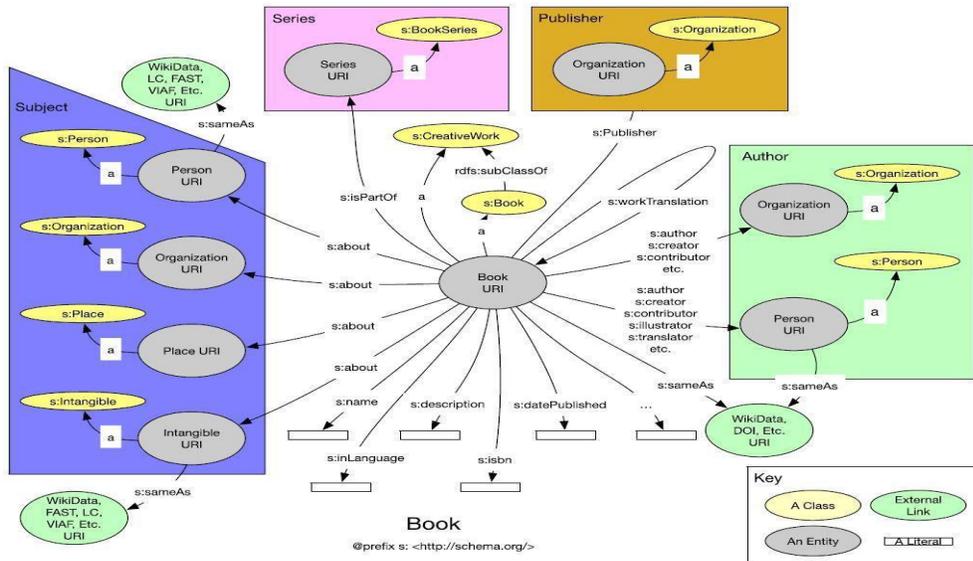


Figure 1. The OCLC model of "Book" expressed in Schema.org

At a lower level of detail, the OCLC model of creative works is enhanced with evidence for the FRBR WEMI class hierarchy [Godby 2013;

Godby, Wang and Mixer, 2015] connected with properties discoverable in the data. This model is shown schematically in Figure 2 below.

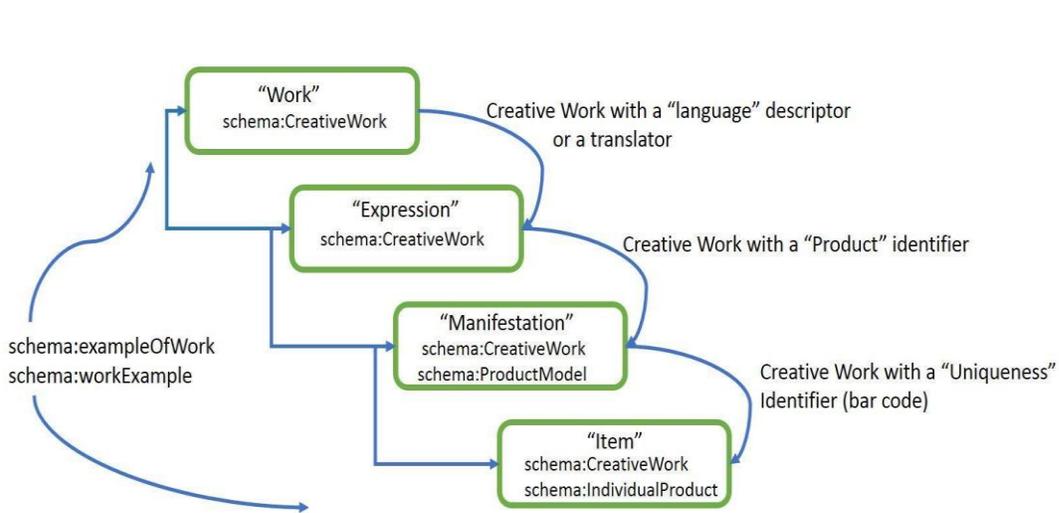


Figure 2. The FRBR WEMI hierarchy

expressed in Schema.org

Though the model is flexible and responsive to details in the input data, the following features have remained constant despite many generations of iterative changes:

- The concepts identified in Table 1 are interpreted as correspondences to FRBR, but are not ontologically identical to the definitions of FRBR classes and properties defined in classic documents because the OCLC algorithms cannot guarantee that FRBR's definitional

properties can be identified from instance data. Thus, the OCLC processes do not make references to namespaces associated with FRBR models or ontologies.

- All four OCLC classes corresponding to the FRBR WEMI categories are assigned the RDF type `schema:CreativeWork`.
- The differences among the objects are interpreted as differences in specificity or detail. Thus, an Expression is more specific than a Work because it contains descriptors pertaining to the language of the content; a Manifestation is more specific than an Expression because it implies a physical object or presence; and an Item is more specific than a Manifestation because a unique object, not a class of identical objects.
- Figure 2 shows that properties associated with FRBR Manifestations and Items trigger the assignment of an additional class type from the ‘`schema:Product`’ ontology, indicating either a unique object (`schema:IndividualProduct`), or Item; or a set of identical manufactured objects associated with a product identifier (`schema:ProductModel`), or Manifestation.
- In the currently published RDF data accessible from VIAF and WorldCat Works, Expressions have the same type assignment as Works (`schema:CreativeWork`) and are not formally distinguishable, but this detail may change as more use cases are considered.
- The Schema.org property ‘`exampleOfWork`’, defined as “A creative work that this work is an example/instance/realization/derivation of”, associates a more specific Creative Work with a less specific one. As Figure 2 shows, `exampleOfWork` points upward in the FRBR-like hierarchy; the reciprocal property ‘`schema:workExample`’ points downward.
- If more detail can be discovered, the generic property is upgraded to ‘`schema:workTranslation`’ or ‘`schema:translationOfWork`.’ The ‘example’ properties were incorporated into the Schema.org vocabulary on the recommendation of the [W3C Schema Bib Extend Community Group](#). The ‘translation’ properties are maintained in the `bib.schema.org` hosted extension, another outcome of the W3C group.
- Except for the properties that are definitional, listed in Table 1 and shown on the right side of Figure 2, any property defined for `schema:CreativeWork` can appear with any description interpreted as OCLC’s empirically derived analog of a FRBR WEMI category.

The most important consequence of the model shown in Figure 2 is that relationships in the FRBR WEMI hierarchy can be expressed, but a hierarchy is not required in the OCLC model of Works. It is only one configuration that emerges when certain facts can be discovered. But others are possible. Since the ‘`workExample`’ definition encompasses all distinctions in the FRBR hierarchy—such as ‘`realizes`’ or ‘`embodies`’—it can be understood as a generic term that includes relationships that may even skip FRBR levels. For example, most WorldCat Catalog records are modeled as Manifestations and are clustered into Works; accordingly, the RDF statements accessible from WorldCat.org contain the statement `<Manifestation URI> schema:exampleOfWork <Work URI>`. But in its simplest form, Creative Work descriptions can be free-standing, as they might be for a unique item such as a handwritten letter, or a daguerreotype photograph. For such resources, it is not necessary to reconstruct the FRBR Expression, Manifestation, and Work “levels” at all.

At the other extreme, the OCLC interpretation of FRBR permits a description of the true complexity of the relationships among creative works and their translations. An example described in [Godby, Wang, and Mixter \(2015, Ch. 3\)](#) is reproduced in Figure 3 below. (In the current version of the OCLC model, properties defined in the namespace ‘bgn’, representing Bibliograph.net, have been absorbed into bib.schema.org.) The figure shows some of the variety of the representations for the fairy tale by Hans Christian Andersen translated as “The Snow Queen” in English. The peach-colored tabs represent Works, or clusters of bibliographic descriptions from which content-oriented descriptors have been extracted, using the methods summarized above and described in more detail in our publications. Each cluster is assigned the RDF type ‘schema:CreativeWork.’ The green tab represents a Manifestation, or a description of a single member of the cluster directly above it. It has the type assignment ‘schema:CreativeWork,’ but is more specific than the Work description because it contains publication information, which triggers the second type assignment ‘schema:ProductModel. The two type assignments capture the intuition that the edition of “Snow Queen” published in 1987 by North-South Books is both a distinct intellectual creation and a class of tradeable physical objects.

Figure 3 reveals other relationships among Works and Manifestations of the Danish fairy tale:

- The creative work published in Danish with the title “Snedronningen” has a German translation with the title “SchneeKoenigin”.
- The creative work written in English with the title “The Snow Queen” was published in 1987 by North-South Books.
- “Snow Queen” was translated from English to German by Anthea Bell.
- Properties proposed with input from the Schema Bib Extend Community Group identify Anthea Bell as a translator; “Snow Queen” as a translation—and, more generically, an ‘Example’ -- of the Work Schneekoenigin”.

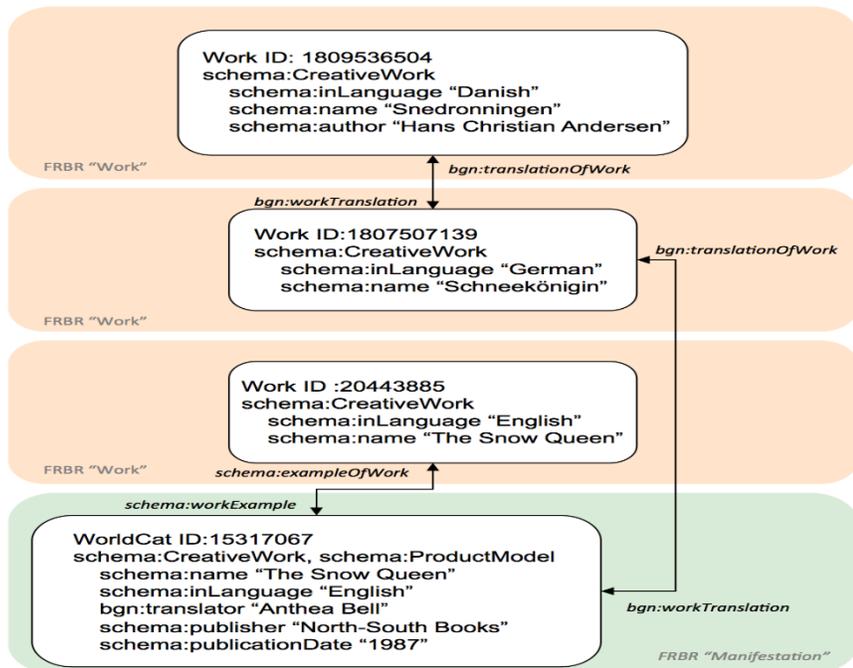


Figure 3. Works and their translations in WorldCat

Since this configuration is derived directly from the data, it is instructive to think about what could change if more information is discovered. First, multiple Work clusters may emerge for the same language pair, which could happen if translations were recognizably different because they were created by different translators. Second, the relationships between Work-Work or Work-Manifestation descriptions could be made more specific if the source

descriptions for translated Works contained more details about the sources and targets. Finally, the configuration could change if errors or inconsistencies were discovered.

At any rate, the configuration represented by Figure 3 shows what can be discovered in OCLC's current data stores. The model is designed and populated with the goal of revealing relationships that support browsing and discovery.

Discussion. If the OCLC model of creative works is as different from the FRBR Group I conceptual model as the remarks in the previous section imply, there are undoubtedly consequential differences in the working definitions of the primary categories. Thus, it is possible to infer the following:

Work. Works in OCLC's model are more concrete and real than the FRBR definition. It has a language and could have other properties. But it is still necessary to resolve the question of what level of abstraction is the most useful. Is a super-Work necessary, which would encompass the original and its translations? The OCLC algorithms can be tuned to produce this result, if use cases require it.

OCLC researchers are currently considering this question in the context of an RDF dataset derived from MARC bibliographic records, which will be used to represent creative works and their translations for users who want to view descriptions in their preferred language and obtain an appropriate copy from a library. Wikidata is one source of inspiration. In the Wikidata model, a translated book has a translator; the translation has a source and a target language; and the

translation is related to an original book, which is tangible. If the translation is recent, much may be known about the original—its language, physical format, ISBN, or publisher. If the original is ancient, however, few of these details may be available. The OCLC model based on Schema.org, which permits the representation of any discoverable detail at any level in the WEMI hierarchy, is consistent with these facts.

Expression. In OCLC's most recent publications, Work and Expression are given the same RDF Type assignment (schema:CreativeWork) and not formally distinguished from one another. But they are presented as separate categories in VIAF because the human interface has access to the business rules listed in Table 1. This inconsistency reflects a genuine uncertainty: are Expressions ontologically different from Works, or does it make more sense to treat Expressions as relationships between creative works that may be realized at any level in the WEMI hierarchy?

Manifestation. The OCLC business rules for discovering Manifestations implies a narrower scope than the corresponding FRBR definition. Not only does a Manifestation have physical characteristics, but it is mass-produced through a manufacturing process. Thus, it is more accurate to say that a Manifestation represents a set of identical manufactured objects, which are typically marked with a product identifier. As a result, a unique handmade object is typed as an Item, not a Manifestation. This narrower definition implies that no important use case is served by inducing a Manifestation description from a unique item.

Item: Conceptually, the Item is the simplest concept in the WEMI hierarchy: it is the thing on the library shelf. The Item defined in the OCLC model of creative works is identical to the FRBR item if it represents a member of a set of manufactured objects. But Items present three problems. First, what about digital objects? What property implies uniqueness—a URL? Second, an artifact of the OCLC FRBR algorithms is that over 50% of the Work clusters contain only one bibliographic record. But it would be a mistake to conclude that most of the single-item clusters are unique items because WorldCat is not a comprehensive inventory of the world's library collections, and the algorithm is tuned to produce false negatives instead of false positives. As a result, the odds are good that the singletons belong to an established Work cluster and could be placed more accurately with cleaner input data and more precise business rules. Finally, it turns out that assertions of Item-hood are strong and controversial because most MARC records describe Manifestations, and it is difficult to establish that something is truly unique. The Item in the OCLC model of creative works is still primarily a theoretical possibility that has not been deeply studied. But Ed O'Neill's classic studies of [last copies](#) interpret WorldCat singletons as Items in a framework that predates FRBR and Linked Data.

But regardless of how these issues are resolved, OCLC researchers have argued that models based on Schema.org describe FRBR with much more subtlety than many in the library community perhaps realize.