Framework for a Bibliographic Future

Draft for discussion, by Karen Coyle, Diane Hillmann, Jonathan Rochkind, Paul Weiss

Introduction

Metadata is a generic term for the data that we create about persons, places, things, documents, and anything else about which we wish to communicate or wish to operate on in an electronic environment. Although it is common to hear that "all data is metadata," it is certainly the case that not all metadata is well designed. Good design increases the potential success of a metadata standard.

The design components proposed in this model are not new. Similar components are used to some degree in standards such as the OpenURL Framework (Z39.88), the Semantic web, and the Dublin Core Metadata Initiative. A framework such as this serves many purposes. In particular, we are interested in producing metadata that is both highly extensible and that will promote compatibility between communities and applications that extend the metadata.

The four components that we propose are: a **model** of basic structures and relationships, a **schema** that defines an extensible set of properties, **guidance** for application of the properties, and **encoding**. The model can be used to create one or more schemas, and any schema can be expressed using one or more encodings. The guidance document is a key element that provides both direction to creators but also describes the semantics of the data elements in a human-understandable way. These four components provide a basis for creation of machine-manipulable metadata that has meaning to a community yet it can be defined in a rigorous way to communicate clearly to any users of the data.

Model

This is sometimes called an abstract model or a data model, although it does not define the data itself. Models are high-level views of the structures and relationships that the metadata will address. In the library community, the entity-relationship structure provided by FRBR is a type of model. It includes basic aspects of the information universe that will eventually be defined by metadata (works, expressions, manifestations, and items, plus the entities such as person and concept that will have a relationship with the primary four). We need to consider carefully how the FRBR model works in the context of other models that are used for bibliographic data such as DCAM and OpenURL Framework. The model that results will be independent of any particular schemas or encodings of bibliographic metadata, but will provide a structure that all implementations of metadata derived from the model will have in common.

Schema

Metadata schemas (sometimes called 'element sets,' 'metadata formats' or 'data dictionaries') define the actual properties that will carry values in the data set, as well as the relationships between those properties. Data elements can be defined at any relevant level of granularity. They can have hierarchical relationships between them or non-hierarchical relationships. The Dublin Core Element Set is an example of a set of data elements. FRBR defines data elements in its attributes, but they must be restructured in a way that allows the development of different levels of granularity and that promotes extensibility of the schema, both over time and across communities. Ideally, the schema would be expressed in one or more machine-readable formats that facilitated its use by both people and computer applications.

Guidance

Guidance is often desired to aid in the creation or assignment of values to data elements in a consistent

way. Guidance may be general or specific, but it usually attempts to address circumstances that users will encounter in the creation of the metadata. Different communities making use of the same data elements may define their own specific best practices that attempt to produce the metadata that is most useful for their purposes, but in general they may not re-define the elements in order to address those needs. The library community has traditionally received its guidance from cataloging rules (such as AACR) and from practices published as part of the encoding of library data using MARC21. Increasingly, specialized guidance for specific communities has been developed that reflects the differences in materials or approach inherent in their tasks: examples are *Cataloging Cultural Objects (CCO)* for the museum community and *Describing Archives: A Content Standard (DACS)* from the archival community.

Encoding

We can assume that any metadata being created today will be expressed and exchanged via a machine-readable encoding. The primary requirement for metadata encoding is that it must be able to encode the full detail of the semantics and relationships intended by the metadata creators; and it must expand as the metadata schema grows and changes. The same metadata can be encoded in different data formats and still be fully shareable, as long as the encoding is true to the data elements and to the overall structure of the metadata model.

Discussion

FRBR

FRBR's entity-relationship model (as defined in Chapters 3-5 of the FRBR Report) is a useful, if not complete or even wholly accurate, analysis of our bibliographic universe. The delineation of the four group 1 entities illuminates an important issue of our legacy: we have been cramming metadata about different bibliographic entities into single descriptions. As just one example, the FRBR report provides an explanation for the ambiguity of dates in bibliographic records: there are at least four dates of creation that apply to each bibliographic resource--those of its work, expression, manifestation, and item. For many resources all these are the same, so there is no need to delve further, but some resources are more complex, and that complexity has led to confusion about dates used in brief displays and search limits.

FRBR, and work by Barbara Tillett, Richard Smiraglia, and others has contributed to an increasingly formalized notion of relationships among bibliographic resources, and between bibliographic resources and associated entities (for instance, FRBR's group 2 entities--persons, corporate bodies--and draft FRAR's families, as well as subject entities). Examining current practices from the perspective of this work on relationships shows great inadequacies in the identification, recording, and utility of relationships.

FRBR does an admirable job of providing one way to analyze the bibliographic universe, though as has been noted by others, it doesn't extend well to museum or archival collections. Although FRBR covers attributes of bibliographic entities, it does not model the metadata itself (that is, none of the entities represents metadata per se).

DCAM

The Dublin Core Abstract Model from the Dublin Core Metadata Initiative (DCMI), on the other hand, takes the next logical step, and models metadata. Its purpose is to "to gain a better understanding of the kinds of descriptions that we are trying to encode and facilitates the development of better mappings and translations between different syntaxes."

The FRBR model and the Dublin Core Abstract Model are not contradictory; in fact, they are complementary. FRBR provides a start at defining properties for RDA and allows the description of

resources using specific relationships that can be assigned at the proper level as well as aggregated for better expression to the user. The DCAM helps us to envision the FRBR entities as a package, allowing the discussion about issues like identity and linking to be posed and discussed in a more useful manner.

Metadata Schemas

Even as we validate the use of FRBR as a model, we take issue with its embedded attributes. One of the things the DCAM and the Dublin Core experience generally tells us is that we need to develop our attributes/properties/elements separately from the model as well as from the values used. Separating elements and their definitions from guidance on determining their values (controlled vocabularies, transcription, etc.) is crucial in order to achieve interoperability and extensibility.

As a first step, the FRBR attributes must be carefully generalized. For example, instead of defining separate elements (including their names, definitions, examples, etc.) for title of the work, title of the expression, and title of the manifestation, there should be one title element reused at multiple levels. The declaration of these elements should include clear specification of where in the FRBR Group I they may be used. This increased generalization promotes interoperability, minimizes a tendency toward complexity, and eases machine manipulation and extensibility. It also requires more rigorous consideration of when attributes at the various levels are really the same thing or not, and can point out inconsistencies that can be rectified. Along with the development of the generalized elements, there should be rules for extension or refinement of those elements, to ensure that appropriate extensions can be made and managed.

Crucial to the proper development of a metadata schema is a clear notion of requirements for technical expression of the attributes, and a plan for maintenance and growth. We have learned much in the library community about the importance of community consensus and how to maintain important standards over time. MARBI is a good example of doing it correctly, and in fact the Dublin Core Usage Board process is based loosely on MARBI.

Guidance for Application

It is critically important that we develop good usage guidance based first on the Metadata Schema attributes in their most generalized form. We must provide this usage guidance in a manner that allows communities of practice to use the general guidance as they extend the basic structure for their own purposes. Traditional library cataloging is just such a community of practice, and should extend the schema and guidance to fit their needs, without the necessity of bringing their special library colleagues along with them. If the general elements, and the guidance attached specifically to them, can be approached as a extensible set, other communities will be encouraged to incorporate them specifically in their metadata and to extend in ways that provide a sound basis for interoperable use and re-use. In this scenario, mapping between library metadata schemas and others, as well as the mix/match capabilities of application profiles, can be made easier. This approach will tend to minimize data loss when information is crosswalked, and improve the ability of machines to act upon the data regardless of its origin.

As part of this development of extended guidance material along specialist lines, we need to recognize that different communities will apply FRBR Group I boundaries differently. Much of the discussion about how decisions will be made about works, expressions and manifestations indicates clearly that specialized communities will tend to make different decisions about where these boundaries lie. This has been seen as a problem, and an impediment to the integration of FRBR principles into actual practice. Part of the rationale for separating traditional library specific instruction from the general RDA, and enabling specific communities to extend from that general base, is that the assumptions and instructions for where these boundaries lie can be made explicit *by community*, and librarians can get out of the trap of trying to herd everyone else into the same decisions. This will make it easier for the communities as a whole to use each other's work--when differences are not exceptions but can be explicitly expressed as policies and appropriately supported with more detailed extensions to the general framework, everyone enjoys easier and more cost effective machine manipulation of data. So

long as the determination of what is a work can be ascribed to the community that made the decision, other communities can predict and cope with the variations.

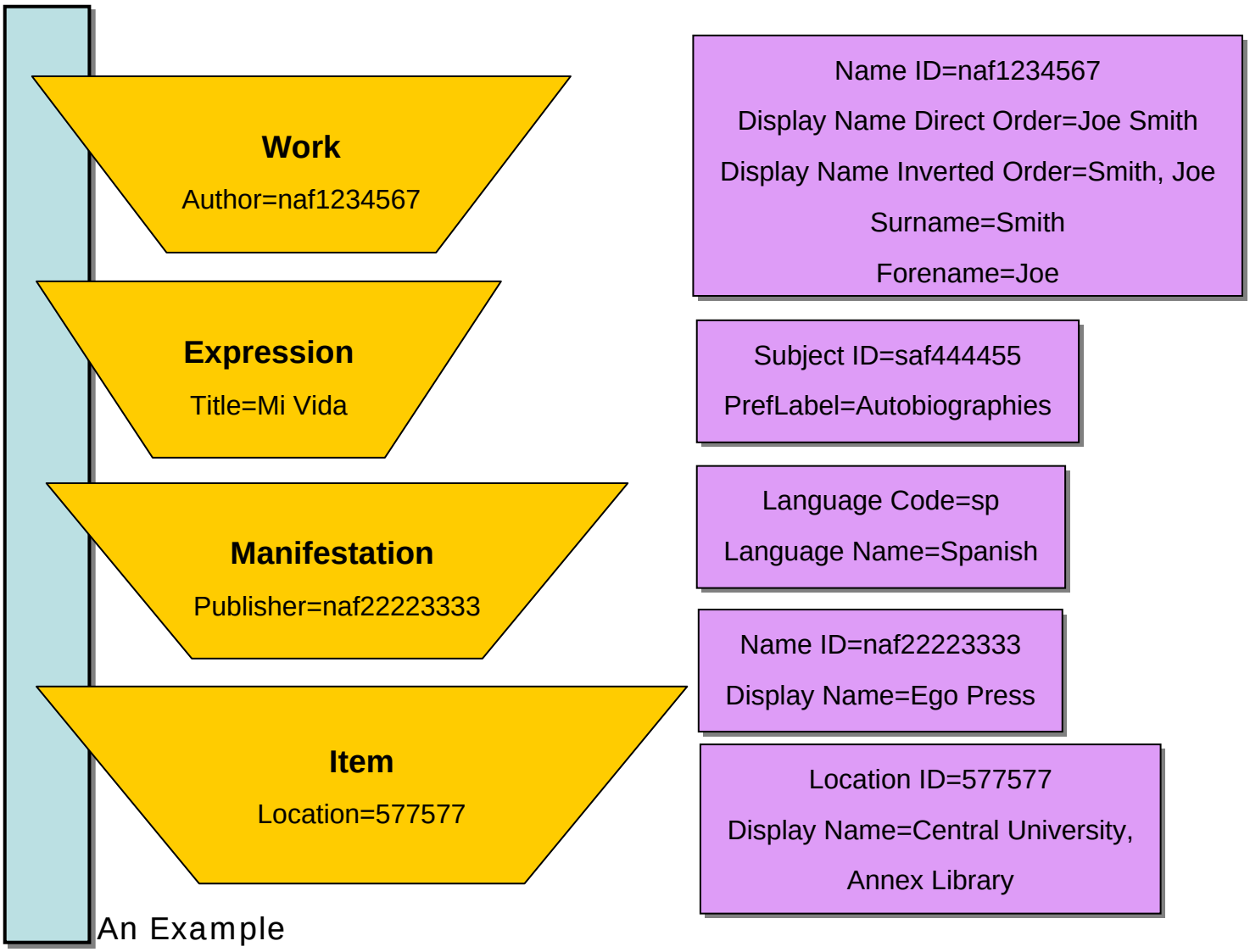
Encoding

It seems unlikely that MARC21 can be sufficiently remodeled to serve as an encoding for a modern metadata schema, but certainly some of the accumulated wisdom and experience embedded in the MARC21 documentation can be repurposed. One issue is that insofar as it supplies definitions, labels and relationships not necessarily explicit in AACR2, MARC21 itself represents a combination of functions that requires significant attention, and perhaps deconstruction, to prise out what should be included in the metadata schema and what remain as encoding.

It should also be recognized that MARC21 encodes more than bibliographic information, and the formats for classification, authorities and holdings might well be more appropriate for future use, given that they operate where competing data structures are sparse. Where they tend to be problematic is in the area of distinctions at the statement level, where specification of language of statement, source, and community of origin may well be necessary.

Encoding for the future must support statement level identification and attribution. Although to a certain extent, this is a 'packaging issue,' it seems important to assert it as a guiding principle, as it supports the notion that the way records will be built in future will be much more iterative, and catalogers are just as likely to start with a re-used description than one created newly for purpose. These catalog records of the future are likely to be aggregations of the work of many catalogers--somewhat like CONSER records are now--and the source and age of particular statements will be critical as we develop applications to make 'decisions' about what statements they will display. Central to this assumption is that, in the shared environment of the future, information may be added, but not subtracted--just ignored if not needed or desired in a particular context.

An Example:



An Example

The figure above illustrates some possibilities for a description set based on DCAM that also includes some of the FRBR entities and shows how they would relate. On the left side are the four Group I entities, with a small assortment of generic properties. In the cases where the value of the properties is contained in another description, the relationship between them is conveyed with an identifier, and the identified Group 2 or 3 description is included in full with the description set. Thus, an application using this description set could presumably pick and chose among the available display values, for the one that suits its goals best. For instance, in the description of the author, there are two identified possibilities for display text for that particular person, one using direct order, and the other surname first.

Note that the linking techniques are the same regardless of what kind of description, whether author, publisher or subject is related to a particular Group 1 entity. There is both a title in the Work description and another in the Expression--the differences between them and their different functions are conveyed not in the property name, but in where it appears, allowing an application to determine how to display either or both. Grouping of expressions and manifestations can be supported using simple linking and

naming strategies, without unnecessary complexity.

Using only the descriptions in this simple example, the following display could be supported:

Author: Joe Smith
Title: Mi Vida
Language: Spanish
Publication: Ego Press, 2005
Extent: 267 pages
Subject: Autobiographies
Location: Central University, Annex Library
Call Number: CT25.S65 2005
Other Language Versions Available: English

Note that the link to an English version is implied by the presence of another description set (not illustrated here) with the same work description and an expression description in English.