R&D: RDA in RDF
OR:
Can Resource Description become Rigorous Data?

Karen Coyle
Code4Lib 2008

Annotated version, March, 2008
This is AACR2*. AACR2 is over 600 pages long, and it is the set of cataloging rules that librarians use when cataloging a book... or a film, or a piece of music, or anything else that will be cataloged.

The cataloging rules are long and complex because the things they catalog are more complex than most of us can imagine. For example, these rules include the proper way to transcribe the names of Thai royalty; and when you have a book written by a spirit speaking through a medium, which one of those is the author? (Note, the answer to that last one changed between AACR1 and AACR2.)

*Anglo-American Cataloging Rules, 2nd Edition
3.5B2. If there is more than one map, plan, etc., on a sheet, specify the number of maps, etc.

6 maps on 1 sheet

AACR2 consists of instructions and examples. Here is a short instruction and the example of what the resulting bit of cataloging might look like.

Note that the cataloging is expressed as text. AACR2 was first issued in 1978 when card catalogs were still the predominant form of catalog. A catalog card, of course, is a text document.
<table>
<thead>
<tr>
<th>025.32 Gorman</th>
<th>Gorman, Michael</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>168 p. ; 23 cm.</td>
</tr>
<tr>
<td></td>
<td>Includes index</td>
</tr>
<tr>
<td></td>
<td>ISBN 0-8389-3494-3</td>
</tr>
</tbody>
</table>


This is a sample AACR2 catalog entry for AACR2.

It is a highly structured, rules-based text, but it is still a text.

(Examples of other structured texts are limericks and the address written on an envelope.)
You can take that structured text and mark it up using MARC21, put it in a database and display it on a screen. It is still primarily a group of text strings that are intended to be read by human beings. In fact, I tend to think of the MARC21 format as a mark-up language, defining the structure of text, but far from a data processing record.
Recognizing that these are different times, that all catalog data will be processed by computers; that there is a whole World Wide Web that is the primary information space for our users, the Joint Steering Committee on AACR began work in 2004 (or so) on the next generation catalog rules.
RDA = Resource Description and Access

It was soon determined that the new cataloging rules would be so vastly different from the ones that came before, that they no longer fit into the AACR tradition. The rules were renamed to RDA.

The principle participants, however, still reflect an Anglo-American nature.

A project of the Joint Steering Committee for RDA

- Library of Congress
- British Library
- Collections and Archives Canada
- Australian National Library

This is the RDA logo. Does anyone else see a tetris game in here?
RDA has some very interesting and laudable goals. To begin with, RDA uses (or attempts to use) the Functional Requirements for Bibliographic Data as its framework. I say “attempts to use” because the is the first implementation of FRBR as cataloging rules, and in a sense is a test of FRBR as a model.
Another goal of RDA was to simplify the cataloging rules. Simplification and modularization will encourage communities to use RDA who before might have found it to be too rigid or too library-centric.
The main “other communities” near the library space are museums and archives. This need to bring them together in a single data format is especially acute for institutions that have multiple roles: museums that also have libraries, libraries that include an archive, etc. They need to create compatible data, but the different functions often have very different metadata needs at some level.
Another goal of RDA, as compared to AACR1 & 2, is to be less prescriptive about display. In fact, the AACRs are very much about *presentation* of information, not its storage or manipulation. This is a symptom of the fact that those cataloging rules were directions for the creation of text displays.
RDA also has a goal of being mindful of the fact that today’s bibliographic data will exist in a computerized, networked world. All of the elements created by catalogers will be processed by computers; some will be presented to humans to be read.
Here are two pages from an RDA draft.

RDA consists of:

- 10 sections
- with 37 chapters
- and 13 appendices

To date we have seen only about ½ of the RDA text (and it is today very much a text). Counting up the available drafts, I found over 800 pages of text, not including any of the appendices.

There are numerous problems evidenced here. One is that such a lengthy text is unlikely to be a simplification of the previous rules, and there have been many complaints from the community that RDA is much, much too complex.
Another problem is that RDA continues to instruct its adherents to create text strings. Although in many cases the actual instruction has changed (and catalogers consider some of these changes to be significant), the resulting output is still: “6 maps on 1 sheet”

A string like “6 maps on 1 sheet” is fine for a person, but if you want to do machine processing on the data, a text string just doesn’t cut it.

I should mention here that RDA is not being designed as a print document. It will be an online service, with the sections and references hyperlinked. This is the explanation for some of the repetition of instructions and definitions throughout the text. The draft itself is an MS Word document with a combination of formatting and print clues to the text’s structure, such as bold fonts and textual numbering. Moving from this text to an online service is a serious challenge.
On the previous page we saw the example “6 maps on 1 sheet.” It may not be clear from the example, but some terms in that statement are based on controlled vocabularies.

There are about 55 separate controlled vocabularies embedded in RDA. This means that the lists are part of the text of the RDA document, which makes it difficult to provide support for these terms in any systems based on the cataloging rules. Each system must create and keep its own version of the list, and updates to the vocabulary lists must be done redundantly in hundreds or thousands of systems.
Similar lists in the MARC21 format have the same problem. Updates are announced in email, and system developers must manually update their versions. For lists that are embedded in the MARC21 standard documentation, it takes a standards update request, often a 2-year process, to get a new value approved for a list, and more time before the new value is added to systems.

By the way, RDA and the MARC21 standard have some lists in common, but more that they do not share.
Not one to suffer in silence, I teamed up with my “partner in crime,” Diane Hillmann, to write an article for D-Lib Magazine about RDA. Our subtitle appears to have been too subtle: “Cataloging Rules for the 20th Century” was actually a scathing comment on the deficiencies of RDA as a modern cataloging code.

In this article, we proposed that a modern cataloging code must not be text-based but must be based on a solid framework that can be supported in a computing environment. Since then, I have become even more interested in how we might make use of the principles of the semantic web to make library data more visible and useful in the networked world we live in.
Essentially, to have functional, coherent, machine-actionable metadata, you need, first, to have a foundation based in IT standards. Above that, you have to clearly define your domain (eg. “metadata for library and archival resources”). Your domain needs clear functional requirements (FRBR has at least some of this). You also need to have formally defined vocabularies (we’ll see more of that in a bit). In most cases, you also want to be able to make use of application profiles. These allow you to define different variations of your data for different users in your domain. So general libraries and specialized libraries could have many elements in common, but each could also extend or reduce the vocabulary set that they use based on their needs.

This multi-layered model is similar to the one that I have in mind, although mine is less detailed, and when I tried to draw it, it was embarrassingly silly looking. So I’ll use this Dublin Core design to illustrate the issues I wish to convey.
RDA is the top box on this diagram, the “usage guidelines.” It provides instructions on how to assign the values used by a community. FRBR may cover the functional requirements and the domain model, in yellow. The “Metadata Vocabularies” in the orange box are implicit in the RDA document, but are not formally defined in the sense intended here. Most of the needed structure, however, is missing from our metadata standard.
In a meeting in London on April 30, 2007 something extraordinary occurred. Representatives of JSC and the Dublin Core Metadata Initiative (some of whom also very active in the W3C semantic web activity) agreed that RDA must have a formal declaration of its vocabulary.

A group was formed to work on this project. Everyone agreed that it was essential to the success of RDA as a modern, web-based metadata standard.

Unfortunately, no one was in a position to fund it, but work is going forward none-the-less.
The essential goals of the project are to create a machine-actionable registry of RDA vocabulary terms that is well-defined and open. Anyone wishing to make use of the terms would be able to do so.
At about this same time I had one of those “ta-da” experiences. It was 2007, around tax time in the US (March or April). I had been reading RDA and had noticed the embedded vocabularies, in particular the “carrier” vocabulary that I showed earlier. This is a list of all of the physical formats for resources, including all of the computer formats.

At the office supply store I ran into something new: tax software being sold on a thumb drive. I looked at the RDA list of carriers, and thumb drives were not there.

RDA doesn’t provide a way to update its value vocabulary lists. I made the suggestion that these terms should be outside of the text of the rules, and should be managed in an online registry.

Diane Hillmann, who had been working on the registry of vocabularies in the NSDL Registry project, took up the challenge and created a sample registry entry for the RDA carrier vocabulary.

http://sandbox.metadataregistry.org/vocabulary/show/id/44.html
Diane entered all of the RDA Carriers into the registry, and coded the relationships between them (broader term, narrower term, etc.)

A key aspect of the registry concept is that each element has a Uniform Resource Identifier (URI). This means that wherever the identified term is used it always has the same semantic value. This is essential to promote interoperability, and it also supports internationalization, since the linguistic value of the term can change as long as the URI remains the same.
Then Diane created an entry for the “new” carrier, the “USB flash drive,” based loosely on my blog post. Initially set as “provisional,” this record showed how a vocabulary list could be extensible, and could also be available to everyone in the interested community online and at the same time. The entry includes information that we do not have today in the RDA nor MARC21 vocabularies, such as definitions, alternate terms, and hierarchical relationships. (Some of this information is available elsewhere, but not by any means at every point where you need to think about the term.)
When accessed by a program (as opposed to a human being), the registry provides a machine-readable record (in this instance in XML, but other formats could be generated). This means that library systems, and other bibliographic systems, throughout the Web can retrieve this data whenever it is needed. Catalogers and users can be shown definitions or related terms, and program requests can get a response that is directly usable as code.
Work on the RDA vocabularies is based on a JSC document that lists the RDA “elements.” Although we have the list of elements there are interesting issues that come up in trying to format these as a formal vocabulary.

RDA Element Analysis:
http://www.collectionscanada.gc.ca/jsc/docs/5rda-elementanalysisrev.pdf
To begin with, many elements serve more than one function in the bibliographic description, and most of these functions are implicit, not explicit. This has always been the case with library data, and it is definitely the case with data that we have coded in MARC format.

For the creation of the vocabularies we have to ask: how many of these functions need their own element?
It’s natural for those of us in the data processing world to look at the publication statement with place, publisher, and date and visualize “publisher” as an entity in itself, perhaps in a separate record that contains the publisher’s address and links to all of the books that it has produced. But in RDA, this isn’t the “publisher” it’s the string representing the publisher’s name that appears on the title page. If the publisher listed there is wrong or fictitious, that’s what goes into the description. Elsewhere, a note might say that the real publisher is “X.”

We have to look very carefully at how the data elements are defined in RDA.
Author names are almost the opposite of the publisher. The publisher name is part of the “surrogate” for the title page. The author name - a horridly ugly thing that we impose on our users - is constructed by catalogers. The interesting question here is: what does this represent? Although the FRBR entity is “person,” the entry in the description isn’t for a person but for a personal name. A real world person could be represented by more than one name (eg. both Mark Twain and Samuel Clemens are each “preferred names” in their own right); a personal name could represent two or more real world persons who write together under a single pseudonym. Yet the entry has birth and death dates that are information about a real world person.

This one is a real puzzle.
Issues/Problems/Puzzles

  
  RDA includes meta-metadata; that is, data about the description and the cataloging decisions. This is important data, but it doesn’t belong in the vocabulary that will define the resource.

● “Publisher statement” = place+publisher+date
  
  Some elements are simply combinations of other elements: the publisher statement exists only as a combination of place, publisher, and date. Should the statement itself be in the vocabulary, or just the individual elements?

  
  Some elements are just text strings, even though they may contain more than one data value (the “6 maps on 1 sheet” is another example of this). Do these need to be divided into separate elements?
I have a particularly hard time with the element called “statement of responsibility.” This is traditionally a key element in library cataloging, but I can’t figure out what to do with it in a vocabulary declaration. It doesn’t stand alone: “by J.R.R...” is intended as a continuation of the title. In this case, it looks like the title is doing at least double duty, which may be the problem. Taken together, the title + statement of responsibility are a surrogate for the title page. But the title here is also the title entry that will be used, without the statement of responsibility, for retrieval and sorting.
Bartholomew world travel series, v. 5

Relationships, like that of a book to a series, are not expressed relationally in RDA but are text strings within a description of a resource, like this series statement. Text strings are “links” only in that humans can read them and search for the related bibliographic item. They aren’t viable links for machine processing of data.

Original title: L’éducation sentimentale

Even the key relationships like Expression to Work aren’t always made clear. While some catalogs may have a “Work title” record in their authority file, it is also acceptable to indicate these relationships with a note. This isn’t a question of right or wrong, but it tells me that we can’t expect to impose a strict entity/relationship model on library bibliographic data.
There is the question of what it means to say that RDA follows FRBR. In general, the FRBR “attributes” are not reflected in RDA. It also isn’t clear how the lessons learned in creating RDA (probably the first real “test” of FRBR) will be reflected back in the FRBR model.

As an example, RDA includes “Families” as agents. This comes from FRAD, but not from FRBR. Is this significant? Is it important to keep these models synchronized? If not, what is their purpose as models?
RDA doesn’t include subject analysis, although it has place-holders in the element table for the FRBR Group 3 entities. This is a very important aspect of library catalog records; how will this gap be filled in, and by whom?
There’s always a lot of discussion about the FRBR Group 1 entities, in particular about the boundaries between them. RDA has recently changed how it presents these. Group 1 has been divided into Content (Work and Expression) and Carrier (Manifestation and Item). There is no bright line between the two entities in each group. This isn’t necessarily a problem, but it does bring up the question of whether the RDA elements can be definitively assigned to one Group 1 entity. If an element can be used either for a Work or an Expression, is that one vocabulary element or two different ones?
FUQ*

- Is FRBR the right model?
  - I don’t know, but RDA could be the real test of FRBR concepts.
- Is RDF the right format?
- Does this replace MARC21?
- Who will be in charge? How will things be decided?
- WHAT WERE YOU THINKING?!
- How can I help?

*Frequently Unanswered Questions*
FUQ*

- Is FRBR the right model?
- Is RDF the right format?
  - We’re working with RDF because 1) it expresses entities and relationships 2) it is being actively worked on by W3C groups so there is support for it 3) people on the project are familiar with RDF. Our data will not be limited to RDF expression. Got a favorite format? Come chat with us.

- Does this replace MARC21?
- Who will be in charge? How will things be decided?
- WHAT WERE YOU THINKING?!
- How can I help?

*Frequently Unanswered Questions*
FUQ*

- Is FRBR the right model?
- Is RDF the right format?
- Does this replace MARC21?
  - No. What we’re working on is not a record format. However, the vocabularies should make it possible to develop a new library data carrier, or multiple carriers, fairly easily.
- Who will be in charge? How will things be decided?
- WHAT WERE YOU THINKING?!
- How can I help?

*Frequently Unanswered Questions*
FUQ*

- Is FRBR the right model?
- Is RDF the right format?
- Does this replace MARC21?
- Who will be in charge? How will things be decided?
  - This is particularly difficult. We submitted a grant proposal to NSF to develop a registry maintenance methodology, including creating the community structures that would be needed. Their reply was: this will change everything in the library world, but it’s not technically interesting. (Most technology is less interesting than its social implications.)
  
- WHAT WERE YOU THINKING?!
- How can I help?

*Frequently Unanswered Questions*
FUQ*

- Is FRBR the right model?
- Is RDF the right format?
- Does this replace MARC21?
- Who will be in charge? How will things be decided?

WHAT WERE YOU THINKING?!

  - I often wonder about that myself. This is a huge task, yet one that some of us think is extremely important. And that leads us to the next question:

- How can I help?

*Frequently Unanswered Questions*
DCMI/RDA Task Group Wiki

This Task Group is for collaborative work to enable broader use of the Resource Description and Access (RDA), building on agreements made at a meeting held at the British Library April 30/May 1, 2007. Participants in the meeting came from DCMI and other Semantic Web groups, and the RDA development effort. The Task Group is led by Diane Hillmann of Cornell University and Gordon Dunsire of Strathclyde University.

The planning for the London meeting was originally located on the DC-Libraries Wiki. Some information gathered for the meeting, including documents, agendas, a meeting packet, and other information of historical interest remains at that location.

Wiki Pages

- Use cases Use Cases
- Analysis task 1 analysisTask1
- Analysis task 2
  - List of in-line vocabularies from RDA PDAVocab
  - List of in-line vocabularies from MARC (for information) MARCVocab

Charter and Work Plan

Follow the work on the wiki ...
http://jiscmail.ac.uk/archives/dc-rda.html

And join the discussion group to participate in the project.
Thank You

kcoyle@kcoyle.net