Ripples and Triples in the Bibliographic Data Pool: RDA and Linked Data

F. Tim Knight

Osgoode Hall Law School of York University, tknight@osgoode.yorku.ca

Follow this and additional works at: http://digitalcommons.osgoode.yorku.ca/librarians

Part of the Cataloging and Metadata Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Repository Citation


http://digitalcommons.osgoode.yorku.ca/librarians/21

This Conference Proceeding is brought to you for free and open access by the Law Library at Osgoode Digital Commons. It has been accepted for inclusion in Librarian Publications & Presentations by an authorized administrator of Osgoode Digital Commons.
Good afternoon everyone …

Today I’m going to talk about ripples and triples in the bibliographic data pool. RDA, for those not familiar with that acronym, stands for Resource Description and Access; and linked data? … well, we’ll get into that a little later on.
I’ll start with RDA and talk about why a change in catalogue rules was necessary. Then I’ll look at some of the practical differences you’ll find between RDA and the older rules AACR2. This will lay the groundwork for a quick look at what linked data is and how it works in the Semantic Web. And, finally a short conclusion suggesting how librarians can help play a role in this transition.

And all in 20 minutes!
We know that RDA replaces AACR2, the Anglo-American Cataloguing Rules, but the reason why this became necessary may not be as well known.

Essentially the bibliographic environment that AACR had been originally designed to work in overloaded the capacity for AACR2 to describe it.

You can think of it this way …
Descriptive cataloguing used to be a relatively simple process that started out to provide a consistent methodology to describe published books and other physical print resources.

However, this basic, physically oriented bibliographic environment gradually evolved into something more like this …
… adding non-print formats like microfiche, continuing resources, cassette tapes, compact discs, computer files, CDs, DVDs, etc. into the mix … and then that environment rapidly transformed into the information space we all live and work in today …
… with both physical resources and their digital equivalents coexisting alongside some entirely new types of online information resources now just a click or two away.
AACR2 emphasized the “physical item.” This meant there were many separate chapters outlining the characteristics for each format cataloguers needed to describe: for example, there was a chapter for books, one for sound recordings, and a chapter for electronic resources, to name three, each with a slightly different approach for describing that particular type of resource.

However, creating descriptions of resources in this so-called “digital world” required a different process. Continuing to describe digital resources as “book-like things” was proving to be problematic and ultimately unsustainable. It complicated the cataloguing workflow and frustrated cataloguers’ efforts to provide meaningful and consistent access to information resources.
Think about it. When we're dealing with a digital environment, *everything*, no matter *which* format, can be reduced to a collection of ones and zeros. It doesn't matter anymore if the resource is a book, a journal, a presentation, a blog, or an audio recording, ...

If it's important to our library users then we need to describe it and make it findable and accessible.

So, in 2005, rather than attempt to fashion something along the lines of a new “AACR3,” AACR was deconstructed and reorganized to create the first draft of a new cataloguing standard, called Resource Description & Access or RDA.
RDA set out to accomplish the following:
• Address those problems that had been identified in AACR2
• And in the process, simplify the rules
• Encourage its use as a "content standard" useful for metadata communities inside & outside libraries
• Encourage application on an international level
• Create a principle-based approach to resource description that builds on cataloguer's judgement
RDA is a “content standard.” Instead of formats it focuses on the resource itself. It provides a set of instructions to describe the intellectual or artistic content embodied within a resource. Formats can still be an important aspect of a resource but it’s no longer considered a primary feature in RDA descriptions.
In the opening remarks of RDA the purpose is clearly stated: RDA aims to provide, “a set of guidelines and instructions on formulating data to support resource discovery.”

This demonstrates a fundamental shift away from the earlier rules where the intention was to build “library catalogues and lists.” RDA moves us toward a general set of guidelines that focus on identifying the relevant pieces of “data” that supports and enables “resource discovery.”

In the process RDA also aligns itself with the library user, an aspect of cataloguing that was never explicitly mentioned in AACR2. RDA’s content centred approach attempts to consider how users operate in this digital environment and, by emphasizing the intellectual or artistic content, makes these instructions applicable to describing any resource.

In Chris Oliver’s book *Introducing RDA*, she reminds us that the central purpose of recording data is to support the so-called “user tasks.”
By “user tasks” she is referring to the four user tasks addressed in RDA that were derived from the Functional Requirements for Bibliographic Records. There has certainly been some criticism of these four user tasks, but when I think about my own research process I think they are representative.

I would equate **find** to entering some terms into a search engine or a library catalogue to see what I get back.

I would then look at the list of results and **identify** those that might look useful to me. If the results are less useful than I expected I might return to **find** and select some different terms.

If the results looked promising I would **select** what relevant resources based on my information needs; maybe picking out electronic journals, ebooks or digital resources that could provide me with immediate access to information.

And then I would **obtain** the resources by clicking the links provided or jotting down the call numbers to retrieve physical resources.

OK, so that’s some background on where RDA came from. Let’s take a look at some of the basic changes you might find in resource descriptions based on RDA.
One of the fundamental approaches associated with RDA is the transcription of information as it appears on the resource. So abbreviations, for example, are only recorded if they are actually present on the resource.

This “take what you see” approach also facilitates the machine processing of information, where, for example, an application might analyse a resource and pull metadata from that resource directly into a discovery engine. This is what’s referred to as “machine actionable data” and is an important aspect of bibliographic control and linked data in the future.
So continuing with the abbreviation example. Here’s a rather cryptic Latin abbreviation used in AACR2 that baffled many of our users. “S.l.” stands for the Latin terms *sine loco* (meaning without place) and “s.n.” for *sine nomine* (without name) which here refers to the lack of that information available for the publisher.

In RDA this lack of information is conveyed in a much more direct way stating, in the language of the catalogue, that neither the place of publication or the publisher has been identified.
Generally speaking, the use of abbreviations like these were a product of the card catalogue era where available space on typed cardboard catalogue cards was at a premium. This is no longer an issue in a digital environment.

Things that we as library professionals have long taken for granted, like the “p.”, “ill.” and “col.” in this example, will now be spelled out in full in RDA which improves the interpretation of the data we provide to our catalogue users.
Here’s another example showing the edition statement now recorded as it appears on the resource.
RDA & AACR2: Transcription

**INACCURACIES**

**AACR2**
- Inaccuracies corrected within the transcribed element using [sic] or [i.e. ...]

**RDA**
- Inaccuracies recorded as they appear on the source; variant access point or note are included if considered important for access

The downside of this “take what you see” philosophy means that any errors found on the resource will also be transcribed as they appear.

But indicating errors using the Latin term “sic” or providing corrected alternatives using the Latin abbreviation “i.e.” will no longer be used in RDA.

If clarification is needed a “variant access point” or a descriptive note can be provided if that information is considered important to access the resource.
In this AACR2 example either “sic” was used to indicate that the typo found in the title was on the title page and not an error made by the cataloguer; or, depending on local policies, the corrected term was shown following the error using “i.e.”

RDA instructs cataloguers to transcribe errors as they appear on the source and to add a variant access point to improve access. In this example it would be necessary to add a “variant access point” to provide access to the correct form “psychoanalysis” in the title.

This is not unusual though, providing what we used to refer to as an “added title entry,” has always been standard cataloguing practice.
The AACR2 “rule of three” has also been eliminated. That rule instructed cataloguers to “omit all but the first of each group” of three or more persons or corporate bodies listed on the resource followed by ... the “mark of omission” and [et al.].

This example has four editors listed. In AACR2 only the first would be named followed by the standard “mark of omission” and then et al. in square brackets.

By default, again following the “take what you see” approach, RDA lists all of the editors and also provides an access point for each editor.

However, there is an optional omission that operates like the old rule of three, again listing the first named followed by a statement indicating how many others were responsible. Here there are three other editors indicated.
Incidentally, while we're looking at the statement of responsibility, the information provided here no longer needs to be abridged as it had been in AACR2. It can now be transcribed as it appears on the source.

This offers the advantage of providing additional keywords that might be relevant for resource discovery and identification. Here, for example, you could search for Osgoode Hall Law School which may not otherwise appear in the description.
Relator codes, indicating the function or role of a person or corporate entity, are now also routinely added to access points in RDA.

These examples show relator codes for “editor” and “author.” These relator codes are something that will be very useful in a linked data context because they provide the role associated with that entity. So for example, this says that Cedric Ryngaert is the editor of “Judicial decisions on the law of international organizations.”
Another relatively major change is the discontinuation of the AACR2 GMD or the General Material Designator. Terminology, like “videorecording” seen in this example, used to appear following the statement of the title proper.

This information is now conveyed in RDA using three new data elements populated with a controlled vocabulary. The new data elements are: Content Type; Media Type; and Carrier Type.

Examples of the controlled vocabulary for this DVD of the movie “Time bandits” are: “two-dimensional moving image” for the Content Type; the Media Type here is “video”; and the video is conveyed on the Carrier Type “videodisc.”
Unlike AACR2 these new data elements are now applied to all resources. For printed books, as in this example, the GMD 'text' was implied in AACR2 and so never added.

In RDA all three of the data elements are used. The Content Type is 'text'; the Media Type is 'unmediated,' which means that no additional device is needed to access the content. You just pick up a print book and read it. And the Carrier Type for books is 'volume' because it’s a volume of text.
This method provides an improved description for electronic books. Instead of using the rather vague GMD ‘electronic resource’, RDA uses the Content Type ‘text’ because an electronic book is still a book. But we can’t just pick up an ebook and read it, we need a computer to view the text, so the Media Type for an ebook is changed from ‘unmediated’ to ‘computer.’ And the Carrier Type for an electronic book is an ‘online resource’ meaning the resource is delivered over an online network.
At this stage in development RDA has really only provided us with a few small ripples on the surface of the bibliographic data pool. In many ways, determining whether or not it has successfully fulfilled the goals it had set out to accomplish is something still open for debate.

But one thing RDA has done is helped the cataloguing community to start thinking in terms of discrete data elements that relate to the intellectual content of a resource leaving display issues like punctuation and lay out for the machines to handle.

In fact, through the use of machine actionable data, the description of resources may largely be performed by computers in the future. This can shift our attention away from data entry to something humans are much better at: classification, subject analysis and fine tuning the relationships that exist between resources.
Before I move on to talk about linked data I should also note that the MARC standard has also hampered RDA from realizing it’s potential. The MARC standard has long been associated with AACR and so in many ways switching to RDA is like pouring a new 21st century wine into an old 20th century bottle.

Our data is left constrained and unable to dynamically connect to non-bibliographic data that lives outside of our library catalogues. This external data could not only enhance our descriptive efforts it could also provide ways for our data to contribute to the findability of other information resources available online.
We are stuck in a data stack that might could be illustrated something like this.

On the top is our bibliographic data. On the bottom the semantic web. In between these two data sources is the MARC standard.
MARC makes it difficult for our data to get out beyond our catalogues. The data needs to be converted into something like MARCXML and then further transformed before it has any chance of connecting to external sources of data.
What we really need is a data environment that would allow us to work directly in the formats that the web uses and ingests. Without that MARC layer we could establish an unrestricted path to enable our data to flow freely and dynamically link up with all of that rich data that exists on the web.

Some of you may have heard of Bibframe which is a pilot project led by the Library of Congress. The first phase of that pilot ended in March of this year and they continue to move forward and apply the knowledge they’ve gained over the past 4 years or so. They are still quite a ways away from a tool that can be used by the whole community but they continue to experiment, and part of that experimentation includes linked data …

So, what exactly is linked data?
On the surface the basic theory of linked data is actually a pretty simple one.

It's the practical implementation and the technological challenges underlying linked data that makes it a somewhat complex exercise. But for now I'll try to stick with the simple stuff ... :-)}
So, first of all you've got a Subject ...
... and the Object of that Subject ...
... and in between those two entities is, what's known in Semantic Web parlance as, the “Predicate.”

The predicate tells us something about the subject, it represents a property or characteristic that the subject has.
Another way to think of it is that there is a Relationship that exists between the Subject and the Object: it's that quality, or property that connects the Subject to the Object.
When a Subject and an Object are connected by a Relationship we have a Triple statement or simply a Triple.
Triples are everywhere. Some common triples associated with legal resources might be:

a Case has a Citation
a Court has a Name, etc.
a Resource has a Format
etc.
If we look at a simple appellate case from CanLII like this one, we can pull out some of the data elements and express them as a collection of Triples.

For example, this entire set of data elements could be described as “Case A.”
So, for example, then:

**Case A has docket number C57108**
Case A has plaintiff Royal Bank of Canada
Case A has defendant Leslie King
Case A was heard in the Court of Appeal for Ontario

And so on ...
If we then list these four triple statements you'd get a collection of triples like this. The Subject for each of these triples is ‘Case A’ each followed by a relationship or property of some sort and a value that corresponds to that relationship or property.

For those of you familiar with relational databases …
... you might think this is starting to look very much like a table with a common key set as ‘Case A.’

And traditionally that’s how data like this has been represented: as a collection of data elements in a database record.

And there’s nothing wrong with that.

However, this data is confined to the relational database within which it was created. Just like our library catalogues.
The beauty of linked data is that it allows the record structure to break down ...
… into a collection of triples.

Triples can connect independently with information found in any number of other data sources that might be out there.

The triples float freely in the semantic web.
You could still bring all of these data elements together as a group with a query that sets out to find our original set of ‘Case A’ data elements.

And we could represent the results of that query with a graph like this one where the circle in the centre is the Subject ‘Case A’ and the spokes fanning out around it are the various relationships and connecting their associated objects.

But once our data is broken up into triples other connections can be made automatically with other data sources which can both enhance our data and enhance the data that others provide to us.
And suddenly, and I imagine quite quickly, once we get enough data to reach a critical mass, our data begins to link up with someone else's data ...
And before you know it there is a chain of resources linked together through their common data points.
Creating what I imagine as these crystalline information structures that connect up in many different ways.
Or maybe something more like this screen shot from Protovis, an early linked data visualization service.

Linked data provides connections that have the potential to reveal new relationships between the many Subjects, Objects and Relationships existing in many different data sets that exist out on the web.
When I was in library school in the 90s we were introduced to the work of Vannevar Bush and in particular to his article "As we may think" describing what he called the Memex system. Although this paper was published in 1945 I return to this concept periodically and I’m amazed by his forward looking ideas. Truly ahead of his time.

Seventy years later it seems like time has finally starting to catch up with his ideas. For example, this passage, citing what Bush calls 'associative indexing,' is particularly relevant to today’s linked data developments and how linked data works.

I’ve taken the liberty of making a minor adjustment to his quote and substituted a couple of contemporary terms so that reads something like this:
“... associative indexing, the basic idea of which is a provision whereby any resource may be caused at will to select immediately and automatically another. This is the essential feature of linked data. The process of tying two items together is the important thing.”

And this process of tying two things together, the Triple, is really the kernel that leads to the potential of machine learning and machine actionable data.
I'm sure many of you have heard this phrase before:

“Law is chaos with an index” often attributed to Oliver Wendell Holmes. Regardless of its origin this is an apt description of what we find in the legal domain.

In a short paper about linked legal data efforts in Finland the authors provide a very nice summary of this potential legal chaos in the opening paragraph of their report.
“Publishing and using juridical information is challenging in many ways. It is produced by different parties, such as governmental bureaus, ministries, different levels of courts, research organizations, and media. The content is heterogeneous and produced using differing tools, data formats, and practices. The links between documents are often informal and/or not made explicit. The law in general is a dynamic, changing entity: for example, it is important to be able to refer to different versions of a law at different points of time. These challenges can be addressed through the use of linked data techniques.”

In addition to these challenges comes an increase in user expectations for access to legal information along with changes to the way lawyers now practice law. Both of these spheres have been influenced by ongoing technological innovations that have been emerged over the past several years.
As the authors suggested, the heterogeneous reality of the legal domain and the increase in user expectations is something that might be successfully addressed in a linked data environment.

But it really depends on whether we can build a critical mass of data sources in a standard format that normalizes the data and makes it more homogeneous. If we can do that it will be easier for both machines and humans to discover and make deeper connections with the information resources that are available to us.

Once the data is up we, as librarians, can place our efforts on the development and consistent use of metadata that describes and guides people to the relevant documents needed to assist them with their research and explore their legal rights and obligations.
That's the motivation behind the KF Modified Linked Data project that Sarah Sutherland and I have been working on for the past couple of years. This linked data version of KF Modified has the potential to create connections between our bibliographic data with other data sources available on the web.
If you have any questions after the session please feel free to contact me

tknight@osgoode.yorku.ca

@freemoth on Twitter

Thank you!