Library Linked Data: an evolution

by Karen Coyle

I am particularly pleased to be here in Florence as part of this seminar on library linked data. I say "particularly" because it was here in Florence, I believe about four years ago, at another conference, where I made an attempt to present these new ideas about linked data, but without great success. In the intervening years I have learned much more about this topic, and at the same time the concepts of the semantic web have spread throughout the information communities, including those of the sciences and the cultural heritage institutions. We are here today to continue our support of this evolutionary development, not only for libraries but for all users of the web who are or who could also be library users.

My goal today is to introduce certain basic concepts that will help to provide a context for the remainder of this meeting. It is not uncommon when discussing a technical topic like the semantic web to focus on particular details, yet for us here today it is essential that we steer our discussions toward areas that are particularly important for our community, especially in these times.

To understand our future we must of course know our past. In the case of libraries, our past is long and I could only give a nod to the centuries of experience and tradition that have brought us here. Many of the presentations that you will hear in these days will give a vision of our future. Therefore, in the few minutes that I have I would like to speak neither of the past nor the future, but of the present. With this I hope to provide some context that will allow us to connect our past and our future.

Our world today:

- is increasingly experienced through computers and devices, like cell phones and iPads, that are connected to the net
- it is enormously interactive; everyone can create (albeit perhaps only a personal Facebook page), can interact, can be seen and heard
- the world is pluralistic in terms of culture, politics and economics; in the analog world power may be concentrated in 1% of the population, but a blogger who belongs to the 99% could have millions of followers and a significant influence.

Our information resources

- are either born digital or are being digitized
- are relatively easily accessible throughout the global network but are also costly to use because they require advanced technology, such as devices, wires and reliable electricity, but also familiarity with this technology

Today's users

- expect to do their research and interact with information without prior training, preferably using a single search box
- interact with the library through software and hardware that is not under the library's control
- To today's users "access" means "obtain a copy," and "obtain a copy" means that the resource is removed from the organizational context of the library or the database or the web site; every user has a hard drive full of documents that have no particular organizational context

Communication today is significantly different from only two decades ago.

- communication is not face-to-face but across distances; if you see two youngsters side-by-side, each sending text messages on a cell phone, there is a good chance they are sending messages to each other
- communication is becoming faster and shorter; it takes years to write a book and weeks to read it; it takes hours to write a blog post and minutes to read it; it takes less than a minute to write a tweet and only seconds to read it
- communication today is based on interaction; one can comment on a blog or respond to a Tweet, or even comment on newspaper articles; a text message is a single entry in a continuous communication; today's youngsters would probably be more at home with a Socratic dialog than with the fixed, inactive, printed book.
- at the same time that the printed word is waning in influence, the use of other media, such as photos and videos, is increasing; these are used not only as mass media but today also as individual communication; and not only as entertainment but as the primary means of instruction – instead of the instruction manual that once came in the box with the purchase of software you now find online "how to" in video form. YouTube receives 60 hours of video every minute.
- communication that in the past was informal and un-captured, like a hallway conversation or a classroom discussion, now may be fixed in a digital form. We have come to treat these formerly informal communications as equal to traditional records, using them in the courtroom or even as the basis for research.

To summarize, the world today is online and interactive; communication that is informal but in digital format now is included in our historical record; the printed word is from another time. Print will not disappear, but it is clear that it is no longer to be considered a modern technology.

The web has changed everything.

Libraries must confront this change; it is a matter of life or death, existence or disappearance. An institution based on the pre-Web civilization cannot be relevant, and we cannot assume that such an institution will continue to exist.

So, what is the state of libraries today?

Our libraries contain a huge cultural heritage. To organize this cultural heritage and to make it available and useful to the public is a complicated and costly endeavor. But the big problem for libraries today is not just the curation of the past; the present provides a huge challenge. Not only has the number of printed books increased in recent years, while the financial support for libraries has decreased, but as we learn with the example above of YouTube, every minute an untold number of new resources is added to our digital culture, and none of these is under the bibliographic control of the library.

Where in past epochs one could consider the library the main source of recorded information, this is no longer true today. This, in itself, is not the problem. We should be pleased with the growth of and use of information and the resulting potential for an informed society representative of an active and vocal populace.

The problem, instead, is that libraries are distant from and unconnected to today's primary information resources, which are on the Web. The push to move libraries in the direction of linked data is not just a desire to modernize the library catalog; it represents the necessity to transform the library catalog from a separate, closed database to an integration with the technology that people use for research and creation of new ideas. Library data needs to be online where it will interact with existing and future information resources. This means that to be visible to today's user, the library catalog needs to cease to be a separate database; it must become data dispersed throughout the web, fully linked to the web of data.

Our job today, as librarians and information scientists, is not to translate library data to linked data; our job is to create a new system for access and use of bibliographic data that is compatible and works within the web.

There are two primary aspects of this development. The first is to make bibliographic data usable on the web. Every person who does research, who studies, who writes and cites, needs bibliographic data, some of which can be provided by libraries. With library bibliographic data on the web, everyone online becomes potentially a "library user."

The other aspect is the use of online data to improve the libraries' user services. By making connections between bibliographic data and web resources one can, for example, place a book within its historical context or demonstrate the influence of an author on his time.

Progress has already been made in some areas, as you will learn from the speakers at this seminar. There are two primary activities that provide the background for the creation of linked data: the first is the development of the metadata elements that one will use for the data, such as "author" or "title"; the second is the gathering of controlled lists of terms that will be used as values, lists like languages, geographic places, and names of persons.

Because library metadata standards already define a number of controlled lists of terms, these have been fairly easily converted. The Library of Congress presents its

subject headings as linked data, as do the national libraries of France, Germany, Japan, and others. Some linking has been created between them, forming the basis for a future web of subject data that is multi-lingual and international.

Name authority data in linked data form can be found in the Virtual International Authority File, VIAF. VIAF, which is held at OCLC, receives name authority records from about twenty different major libraries. It clusters the records for the same person and creates an identity for that group. Where possible, a VIAF cluster links to the Wikipedia article for that same entity, and in some cases there is a reciprocal link from Wikipedia to VIAF. Again, this is the beginning of a web of data.

There is a certain amount of experimentation in the translation of traditional bibliographic schemes to linked data: in particular, ISBD, FRBR, FRAD and RDA have been coded using semantic web standards. However, these are not connected to any Web-based data, and this is a very important point to make.

A key part of the semantic web that differs significantly from metadata practices of the past is that of linking, and in particular linking between metadata elements from different communities. It is only through this linking that we will make the transformation from a closed world of library bibliographic data to the open world of the semantic web.

This means that we need to make connections between library data and data that has its origins in other communities and resources, whether these come from scientific research, government data, commercial information, or even data that has been crowd-sourced. If we must understand one key thing about the semantic web it is that it is an information environment that is highly heterogeneous, both in its breadth but also in quality. The closed world of bibliographic control that we have enjoyed up until now will not be part of our future.

To conclude…

We must ask ourselves if linked data is going to solve all of the libraries' problems, and the obvious answer is: no, of course not. But the bottom line is that we cannot move into the rich and dynamic information environment of the 21st century with data that is based on 19th century principles.

It is possible – no, it is *probable* – that we will need a profound change to library data to meet today's needs. In the end there will be a significant difference between today's library catalog and the access and view of library data that integrates with the web.

We must no long create bibliographic data that is intended only for library use. Our users are not limited to those who interrogate the library catalog but are all persons who seek information and create new resources, whoever they are, wherever they are.

We must be not only *on* the web, but *of* the web. We must use the standards of the web, the structure of the web, and the services and applications of the web.

The biggest risk is that we will change, but we will not change enough.

The original goal of RDA was radical: it intended to break with the cataloging standards of the past and create a new view of library bibliographic data that was open, flexible and extensible. However, as the work on the standard went forward many in the field questioned our ability to make this change, and the committee retreated to a position of guaranteeing that RDA would integrate well with current library data. Unfortunately, no analysis was done of possible systems solutions for transformation of the data. We have let our past anchor us in place, and to keep us from moving forward. The result is that when we adopt RDA in 2013 it is possible that our data will be nearly indistinguishable from that of our current catalogs.

It is not just the machine-readable format of our data that needs to change, but the content of our data. We will not become relevant by recreating ISBD or MARC in RDF. The library bibliographic record today is essentially a marked-up text, using natural language to describe resources, and is not suitable for machine actionability. We continue to create headings whose function is directly relevant to the linear catalog and alphabetical order. This is not only no longer useful in today's world but it actually makes it harder for us to exchange our data with communities whose data is structured for machine-applicability.

We can no longer view the goal of our data creation to be a library catalog that looks much like the catalog we have today. And we can no longer view our catalog as a destination that is separate from the open web. The time of the library catalog is over, as much in the past as the time of the horse and carriage. Instead of insisting that our data cannot change because it has always been like this, we have to turn our attention to ways that we can re-utilize this data: to the transformation of our data using the computing power that exists today as well as the computational capabilities provided by the web itself.

All this said, I want to end with a call to all of you, in the days that follow and beyond this particular meeting, to consider the idea of a library *of the web* as worth exploring; as one possible future, but not the only one; to be willing to consider that library data will take an entirely different form from what it is today, and that this will not lead to the destruction of the library as we know it but to its evolution for future generations.