2020-10 VAOH Session

Presentation summary

Linked data, and the road to learning about it

Nathan Putnam & Cynthia Whitacre introduced the sessions, informing participants that that day's topic, and the topics for the remainder of this year's sessions, were selected from topics suggested from past user survey responses. The presentation provided by Charlene Morrison covered the basics of linked data; how OCLC has been and continues to work with linked data, as well as how Metadata Quality staff are participating. Then concluded with some resources for furthering your understanding of linked data within the library community. Joining them during the Question & Answer section were Becky Dean, Robert Bremer, and Shanna Griffith.

URLs mentioned during the presentation:

2020 LD4 Conference
https://www.youtube.com/watch?v=cTibbkBPKG0&list=PLx2ZluWEZtIAETxLY-TaqJWsRMNY59r9v

21st Century Indexing: Learn how FAST (Faceted Application of Subject Terminology) can help libraries and other cultural institutions to assign subject headings https://www.oclc.org/en/events/2020/21st-century-indexing.html

Abstract Wikipedia project https://meta.wikimedia.org/wiki/Abstract_Wikipedia

ALA Fundamentals of Metadata http://www.ala.org/alcts/confevents/webcourse/fom/ol_templ

Coursera, Web of Data https://www.coursera.org/learn/web-data

FAST (Faceted Application of Subject Terminology) https://www.oclc.org/en/fast.html

Linked Jazz https://linkedjazz.org/

OCLC and linked data https://www.oclc.org/en/worldcat/oclc-and-linked-data.html

OCLC Bibliographic Formats and Standards: Control Subfields
https://www.oclc.org/bibformats/en/controlsubfields.html

OCLC Research Linked Data https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-outputs.html

PCC Task Group on Linked Data Best Practices / Final Report
https://www.loc.gov/aba/pcc/taskgroup/linked-data-best-practices-final-report.pdf

RDF formats and serializations: https://ontola.io/blog/rdf-serialization-formats/ ; https://en.wikipedia.org/wiki/Resource_Description_Framework#Serialization_formats

SEMI (Shared entity management infrastructure) https://www.oclc.org/en/worldcat/oclc-and-linked-data/shared-entity-management-infrastructure.html

Serialization https://en.wikipedia.org/wiki/Serialization

URI FAQs / PCC URI Task Group on URIs in MARC / September 26, 2018
https://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI%20FAQs.pdf

Wikidata https://www.Wikidata.org/wiki/Wikidata:Main_Page

Member questions

2020-10-06

Please define "serialization."

Serialization is basically a way of formatting the data. It's kind of the way that you code it in the background.

https://en.wikipedia.org/wiki/Serialization

Participant response: Simply put RDF data can be output in different formats, like JSON, JSON-LD, turtle TTL, etc. for Web services to ingest and process.

Machines negotiate content for delivery in HTML so humans can digest. It depends on the receiving service requirements; structural data can be queried, and output based on its requirement for processing to meeting the need.

JSON, JSON-LD and TTL are supposedly more human legibility friendly. Alternatively, some browsers may have plug in to read in the data and output friendly to humans, e.g. Sniffer

For non-coder catalogers, this will be behind the scenes and we'll be able to interact with user-friendly interfaces--correct? I hope so

Catalogers will be using a user-friendly interface as opposed to looking at all that coding.

It would be something like when you look at Wikidata entries. I highly recommend joining in the LD4 Wikidata affinity group, it's a good way to get an idea of what an interface might look like, that a cataloger would work in. But you wouldn't necessarily be working on the serialization side

of things. In a sense it's like the MARC format, in that Turtle code, RDF, RDF/XML, etc. is just a way of coding the information so that the computer can do its work in the background. It wouldn't necessarily be what is displayed for the end user, or even necessarily what is displayed for the cataloger within the linked data.

Can the presenter talk a about the (current or future) relationship between MARC cataloging and linked data? i.e. how you see the relationship, etc.

What we've been doing is prepping that MARC data to be used within linked data and as we transition more towards the linked data world, we eventually could leave MARC behind completely. But it will take some time as we transition and again, the subfield $1, subfield $0, those will help us translate that MARC data into linked data much easier.

Then focusing on the authority records some of the things we're learning is all of the information that's used to create authority records is in no way comprehensive about the "thing" being described. So, what does comprehensive really look like in terms of linked data? If you're describing a person, how much about that person do you want to know? And we're looking at it really from two ways: 1) what do you know about a person, e.g. their birth date, their death date; where they may have worked, etc. But 2) we're also looking at it from the bibliographic side of: Do we know everything they wrote? Do we know everything that was written about them? So, in terms of authority-ness , we're really limited in some ways, only to those people who wrote things. We're very light in the authority file on subjects for persons. We have the prominent people, but if you really look at… For instance, one thing I enjoy listening to is CBS Sunday morning. They always have this fascinating segment that's called A life well-lived. And I got really curious one day and started trying to find out, for all these people who had lives well-lived, their fascinating history. None of them were represented in the authority file. Well, that seems really odd, and yet many of them were in wiki-data. So, there were works that they contributed to, in terms of their life, but we're not able to capture that in authority records. There's a lot of room for growth, in terms of what we can do to find relationships between people who were not authors and works that represent their life well-lived.

The major block to adopting LD (linked data) in libraries is the huge amount of legacy (unlinked) data we already have. Are there strategies being developed to address this issue, especially for libraries with small staffs that couldn't possibly update their existing data manually?

A lot of work in linked data so far is in the academic library world. How do you see how public libraries will benefit from linked data?

There are clearly some people in the chat who are advanced at linked data policies, and procedures, and just how it all works together. One of our concerns with presenting this topic, though, was how do we reach the person who've been aware of the linked data stuff, but not necessarily paying too close attention because there hasn't been an effect on their day to day work. And OCLC, and several of the other vendors, there is work being done on creating the interfaces for you to use. So, you don't need to be a programmer or know RDF or that kind thing. And you'll make use of linked data because the underlying structure will change and create all of

these or allow you to transverse all of these relationships. The LD4 community and the LD4 wiki affinity groups are looking into these interfaces. OCLC is looking into the user interface in terms of how a cataloger would actually use all of this data. The stuff to do right now, if you are in a place to do that is to, for example work on cleaning up the subfield $0 the subfield $1 in your MARC data, if it's at all possible. If it's not possible, we're still open and sharing our data. Yes, we have a subscription, so I'm not going to downplay that at all, but we're still based on the fundamental cooperative cataloging model so no one will be left behind, so to speak.

Are the major ILs vendors, especially those that serve the public library market, involved in this work?

I believe that Innovative Sierra is working on Linked Data. Xlibris is also working on incorporating linked data into Alma and Primo.

What is the level of adoption of Linked Data in libraries? How are they using LD (linked data) to meet their missions?

I know that we've worked with several groups that are doing that. There are several libraries involved with the LD4 groups, that are working with the different standards, things like that. Right now, a lot of it is just experimentation and developing the infrastructure to be able to use the data.

How do you respond to those in the cataloging community who are skeptical about RDA, BIBFRAME, and the actual implementation of linked data?

What you can say about RDA is that it's an evolution in our cataloging instructions that is better designed for transitioning to linked data in the future. Because one of the changes was an emphasis on actually coding relationships. Under AACR2 we didn't supply, in terms of MARC, subfield $e, with relationships between authors as access points in the same way that we do now under RDA. So, with all of that data specifically coded in our current environment, it's the kind of thing that we'll be able to map forward in the future into a linked data context so that it can operate on the web.

With BIBFRAME and implementation of linked data, some of this is still remains to be seen how it plays out. Some healthy degree of skepticism is good, because it will get those questions answered that need to be answered in terms of How does this affect me? How can I help my systems? What's the benefit to me? etc. So, having questions like is always good.

I think the skepticism comes from the fact that this transition seems to be taking a lot longer than I thought it would. Any ideas on why that is?

I think it's taking longer because it ended up being harder than we originally thought it might be because this has been in the works for 10 years plus probably. We've made a ton of progress in the last 5 to 10 years and I think there are going to be breakthroughs with actual practical use. OCLC, the LD4 community, the other vendors that are working with us in the next year or so as, as we really home in and move this all forward.

Or is it taking longer because discovery services we use are lagging behind?

That's interesting, though I don't know how well we can answer that since most of us are catalogers and our primary focus isn't the discovery system. There's definitely that piece of it: how the discovery system works for the end user, not the librarian, not the cataloger, but the students or the public that come into our institutions. How are they ultimately going to use this and make those connections? Those are questions that still need to be answered, in the grand scheme of things.

Participant response: I think it has to do with legacy data.

There is definitely something there. The goal with BIBFRAME was not to leave MARC completely behind and start fresh. There are aspects being ported over because that's how it is in MARC. Once we get into more of a linked data environment, it'll be good.

I also wonder if a part of the challenge isn't just as a community. We are limited by budget. We are limited by training, et cetera. Moving to a completely new dynamic infrastructure is a big shift. Not only do we have to understand it, but we've got to persuade budgetary constraints that this is a good use of their funds, knowing that we just don't have unlimited funding anywhere.

Participant comment: It is hard to "sell" the idea without being able to demonstrate it effectively in a discovery/user facing system.

Absolutely agree. And I think that's where some of the smaller libraries who aren't involved in the establishing rules, standards, etc. are struggling because they haven't been able to see the progress etc.

Participant comments:

Good point in that the reduction of cataloging staff during the past 5-10 years has been significant in all libraries and sectors.

My library participates in the Library.Linked project with Zepheira, which is a good first step for public libraries to get into linked data.

Wikipedia and Europeana effectively uses LD (linked data), at least that is my understanding.

Those are good communities to check out.

How useful can LOD be for increasing and diffusing knowledge? Having structure data readily available for query services, e.g. SPARKL, when there is a need to generate Web pages for a trending topic, e.g. CoViD-19, the process is much faster. Output data is constantly updated as the queries are being conduct when a user clicks on the link. https://sites.google.com/view/covid19-dashboard PS: The site was up in April 2020

Someone in chat points out that having the structured linked data, the linked open data, is increasingly good for diffusing knowledge. Being able to have these query services and the SPARKL end points, especially in our current environment, where most of us are working from home and lockdown, away from our normal infrastructure this sort of processing is so much faster than having MARC data locked up in our different MARC repositories, our different MARC silos.

She also points out that the output data is constantly being updated as the queries are being conducted so you don't have to worry necessarily as much about as stale data.

Our bibliographic and authority data is incredibly nuanced, especially in communities like the Rare Books community. Being able to map data between MARC and LD, and back again, without data loss is proving very complex.

Going back to why this is so incredibly complex, someone mentions being able to map MARC and linked data, especially outside of the normal monograph book cataloging, and specifically with the rare book community being able to map data between these and back again, without loss is proving to be very complex.

This is true of archival material as well since the MARC record doesn't deal well with collection level information. Making sure that that information doesn't get lost and the contextual notions, ideas that are available within the description of those records makes it challenging to put a linked data wrapper on it.

How then will legacy data be "updated" or move to linked data environment? or will we have 2 systems running parallel to each other? AND BTW, a lot of people still not clear about linked data.

Absolutely, people are not necessarily clear about, or sold on, linked data, but I think as OCLC and other groups, like LD4 and PCC (Program for Cooperative Cataloging) as they continue their investigations for this that it will all become clearer.

As far as how legacy data will be updated or moved to the linked data environment: it has been something we've talked about a lot. There are certainly challenges. Knowing that when you look at a field in a MARC record, it's comprised of different subfields when you express those different subfields in data, you're looking at different properties. You've got to be able to pull apart the pieces and then be able to dynamically update them going both ways. It's certainly something we've been looking at. We understand some of the challenges. We have not solved the entire puzzle yet, but we certainly understand the need that those two expressions, if you will, need to have a relationship and how to maintain that is definitely going to be a challenge.

Of course, well-coded MARC data is something that will transition to linked data much better than the case of MARC records that are poorly coded and incomplete.

One reason we're having a hard time moving forward is that many are still trying to understand linked data through a MARC lens. Thinking out of the box and getting away

from a "record" concept might help. One of my favorite linked data sites is here: https://linkedjazz.org/network/

There seem to be a lot of systems/interfaces for linked data now and lots of experimenting is happening now. Is it likely that just one will eventually surface and we'll all use it, like MARC?

In terms of linked data, aren't all the instances where an authority has been linked to a bib resource also part of the LD (linked data) information about that entity?

We do look at this through the MARC lens and part of that is we don't want to necessarily lose all of this very rich data that we have trapped in MARC. But at some point, that does inhibit us to some degree. The linked jazz network is good for doing some exploring, to go through and see how Jazz is all linked together with all the different people. It's a very nice interactive website.

This is one thing that helped me get out of the MARC lens so dramatically was when I was asked to look at MARC data, because I was very focused on understanding these are the elements in a MARC record. What would they look like in linked data? What are the properties? etc. And when I got done someone I was working with at the time, looked at me and said: What question are you trying to answer when you look at a MARC record? What are the questions you're asking yourself when you look at it? Are you asking: What is the title? Who was the author? And that sort of helped me break the MARC-ness bias I have, no matter what the approach, it's still, what question am I trying to answer? If we can start thinking about linked data as not necessarily how does it equate to a MARC bibliographic record or a MARC authority record, but what do we need to know to answer the question that were are posing? What are you trying to do? Perhaps, if we start looking at it like that, it might help us break out of that detail level of subfield $a, subfield $b, subfield $c. It gives us a better way of looking at what are we trying to do to help our users.

Participant comment: LC is creating BIBFRAME (which is LD [linked data] centered), and they have been able to convert MARC to BIBFRAME, and they are trying to convert MARC to BIBFRAME.

If you're interested, it is a good site to explore. They actually provide a good visualization of how BIBFRAME would look. How a MARC record would look in BIBFRAME. And then going from the BIBFRAME record to MARC. So, it is a really good tool to check out and play around with.

Participant comment: I believe that LC is working on their backwards conversion so that we can work in two systems, per se

Where is that PCC document about $0 $1 in authority records?

The FAQ for the URIs in the presentation notes breaks it down and it does a pretty good job of explaining the difference between the two and the purpose of each one. Bibliographic Formats

and Standards has information about the control subfields, including examples. The MARC documentation also does.

I find records with wrong 520 fields and youth headings attached to non-youth materials. I've been deleting them and updating the WorldCat record, but should I report them instead?

We would appreciate you reporting them, that way we can see if there's a bigger problem, and then find other records that may be involved with that.

Do we have any sense as to why this is happening right now?

They could be coming from merges when fields transfer, or they could also be coming in via Ingest.

If we report records with invalid 007 values that block replacement, would reporting them prompt a batch-based correction? when appropriate

Yes, we'd like to know about those kinds of situations where there's a problem that's widespread, because that is the kind of thing that that would lend itself to some automated fix. Some are easier than others, but certainly cases where you have messed up 007 fields that are getting in the way of being able to do replaces on records, that's something that we would like to take care of across the board. So, yes, please do report that kind of thing.

2020-10-15

Will OCLC be offering a cataloguing interface for cataloguing in BIBFRAME? Something like Record Manager?

What we're working on is the SEMI project that was described in the presentation. We're working on the interface for that. Its relationship to BIBFRAME is something we can take to others. The relationship of SEMI to Record Manager is certainly something we're thinking about, but right now we are trying to keep them separate in our approach to ensure that we address the needs and the user stories associated with linked data. And then we can look back at Record Manager to determine similarities, differences, and that type of thing.

OCLC may or may not implement BIBFRAME?

Everyone is talking about BIBFRAME as we're looking at SEMI, so, in no way are we saying we will, or will not implement all the discussions related to BIBFRAME. We are using all of the knowledge in the community from BIBFRAME and other sources and discussions to guide us in our thinking and our understanding. And there's a lot of information on the community site as we're working with the User group for SEMI to ensure that we are engaging the community and understanding their needs. And a lot of them are also involved with SEMI and other projects. So, there is a close relationship between what we are trying to build, and the standards that are under discussion in the community.

Participant comment: ALA Fundamentals of Metadata had a good introduction to metadata and some discussion of linked data.
http://www.ala.org/alcts/confevents/webcourse/fom/ol_templ

Will the result of linked data be open access knowledge graphs?

That too is under discussion. There are a lot of conversations going on regarding accessibility and what it means to have linked open data for WorldCat, understanding subscription models, and that sort of thing. So those are ongoing discussions. And again, I think a lot of that information will be made available because that too is a topic that has been put forward to the advisory groups on the SEMI team, and its users to get input on how people are thinking about who should be able to see what, what should be linked open data, what should be more guarded for OCLC members. So, a lot of discussions about that are in play.

SEMI, to reiterate, is the Shared Entity Management Infrastructure project that is funded by the Mellon grant, and it is underway now. The end of that grant will be at the end of December of next year, 2021. So, we expect by the end of that grant to have an interface for entities.

So, it won't be an interface for a bibliographic record, it'll be interface for just pieces and parts that are entities. One thing we're looking at is the different entities. We are creating what we're calling a minimum viable entity description. We're using the properties and classes to guide us in determining, not unlike some of the forethought that was going into Bib Formats and Standards, to determine what fields are required, that sort of thing. We're taking a very similar holistic approach to understanding what properties we think are needed to be able to describe a particular type of entity and then growing the interface around those rules and thinking.

Any comments on ORCID? I have worked on that in a remote project.

Certainly, we're aware of ORCID a lot of research folks have been involved with that project. And it is one of the identifiers that we're looking at incorporating as a property.

Can you address who would have editorial control over linked data entities?

And again, I think that goes back into the interface that we're building, looking at that subscription type approach to modeling of who would have access. We're not trying to build the bibliographic infrastructure, and I really want to make sure that people understand, in no way are we looking to rebuild that. We're looking to that as guidance as we make decisions. All of that is still very much under discussion. One of the things that I think we all know from Wikidata is that there are a lot of people in the world who know a lot of things about particular types of information who could easily add statements and claims that they just know because of their education and their familiarity with the specific topic. We're trying to ensure that we provide a way to ensure that everyone can contribute to OCLC's linked data in the same way that we've seen the community build the Wikidata. So again, those conversations are very much under discussion, but we are definitely looking at ensuring that people can contribute claims and statements as they are aware of that knowledge and making sure that we keep that open so we could share that information. That's the whole point of linked data, to share what you know.

If you're using Wikidata now and you are adding statements to Wikidata, I think adding to our entity data, once it's ready or once it's ready for people to edit, will be very similar. We are definitely using Wiki-based infrastructure as our underlying technology, so many of the same concepts and some of the look and feel is very much like Wikidata. But we are trying to ensure that we fit it to meet library needs and the community.

Any plans for Webjunction courses on LLD [library linked data]?

Please see their website: https://learn.webjunction.org/

My understanding is OCLC has "published" linked data, what does that actually mean, and does OCLC know how anyone has used that published linked data?

As far as I am aware, publishing linked data means making it publicly available.

One of the ways is the FAST Linked Data Service. WorldCat.org has some published linked data. Viaf is considered published linked data as well.

We know lots of people are using FAST, there's evidence that people are using the linked data OCLC has published and linking to it.

Why is OCLC not going to BIBFRAME till now to get more benefits from Linked Data ?

A lot of it, as noted in the presentation, is that OCLC has spent a lot of time working with linked data, and again, we are certainly not ignoring BIBFRAME, but we're also drawing on our own research as we explore the work with identities, with ContentDM. There's a lot of knowledge there that we are using to help guide us, and all the user communities' feedback that we have from those projects. So, again, we are in no way ignoring BIBFRAME. We're just trying to include everything that we have learned as we look at the other standards and discussions going on in the community.

One reason is that BIBFRAME doesn't actual handle RDA well. It is less full than RDA in RDF. It doesn't follow the LRM model.

There's still a lot of development going on with BIBFRAME and the Library of Congress is still experimenting with it as are many other people. OCLC has pledged to have a way to, in the future (no dates associated with this) to ingest BIBFRAME data. We will talk about that widely once we're at the point of figuring out what we're going to do with that.

I believe there are more than one format (style) for linked data. What is the best one to start with? RDF/XML

They all bring their own set of pluses and minuses. The resources listed in the slides help to explain it better. It really depends on how comfortable you are with coding and what style you like best: Turtle, JSON, RDF/XML, N-Triples, etc.

I continually get turned down for courses that have the words "fundamentals" or "introduction" and "metadata", so I was wondering if it would be helpful to someone who has been cataloging for 30 years, but in MARC format, not BIBFRAME or RDF. My library administrators consider me an expert in metadata, so I don't get to take these kinds of courses.

There is a great linked data course on Coursera, Web of Data - quite advanced, I enjoyed it

https://www.coursera.org/learn/web-data

How can OCLC help subscribed libraries to transform their data to linked data in the WorldShare Management Services (WMS) ?

We're not familiar with any discussions, but it certainly doesn't mean there aren't any. We'll just assure everybody that OCLC intends to support MARC for years to come. We don't have an end date on that. And we think the evolution to linked data will be just that, an evolution, rather than a revolution. So, there won't be a hot cutover at any point. We'll keep you posted as we have new developments.