"in the Early Days of a Better Nation": Enhancing the power of metadata today with linked data principles

John Mark Ockerbloom, University of Pennsylvania
Good afternoon, and thank you all for joining me here.

I’d like to talk today about linked data, and how the principles of linked data can really improve and amplify the information and services we provide in libraries.

I’ve often heard people say lately that linked data is the future of libraries. Actually, I’ve been hearing that for a few years now; and I’ve heard colleagues wonder whether it always will be the future of libraries – and never the present.

But I’d like to show how the principles of linked data are increasingly becoming the present of libraries. And I’m seeing that not just the adoption of linked data technical standards, but in the adoption of the principles behind linked data, and the increasing level of sharing and aggregation of the knowledge embodied in libraries that those principles make possible.
The title of my talk is taken from a quotation on one of the outer walls of the Scottish Parliament building. It’s a motto frequently used by the Scottish writer Alasdair Gray, “Work as if you live in the early days of a better nation”. The Scottish Parliament itself is a testament to the power of that idea. The very existence of the Scottish Parliament, founded in its present form in 1999, is a sign of the growing autonomy and identity of Scotland as a nation.

As an American, I don’t have a strong opinion about whether Scotland should ultimately stay in the United Kingdom or go its own way. But the motto makes it clear that Scots aren’t waiting for the resolution of that question to build the nation they want to live in. Even if the formal basis of their government isn’t now where folks like Alasdair Gray want it to be, they’re still implementing the principles they want to live by today.

I see the same sort of principled implementation happening in libraries. Not all of them, yet, but in many of them, and many others can get involved too if they want to; and I hope in the next 20 minutes or so to give you some idea of how.
So, if you’ve gone to a conference or a webinar where they’ve talked about linked data, you’ve probably seen this diagram before. It’s a diagram of a number of services that implement linked data technologies, and how they connect with each other. A fair number of the bubbles in this diagram have to do with libraries, or publications, or research, or government, or related information that researchers and their libraries might take an interest in. This diagram is now about 3 years old.
There’s a new version of the diagram that came out at the end of this summer, and it’s got nearly twice as many bubbles, and they connect quite a bit more densely in the middle. In one respect, it’s an impressive array of data and services. In another respect, it gives me some pause that you can still represent the world of linked data, down to the level of individual services, on a single slide. I couldn’t imagine doing that for the World Wide Web, even a few months after the first graphical web browsers started to appear.

Now, I admit I’m not being entirely fair here. The people who made this diagram make it clear that not all linked data instances are represented on it. They have a certain threshold for inclusion. And indeed, there are a lot of linked data projects in the world of libraries that don’t show up here.
We get a better idea of the extent of linked data use in libraries from a survey that OCLC Research did this past summer. They put out a call online asking about linked data projects in libraries, and they got back replies about over 100 such projects. People went into detail about 76 of them. The vast majority of these projects are consuming linked data from various sources, primarily to enhance the data they already have in their catalogs and elsewhere, so as to provide a better experience for their users.

Most of those projects also publish linked data. Sometimes it’s just for demonstration: “Look Ma, Linked Data!” But it’s also often done to better expose things the library has to offer, by putting that data into a wider variety of contexts. With linked data, users don’t have to purposely go a library’s own website to find out about resources it can offer them.
And that’s possible because the purpose of linked data is to make it easy to share data, reuse it, and have machines be able to understand it and apply it or repurpose it.

Tim Berners Lee, the inventor of the Web, set out principles of Linked Data some years back. From a technical standpoint, they’re pretty simple to summarize: Use URIs as names for things; use HTTP URLs – that is, ordinary Web addresses – so you can look up information about those things. Make that information available in a form that’s comprehensible to machines, “using the standards” – and here he explicitly mentions the RDF data model and the SPARQL query interface. And link to other URIs so that more things can be discovered.

Now while this is simple enough to summarize on one slide, it’s also presenting linked data primarily from a technical perspective. And in some ways – particularly in the bits mentioning RDF and SPARQL – I’d argue it gets more specific than it needs to. There are in fact a number of other standards that are gaining traction in the linked data world: JSON-LD and LDP are two such standards.

I think it’s worth stepping back a little bit, getting a little more fuzzy with our specifications, and examining exactly what we’re trying to accomplish with linked data.
Here's a more functional and fuzzy take on linked data. You want to have unambiguous identifiers for things that you can look up on the Web. A very straightforward way of doing this is to use HTTP URIs for identifiers, as Tim Berners Lee’s principles recommend. But as we’ll see, there are other ways as well.

We want the lookup to provide data that a machine can usefully interpret and process, without needing a human to intervene. That’s what distinguishes linked data applications from things like copy cataloging. In copy cataloging, you get data from another source, but you still use a person to retrieve it, or look it over, maybe tweak it a bit, and then stick it into a local catalog. But to integrate data at a large scale, we need to automatically work with it without direct human intervention. And we need to be able to do this with multiple sources, not just one, so we can create useful new information from the combination and analysis of existing information. I'll show you some examples of that in a bit.

There are a couple of other important points. Data can go a lot further when it’s open—when the legal and technical barriers to accessing and repurposing it are as low as possible. Also, data becomes much more useful in practice when there’s a commitment to actively support it for a community of users, and keep it useful and up to date.
Linked open data: Not a 1-step process

And the World Wide Web Consortium recognizes these things at some level. Their “5 star linked data” mug implicitly recognizes that there are incremental advantages of various ways of putting data on the Web, even if it’s not in the linked RDF paradigm that they see as the “5-star” standard. And they’ve also recognized the importance of open licenses accompanying the technology.

Let’s talk about those open licenses.
Generally speaking, I think it’s in the interest of libraries to make data about their resources as open as possible. The most straightforward way to do this is to put the data in the public domain, through a dedication like CC-Zero. CC0 relinquishes any copyright-like claim you might make on the data. Of course, folks who want to be seen as credible or scholarly will still want to cite their sources.

To legally require attribution, you can use a license like ODC-BY or CC-BY. With those, anyone republishing your data, if it’s protectable as intellectual property, has to cite you, Or Else. As data get reused and remixed downstream, it might need an ever-increasing set of notices attached. But there are ways to mitigate that. OCLC, for instance, says that if it’s impractical to attach a prose attribution to data that’s derived from theirs, just using their original URIs, which implicitly credit OCLC as their origin, is good enough for them. Or just credit at the collection level. The University of Illinois’s open-licensed catalog records include an original records file under CC0 and a “WorldCat-derived” records file licensed ODC-BY.

You’ll sometimes also see ShareAlike licenses for data. Wikipedia uses CC-BY-ShareAlike, so data sets deriving from Wikipedia, such as DBPedia, tend to use it as well. It’s more cumbersome to work with, but it is an open license, in that anyone can use the data for any purpose they like, as long as they keep attribution and the same license for any derivative data they publish.

Creative Commons also has Noncommercial and No-Derivatives licenses. These are not open licenses by most definitions, and I don’t recommend using them for linked data.
Now, there’s lots of ways you can make data available for consumption, some involving linked data standards, and some not. And the standard isn’t always the way you want to go.

At Penn, we run a linked data service on researcher activity called VIVO. VIVO has RDF, a lot of popular RDF ontologies, searching and browsing facilities, and a SPARQL interface. It’s definitely 5 stars on that linked data coffee mug, and it’s great for a lot of things. But if you’re a stranger who wants to download our whole VIVO data set, you’ll find that hard to do. Like many VIVO instances, we don’t leave our SPARQL interface open to the world, because we can’t easily control what queries get run from that interface. Some of them can consume a lot of resources. So the way that you get a full set of linked data from our VIVO instance is to issue a lot of RDF REST queries. And that can take a while. About a week, one consumer told me. It’d be much faster if we just dumped all our VIVO RDF data into a big file for download.

Dump files in fact are often the most convenient way to share data in bulk. If you have a simple data set, sometimes publishing a dump file is all you really need to do. There’s a great web site called simpleopendata.com that has some useful tips on bare-bones open data publishing.
The linked data in our VIVO instance ultimately comes from some local databases that manage data in more conventional tables. And those databases can deliver data very quickly and conveniently. An internal system we have called Symplectic Elements will eventually manage nearly all the data we publish in our VIVO service. Authorized users can download spreadsheet reports on just about all of the data in seconds, or minutes at most. The people who want those reports don’t know their way around SPARQL queries, but they are experts at slicing and dicing spreadsheet data in Excel.

If users like these are a high priority for you, you might want to look beyond the usual linked data standards to support their needs. You can still export RDF for applications that work well with that, but there’s a lot of data published in RDF that’s not managed in RDF.
Not only do you not need to recast all your underlying data in RDF, but you don’t always need any extra files or interfaces to take advantage of linked data. It can go right into your web pages. Schema.org, for instance, defines a set of markup tags and vocabularies that you can add to ordinary web pages that helps machines understand what your web pages are about. The markup can be expressed in an RDF format, or in other markup standards.
Here’s a web page showing a record for an open access book in my catalog of free online books. To a human, it looks like an ordinary web page. But there’s schema.org markup in the underlying HTML that tells machines that know those conventions—and that includes all the major search engines—that this isn’t just an ordinary web page with some text.

It’s specifically a page that describes a book. And the part of the page that says “Willinsky, John, 1950-” identifies the author of the book. And the part that says “The Access Principle” is the title. And so on. So if a search engine that’s crawled this page later gets a query from someone looking for books by John Willinsky, it knows that this is a result page that’s highly relevant to that query.

A growing number of open source and commercial library catalogs and discovery systems can now automatically add schema.org tags to search results. This is helping many libraries make their resources more discoverable.
There’s ongoing work with the schema.org vocabularies to express more about what libraries provide. One working group has been working on vocabulary for library holdings, for instance. So when someone asks about a book at a service that understands the vocabulary, it can respond “I know where you can find THIS BOOK at THIS LIBRARY!”

“And by the way, that library closes in an hour, but if you go by this route you should be able to drive there in 40 minutes in current traffic.”
Here’s another example of open data, that I’ve published on the Web. It’s a data set that specifies how to construct links to search various kinds of library catalogs and discovery systems. This file gives the patterns for URIs that launch various kinds of searches in 360Search, Aleph version 1, and dozens of other kinds of systems. I publish this file on Github, along with some open source code for interpreting the data in the file, and I give the data file a CC0 license. This isn’t “5-star linked data”, but it has many of the benefits that linked data provide. It’s got an open license, it can be automatically downloaded from a well-defined location, and it’s in a format that’s fully documented for machines to interpret. So that’s all good.
Here’s another file related to that one. This file has data on hundreds of individual library systems. For each one, it identifies its specific catalog or discovery software, and any library-specific information one might need to search it. It uses the same open license and basic format as the other file I just showed you.

But the data in this file overlaps with another data source out there, and that’s the WorldCat Registry. OCLC’s WorldCat Registry has data on thousands of library systems—a lot more than this file covers. It has some information on the catalogs that those libraries use, though it doesn’t go all the detail you need to build targeted search links. (It also isn’t open-licensed yet, but that’s reportedly in the works.) You can imagine how it could be quite useful to combine the data in this file with the data in the WorldCat Registry. And when you start getting into combining data sources, it becomes more and more useful to follow common conventions and standards.

This file does use some common conventions. It identifies libraries with OCLC and ISIL identifiers whenever possible. Since those identifiers are also used in the WorldCat registry, it’s possible to match up data records in this file with data records in the WorldCat Registry.

The WorldCat registry publishes its data using RDF. If I also published in RDF, then someone could easily match up data from both sources with a standard RDF processor, instead of having to use an RDF processor to parse the WorldCat file and a custom processor to parse my file. So I should probably offer an RDF conversion when I get the chance.
It’s in situations like this, where you’re matching and mixing data, that you really start seeing the value of points of commonality for tying data together. I call these points of commonality “hubs”, and they’re a crucial part of the linked data world, and I think they’re a crucial part of data sharing in general.

Hubs exist on multiple levels. Some hubs provide a common vocabulary for data. Schema.org, which we saw earlier, provides one such common vocabulary set. BIBFRAME will hopefully provide another one. And standards themselves are hubs of a sort: RDF Is one of the main hubs for linked data at the level of standard, but JSON-LD is another.

Some hubs provide common identifiers allowing different groups to clearly describe the same things. ISSNs, DOIs, LC Control Numbers, and OCLC IDs are ways we unambiguously refer to specific works, or people, or subjects. Traditional authority headings, like authorized names and subject headings, are also common identifiers. They’re less opaque than coded IDs, but they do have the problem that they sometimes change. That’s not an insurmountable problem, particularly if they’re maintained in a system that has good cross-references and change tracking, but we could do a whole talk just on that.

Some hubs provide comprehensive data sets that can be used in all sorts of places. That’s what makes resources like HathiTrust’s catalog of digitized books, the various OCLC databases, and the various Wikimedia projects so valuable. Wikipedia and Wikimedia Commons aren’t built on linked data as such, but the ease with which users can both extract their information and contribute new information makes them valuable information hubs.
These hubs make it possible for me to build a browsable collection of millions of free online books, as a part-time project.

My Online Books Page fundamentally depends on multiple hubs to function. Most of the catalog records in this collection are automatically imported from HathiTrust— not as linked data, but as open and readily downloadable data— and from other open bibliographic data sources. The description you see here of open access publishing, and its relationships to other subjects, comes from id.loc.gov, which is the Library of Congress’ linked data service for its authorities and vocabularies. The links to search for books in other libraries are based on the library catalog search data files that I showed you earlier. The link to the Wikipedia article on “Open Access” is derived from an crosswalk that’s in turn derived from other open data sources like OCLC’s VIAF as well as dump files that Wikipedia provides.

We’re now using some of these same data hubs to augment and improve the main catalog of the Penn Libraries as well. Recently, a bunch of HathiTrust books appeared that weren’t there before, and there are more changes on the way.

We take linked and open data from multiple sources, and we also give back. Besides the schema.org tagging I mentioned earlier, I also publish all of my original catalog records for The Online Books Page with a CC0 dedication. The mappings I create between LC Subject Headings and Wikipedia article titles are also published CC0 on Github. And I regularly send corrections to HathiTrust, and contribute library links and other information to Wikipedia. This sort of give and take helps keep these community hubs strong.
With all the various hubs out there, libraries have access to a lot of linked and open data they can put to good use. And we can contribute a lot more from our own libraries, whether it’s by putting schema.org markup on our web pages and catalogs, or contributing to a suitable open information hub, or publishing our own open data, whether it’s 5-star RDF linked data or some other suitable form.

The important thing is: if you want to share data with the world, make it as easy as possible for users of that data to find it, process it, and remix it. Standards play an important role in this, but a commitment to provide data openly so that a community will readily work with your data is even more essential. You haven’t seen a single line of RDF or SPARQL in this presentation, but that’s not what you need fundamentally to get started. I also haven’t really discussed some major linked data initiatives now in development, such as the Fedora 4 repository, the Harvard-Cornell-Stanford Linked Data for Libraries project, or BIBFRAME; though if you stick around for Jeff Penka’s talk I think you’ll hear more about that.

But we don’t have to wait for BIBFRAME to overthrow MARC, or some other future technological revolution. We can use the principles I’ve discussed in this presentation to build better library services and outreach right now— in the early days of linked open data.

In summary

• Lots of open data out there libraries can use
  – Lots more they can contribute
• Making it as easy as possible for interested users to find & reuse your data makes big difference
  – Optimize the top priorities for your users
  – But support broad audience (standards help here)
• Find, and work with, your hubs

• Commitment, openness, & community ties key
Questions!

- John Mark Ockerbloom
  - ockerblo@pobox.upenn.edu
  - @JMarkOckerbloom

All content original to this presentation is licensed CC-BY

Thanks for hearing me talk, and now I’d love to hear from you. And if you want to talk with me after this session, here’s how you can get in touch with me.