JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12762

JLIS.it

Follow me to the library! Bibliographic data in a discovery driven world

Richard Wallis[a]

a) Data Liberate, http://orcid.org/0000-0001-8099-5359

Contact: Richard Wallis, richard.wallis@dataliberate.com

Received: 7 June 2021; **Accepted:** 7 June 2021; **First Published:** 15 January 2022

ABSTRACT

Libraries are generally welcoming organisations and places. Engaging with communities, inviting all comers to immerse themselves in the information rich environment curated for the benefit of all, from the entertainment seeker to the educational specialist. Traditionally this immersion would take place in open welcoming impressive buildings at the heart of the town square or university campus.

However, as witnessed by the phenomena of the declining town centre and the lockdown Zoom culture of 2020, traditional routes to resources are changing rapidly. In the online discovery and delivery world that has emerged, metadata especially quality metadata, about resources and information is key. Without a detailed understanding of available resources, it can be difficult if not impossible to direct them towards those that might benefit from reading, watching, analysing, interacting with, or purchasing them.

KEYWORDS

Bibliographic data; BIBFRAME; Schema.org.

Hello everybody. Thank you for the organizers of this conference and for inviting me. I hope you'll find this interesting. So here we go. First, if you've never met me. I'm an independent consultant. I have been around computing, far too long to own up to, but involved with cultural heritage technology for a significant period of time and with the Semantic Web and Linked Data since they were first introduced.

I have been involved in the W3C consortiums heading up community groups mostly around bibliographic, archive, financial data etc., and the standard schema.org – of which I'll come on to in a minute.

I work with various organizations. I work with Google (not for Google) helping them to contribute to the open schema.org[1] vocabulary project. I have a lot of involvement: making sure the site still runs in that area, extensions, documentation, community engagement etc.

I have worked with my previous employers (OCLC), financial industry people and with various clients that are relevant to this conversation. I've worked with The British Library, Stanford University Law, Europeana, and National Library board of Singapore.

The reason I am here is to talk about the way we may have to change our approach.
I am using the analogy of libraries all the way through this conversation, but it could equally be archives, it could be museums, it could be aggregators like we heard about it in the previous presentation.

Libraries have a reputation of being welcoming places usually in settings of imposing or inviting buildings in town and city centres or having an imposing but important place on university campuses. Within the buildings offering the right sort of environment for people to read and study etc. That includes more social spaces, often found on university campuses. We reach out to plus possible users at an early stage inviting schools into libraries etc. Once people are into a library. Once through the door it becomes a little intimidating when you first come in, because your first challenge is to find stuff. Traditionally this was done within impressive wooden sets of drawers with catalogue cards in. Those catalogue cards evolved into some standard formats that were used – often a little obscure to the users – but the librarians were usually quite pleased with these.

When technology turned up that could help us, in the 1960s, libraries adopted it very rapidly; which led to the arrival of the MARC record card this introduced cataloguing data standardization.
We needed standardization so that the computers could work across the data and understand it; and let us build some systems; and those systems enabled us to roll out the catalogue for people to search and interact with, it well beyond those little wooden drawers across the library and very often into the outside world.

[1] https://schema.org/

Talking about the outside world…
During this period our world has changed. You only have to interact in the outside world, and you find that we're moving away from the traditional town centre, or city centre, and moving towards shopping destinations or entertainment destinations etc.

We saw the growth of out-of-town shopping centres, which introduce efficiencies for the retailers and a destination for the shoppers. This has evolved even further to online retail, which has delivered massive efficiencies for the retailers, and kind of removed the environment that the library used to interact with and led to what people have christened the death of town centres. The inevitable move to an online culture has been what we've been readily aware over the last 12 months. We have moved into the Zoom society, where people interact for business meetings or family occasions etc. This has been exacerbated by recent lockdowns.

Libraries started to react to this move by starting to reach out even more, and become attractive – be it in the public arena or in the academic arena – providing social spaces and traditional non-library spaces (often becoming as concerned about the quality of the coffee as the quality of the reading materials). This translated online as well. So, the initial computerized catalogues were fairly dry affairs that emulated the original card catalogue. We started to add book jackets and links to other resources. Following on with, what traditionally got called web 2.0
Standards, where it became far more graphical and started to emulate the online retailers. Not only in the look, but in the searching capability. Moving away from the traditional keyword within title searching that was available, towards entity-based searching. So, searching often returns now sets of authors or organizations or works or articles or whatever.

Equally following the evolution of the look and feel, of the rest of the world, it started to move in some libraries – this is the public library of Oslo[2] – to a much cleaner much more graphical environment. As our potential users started to understand these environments, because they're using them all day every day for their social networking and other things, and if you take this sort of approach you can start moving into a very visual environment that would operate on a an iPad or a mobile phone of some sort.

What's going on behind the scenes to enable all this?
There was a set of developments where we started to inherit that the technology for the rest of the world. The move from MARC to MARCXML enabled us to use standard programming tools to start working with bibliographic data. Starting to use in some circumstances simplified vocabularies to describe. We introduced textual indexes to enable the interpretation of the material beyond what were the capability of relational databases, and with the introduction of Linked Data RDF (the Resource Description Framework) started to enable us to move in a Linked Data direction.
The introduction of BIBFRAME[3], from the Library of Congress, in about 2011-2012 which was

[2] https://deichman.no/sok/abbey
[3] https://www.loc.gov/bibframe/

the first approach of the very detailed bibliographic vocabulary to describe bibliographic resources as linked data.

RDA picked upon the linked data theme and started to introduce a Linked Data version of RDA along the way and, more recently, BIBFRAME 2.0 came out which reflected some of the challenges that were encountered with the initial BIBFRAME, making it easier to operate within a linked data world.
So that's what's been going on in the library world. What's been going on in the wider world?... Well, the wider world has been taking on something that they call Structured Data.
Why are they doing that?...
Well, there's two driving forces for it. There's the search engines, Google and their colleagues, who have come to the end if you like of the capability of being able to confidently mine textual data on the web page, to work out what the thing was and its attributes that the page was describing. Equally the publishers of websites had hit the point where they wanted to be
able to more accurately describe their resources, or things, to the search engines.
Early attempts included: calendar and business card formats, that could be embedded in the page, Google had an attempt at an open vocabulary called data-vocabulary.org. Eventually in 2011 Schema.org arrived on the scene. Backed as an open project by Google, Bing, Yahoo! and Yandex. This introduced the standard that has since been taken up by others like Facebook, Alexa, Apple, Pinterest etc.

So, what is Schema.org?...
It's an open vocabulary for the web, a Linked Data vocabulary (although though they don't shout about it too much), RDF based. It's got well over 2000 terms in it (778 types, 1383 properties in the last release). It releases every two or three months, so it evolves. Basically it means, in the wider world, most things have got a type, in Schema.org, that can be used. Things like creative works (CreativeWork), persons (Person), volcanos (Volcano), libraries (Library), medical procedures (MedicalProcedure), books (Book), etc. And this wide vocabulary and ease of use has enabled it to deliver a significant penetration
There is an open crawl project that happens every year, in 2020, round about September time, they did a crawl and they crawled 3.4 billion urls that were held on 34 million domains – so quite a significant chunk of the web. They identified that 44% of those domains had structured data embedded on them (mostly Schema.org), and 50% of the pages.[4] So 50% of three point four billion pages, had Schema.org or similar embedded on them.

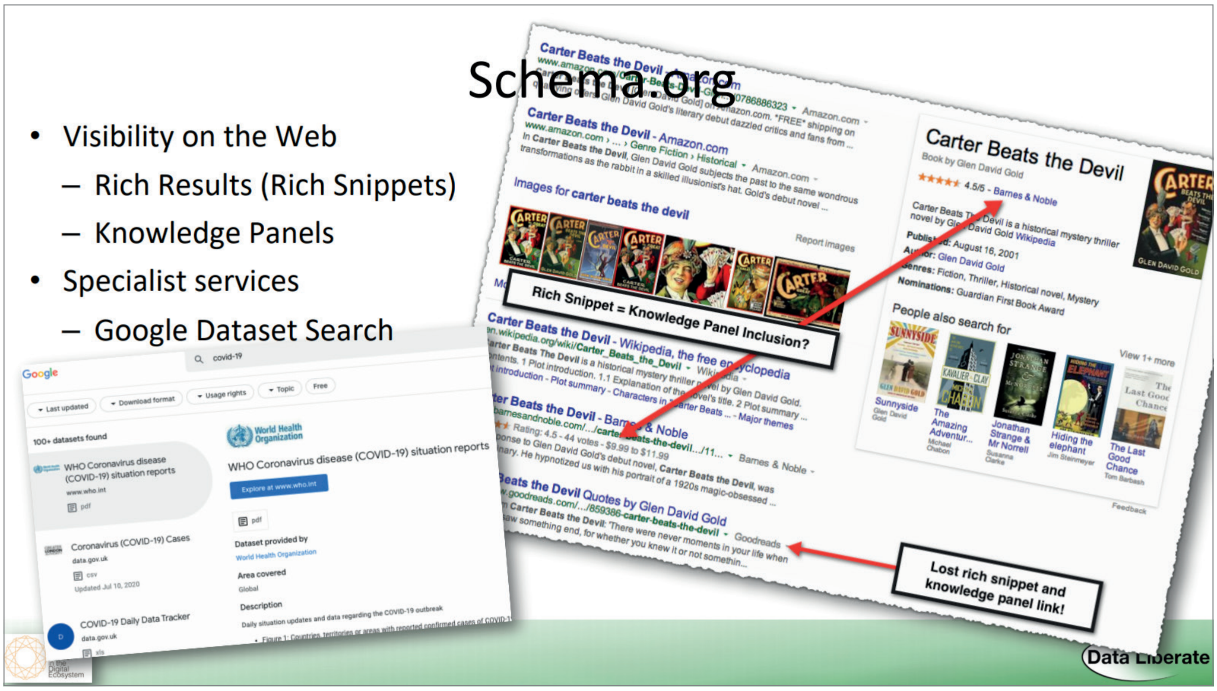
The key is it's embedded in the HTML so the publishers don't have to do anything technically clever. They don't need specific endpoints to query the data. They just embed it in their normal HTML website in one of three formats to choose from (Microdata, RDFa, and most popular nowadays the JSON-LD)[5].

It gives you visibility on the web. What does that mean?...

[4] http://webdatacommons.org/structureddata/2020-12/stats/stats.html
[5] https://en.wikipedia.org/wiki/JSON-LD

On a search engine, you can obtain rich results or be part of a knowledge panel. This screenshot from Google (see Figure 1) is showing you that the bottom but one result is a rich result so it includes things like ratings and pricing information. They often include an image etc whereas the one below it for the same item is just a boring ordinary listing. Whereas in the knowledge panel, the data harvested from all sites describing this thing is an aggregate representation of what that 'thing' is about and what it's related to.

It also drives specialist services. I'm only going to pick up one now because we haven't got a lot of time. That is Google's Dataset Search (see Figure 1), which is a specific search and looking for data sets that are openly shared on the web:[6] You see a lot of Covid-19 data sets are being shared at the moment, academic data sets etc. The key to this is, unless you embed Schema.org in the page that describes that data set, and where to get it, you almost certainly won't end up in Dataset Search.



Schema.org

- Visibility on the Web
  - Rich Results (Rich Snippets)
  - Knowledge Panels
- Specialist services
  - Google Dataset Search

Rich Snippet = Knowledge Panel Inclusion?

Lost rich snippet and knowledge panel link!

Fig. 1.

So, what's going on under the hood on the websites that are doing this?...

They want to describe their 'things', they want to describe their products, their events, their services, offers, articles, persons and organizations – that are for sale or to lend or whatever. To do that they have to mine their data. Quite often that's done by just updating the software that builds the HTML page to encode the data, that the page is about, inside the HTML. Or sometimes it's extracted from databases, and APIs are used to create a Schema.org structured payload that gets embedded in in the page. The search engines and others – it's open for anybody to crawl – extract the HTML from the pages, and from within that, extract the Schema.org structured data for use.

[6] https://datasetsearch.research.google.com/

So, we have got different data practices:

- *libraries* use Linked data -- the *web* use Linked data;
- *libraries* use detailed standard vocabularies (RDA,BIBFRAME, etc) - whereas the *web* is using a common global general purpose vocabulary, schema.org;
- *libraries* quite often use this structure [Linked data] to make it easier for them to link often externally -- the *web* is, almost by definition, totally externally linked (even within your own website the linking is identical) So whether you're linking to a Wikipedia article or another page on your own site it's the same principle. [an entity linking oriented];
- the *libraries* have used this to start delivering enhanced discovery service interfaces, entity based local searching (such as I exampled earlier on), improved detailed display (so it's there to improve the discovery interface) -- whereas on the *web* output in schema.org gives in enhanced data for search engines, rich results display, representation in knowledge graphs. Which almost by definition means you're far more likely to receive accurate links from the search engines into your resources.
- for *libraries*, the standards and the uses are for libraries and partner libraries only (or things like aggregated catalogs which we saw earlier on) -- whereas out on the *web* the structured data is for growing global representation and linking.

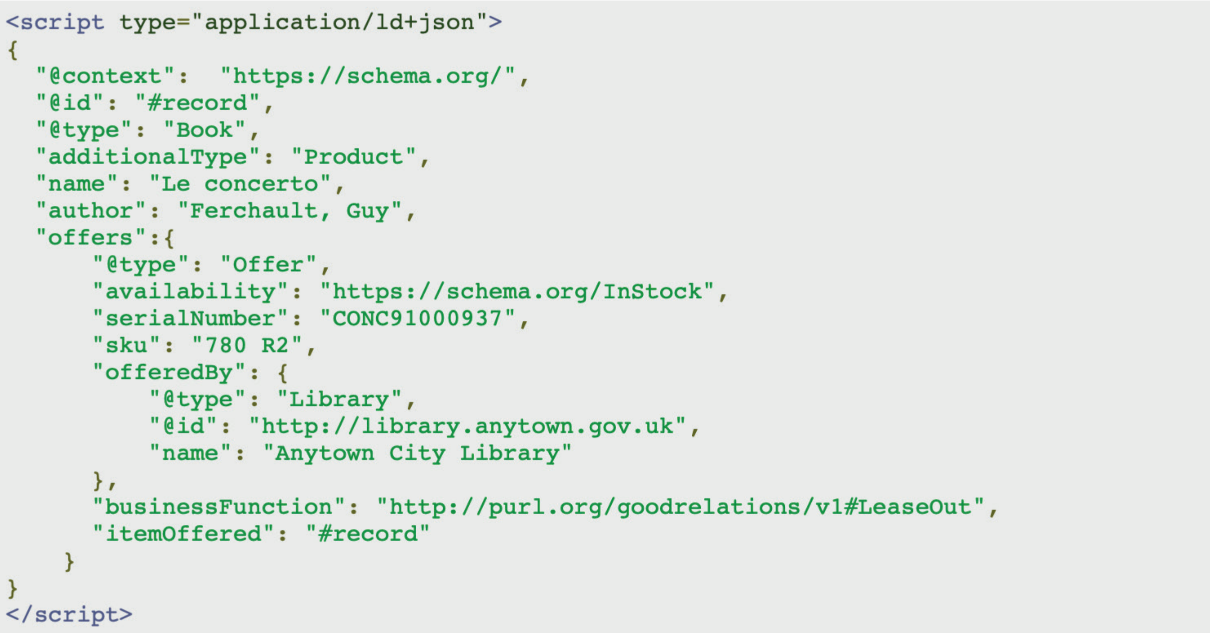
So, what are libraries doing on the web at the moment?...
Well mostly (there are some exceptions), basically the web knows about your discovery interface (hopefully) or (more likely) the homepage of your website. Not the things you can discover using that interface. Users do not start their discovery journey in your interface, they're not looking for you (the library), they're looking for the resources that you can provide.

So how do we get our resources visible in the web?...
The answer is quite simple: start sharing Schema.org data from our discovery interfaces.

The answer might be simple, the implementation might be a little bit more of a challenge but more of that in a moment. Let me show you an example of bibliographic data on the web[7]. This is the National Library Board of Singapore (I have worked with them for many years). Here what they have done is, taken every catalogue record from their library system, and produced a static web page that describes it – very, very catalogue card-ish I would suggest. They've enhanced it fractionally by adding an image (if they've got one), and a link to the library system it came from. There is no search interface for this. There are just thousands and thousands of static web pages on a website. But, embedded in those pages is Schema.org. This is the structure that's in there describing the entity, and if you want to actually look at the JSON-LD that's embedded in the page – if you're that way inclined this is what that JSON-LD looks like[8].

[7] https://www.nlb.gov.sg/biblio/12343857

[8] Example from https://schema.org/Book



```
<script type="application/ld+json">
{
  "@context":  "https://schema.org/",
  "@id": "#record",
  "@type": "Book",
  "additionalType": "Product",
  "name": "Le concerto",
  "author": "Ferchault, Guy",
  "offers":{
      "@type": "Offer",
      "availability": "https://schema.org/InStock",
      "serialNumber": "CONC91000937",
      "sku": "780 R2",
      "offeredBy": {
          "@type": "Library",
          "@id": "http://library.anytown.gov.uk",
          "name": "Anytown City Library"
      },
      "businessFunction": "http://purl.org/goodrelations/v1#LeaseOut",
      "itemOffered": "#record"
    }
}
</script>
```

The search engines are pinged to say these pages are available. They provide a sitemap[9] to tell the search engine where to crawl and then they leave it to the search engine.

So, what's the effect?...
Well here we're seeing the effect. This is a snapshot out of the Google search console which is reporting traffic to that site. It's a 28 day period and in that period 1.58 million times one of those pages appeared in a set of search results. 61,000 times somebody clicked from one of those search results through to the site, and many of them clicked on to find in the library etc.
So that's a 3.9% click-through rate which – if you speak to any SEO expert – is not bad actually, especially for a static page – that hasn't got any kitten videos or similar.
So, this is something I believe most libraries would like to do. But this delivers a bit of a dichotomy if we use BIBFRAME.

BIBFRAME is a library standard, it's replacement for MARC, it's led by the Library of Congress, system suppliers are investing in it (at least importing or exporting BIBFRAME), it benefits the local interface, and libraries implementing this kind of thing see it's a significant step forward. A step forward, that is a development goal for many libraries and aggregators. So that you could almost do a very similar list, about most of the new technologies that are being worked on in the library world.

Whereas Schema.org is a global standard, it's not a library standard, it's backed by the search engines (and others of course), library suppliers are kind of looking at it but not really investing heavily, it benefits the global discovery and linking of your resources (not the local interface). It's a different step forward, and at best appears to be on the agenda in most libraries as a 'nice to have'.

[9] https://www.sitemaps.org/it/protocol.html

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12762

So, when I'm talking to libraries about trying to attempt to do what I am describing here I get answers like: "*well, BIBFRAME is taking our current focus*", "*schema.org is a different data model*"; "*we can't do both*" – well maybe we can.
As a world we're investing in linked data, it's the subject of many, many presentations on conferences like this and BIBFRAME tends to be the default Linked data standard for sharing your library data (there are others, don't shout at me in the chat).

We can build on this investment: not replace it but add in something like Schema.org on the back end of it. This is the subject of a W3C community group entitled Bibframe2schema.org which I chair. The objective is the creation of a reference mapping from BIBFRAME 2.0 to Schema.org, and the development and sharing of reference software implementations for people to copy. This site is very small at the moment but, if you look on the comparison viewer it will bring in BIBFRAME 2.0 records and demonstrate an initial prototype conversion to Schema.org – so that we can see what the effect would be. So if we could reproduce this, if we could produce Schema.org into our discovery interfaces, so that the search engines and others can crawl it; we have the digital way to share digital breadcrumbs across the web, to draw people to our resources.

They don't have to find us first, and then learn how to use our specific interface. Their day-to-day tools, their questions to Alexa etc, should be able to pick up these breadcrumbs wherever they may be. To deliver the value of your resources, that you're spending a lot of time in an effort in encoding, and building standards around them and sharing in your own interfaces. Most of your users want to be able to get that at them from where they get up in the morning if you like. So, to be visible on the web we need to get our internal Linked data right.

BIBFRAME is a good candidate for this (not the only one). But we mustn't expect the rest of the world to use our vocabularies. Having a fully Linked Data catalogue is not going to do a lot, for people finding your resources across the web. Outputing the global de facto standard vocabulary Schema.org, *as well as* our relevant detailed vocabularies make this, as a community, easy and consistent for developers and implementers.

So, as the BIBFRAME world implemented MARC2Bibframe, which is a piece of software that you can use which will take a MARC21 record and produce Bibframe 2.0 data; equally we should be able to take Bibframe2Schema.org outputs and produce software that will take, the already produced BIBFRAME data and translate it into Schema.org terms which we can then fairly simply embedded in our user interfaces.

And to make this work we need, as a community, to participate in the community groups, participate in the discussions – participate in the web so our users can find the resources that we actually have on the shelves, and on our disk drives, for them.

Thank you very much.