Linked open data on its way into next generation library management and discovery solutions

Axel Kaschte

Ex Libris is known as an innovative company - we really embrace new technology. And linked data is a very new revolutionary technology. We have seen in the agenda of this event that there are many aspects of linked data and I have tried to put some light on what a commercial company like Ex Libris looks at in this new field. Whenever I look at something new, I remind myself of the Little Prince. He has this special capability of looking at certain things with fresh eyes. You probably remember this question, «What do you see here?»while the Little Prince is showing a drawn picture. Most people will answer immediately that this is a hat. But we all know (from the book) what the answer is – a snake that has swallowed an elephant. Now the next picture I show to you is this curve, looking very similar to that hat, and as you can imagine, I now put the question to you, «What do you see here?»It's not the snake, this much I can tell you. It is a bell curve. To be more precise, it is the bell curve of the adoption of innovation from Rogers Everett (*Diffusion of Innovations*) who invented this way of presenting this information, it demonstrates how many usages of new technology over time will happen. So you see in the beginning of a new tech-

nology, we have 2.5% of innovators. These are libraries that get new technology going, so really they are trying to invest in new things. Early adapters are the next phase with 13%. It's already a remarkable percentage but it's still called innovative. Then we see the early majority, the late majority and so on. Just to give you some examples of products we know in these internet times, there are a TV and a newspaper. If you are on the right hand side of the bell curve of technology adoption it doesn't mean that it is not used - quite the opposite. It means that everybody already has it and is using it. The bell curve refers to the growth rate of how many users are added each month. Facebook is on top right now, which means that the growth rate is still dramatic but there will be less and less new people coming on board. I mean, there are already 600 million users already. You also see also new websites such as Vimeo, and technologies like the iPhone. Blackberry is a little more on top; the iPhone is newer. Amazon Kindle, the e-book reader, is only in the early adapter phase. It has been on the market for the past three years but the adaption curve shows how much is still ahead for this product. You see also there is profit opportunity mass market. The point is now, as we a commercial company, of course we are interested in money, but also we are interested in serving customers like you with commodity services. Whenever there is a technology and the aim is to make it available to many, to really make it available as a cheap solution, then it is a commercial company that has the best model. How does this translate to libraries? Integrated Library System (ILS), integrated library systems, were invented a long time ago, some 30 years or more, and you see there are still some libraries, especially in Asia, who do not yet have a library system. This explains why the growth rate of new libraries is slow but these are still state of the art technologies and they are still in demand. Then you have things like meta-search. You have things like link resolution.

It's very interesting that, if you look at our statistics, such a large number of libraries are still purchasing link solutions. There are many who have not yet entered into the area of electronic resources but are either in the process of doing so or will in the near future. For that purpose they will need a link resolver, which explains why it is still in this high area of the bell curve. Let's have a look at Discovery, which is the next generation of OPAC, where you see search engine technologies coming into play: discovery is a little bit before the top of the bell curve. We just look at our statistics: Primo, as our discovery solution, is growing rapidly but is not at the top yet.[1] And you can also see central e-resource indexes like Primo Central, a mega-aggregation of electronically-available articles for research, for scholars, and so on. This is something which started just two years ago and it has been adapted very quickly and we are in the phase of rapid growing. These are just product categories. I will now put the technologies next to them. ILS is a technology, which is not growing anymore. Search Application Programming Interface (API), Open URL, search engines, and then cloud, as an e-technology which enables this kind of service to have one central index for every library and also offers increased cost-effectiveness. This highlights why commercial companies are so good at these models; if they can provide a solution very cost-effectively to many libraries, this is the model libraries should use. In this model there is also some more details to share with you. Geoffrey Moore (Moore) introduced the concept of a chasm. There are many products who are very much in the innovation phase and who will never make it as mainstream products. There fall into the chasm. Two of the things Ex Libris invented fell into the chasm: an ERM solution - probably many of you have heard of Verde - and a digital asset management system. If you look, it's not just Ex Libris who failed to deliver these

---

[1]http://www.exlibrisgroup.com/category/PrimoOverview.

to the mass-market. There are a few hundred using it, compared to four thousand using Aleph worldwide. It's not growing anymore. The chasm is just a way to present the certain technology which is not meeting the needs of the library in the best way: it's never reached to be a best practice. I'm just being very honest. You have to analyse as a commercial company and be able to say, "OK, that was a mistake. It was money spent that was not good for all of us, neither for you nor for us". What we are now introducing, and the whole market is following this idea now, is a solution to the problem of automation in silos. If we go away with the silos, if we go unified - meaning there is one solution for your print management and your e-resource management and probably also your digital assets management - meaning if you introduce one single environment, we find that this is very much what libraries want today. And this is what we are in the initial phase of doing right now; our first customer will probably go live with it next month. It's the software called Alma in our case, and it is cloud technology which allows it to be done. I showed you all of this because I want to bring to you the idea that commercial companies have to look very close when it is the right time to get on board with a new technology in order to make a mass product out of it. Just look at certain other technologies which are established in ILS, we have heard about them today: AACR2, MARC21, Union catalogues, authority files, they have been around for quite a long time. We have also heard a lot about the emerging technologies like RDA. Alan Dunskin from the British Library talked about it and we listened to him asking, «Can you please help us to close the gap and get it used; provide the tool set where the cataloguing happens». In other words, it is the right time and Ex Libris will look into providing the necessary tools in its applications soon. Then we have Resource Description Framework (RDF), which the whole seminar is talking about. And

we have open data sources. All of these new technologies are very far on the left of the bell curve. These things are really still in the research phase, as are many aspects of linked open data. We have seen these schematics of objects and their relations; we have books, we have paintings, we have authors and painters. We have these objects and the creators of it – and many more relations. Library data was already highly linked in the past in certain ways. It was not open, it was not using URIs but inside of the systems it was already linked. So if you look at solutions like discovery systems (e.g. from the Austrian Union Catalogue in Vienna) you see things where you can click at the author and get all the manifestations of the works from the author, you have the same for subjects. So you see links are there and you can navigate them but you only stay inside the environment of the library. They also have already a permalink to the manifestation. With this link you will always get to the same point. It's not yet a URI because you don't get to the data, you just get to the same page but it's at least this permanent way to get there. Building an API to just give you the data in a structured form is just the next step. But now there is one point more you can see – this is what they have done in the Austrian Union Catalogue by including Wikipedia as a data source outside the library metadata and they use the authority record with its identifier in there to link into Wikipedia and if you click on it you get some information from Wikipedia. A very simple example, it seems. There are several such examples in various discovery solutions in various libraries. These are not yet using the true URI mechanism. The links are constructed on the fly and it's something which just works because the discovery platform and the data underneath allows to present this to the end user. With this "experiment" in place we can have a look at the acceptability by the end users. Is it something they actually want? Is it something they actually click on? And if they don't, we don't bother. So you see

this is the kind of thing we try to do here in this research phase. So what is actually new with linked open data from our point of view? It's not new that you can do linking between manifestations and the authors. We also have subject linking. We can introduce links between many types of library data but the data structure is highly specialised, no one outside the library can actually read it and it's very difficult to exchange and interact with outside the library. So from my point of view if I have to summarise to an outside person who is not from the library business, what is the important thing for libraries in the linked open data theme, it's making the library data available to the outside and maybe even more important, take library data from the outside. In other words, make the library world part of the all-embracing World Wide Web. This is reflected by the work of the World Wide Web Consortium, W3C. It has established an incubator group to look at linked data in libraries and related in software developments in May 2010. This group was looking at real use cases and submitted theLibrary Linked Data Incubator Group Final Report.[2] These use cases are about getting library data into the linked data world. We have seen several of these cases in presentations during this seminar and I will try to summarise them here by putting the various different use cases into a very simple pattern of 3 areas of work. But first we follow what the incubator group did, they categorised all use cases into 8 groups. The first group is about the handling of bibliographic data, bringing it over to a linked data scenario. For example, British National Bibliography, Bibliotheque National France, Bavaria State Library, the Open Library. The second group is about authority data, same institutions but different data. This is just repetition of things you know to get you to a point where you see a pattern here. Third group is work on vocabulary alignment. Many presentations have been done during

---

[2]http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025.

the seminar on that. Fourth group is about archives working on getting their data into the linked data cloud. Europeana – interestingly enough – is mentioned in the archives group although they are probably working in all the groups. Fifth group is about citation of scientific data sets being expressed in link data, a very new thing, it wasn't done so far. It is now helping to enhance publications, which means that publications come already with metadata about the research data used. The sixths group is about digital objects in the library world. The goal is to provide a digital text repository as linked data so that the metadata, the text and the extra objects the text is referring to are provided in one comprehensive format. Here we have use cases from outside the library world, the UK open government data initiative. It provides many examples where you can draw data from and see how they are interlinked. The sevenths group is about collection building. Librarians have talked about Functional Requirements for Bibliographic Records (FRBR) structures, in which the work level is the highest level. But what if you go even to a higher level and start describing collections. There are already use cases trying to define collections in data sets and more. The eights (and last) group of use cases is about social networks and cross linking environments. The uses cases in this group seem to be not be related to a classical library view but the typical users of libraries are very active in these areas exchanging information especially about the literature they are using, like e.g. via Mendeley. All of this work in the 8 groups, when looking at it from a little distance, translates into three main work areas.

- One area is the data preparation. Creating the data which needs to be there in linked structures to be able to use it. This area of work is about creating tools to be able to handle mass transformation and mass storage with high performance.

- At the same time there is the area of the definition of the rules

to transform the data. It's kind of an interaction between the two. There are projects which have tried to put data into link structures but then they get some experience and they need to say "no, you have to change the transformation rules, we have to do it again", so we are in a very frequent iteration process right now.

- At the end, all that matters is what really comes to the end-user interface and how they can they make more use of it than today, which is the third area of work.

Somehow it seems that there are far more projects about getting data into the linked data cloud than there are projects about what to actually do with this data what could not be done before. This and the high frequency of changes to the definition rules of transformation result in our conclusion at Ex Libris that linked data is still in research mode. One of the research examples is Europeana, we have heard about it just today during the seminar. Europeana have a website, it's a productive site - why do I call it research? By looking at the problems they face. The central Europeana portal is not able to deliver state-of-the art performance. The problem arises by sticking to one of the main ideas of linked data – to link data of various sources together. These sources are in fact data silos and to make them discoverable from one central place one has to do a federated search. We do have the experience of the last decade of doing meta-search in library databases, in order to create performing solutions one has to create a central index. Central indexes means harvesting from the various sources and that actually means a variety of source formats, and most important versioning. In the RDF world this is highly problematic and in my view an area of research. This is all known to the very knowledgeable people working at Europeana. In order to make progress in this area a new a European Community funded project has started just this March. It's called

DM2E Digitised Manuscripts to Europeana. The major part is to digitise more materials and to get it done quicker and to easily create metadata. However, work package 2 is about interoperability infrastructure. Because many of the institutions who do the digitisation are libraries, they have library systems and use them to collect the metadata in classical library format like MARC. Because Europeana uses link data structures, a robust RDF transformation toolset will be created as part of work package 2 (WP2). Ex Libris is a partner in WP2, taking part in this research and actually creates products which will be open source and which will be possible to add to an existing library system. This tool will take e.g. MARC-XML and transforms this into RDF. We have already talked about similar examples like the British National Bibliography, we hear this afternoon how the Bavarian State Library has done it. However, these examples are not using common technology; it is something which is in an experimental phase. The tool which is created in WP2 will allow various input formats like MARC21, UNIMARC, DC, MODS, transforming this into a RDF presentation, which in essence is just a different transport format. As a second step, a transformation into the Europeana data model will be done. Both of these steps are based on mapping rules and actually the task is here to make it very easy to change these mapping rules because we are in the phase of defining the vocabularies and that's why we would like to play with it. Currently we still do not know the definite vocabulary that should be used, in every project mentioned in the use case report mentioned above, they use a different ontology. Creating a tool which allows to play with the ontologies is our contribution to the current research phase. To summarize – Why should we as Ex Libris start investing in products using linked open data technology? It is because of interoperability, especially with other domains, in the discovery sector. It's probably reshaping metadata management –

cataloguing will most likely look totally different, it will be more about including external resources as links, as it is typing data.

References

Moore, Geoffrey A. *Crossing the Chasm*. New York: Harper Collins, 1992. (Cit. on p. 315).

Rogers, Everett M. *Diffusion of Innovations*. Glencoe: Free Press, 1962. (Cit. on p. 313).

AXEL KASCHTE, Ex Libris Group.
axel.kaschte@exlibrisgroup.com

ABSTRACT: Library Linked Data Model is an important topic for librarianship and it is equally of interest to the many organizations that provide products and services to that community. Ex Libris, as one of those organizations, frequently gets asked: "Where do we see it fitting into our plans?". In order to be able to answer questions like this, we need to ask: "What exactly are the problems being solved for the profession by this technology that can only be solved with the Library Linked Data model?" What most developers/providers of products analyzing the potential of library linked data would see is that at this stage, this technology is very much in the research stage. The presentation talks about the research ExLibris is involved in and how this can be utilized by innovative libraries to help defining the actual use cases in which the potential of the Library Linked Data Model is indeed exploited.