SIMPLIFYING SEMANTICS: LINKED DATA RULES OK!

Fran Alexander, BBC Information and Archives

1. Introduction

This paper is based on a presentation given to the Organizing Knowledge Taskforce at the 42nd IASA annual conference in Frankfurt in September 2011. It gives an overview of some of the ideas and principles that underpin Linked Data, Open Data, and the Semantic Web. It offers some case study examples and suggests some ways to start thinking about Linked and Open Data projects.

2. Evolution of Knowledge Organisation Systems (KOSs)

Metadata is nothing new. Fashions for declaring the author and title of a work have changed over the ages, but for centuries archivists, librarians, cataloguers, and indexers have been working out how to make sure they know what content and assets they hold and how to make those works or artefacts findable. We catalogue, we describe, we classify. The Semantic Web and Linked Data take those principles into the world of computers and automated processing.

Very early on in the history of human literacy, the usefulness of the **label** became apparent. If you fail to add labels to things, you lose them, or you forget what they are and what they mean. Then people realised labels could be grouped into **lists** (and eventually controlled vocabularies, keyword lists, tag lists, and folksonomies). Some of the oldest forms of writing known to history are lists and catalogues from ancient Sumeria.

Once lists started to get long and unwieldy, people broke them up into sections or categories and formed the first **taxonomies**. Taxonomies were known to the ancient Greeks, although it was the biological taxonomies of Carl Linnaeus (1707-78) that made them a key scientific tool. Taxonomies are easy for humans to understand, especially to provide vertical "drill down" navigation. This means they are used for folder structures, website navigation, and for applying tags. Related taxonomies can be joined together for richer information structuring as **faceted** or **polyhierarchical taxonomies**, labelling different aspects of a concept with different facets, which are useful for refining search results.

More recently it was noticed that if you specified and defined the relationships between the facets (or terms and concepts) as **ontologies**, you could use computers to perform very sophisticated processing of complex queries, including processing Linked Data. Ontologies are often shown as diagrams that look more like topic maps with lots of links and connections rather than a simple tree diagrams.

3. What is Linked Data?

Archivists, librarians, and knowledge managers already work with computerised systems and already use "linked data". Cross references are links. Multiple types of indexes enable us to find resources in different ways, by looking up from different starting points (by title, by author, by date of purchase). There already exist standard formats for creating, coding, and publishing data so that it can be shared, either within organisations or with the wider community, for example MARC (MAchine-Readable Cataloging) or Encoded Archival Description (EAD) — an XML format. Linked Data is another way of making records machine readable and interoperable. Linked Data uses a format called RDF (Resource Description Framework) for publishing or exporting data. Linked Data can be shared easily because it follows several key principles:

- **Use identifiers:** ISBNs identify specific books and primary keys in databases identify particular records. Linked Data identifiers are known as URIs — Uniform Resource Identifiers. A familiar form of URI is the URL — Uniform Resource Locator — used for website addresses.

By using public identifiers, for example to identify concepts in a taxonomy, instead of words, it is clear what you are referring to. An example would be the disambiguation of jaguar the big cat from jaguar the type of car. More significantly, computers can automatically associate every time you have used the concept jaguar (the cat) with uses of jaguar (the cat) in somebody else's data, because the URIs will match, so avoiding the problems of understanding ambiguous words or resolving differences in spelling.

- **Provide descriptions:** When you publish your data, you should explain what your data means. This usually takes the form of public identifiers and an ontology showing how the concepts you have used relate to each other. If you use URLs as identifiers, you can provide extra information about the concept yourself. If you use a published open URI, you simply re-use the work that others have done.

- **Include links:** Providing hyperlinks to your sources, to other similar data, and to other sites that are using your data, and any other resources that might be interesting or useful will also help others to put your data in context, understand it, trust it, and use it.

4. What is Open Data?

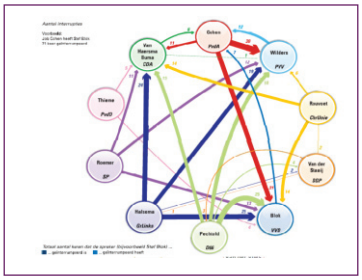
Open Data is data that is available on the open web for others to reuse. It should be published under an open licence (and it helps to make sure that the licence terms can be found easily). The Creative Commons licence is an example of a flexible, open licence. Open Data also needs to be in an open format, so that others can reuse it without having to buy specialised software in order to access the data.

Linked Data only becomes Open Data when you make it available publicly. It is possible to use Linked Data to share data entirely within the firewall of an organisation. Open Data is only Linked Data if it is in Linked Data formats. This is why people talk about Linked Open Data, or Linked and Open Data.

5. Case studies

A growing number of organizations around the world are publishing data in Linked Open Data formats. There is a lot of activity in the academic life sciences community and a large number of libraries are using Linked Open Data formats to publish their bibliographic and cataloguing data. Libraries have a long tradition of sharing cataloguing records but Linked Open Data can provide opportunities to do more interesting things with data than simply merging records databases.

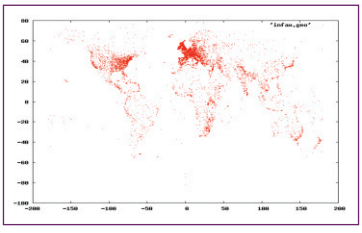
The teams at the Dutch Parliament (Tweede Kamer) realised that their data is ideal for a Linked Data approach, because parliamentary procedures are highly defined and well structured. In collecting the data, they gather a lot of associated information. For example, of every word spoken in parliament they know when it was spoken, by whom, on behalf of which party, in which debate, etc. One of the most popular ways they have visualised the data is into "attaquograms" — colourful diagrams illustrating which MPs interrupted others the most during debates and which MPs were interrupted the most. The Guardian newspaper in the UK used a similar process to produce "League Tables" of MPs' performance, voting, and attendance records.



The UK government supports Linked Open Data as a way of fulfilling Freedom of Information requests, providing public access to public data and governing transparently, and minimising costs of data management and curation. It also hopes to promote the development of small businesses based on making use of the data. The UK government's Linked Data portal http://data.gov.uk contains over 1,100 datasets and has over 1,500 registered developers.

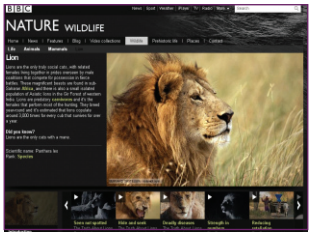
An example of the way people have used government data is a map of bicycle accidents in London (http://citybeast.com/londoncyclists.html). Data from the Department of Transport was plotted on a map, which cyclists can use to plan their routes to avoid dangerous places.

Maps are an intuitive way for people to understand data and can provide a novel way of arranging content for people to find. For example, the BBC Research and Development team took location data from BBC Archive cataloguing records and "mashed it up" with a map application. This plotted the locations of BBC Archive assets on to a map of the world. It shows (unsurprisingly) that BBC assets are clustered in London, the UK, and the English-speaking world. However, it also means that people can choose a place they are interested in and take that as an entry point into the archive, searching for assets associated with that place, as an alternative to using the title or transmission date as the only way to find a programme. Providing different ways of finding assets supports users who are not sure exactly what they want, and is especially good for browsing and exploring online, allowing users to wander about the data without a pre-set retrieval plan.



The *Wildlife Finder* website (http://www.bbc.co.uk/wildlifefinder) was built using Linked Open Data and ontologies. This means that the website is updated dynamically, pulling in data from different sources, such as Wikipedia and the IUCN's list of endangered species in an automated process. By collecting the data from the sources directly, every time the website is refreshed, the updated and latest version of the content is obtained automatically from the source. This means that editors do no have to check for updates themselves. It would be very expensive

to pay editors to write articles about every animal, and so syndicating the free content that Wikipedia and other sources provide is a cost-effective way of giving users extra information. The website carries a disclaimer so that readers know that the content is from Wikipedia, not the BBC:



"This entry is from Wikipedia, the user-contributed encyclopedia. If you find the content in the 'About' section factually incorrect, defamatory or highly offensive you can edit this article at Wikipedia. For more information on our use of Wikipedia please read our FAQ."

6. Challenges

Linked Data and Open Data are not magic bullets. They cannot solve all your information management problems, but they are useful tools for achieving particular goals. The key to using them effectively is to devise a project with a clear scope and purpose that suits a Linked Open Data approach.

Linked and Open Data standards were devised for publishing, not for archiving, preservation, or long-term access. This means that you should consider how to preserve core archive data as an entirely separate issue. In general, the simplest most technology-agnostic formats tend to be the best for long-term preservation purposes. However, making sure that you can export or publish core archival metadata in Linked and Open formats means that you can gain the benefits of joining the Linked Open Data community, while preserving your underlying data sets.

Open Data needs to be data that is free from rights, licensing, data protection, and other legal constraints. However, by thinking through the sources of data that you have, you can usually find some data that you are free to release. Rights and licensing may cause problems for content, but often not for metadata. So, you may be able to publish your catalogue, even if you cannot easily make the assets themselves available. You could not publish the names and addresses of your visitors, but you could publish anonymized data about numbers of visitors or most frequently accessed collections. If you start to think about your metadata as a form of content in itself, you may find you have much data that you use operationally and are free to release. You may not be able to think of a use for it, but if you release it into the Open Web there is a chance somebody else will be interested.

Linked and Open Data and Semantic technologies alone will not solve your metadata curation and governance problems. They may in practice make such issues worse. If there are flaws and inconsistencies in your underlying data, publishing the data will reveal those flaws. For example, if you started collecting date information in US-style month then day, then a few years later switched to European-style day then month, then a few years later to a fully numerical format, your date information will be inconsistent and automated processing of dates will generate errors.

Linked Data depends on using identifiers so that data can be mixed and matched, but making sure that those identifiers are kept up to date and consistent requires some curation and governance effort. If you want to merge or mash up your data with someone else's you will need to check that they have used terminology in the same way that you have. For example, suppose you have some data on the number of international visitors to archives and you find a data set about the total number of tourists who visit European

countries, you could make an interesting comparison to show whether there is a correlation between total numbers of tourists and visits to archives. If both data sets include "Northern Europe", before you identify your concept of "Northern Europe" as the same as "Northern Europe" in the other data set you should check that both have included the same countries or parts of countries under "Northern Europe". If the definitions do not match, the figures will not be comparable and mashing up the data will generate nonsense. So, you cannot assert that your identifier for Northern Europe is the same as their identifier for Northern Europe unless you have checked that they really do refer to the same things. If you subsequently decide to include a different set of countries in "Northern Europe" in a new version of your data, you should also adopt a new identifier as you have effectively created a new "Northern Europe". This sounds complex, but it is a natural continuation of the problems of maintaining consistency in indexing and vocabulary control that cataloguers have been grappling with for centuries.

Another challenge is finding technical staff to help you put your project into practice. However, the technical issues tend to be easier to resolve if you have a clear idea of what you would like to achieve with your data. Existing technical staff may have enough knowledge already or be willing to find out what they need to do and there are many free tools and sources of support within the Linked and Open Data community. There are consultancy firms which can help with specific projects, and individual freelancers or consultants are another option, but it is still useful if you start with an idea of what you would like to achieve. Finally, if you have no funds, publishing your data in whatever format you can and putting out a call for help is worth a try!

A Linked Open Data project is a publishing project, and so needs to be considered and assessed as such, and as with any project the key to success is to define a clear scope and aims. A small, well-specified pilot project is often a good way to start.

7. Opportunities

Although Linked and Open data projects require some thought and effort, they provide interesting opportunities without the need to spend lots of money on software or infrastructure. Even if you only want to release your data in a constrained way, to share amongst a group of universities, or even a group of departments within your organisation, Linked Data principles may help break down content silos.

By openly publishing your data, you promote your organisation and use and re-use of your assets. In the commercial world, catalogues are given away free because they are seen as marketing. Shopkeepers do not consider charging for their catalogues, because their catalogues and metadata are the way they draw customers into their shops, let people know what is available, and advertise and market their wares. Libraries and archives are beginning to see how releasing their cataloguing metadata can work in just the same way, showing people what they can find and encouraging them to engage. When that metadata is published in Linked and Open formats, networks of catalogues can be built up and associated so that when someone searches for something in someone else's catalogue, they can be given a link to related things in your catalogue. By co-operating in this way you provide a richer service to researchers, some of whom may not have known to look for you directly. In the past, when you had to physically enter an archive or library to access content, there was perhaps less appeal in reaching out to a non-local audience, but online access puts you in reach of anyone anywhere in the world. Now, a researcher in Poland might be delighted to find their local archive's data linked to that of an archive in New York or Sao Paulo or Sydney.

As well as making your data available for others to use, it is worth thinking about how using the Linked Open Data sources that are freely available may help you. One example might be that you would like to add biographies of authors to your catalogue. Instead of commissioning writers and editors to research and write a set of biographies for you, you could look for a Linked Open Data source.

Other benefits of publishing your Linked Data openly are that you can use comparison with other data sets as a way of discovering buried inconsistencies and quality problems in records and benefit from free "crowd sourced" help in spotting and correcting errors.

Crowd sourcing is a recent term for another familiar idea — accepting comments and corrections from your readers, visitors, and other members of the public. There have always been people who have written to archives, libraries, publishers, and museums to point out errors and omissions or make suggestions and volunteers who have offered to work in the archives for their own interest and desire to participate in something culturally valuable. In the past, they had to do this by working in the archive in person or by writing letters, which required quite a high level of commitment. Online, it is very easy to fill in a form or send an email, so it is easier to get people to contribute. Online, anyone who has access to a computer is a potential visitor, so this increases the number of people who are likely to find and use your data and it is easier to gather communities of specialists, as you are not restricted to only those people who can visit in person. In addition, people enjoy contributing and if you can find ways of making contributing fun — such as by turning improving your data into a game — you are likely to attract more helpers. An example of such a project is the Transcribe Bentham project at University College London (http://www.ucl.ac.uk/transcribe-bentham/) where volunteers type and proofread sections of Bentham's handwritten letters.

Finally, it is worth understanding how new techniques and technologies, such as semantic technologies work, so that you can decide whether or not they would be useful or helpful in other business processes. For example, as speech-to-text processing software improves, it may provide a way of generating transcripts of audio content that has not been cost-effective in the past. Such techniques will result in large text repositories that need indexing and linking, and automated semantic techniques may be the easiest way to do this.

8. Some questions to start you thinking

In order to set up a successful project you need to think through what the challenges and opportunities mean with regard to your particular organisation and circumstances and then to begin thinking creatively about the data you gather and store as an asset in itself.

Here are some suggestions of questions that are intended to help inspire ideas. They are not intended to be a definitive checklist, as no two projects or two organisations are the same, but they highlight general areas that are often worth considering.

- **Who are our key customers?** By thinking about your customers, readers, users, and audiences, you can think about the sort of data that they need and use and what other data they might find useful. Are there questions that are frequently asked? Are there any questions that they ask that you find hard to answer? Would they like more maps, or biographical data, or different kinds of indexes? Is there a particular institution, publication, or source that they use alongside the sources that you provide?

- **How could we present our data in different ways?** Do you provide lots of alphabetical lists? Could those be re-sorted by theme, or chronologically, or vice versa? Could some of your data be visualised on a map or a timeline? How can different types of data be connected?

  Many of the techniques that are used when creating exhibitions and special collections apply just as much to metadata as to the assets themselves. Do you have information about birthplaces of authors stored separately from information about events that happened in those places that you could link to place authors in historical context? Could you bring together catalogues of different types of assets and find connections based on indexing terms, or date of acquisition, or how often an artefact has appeared in exhibitions?

- **How can we revitalise legacy content?** Do you have indexes and catalogues or other sources of data that are hardly ever used or referred to? Why is this? Is it because they are hard to access? Are there collections that are hidden because they are not well indexed and so worth opening up to the public to see if anyone actually is interested?

  One of the benefits of choosing some little used or neglected legacy content in a pilot project is that it represents a low risk, but potentially big reward if the project is successful.

- **What data is out there?** Simply finding out what is available for reuse, which other organisations are publishing Linked and Open Data, and what projects are being undertaken can serve as a source of inspiration. For small organisations, offering some data to a larger organisation to include in a specific project could be a very cost-effective way of getting technical support and practical help as well as gaining knowledge and skills.

9. Conclusions

The world of Linked and Open Data is one that we have been working in for years. The difference now is that we have new standards and formats and new ways of using our data with the aid of powerful computing. Although there are pitfalls to avoid and obstacles to overcome, if you find out what Linked and Open Data can and cannot do, you can start to think of projects that could benefit your organisation. As information professionals, we should be the ones who best understand our metadata, best understand how to evaluate external sources, and so are best placed to devise innovative and interesting projects to promote our content, our data assets, and our organisations and institutions.

Fran Alexander is Taxonomy Manager, BBC Information and Archives. In 2009 she was awarded a Master of Research degree by University College London, her dissertation proposing a framework for assessing the subjectivity and objectivity of taxonomies, based on original research into 15 major commercial and academic taxonomy and classification projects. She blogs at http://www.vocabcontrol.com/.

Some further reading and resources

An Introduction to Linked and Open Data for Information Professionals:
http://web.fumsi.com/go/article/share/64146
Linked Data is blooming:
http://www.readwriteweb.com/archives/linked_data_is_blooming_why_you_should_care.php
Nodalities blog: http://blogs.talis.com/nodalities/
Linked Data.org: http://linkeddata.org/
Are you a semantic romantic? http://www.vocabcontrol.com/?p=213
Linked Open Data in Libraries Archives and Museums: http://lod-lam.net/summit/
Semantic Web Conference (academic): http://iswc2011.semanticweb.org/home/
SemTech (technology focus): http://semtech2011.semanticweb.com/
Museums and the Web Conference: http://conference.archimuse.com/mw2011/about
Freebase: http://wiki.freebase.com/wiki/What_is_Freebase%3F
MusicBrainz: http://musicbrainz.org/
DBpedia (derived from Wikipedia): http://dbpedia.org/About

Case Studies

Tweede Kamer (Dutch Parliament): http://www.fed-parl.be/parliaments_net/pdf/_Nelleke-Aders.pdf; http://staff.science.uva.nl/~marx/pub/adersgielmarx/
UK government and The National Archives: http://data.gov.uk/linked-data
BBC Wildlife Finder: http://www.bbc.co.uk/nature/animals/

BBC programmes ontologies:
http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml
NoTube – Semantic TV: http://notube.tv/
The New York Times: http://data.nytimes.com/
The British Library: http://www.bl.uk/bibliographic/datafree.html
Civic Apps**:** http://www.civicapps.org/
Walking Through Time**:** http://www.walkingthroughtime.co.uk/

Glossary

API — Application Programming Interface, a "gateway" or interface through which computers can exchange information.
Dereferenceable — you can look it up. A dereferenceable URL can be followed to find or access further data, documents, etc.
Instance — an object of interest — for example a concept that has a URI tag (e.g. lion). The data about instances is set out as triples — e.g. lions eat antelopes. The ways triples can be created is set out in an ontology — e.g. there are lions and there are antelopes and what happens is that lions eat antelopes. Ontologies can be presented as diagrams.
Ontology — a knowledge model or "world view" that a computer uses to process metadata.
OWL — Web Ontology Language — a computer language that ontologies can be written in.
RDF — Resource Description Framework — a computer language that triples can be written in.
SPARQL — a computer language that is used to process RDF to answer questions — e.g. tell me the names of all artists with paintings in the museum painted after 1975.
SPARQL endpoint — a "gateway" or interface to a set of data that accepts queries in SPARQL. A SPARQL endpoint can be public, to allow free and open access to data.
Triple — semantic metadata statement made up of three parts – subject, predicate, object — e.g. lions eat antelopes, Goethe is the author of Faust, Faust is a poem, etc.
Triplestore — a data store designed to hold triples, rather than relational database tables.
Unique identifier — a "key" usually a number or a mixture of numbers and letters that uniquely identify something. URIs are unique identifiers; some are dereferenceable (http URIs) and some are not.
URI — Uniform Resource Identifier – a "key" that identifies a specific concept or relationship. Http URIs can be linked to (they are dereferenceable). URIs can be URLs — Uniform Resource Locators — which we are familiar with as website addresses.
XML — Extensible Markup Language — a computer description format. XML is one of the formats that can include RDF and OWL.
Web services — (in some contexts) "gateways" such as APIs that can be provided so that data can be accessed easily, usually from a website, but sometimes from an internal data store. Web services could be any service provided over the web (e.g. web-based email, cloud computing).