Practical production of linked data

Rurik Thomas Greenall, NTNU Library

Am I going to bore the non-techies senseless? No.
Am I going to bore the techies senseless? Probably.
However, I welcome feedback at the end of the session.

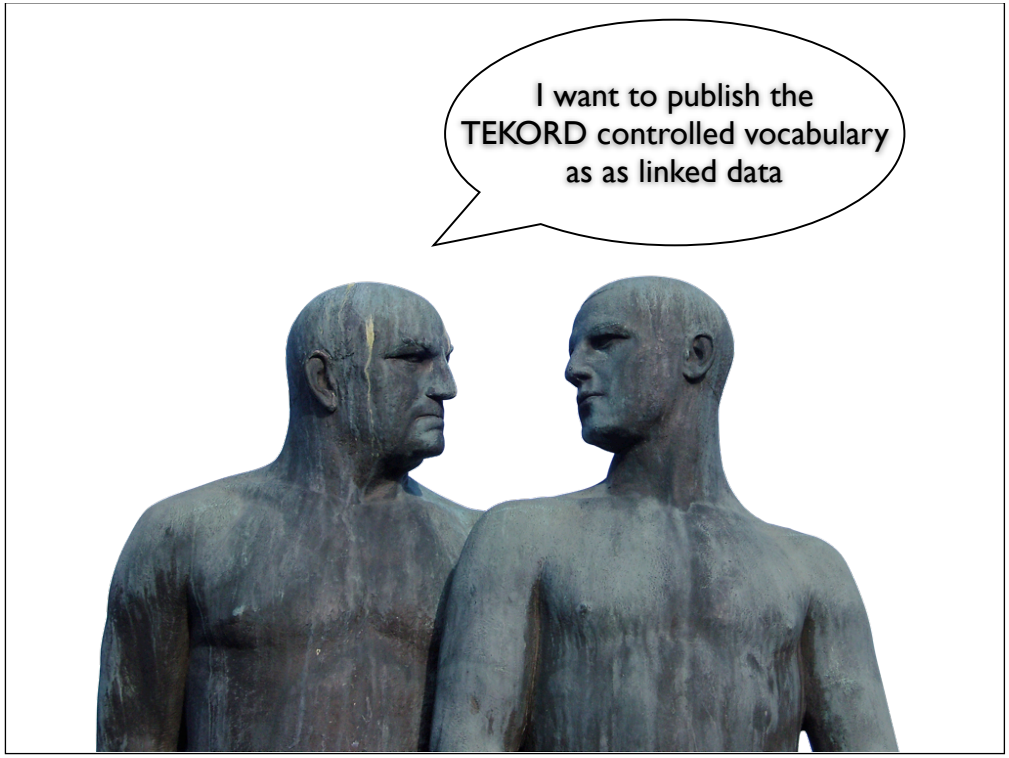
Buyer beware!

- This is a presentation for library people who want to try linked data
- I do not hold a degree in library science
- I do not hold a degree in computer science

I'm probably getting the wrong end of the stick…I will iron these problems out as I go along, you will too…
Let's start by looking at our problem:
Linked data is great — it does everything we want and gives us limitless possibilities. (Maybe not entirely true.) Let's rephrase that: You've got some data and you want to publish it so that it is available to the general public, or you just want to publish your data full stop.
There is a lot of talk about linked data, and you think that this might be a good way of publishing the data.
You've read the technical documents, you like the idea that search engines are getting linked data savvy, or you just plain want to try it.

So what I will be talking about is how you can use online tools, standard equipment and a few tricks to publish usable linked data.
If you don't know much about linked data, I hope that this talk will inspire you to take a closer look, if you already use linked data, I hope that this will inspire you to help the academic library community to play catch-up by sharing your experience.

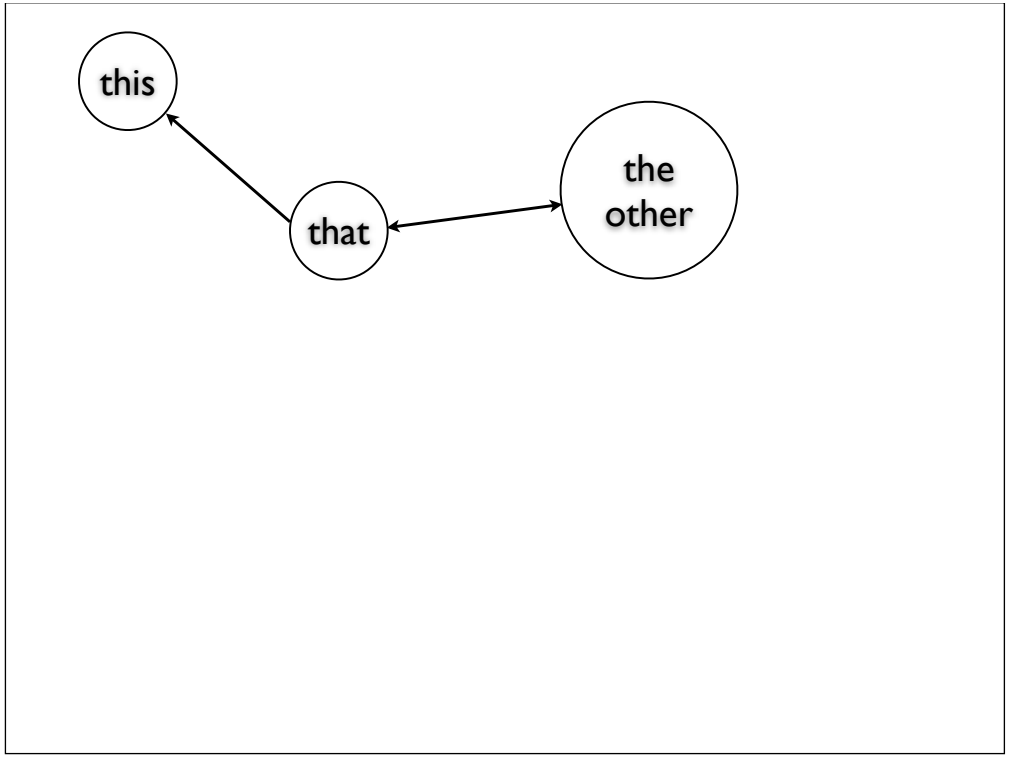
So you need some data. This sounds trivial, but believe me it isn't. The data will have to be something that you or your employer owns, or something that you can get signed over to you. This is not so easy to achieve: for a start people don't like just giving data away; it might not be worth anything or be used by anyone, but there is an ingrained culture of keeping things to ourselves. Let's take an example: millions are spent each year on data about serial publications. This metadata is typically available for free (in the wrong format) from publishers' websites, but it isn't "for sale" or "open", so you don't know what you can do with it.
In many countries, it is not the case that copyright law applies to metadata, and this has specific implications: since you don't hold a copyright, you can't release the data under a liberal copyright license like creative commons.
More bizarrely, it is not uncommon for an institution to just not have a policy about data: what do you do then? Here's our story.



Why Tekord?

I asked the coordinator the cataloguing dept about this. They passed me on to the head of the cataloguing department who they reckoned was the person to ask. They said “that position hasn’t existed for ten years, so I’m not the person you want to speak to”. Who is responsible then? “Not sure, maybe the coordinator of the catal…” They say you are the one who had most to do with it… “hmm”. OK, who has the most to do with it “that would be me”.
In the end, it took one of the managers to decide that we could publish the data, and we did so under an Open data commons license.
So there’s the first, boring part out of the way: getting permissions. You might like to start this process before you start processing the data, and then continue it while you’re doing the work I’m about to describe.

You can’t get away from the fact that you’re going to have to know something about RDF. The majority of people aren’t that familiar with the concept, and even if you are…read more about it.



So JørnH wins his bet

Tools

- Text editor
- Triplestore (just playin'? RDFQuery)
- RDF/N3->RDF/XML converter
- Validator (http://www.w3.org/RDF/Validator/)
- A scripting language
- Regular expressions
- A spreadsheet

The only thing you actually need, and cannot live without is a text editor, ay old text editor will do, but a fancy one with regular expression loveliness, code highlighting and possibly built in semantic web stuff might make your life better…or not depending on what you choose.

The second thing you probably need is somewhere to store your data that also allows you to query it using sparql, if you have very limited data and needs you can try RDFQuery which is a cool RDF-plugin for the Javascript framework JQuery. We use a developer account on Talis Platform because this is very easy and has a friendly developer community, although it is perfectly feasible to download and install a number of open source triplestores.

If you're like me, then you're not going to want to write XML, but you're going to want XML as a storage serialization. XML is nasty to write, but nice to work with — there are a lot of good XML tools out there. I write RDF/N3, which I then convert to RDF/XML, I use an online converter based on the CWM (Closed World Machine) program, which saves me the trouble of using it locally — which might be a good idea too.

The output of this isn't necessarily going to be valid, so use the online RDF validator

Next: you need a scripting language, something that allows you to take some data that is formatted in one way and then bend it into the Linked data model you've created with the tools I outline above. Personally I use PHP, but Perl or Python or whatever is probably a good choice. Add to this a good understanding of recursion and the key to data gleaning: regular expressions. Or XSLT :D
The final tool you need is a some way to view the raw data in a structured way, and potentially as a delivery format. I often choose CSV. Unless you're working directly with MARC data (in which case I'd use MARC/XML)

TEKORD: Our story

- Controlled vocabulary of technical subject headings
- Around 16,000 entries
- Structure: Term, BT, NT, SEE,USE, UDC number
- Data from BIBSYS delivered as XML by @OleHusby

Why tekord?
Controlled vocab
16000 entries
Structure: Term, BT, NT, SEE,USE, UDC number
Data from BIBSYS delivered as XML by @OleHusby

TEKORD: Our story

- Conversion
- Model based on SKOS (Simple Knowledge Organization System)
- PHP script containing regular expressions that reformat the data
- No external links…

Conversion Model based on SKOS (Simple Knowledge Organization System)
PHP script containg regular expressions that reformat the data
No external links…although could do this to a sql search of BIBSYS SRU

TEKORD: Our story

- A work in progress…
- Finding other sources of links — UDC
- Finding other things to link it to…
- Finding a use for it…

Because the data is in Norwegian, it is interesting to link to other data using UDC in order to get at least an English counterpart and other useful information, however, not many such resources exist on the open web cf. question on semanticoverflow.

What did we find?

- The SKOS representation is interesting
- It isn't difficult technically speaking
- It is difficult to get the institution to release the data if there is no policy beforehand; now is the time to create one
- It is useful to produce open data (UDCC used it to add Norwegian top terms)

The SKOS representation is interesting because the field for "TERM" is ambiguous in the non-linked data controlled vocabulary: it is either something that should or should not be used, whereas in SKOS it is either a preferred or an alternative label, which are separate semantically.

copyright & data:
http://bit.ly/cOBXbP, http://bit.ly/d8LNns,
http://bit.ly/ohhXd

RDF:
http://bit.ly/3GHNUx

Linked data & libraries
http://bit.ly/92a04i