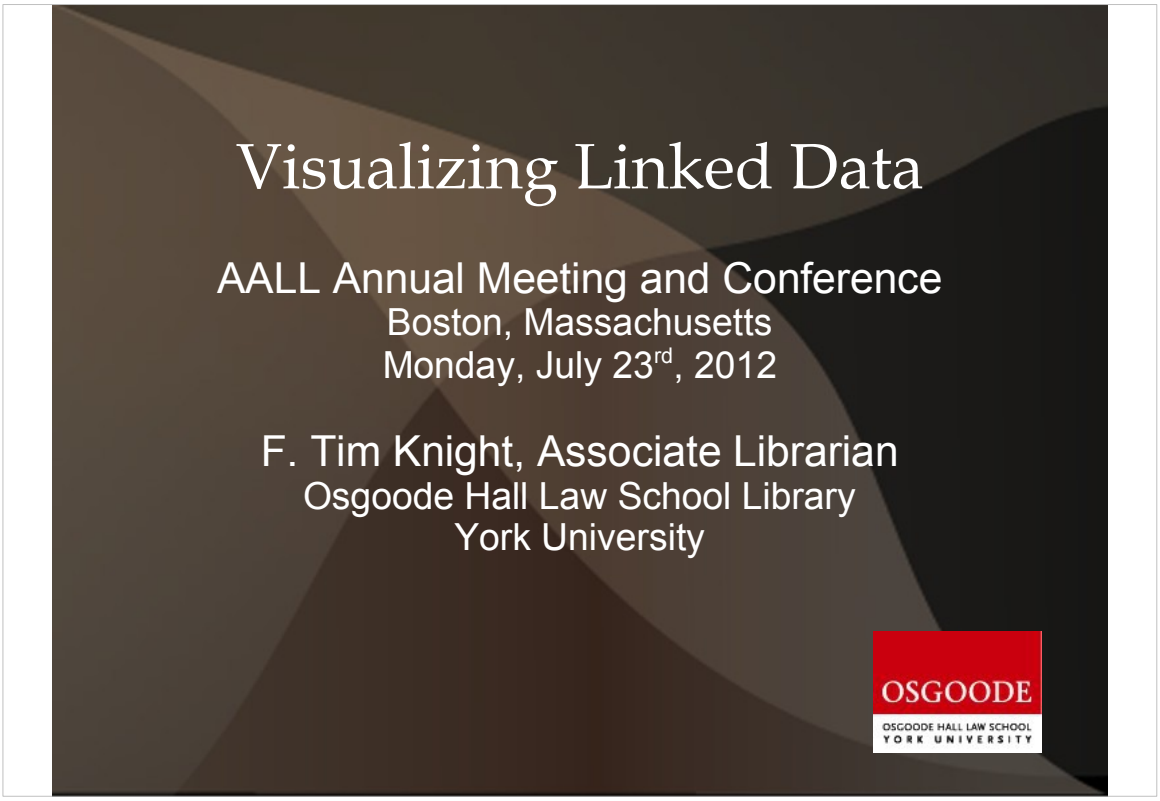
7-23-2012

Visualizing Linked Data

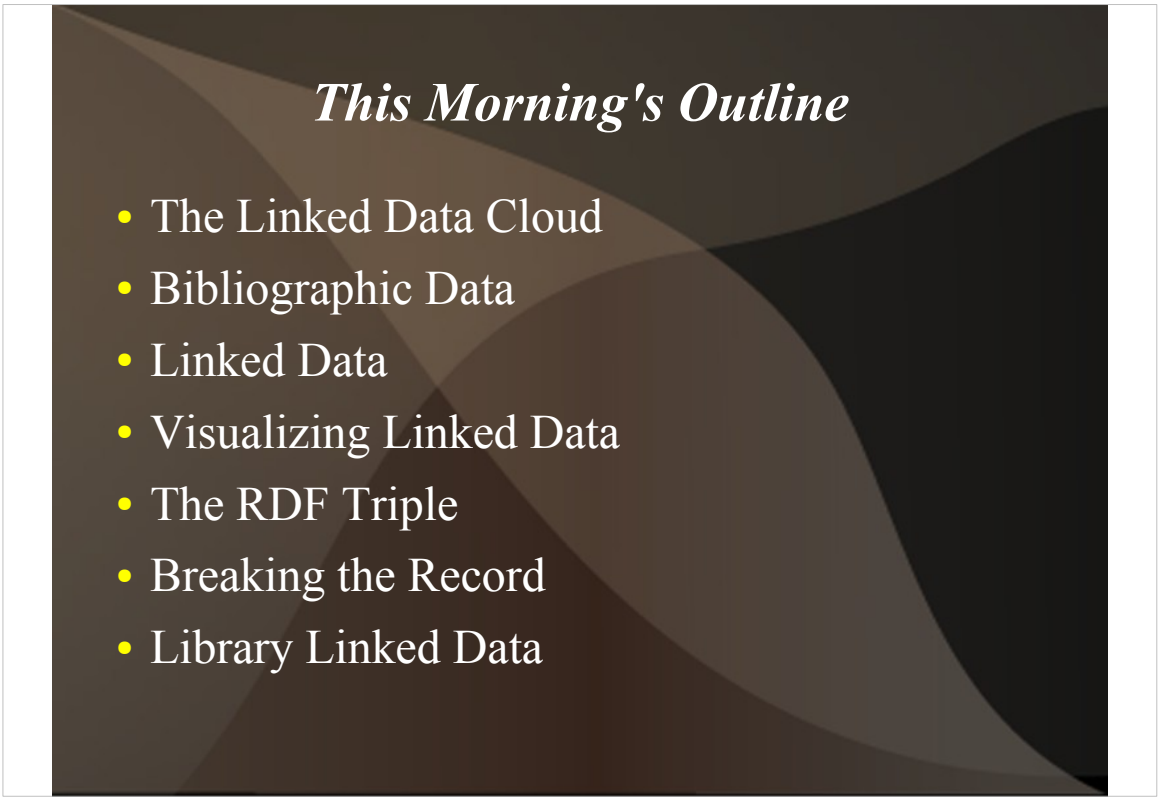
F. Tim Knight
Osgoode Hall Law School of York University, tknight@osgoode.yorku.ca



Good morning everyone and thank you for coming out so early on this Monday morning.

This morning I'm going to provide a little context that will set the stage for my two fellow speakers. I'll provide an introduction to linked data considering ways to visualize and get a better understanding of this relatively new concept.

Linked data is a different way for our bibliographic data to interact with the data available on the Web.. It will hopefully **enhance** the information we already provide, make it **easier** for us to **provide** that information and allow libraries to **integrate** themselves better with other information sources on the web.



I'll start off by taking a look at the Linked Open Data Cloud and its growth over the last few years.

Then I'll take a quick look at how we have traditionally handled our bibliographic data framing it in temrs of what many commentators have referred to as our “record-centric” environment.

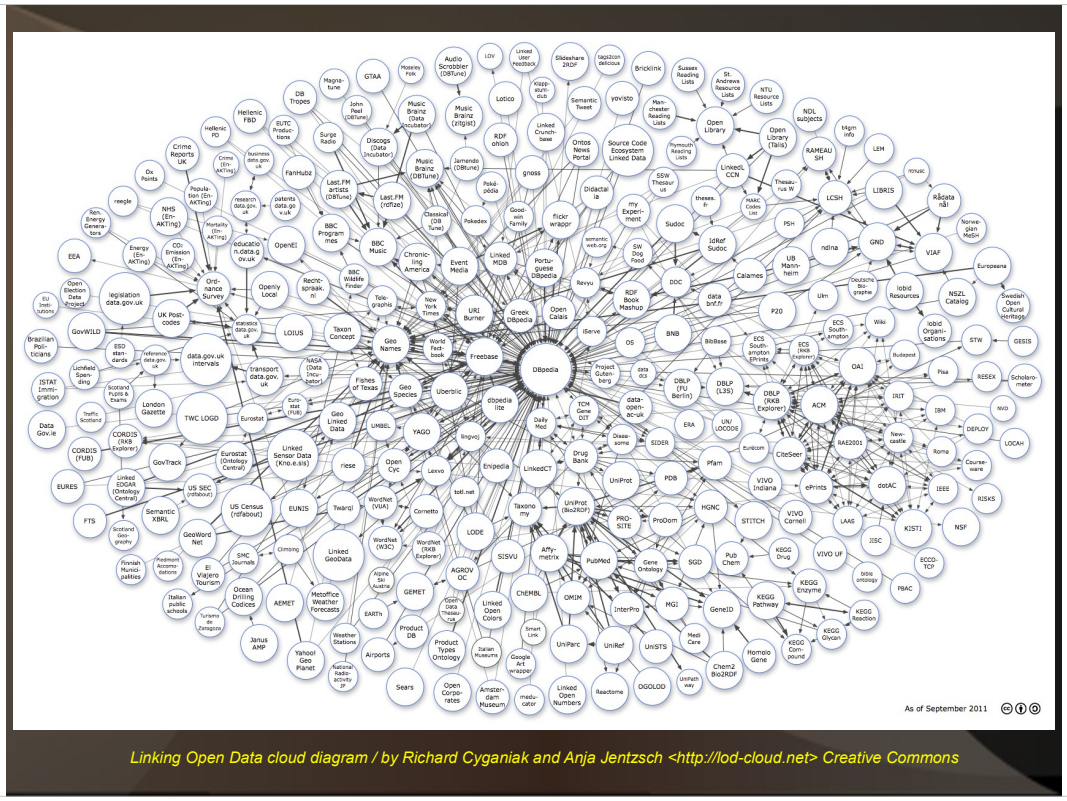
Then I'll take a closer look at what Linked Data is including a reference to a short video that Eurpeana.org released last year which provides a great way to visualize linked data on the web.

Then I'll zoom in a bit further and look at the RDF triple, the data statement that has become the fundamental building block of linked data and the semantic web.

I'll start winding down by introducing the idea of dismantling or breaking down the bibliographic record and the potential of **library** linked data on the web.

And finally, if there is time, I'll touch on a couple of the recommendations from a report on library linked data issued by the W3C last fall.

So let's get started ...



Linking Open Data cloud diagram / by Richard Cyganiak and Anja Jentzsch <http://lod-cloud.net> Creative Commons

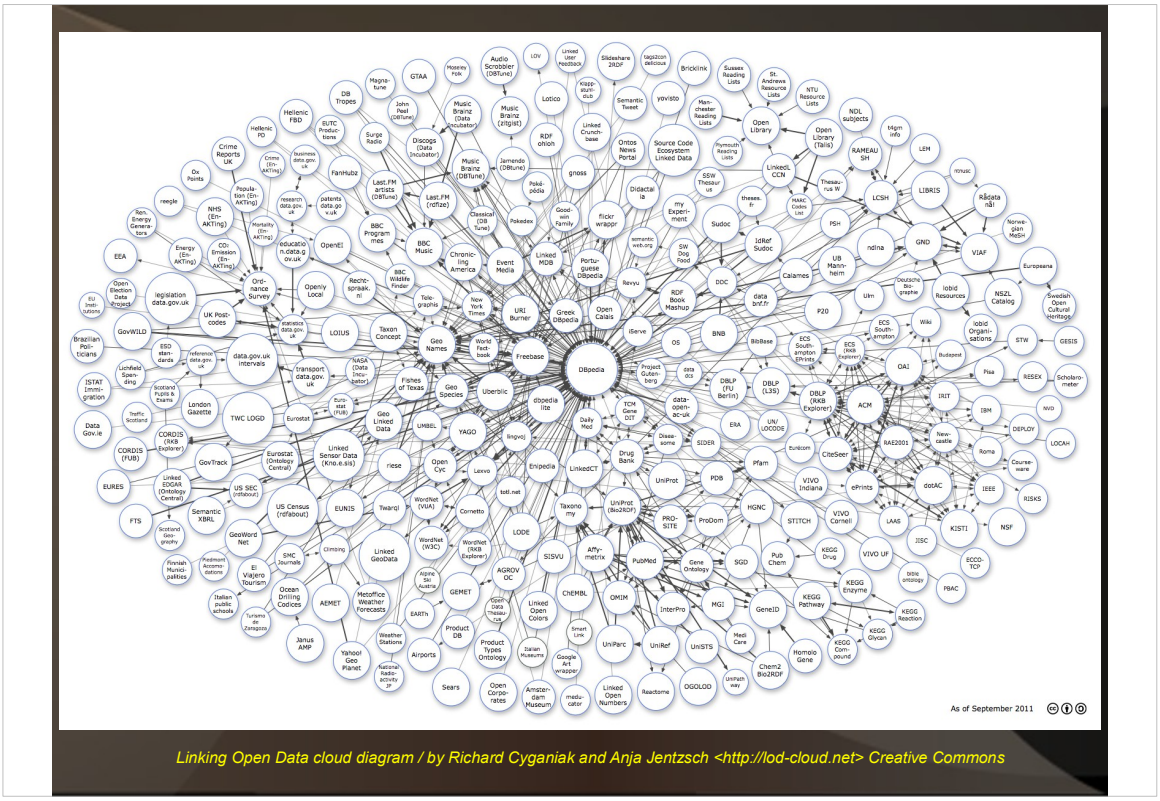
When I think of “linked data” this is the first image that comes to my mind.

Well that's not entirely true ...

tinker toy molecule no. 1 / by Otto Lange
<http://ottolangegallery.blogspot.ca/2012/03/tinker-toy-molecule-no1.html> Used with permission

This is what I really think of when I think of linked data.

Yes, good old Tinker Toys. Do you remember Tinker Toys? Very popular with me when I was about 6 years old.



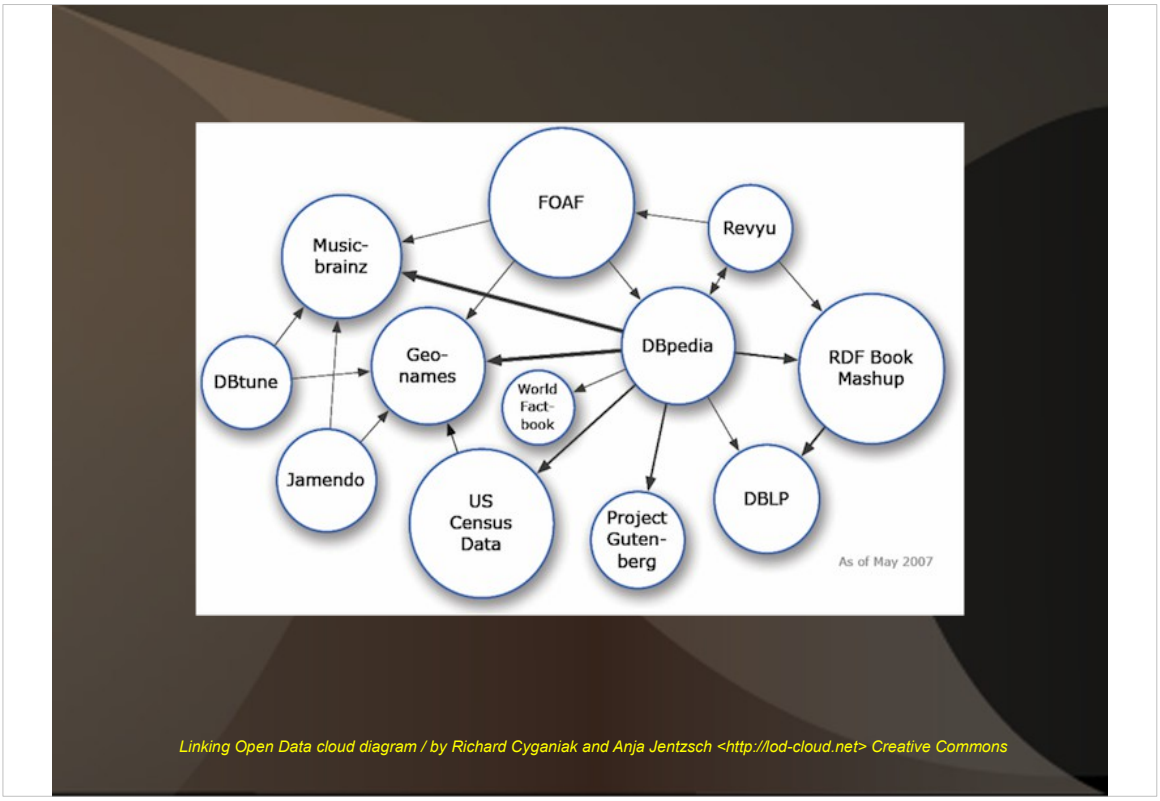
Linking Open Data cloud diagram / by Richard Cyganiak and Anja Jentzsch <http://lod-cloud.net> Creative Commons

But let's return to this image. How many of you are familiar with this representation of the linked data cloud? It's pretty popular so I'm not surprised to see that many of you have seen it.

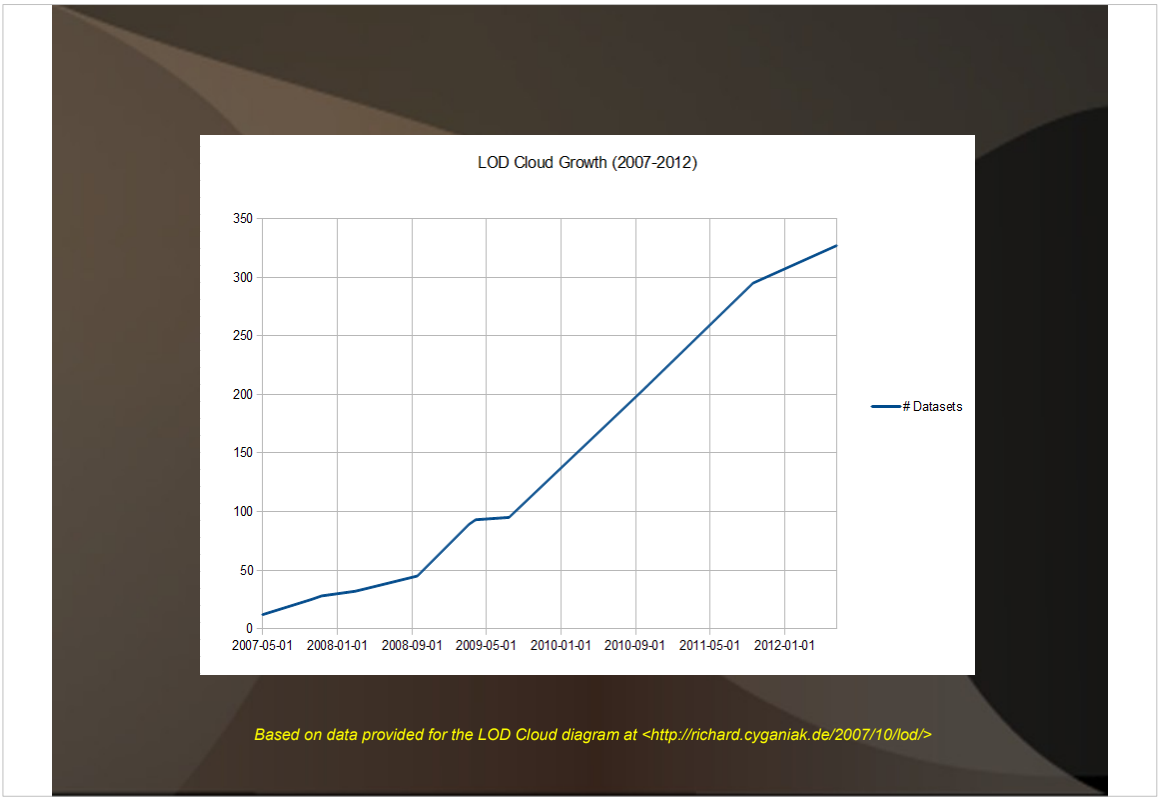
This is an image created by Richard Cyganiak and Anja Jentzsch and is based on metadata from the Linked Open Data Cloud group on CKAN, the Comprehensive Knowledge Archive Network, described by Wikipedia as a “web-based system for the storage and distribution of data.”

This is a cloud of data sets that are currently available on the Web as linked data and which contain links that point to other linked data sets.

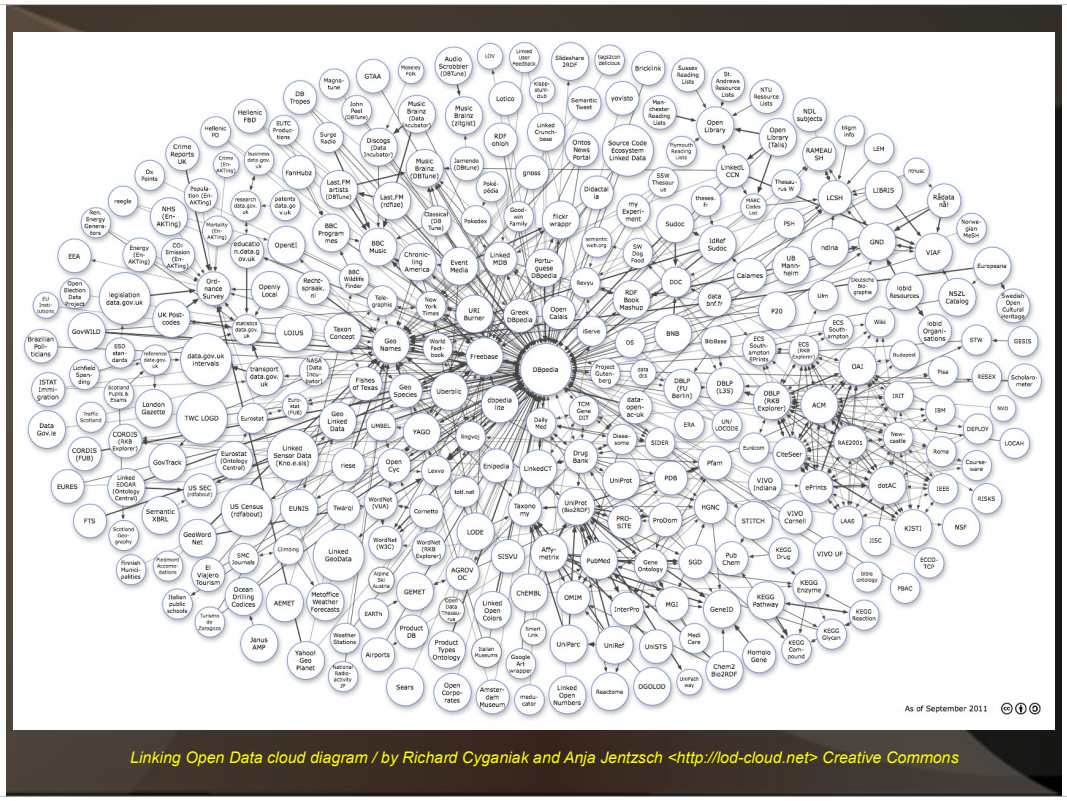
This particular diagram is billed as the 'latest' but it's quite a few months old now originally generated in September 2011. This “cloud” represents 295 data sets. The current number of linked data sets recorded at CKAN is currently listed as 327.



The first version of this was generated in 2007 when, at the time, there were only 12 datasets connected.

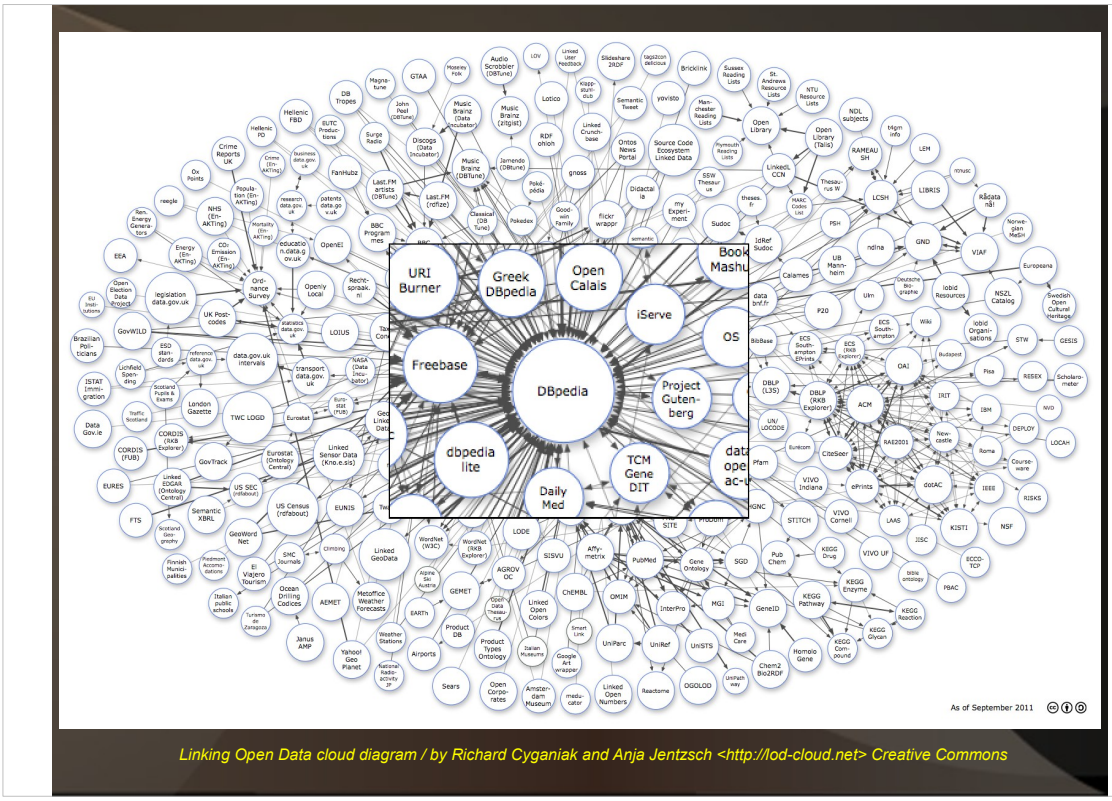


So as you can see here, there's been quite a bit of growth in the cloud over the last 4 or 5 years.



Linking Open Data cloud diagram / by Richard Cyganiak and Anja Jentzsch <http://lod-cloud.net> Creative Commons

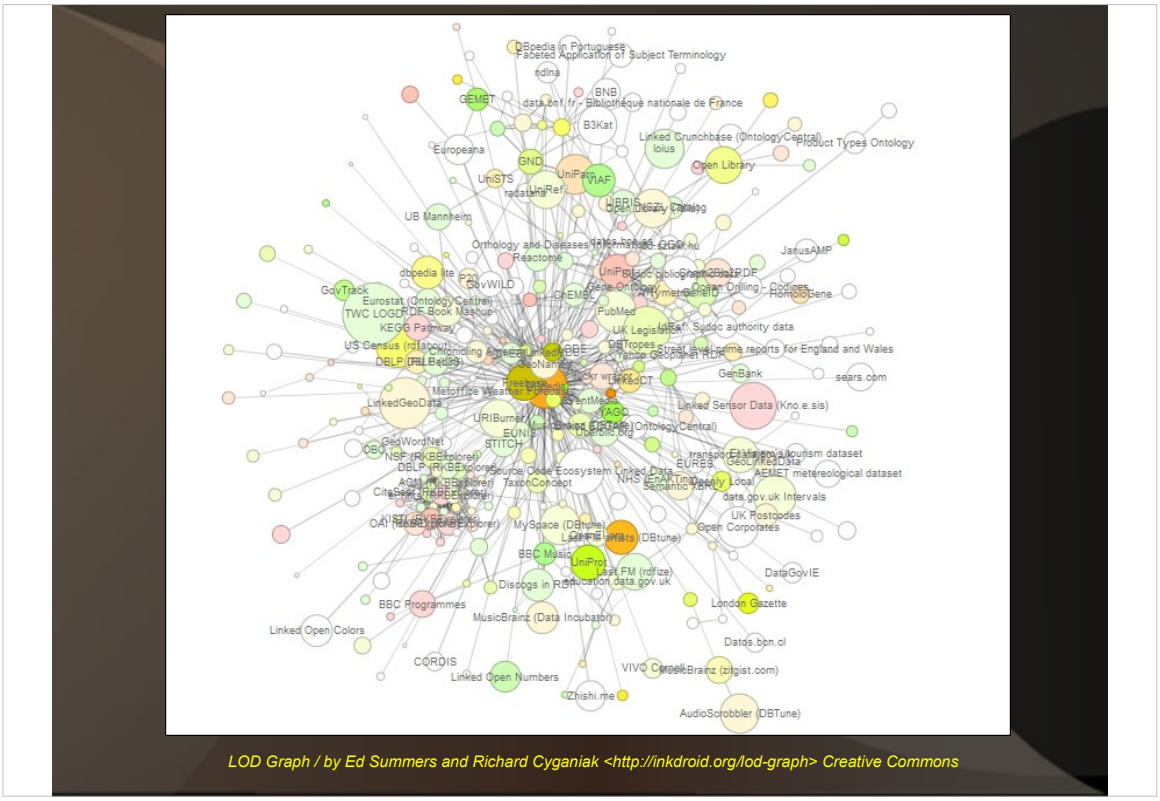
If we return to the larger more current version of the cloud diagram ...



Linking Open Data cloud diagram / by Richard Cyganiak and Anja Jentzsch <http://lod-cloud.net> Creative Commons

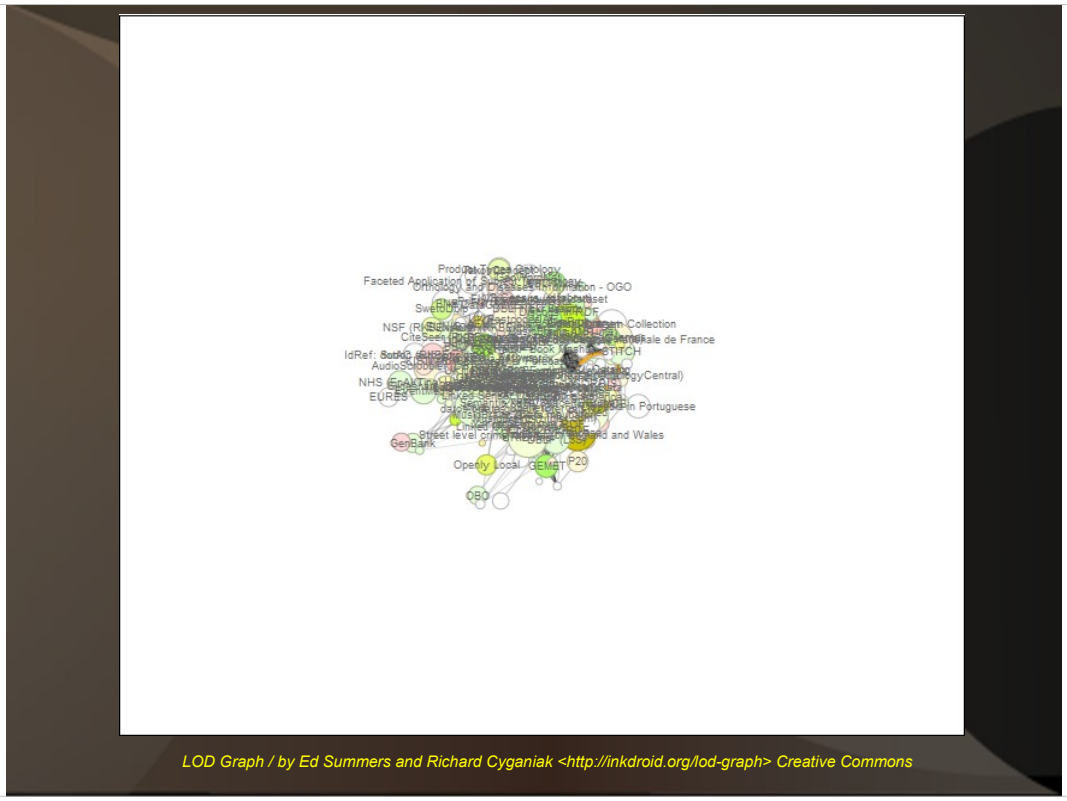
... you can see that **DBPedia** is prominent here, depicted in the center of this cloud. DBPedia is the “community effort to extract structured information from Wikipedia and to make this information available on the Web.”

And it is here that many data set providers start to link their data. And because os this DBPedia is fast becoming the center of the linked data hub.

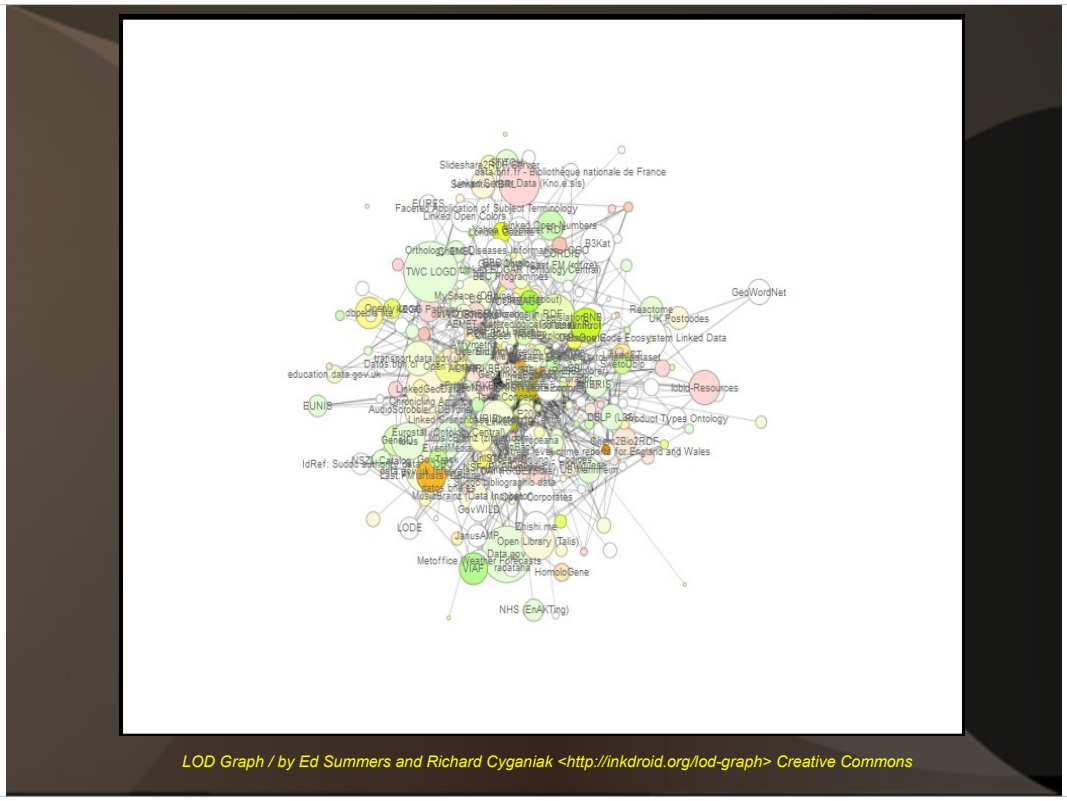


Here's another illustration using the same metadata provided by CKAN. This visualization was created by Ed Summers and Richard Cygniak using a now unsupported software platform called Protovis.

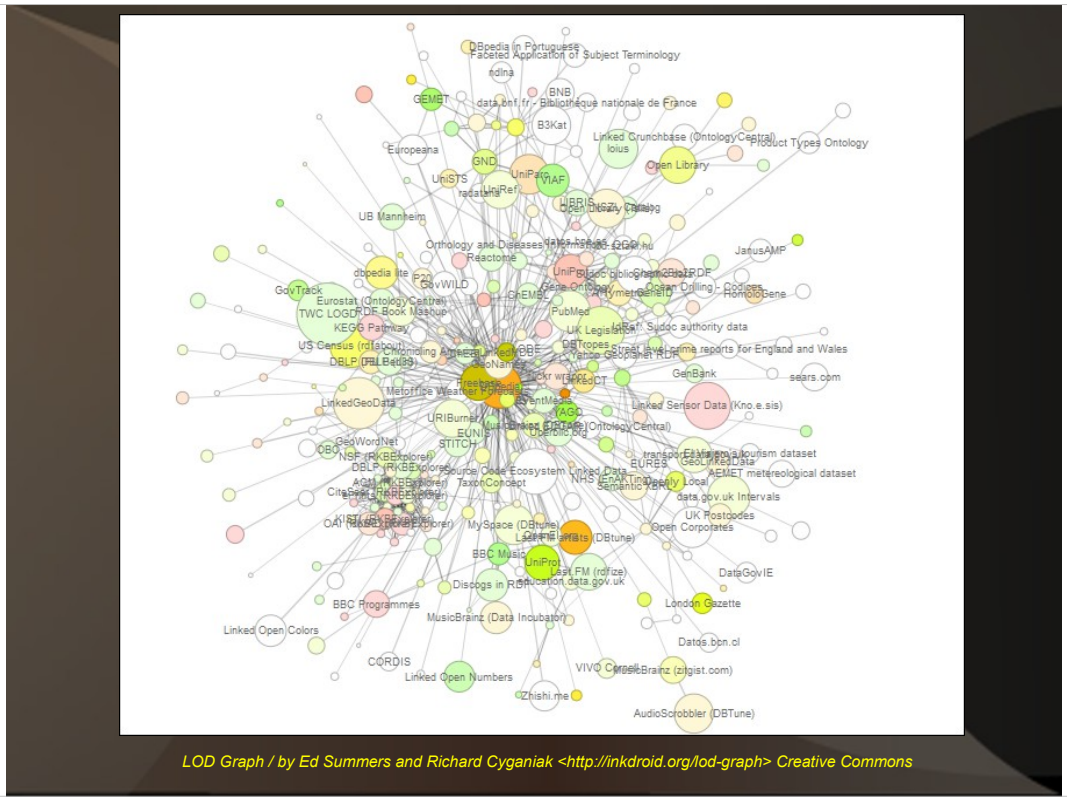
This is a dynamic representation which is generated, I think, in real time and so provides a current representation of the data sets in the linked data cloud.



When you first go to this site you will see a clump of circles and text that gradually starts to expand over time

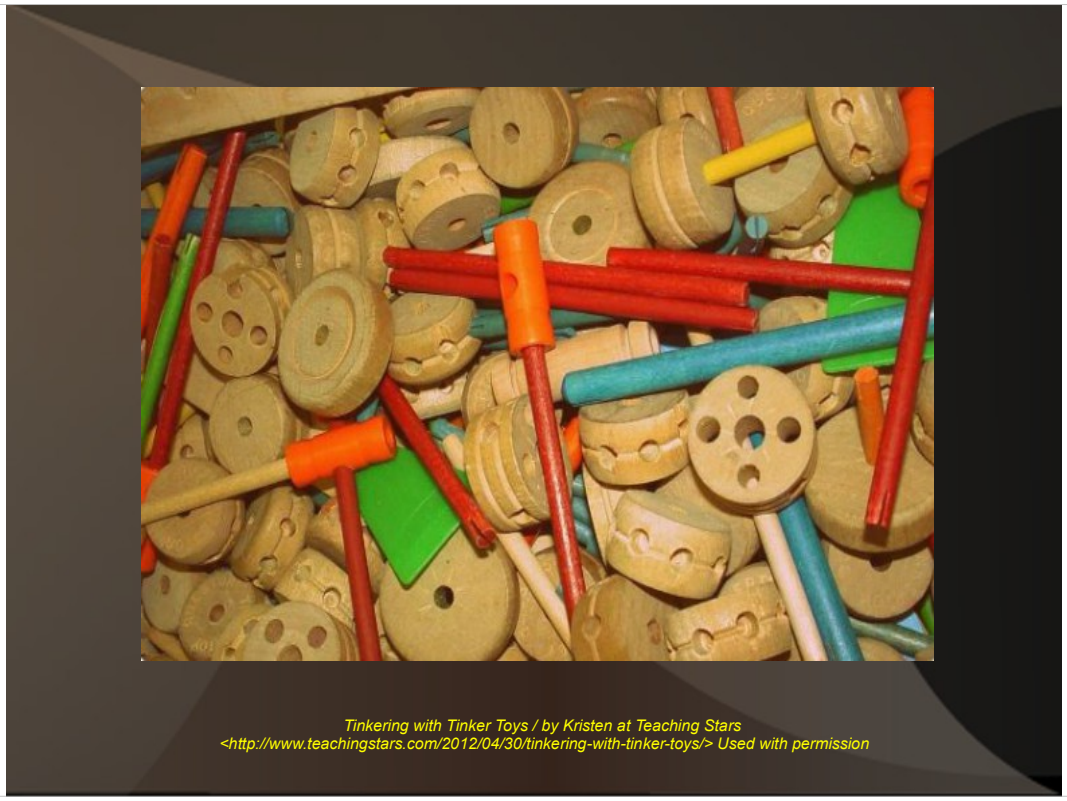


It's very cool to watch this unfold and I encourage you to have a look if you haven't seen this before.



LOD Graph / by Ed Summers and Richard Cyganiak <http://inkdroid.org/lod-graph> Creative Commons

It's certainly a little messier than the other diagram, but as I think you may be starting to appreciate, linked data is a messy business.



Tinkering with Tinker Toys / by Kristen at Teaching Stars <http://www.teachingstars.com/2012/04/30/tinkering-with-tinker-toys/> Used with permission

So let's take quick a look at how the library community has traditionally handled our own bibliographic data mess ...


Bibliographic Data

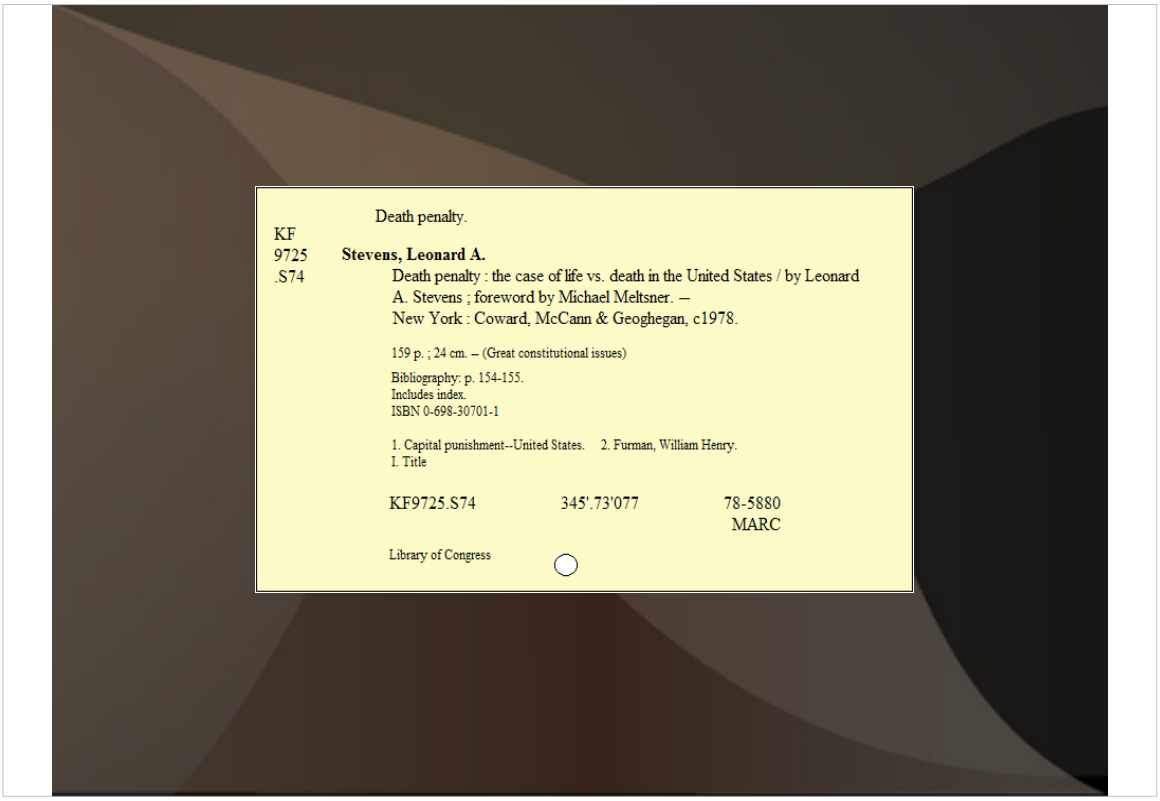
852: Rare – Harvard, It's in the Cards / by Mindy Johnston <http://etseq.law.harvard.edu/2008/10/852_rare_october_2008/> Creative Commons

Most of us are familiar with the card catalogue. A rack of drawers full of surrogate resource descriptions that gave users a quick and efficient way to explore and find materials in the library collection.

Many library users became quite good at navigating through the card cataloguge. In the book 'Hunting and Gathering on the Information Savvana', the authors suggested that this might be because of the "tactile delight of dealing with card drawers." (p. 7)

The card catalogue also provided library users with a reasonable representation of the information space and their place within it.

From the same book: "... [as] they pulled drawers from cabinets, they [know]: I am in the C's now and the W's are over there." (p. [137]-138)



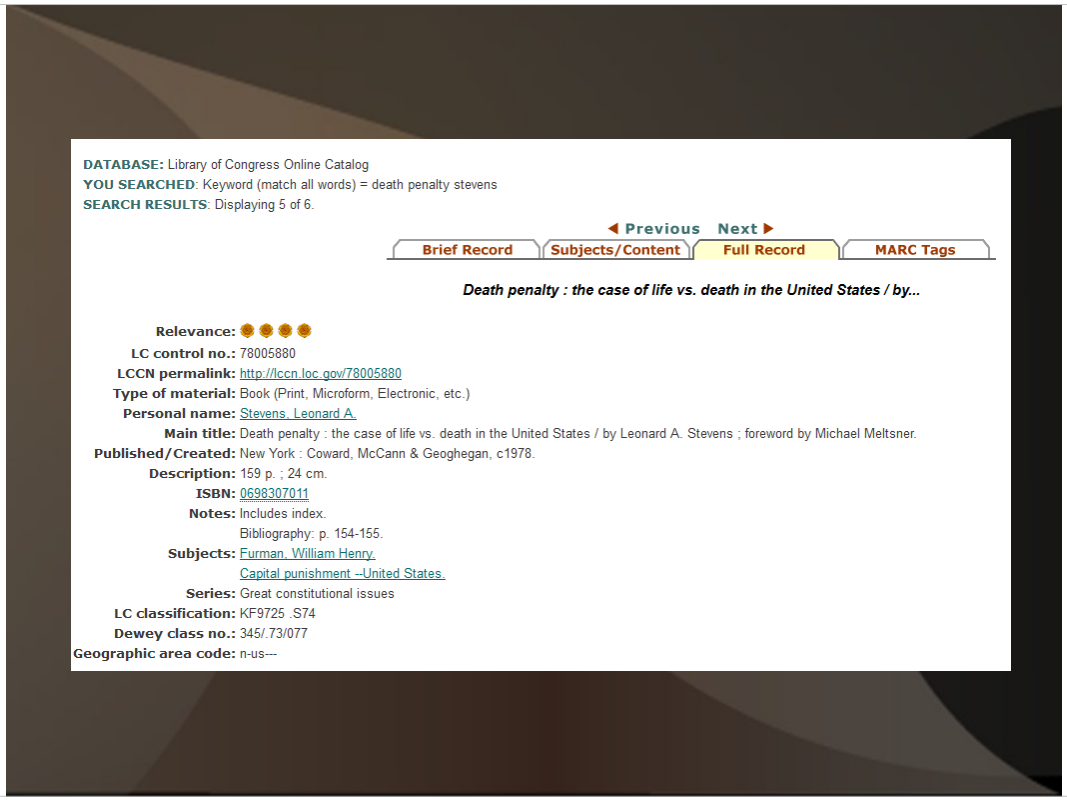
Each of the catalogue cards identified the important data elements for each resource in the library's collection. And the layout used ISBD punctuation to help clarify and delineate the information displayed on the card.

Paper copies of these cards were distributed to libraries at first ...

IBM 2400 Series Magnetic Tape Drive Units (1966)
<http://www.chipsetc.com/ibm-international-business-machines.html>

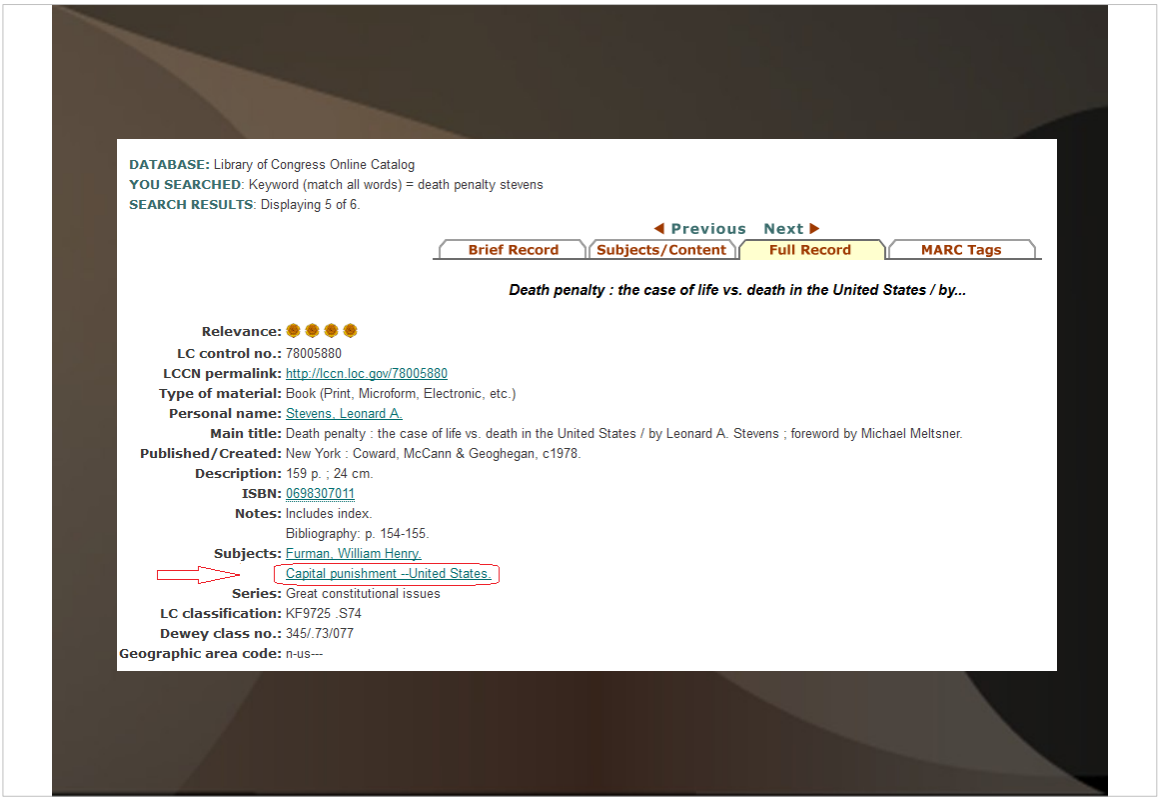
... with the data later shared using computer tape and a shared communication format known as MARC.

The MARC format ultimately led to the evolution of the Online Pubic Access Catalogue that began to appear toward the end of the 1980s.



The catalogue card was transformed into something seen in this display here from the Library of Congress catalogue.

It's very similar to the original paper card layout, but now the data elements are also keyword searchable ...

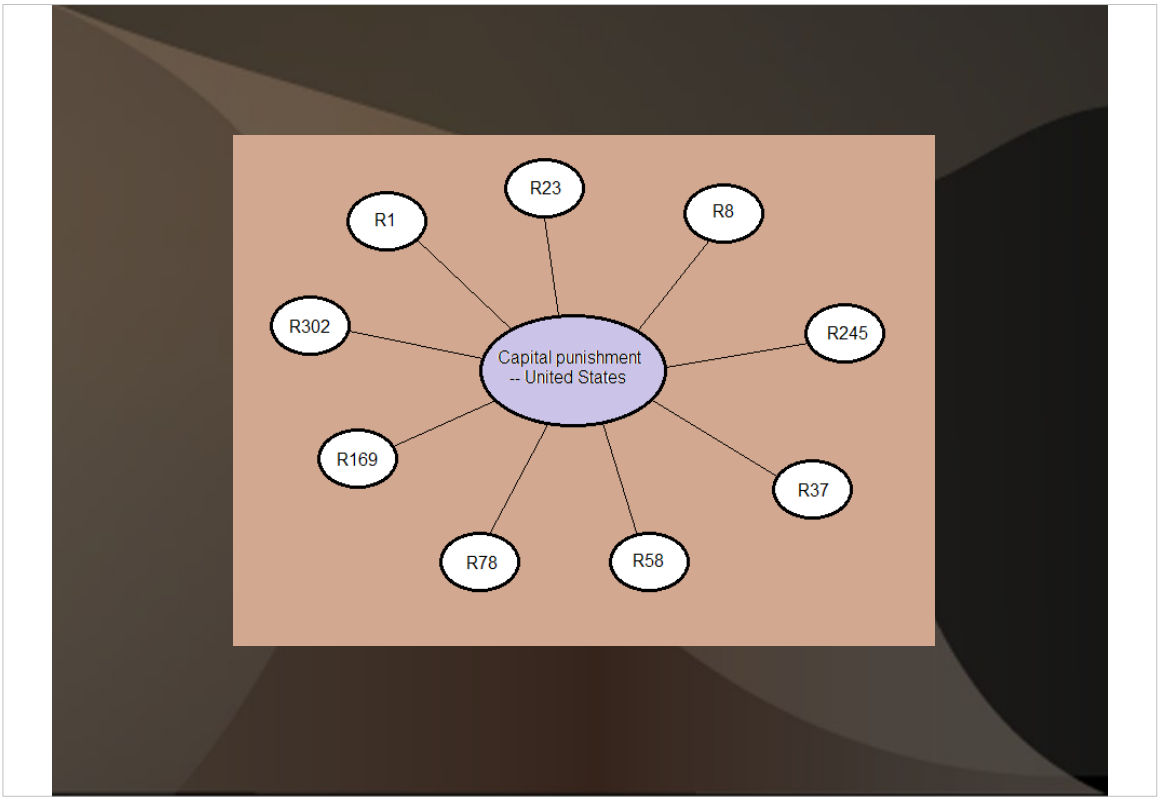


... and a new hyperlinking feature was available for things like subject headings.

So when you click on a subject heading like 'Capital punishment—United States' found in this record, you are linked, in this case, to ...

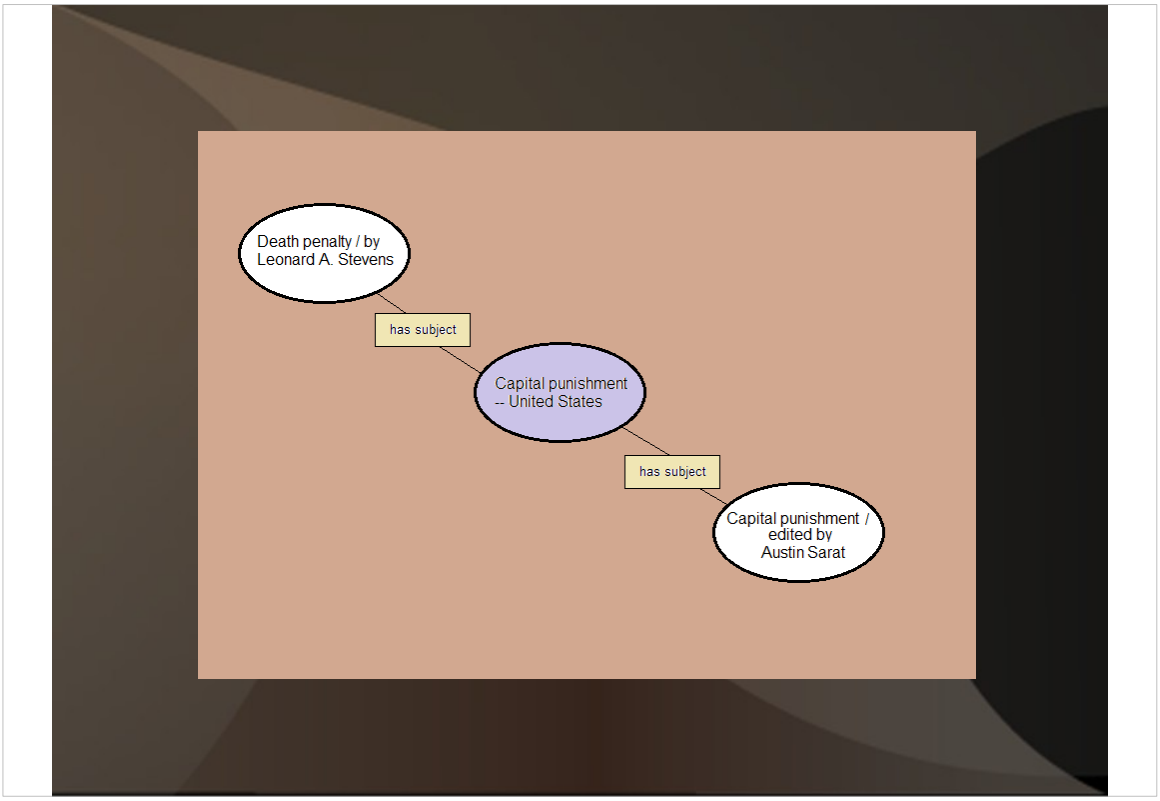
... 326 other library resources with the same subject heading.

I like to think of this as the library community's first realization of "linked data."



This could be shown in a diagram that might look something like this ... where each of the white ovals here represents one of the resources in the library surrounding the subject heading 'Capital punishment—United States' located in the centre.

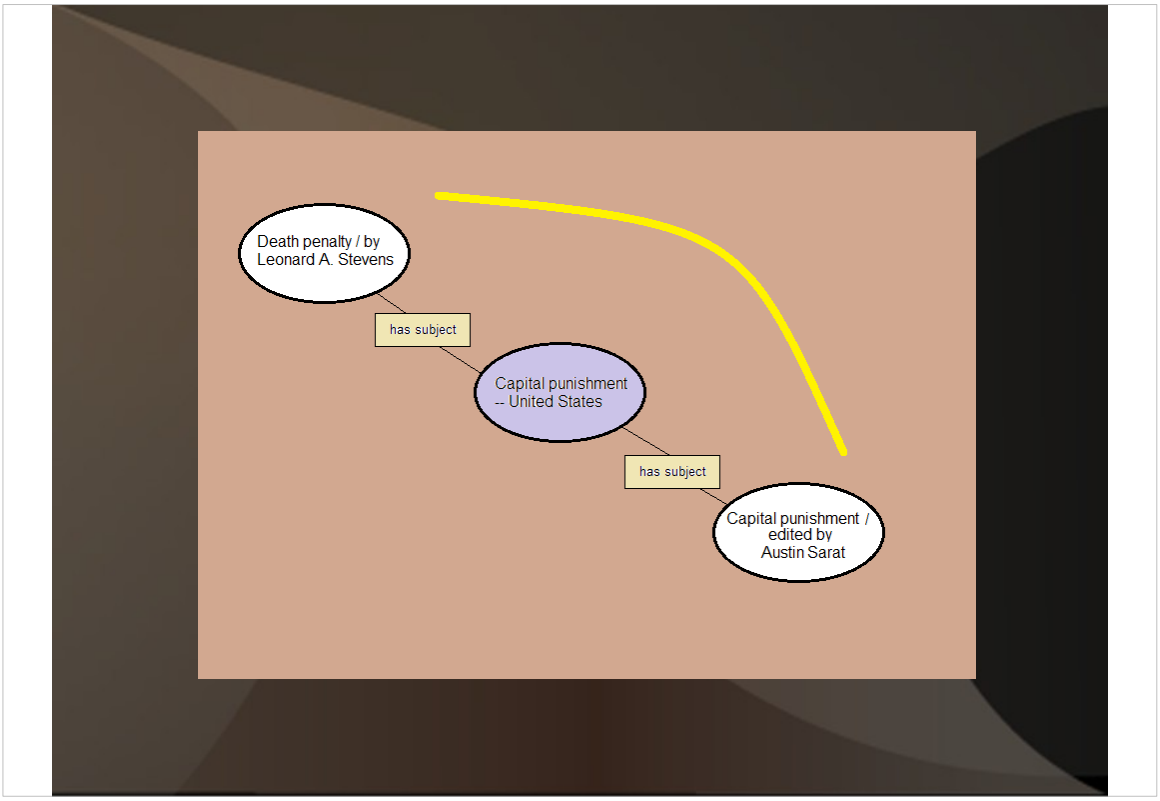
Which reminds me again of the Tinker Toy analogy.



Or to look at it another way: something like this ...

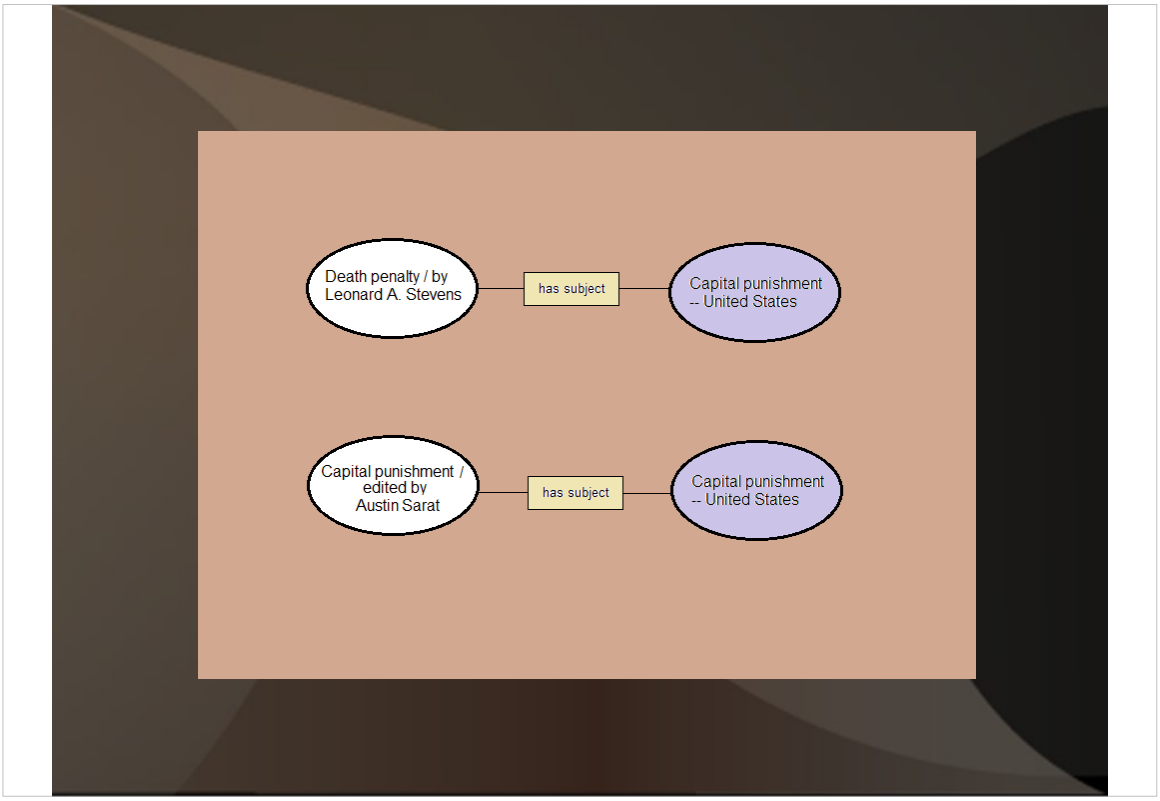
Here we have the original title, 'Death penalty' by Stevens leading us through the subject heading 'Captial punishment' to another book with the same subject. This other book, a collection of essays called,"Capital punishment," edited by Sarat.

So what's effectively been done ...

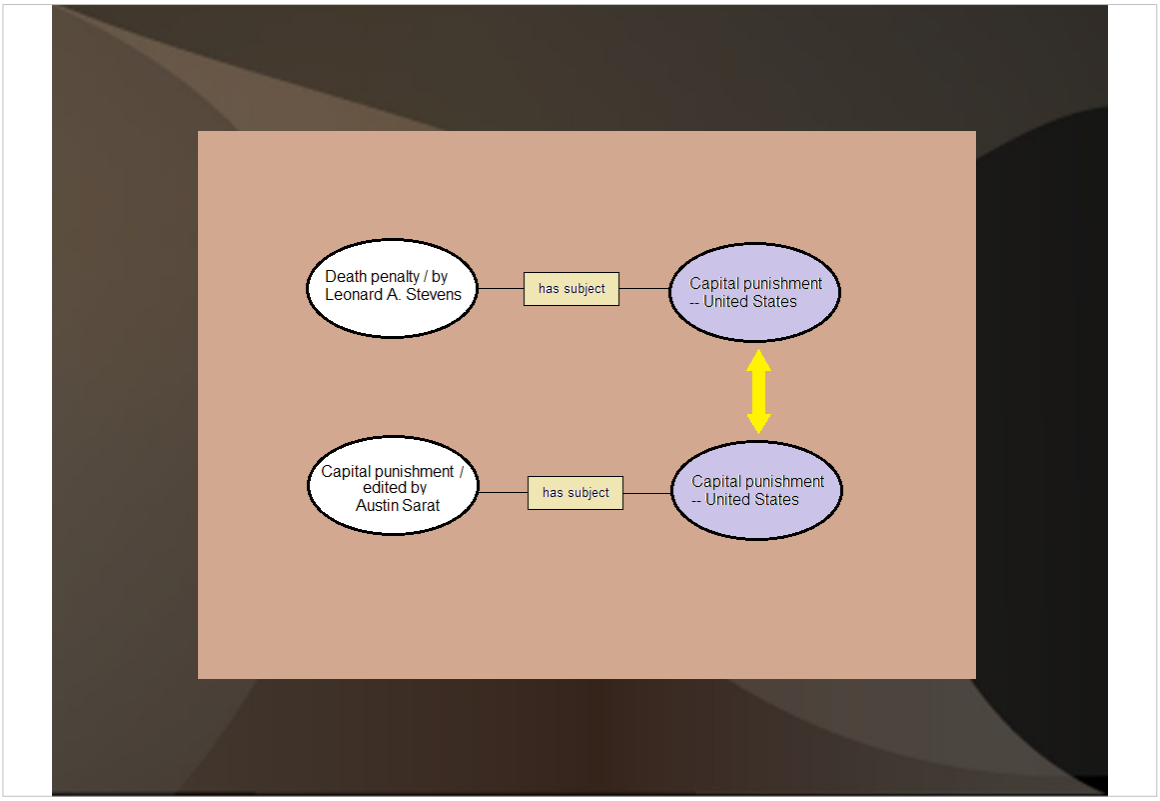


... is these two titles have been linked together for the user in the library catalogue.

Let's use that idea as our starting point for thinking about linked data as we move forward.



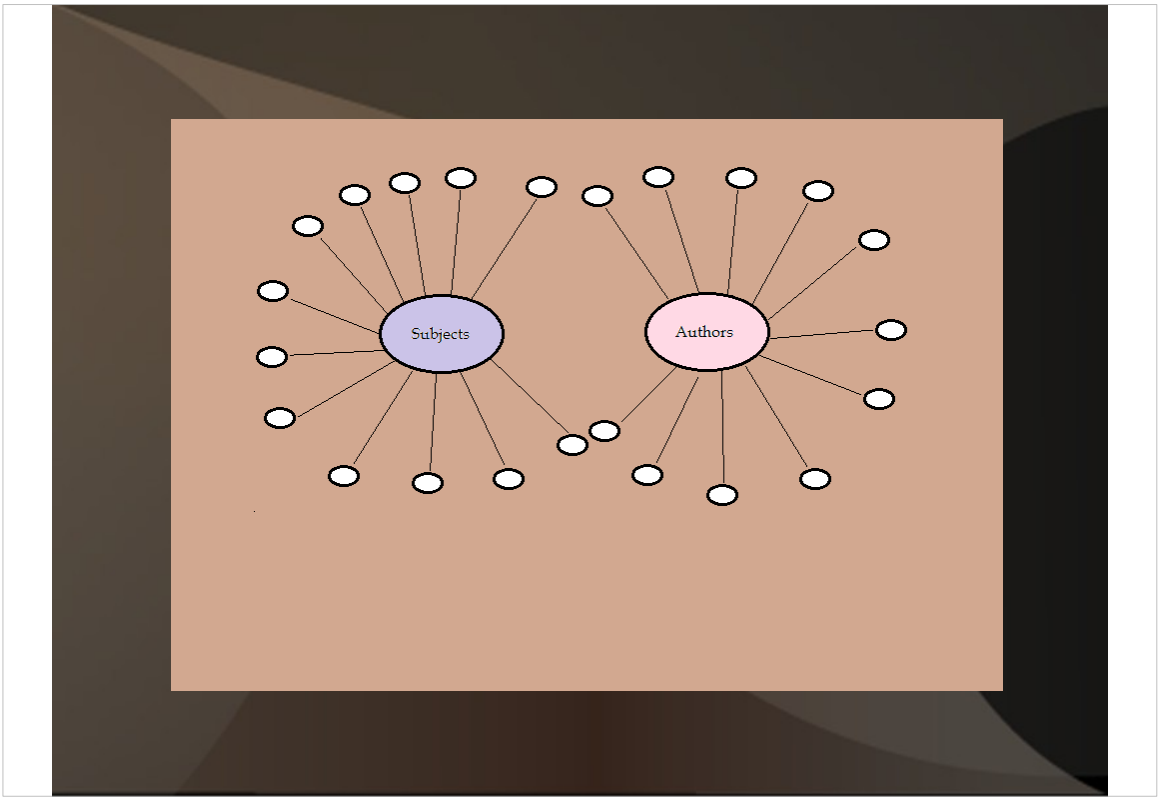
That earlier diagram can be redrawn so that each of these resources has the subject 'Capital punishment—United States' creating a couple of data statements like this.



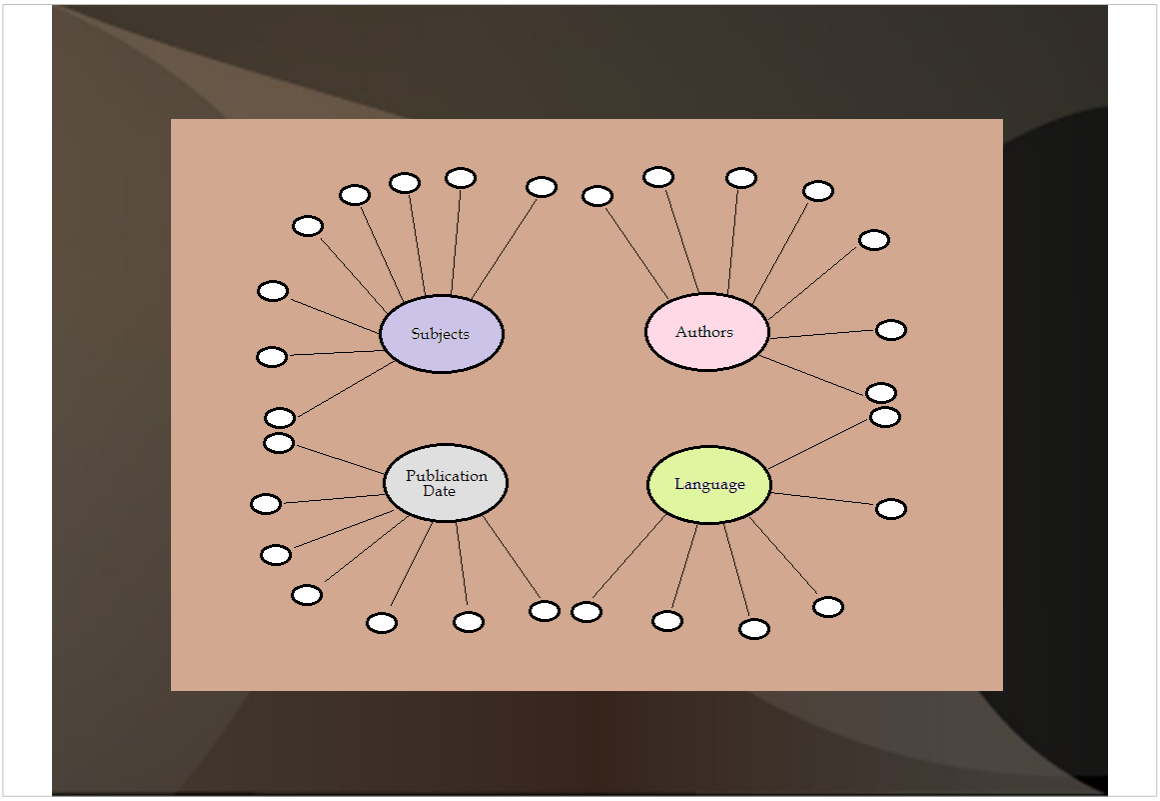
And because each of these resources have this subject relationship in common we can consider them to be related or linked to each other because they cover the same subject area.

This kind of data statement, expressed in three parts like this, has been called a 'triple' in the Resource Description Framework and this is the data model that semantic web developers are working with when they create linked data.

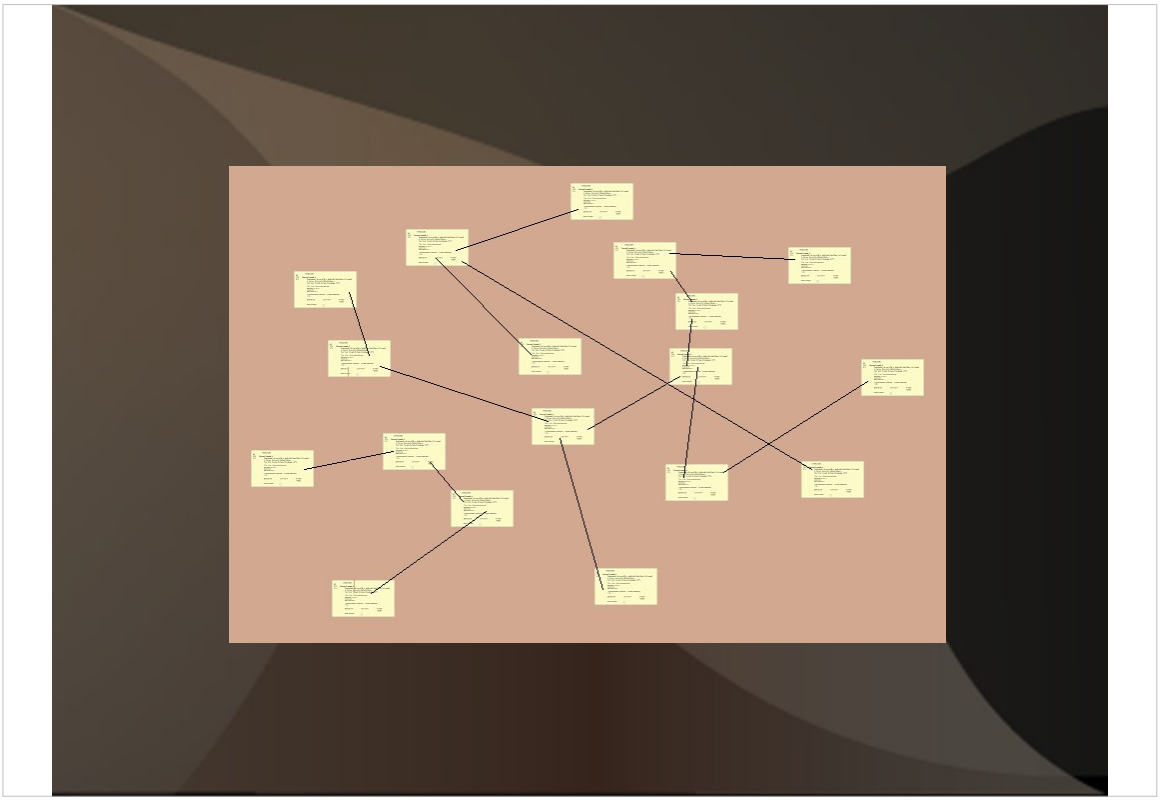
I'll come back to RDF triples in a moment.



So typically in today's catalogue we have these hyperlink connections mostly between subject headings, name authorities and uniform titles, that act like linked data.



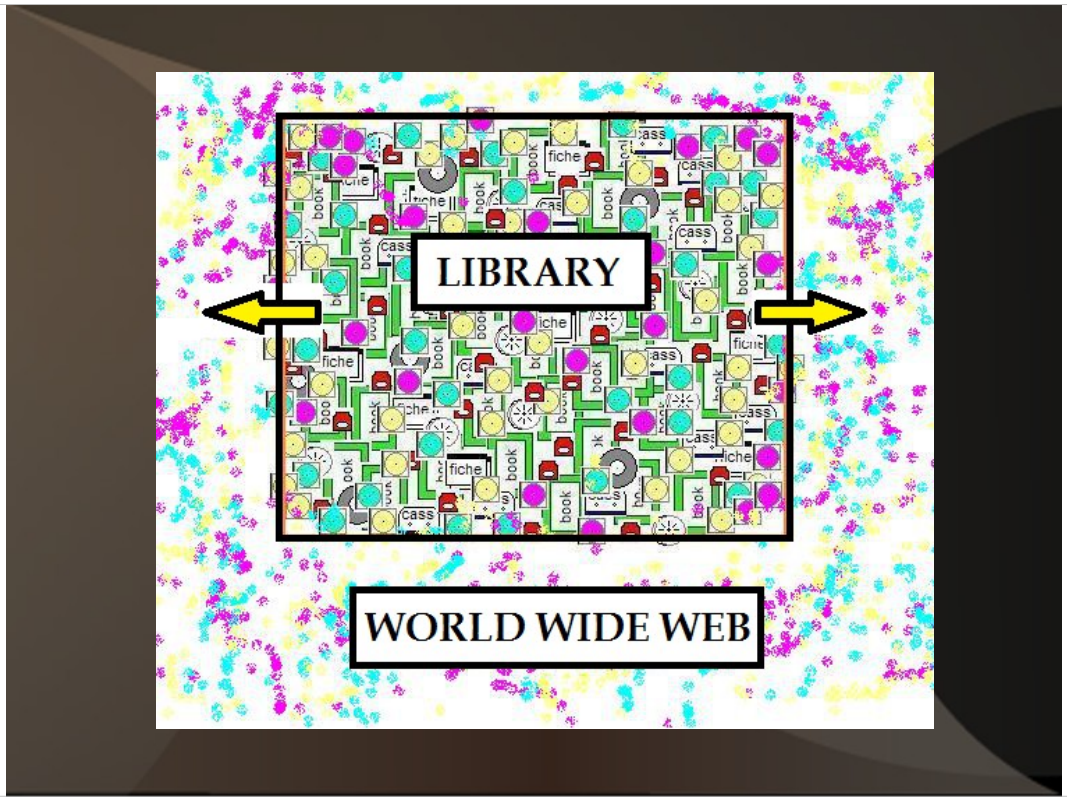
And our discovery layers have also added additional functionality through facets like publication date and language which apply search filters or limits to our search results.



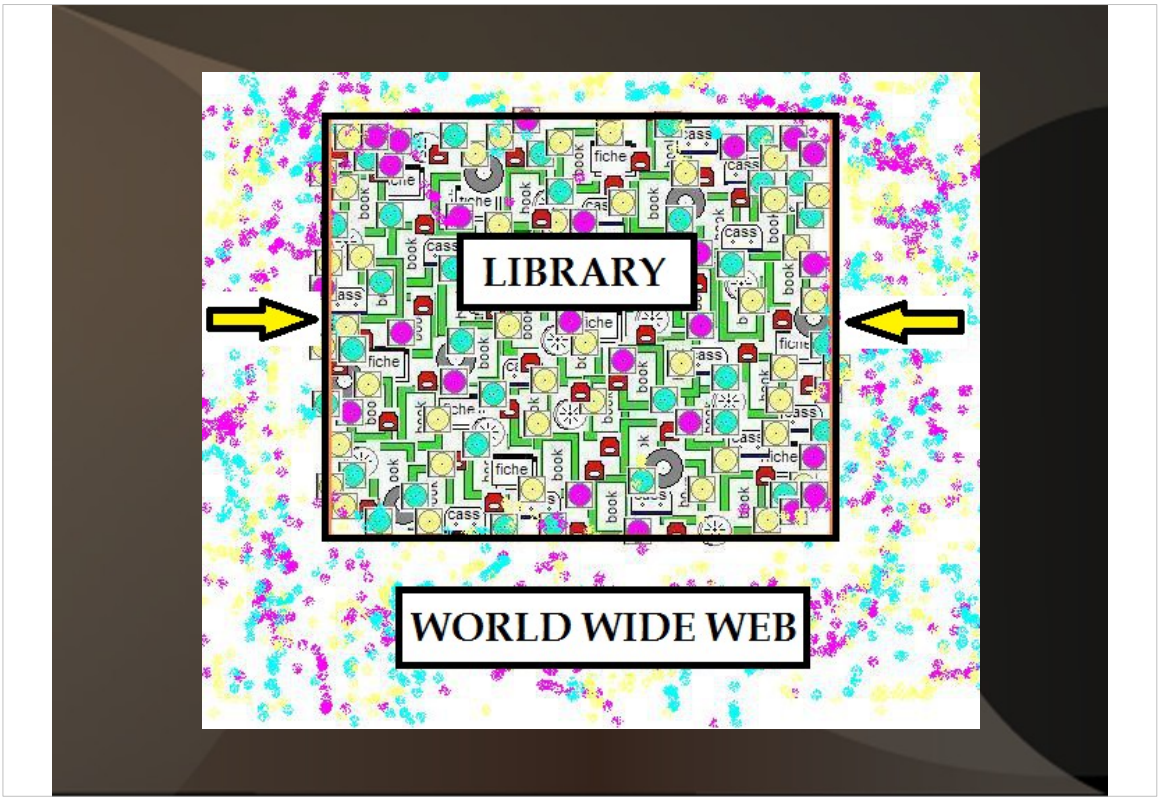
So within our catalogues we have succeeded in linking together the electronic versions of our catalogue cards and created a “card catalogue cloud” if you will.

It works very nicely and has served us and our library users well. Because of the MARC format we can use the MARC coding to tell us that the link is a subject relationship. If the data appears in one of the 6XX fields in the catalogue record our ILS can generate a nice label to let us know that this is a subject link.

But there are other relationships that exisit in our catalogues that we are **not** seeing. And, we are not seeing **any** relationships that might exist between our resources and resources that live outside of our library catalogues.



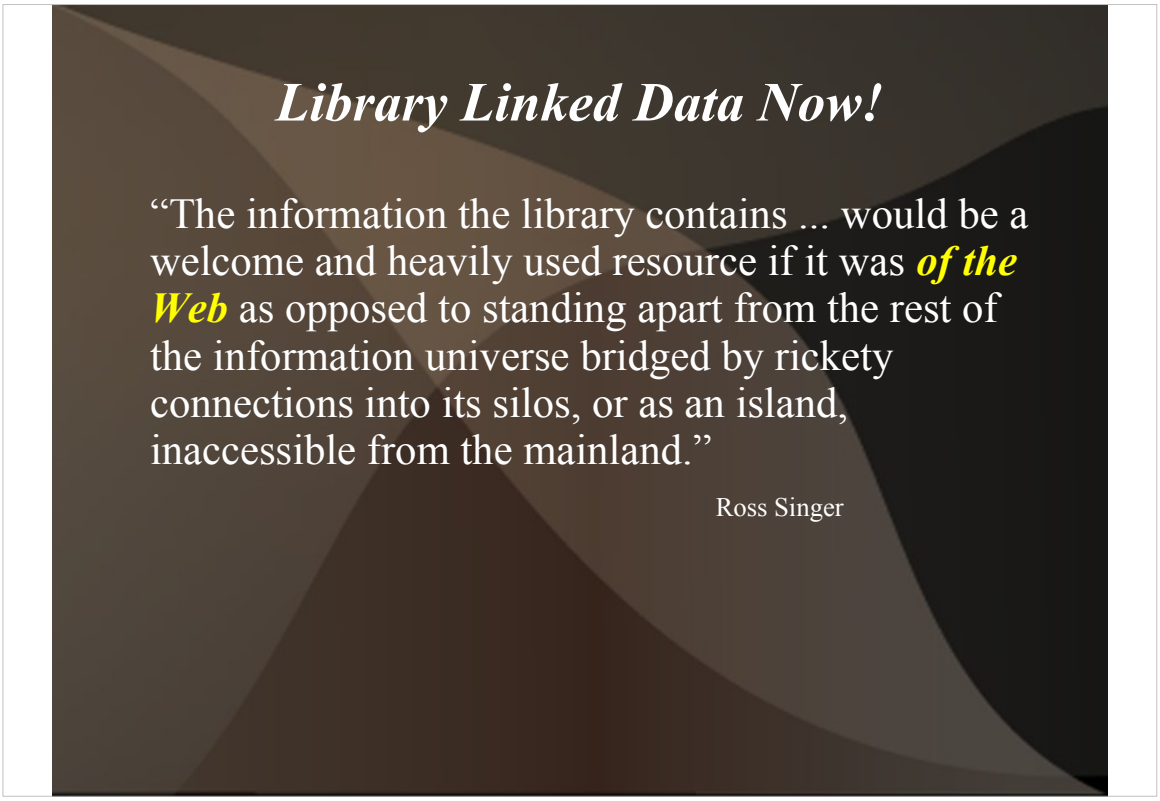
We do provide links out to resources on the web when we can, but other than that there is no connection between things that appear in our catalogue and things that don't.



Since there **is** a lot of information that lives outside of our catalogues it's increasingly important that we try and make these connections.

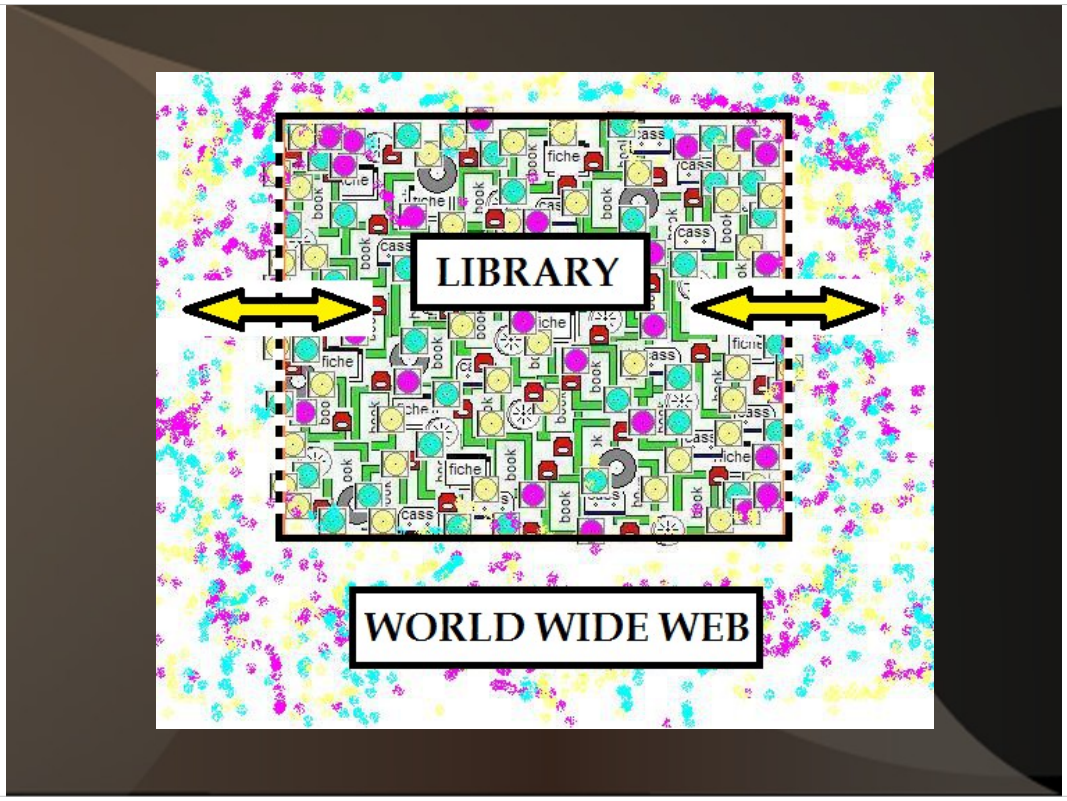
The library catalogue is no longer, if it ever really was, the first place that information seekers come to find information resources.

As we know, most will start on the web somewhere, and if we are not there with them, we will be overlooked as a resource.



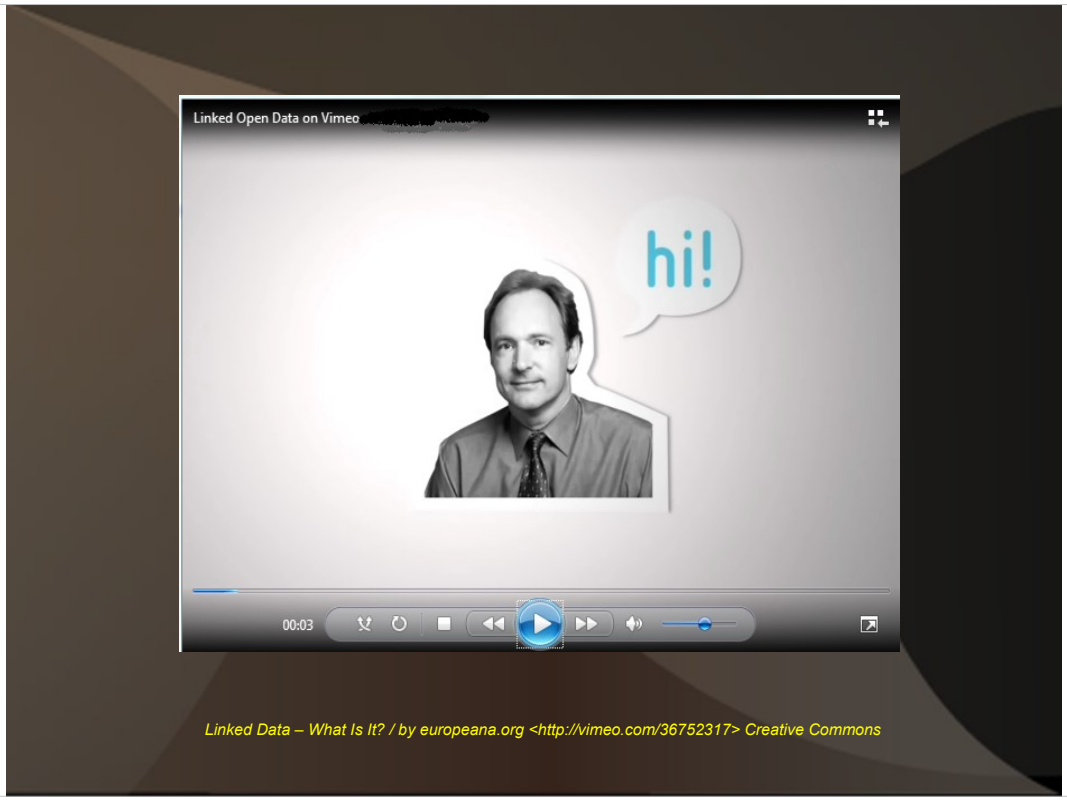
As Ross Singer stated in his 2009 article Linked Library Data Now!:

"The information the library contains ... would be a welcome and heavily used resource if it was *of the Web* as opposed to standing apart from the rest of the information universe bridged by rickety connections into its silos, or as an island, inaccessible from the mainland."



And this is where linked data can play a role; breaking down the barriers that exist between our bibliographic data and data out there on the Web.

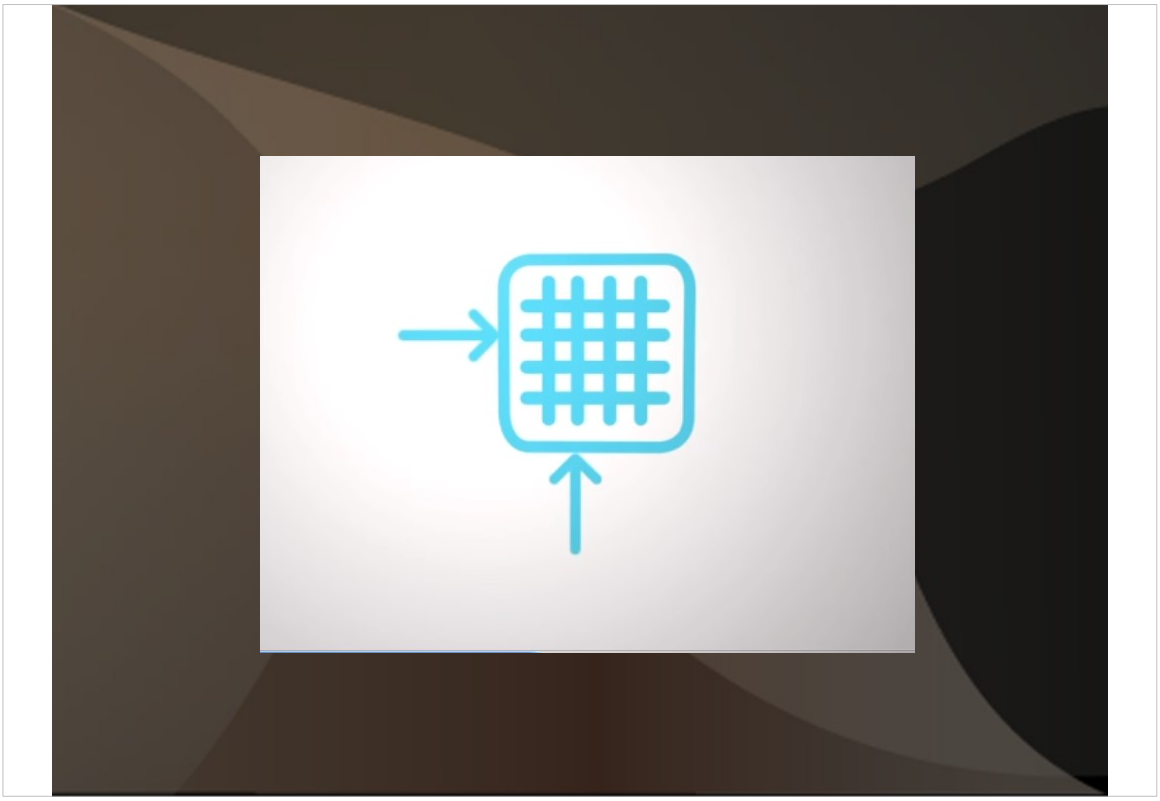
So let's see if we can conceptualize linked data as it is used in the context of the semantic web.



Originally I was going to play a short excerpt from the Europeana.org video that was released not too long ago. I don't really have time to do that today but would encourage you to have a look if you haven't already seen it.

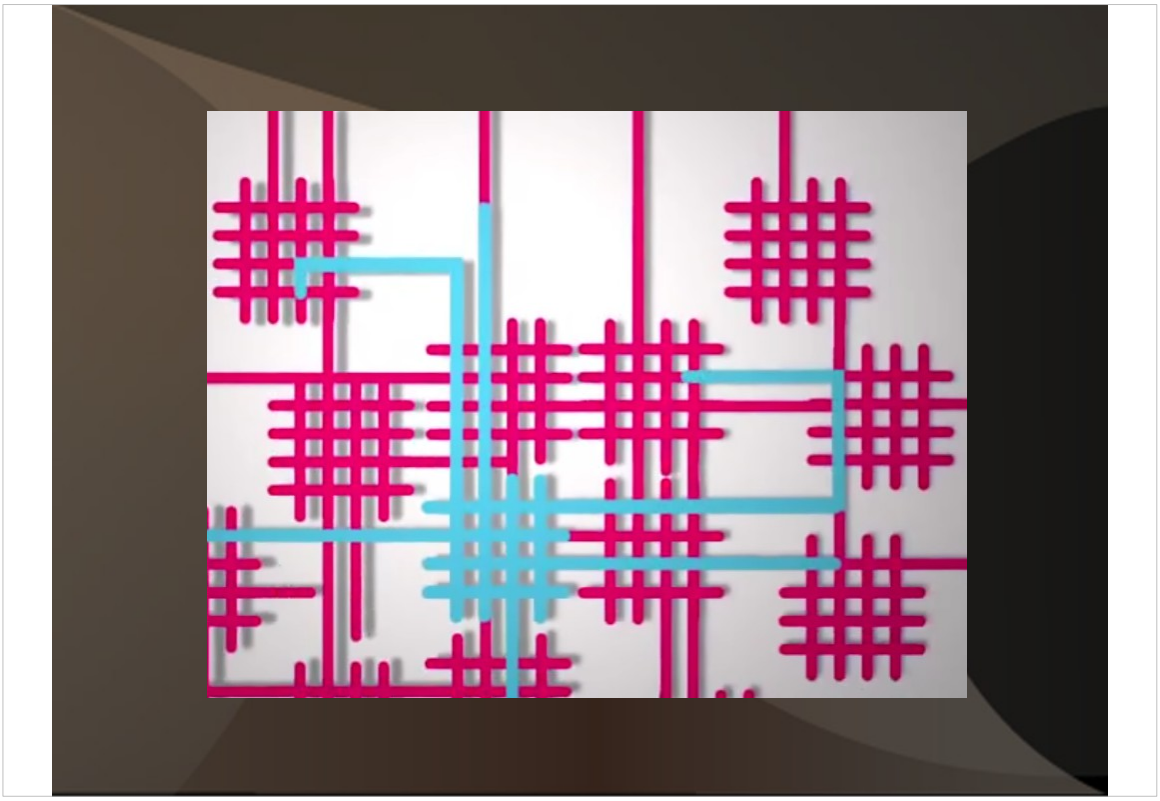
The video provides some great visual representations of the potential of linked data. It's also short, gets the point across really well and is quite an entertaining piece of work.

There are however a couple of key points and images that I want to share with you that come from this video.



The first thing is that we need to open up our data. Currently, our records and our catalogue software effectively box our data in and make it very difficult for others to use the rich metadata that we have created.

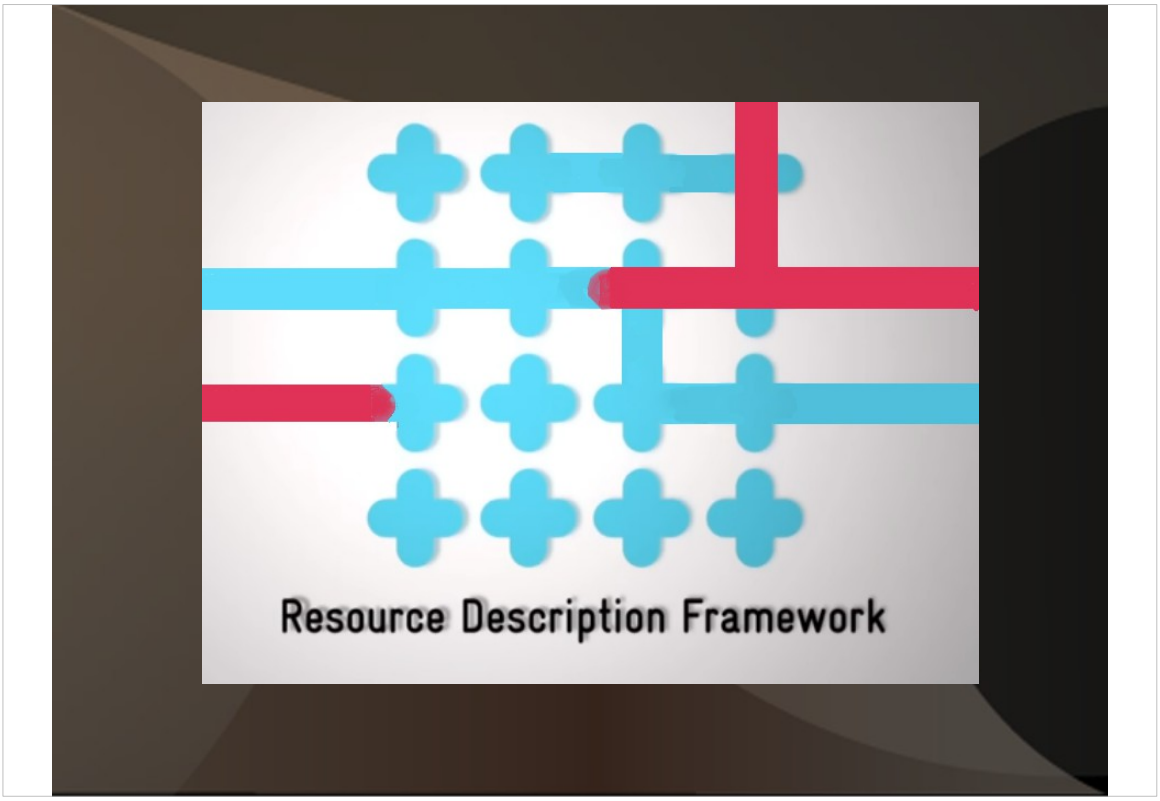
Although things are changing, our bibliographic data is really only available internally and we need to work to remove these barriers.



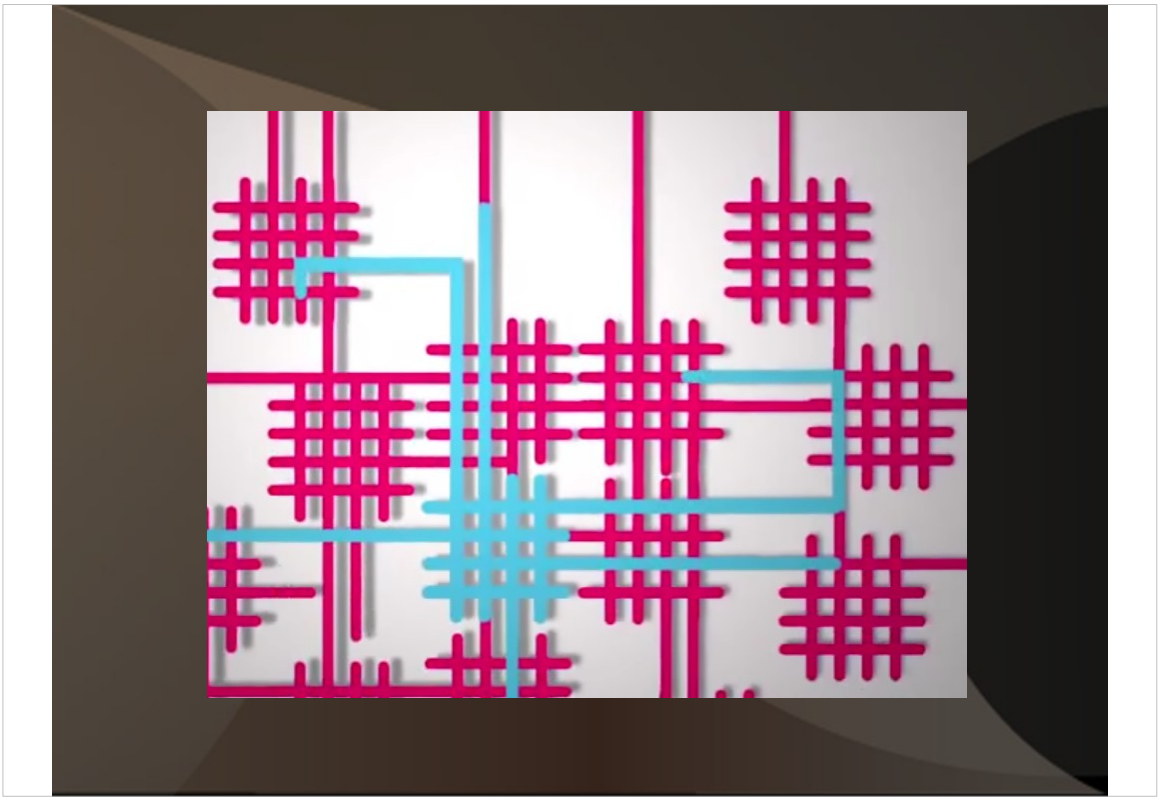
If we can succeed in doing that then our data can more easily link up with other data sets on the web providing connections to resources and revealing connections we may never even knew existed.

And the second thing, is that we need to express our data in the language of the semantic web. Which at this point is the Resource Description Framework or RDF.

If we use the RDF data model it allow us to provide better internal connections within our own catalogues **and** it will improve our ability to link to data sets that exist outside of our own collections.

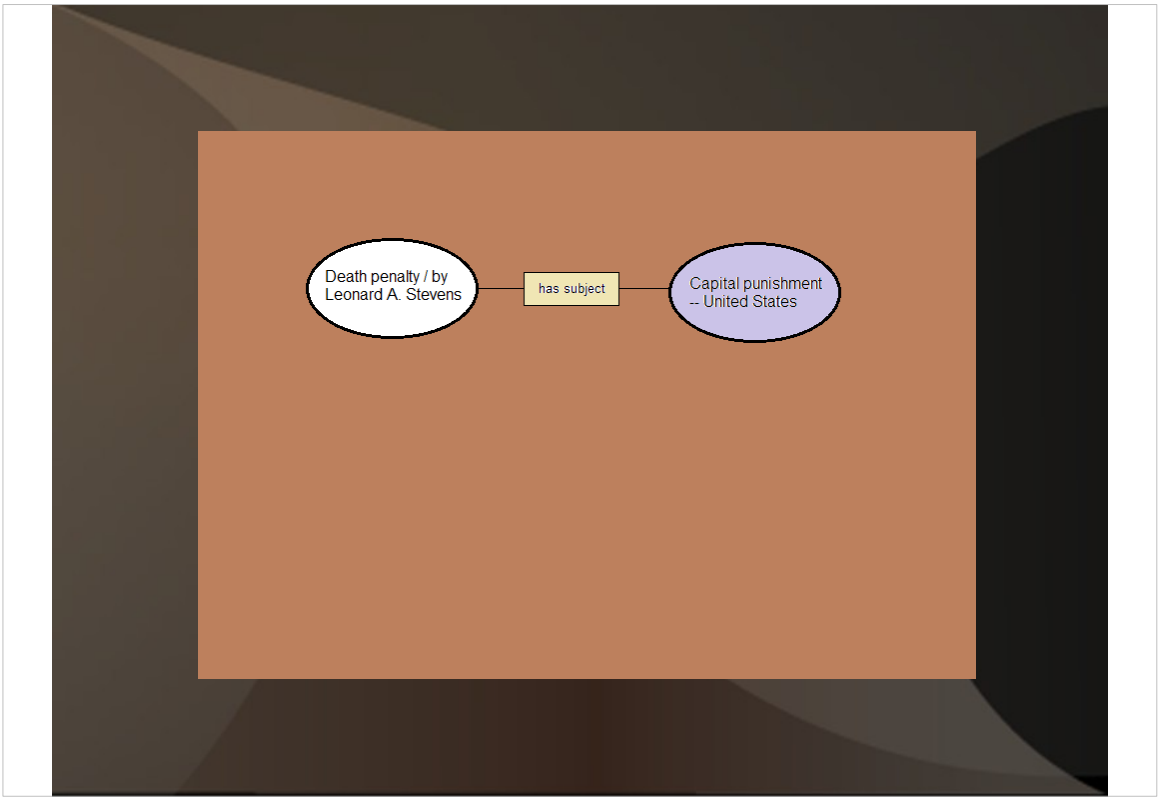


And this framework will also make it much easier for others to link into our data ...



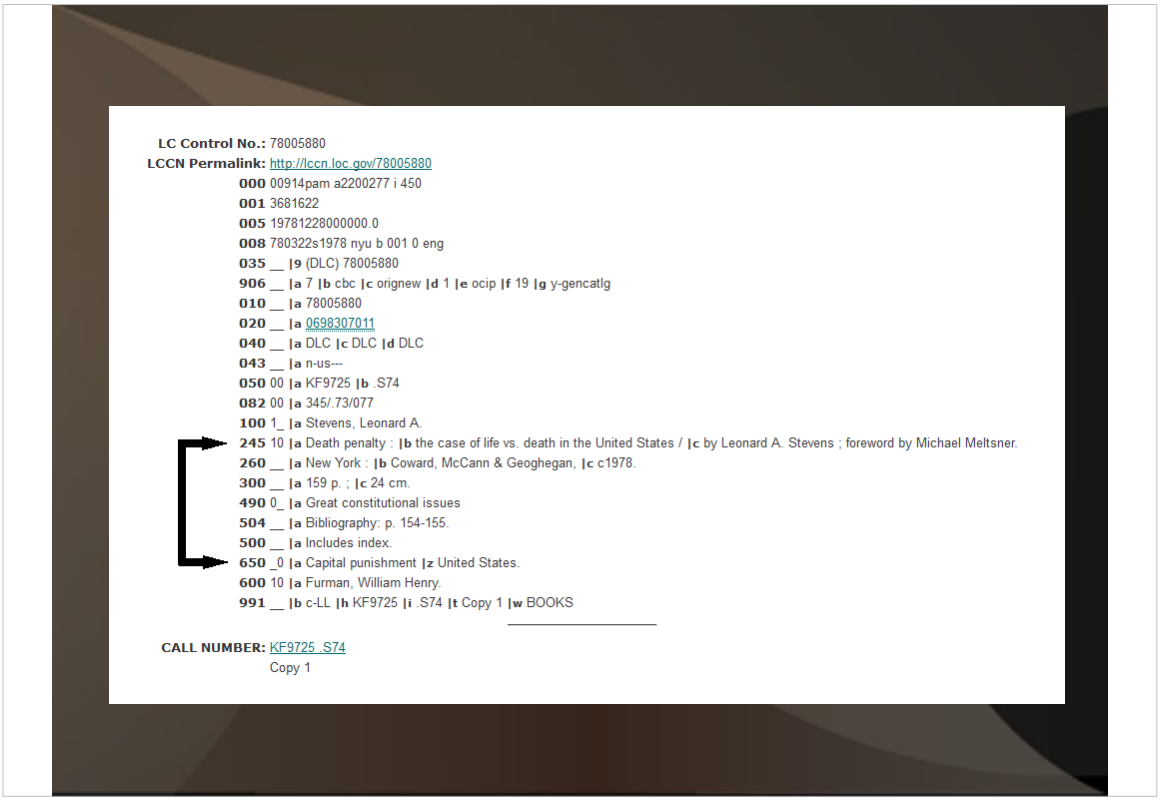
... so that we can become part of this emerging web of data.

The basic building block used in RDF is the 'triple'.



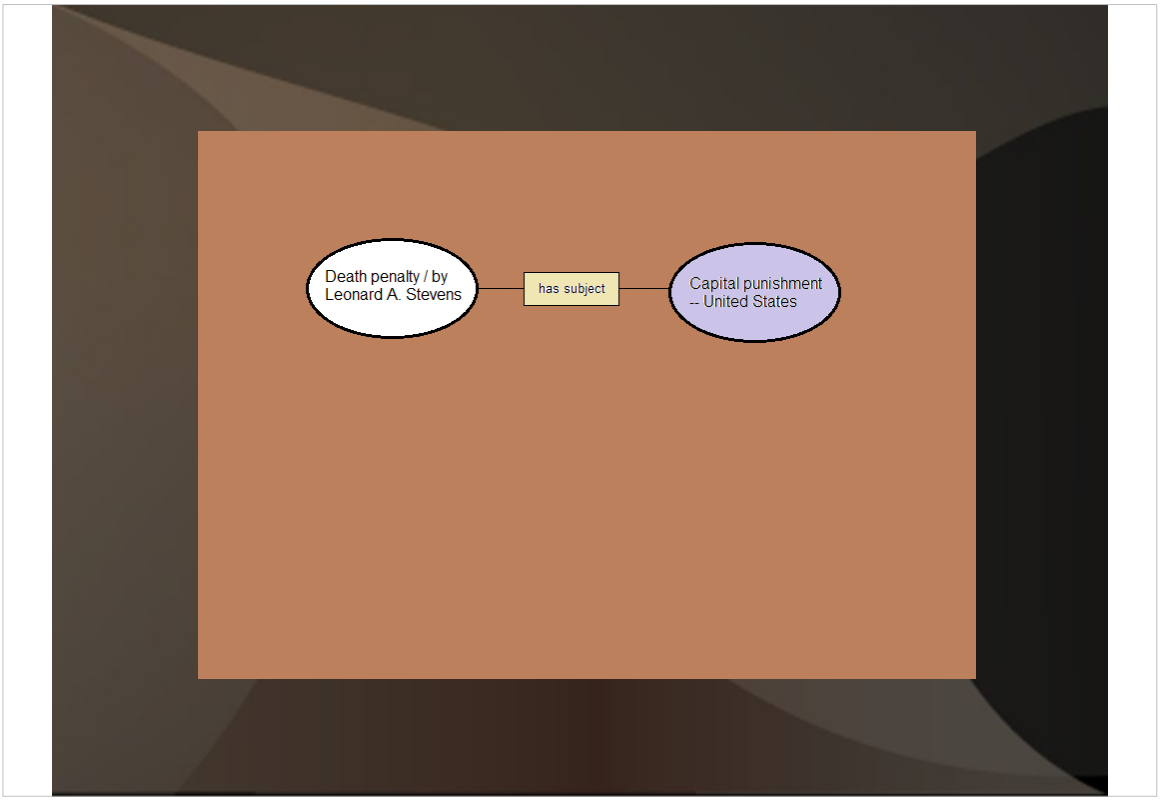
It's called a triple because there are three parts in the data statement. Here you can see that: this book -- has -- this subject.

We've got the parts, we just need to express them in terms that other data creators can recognize and use.

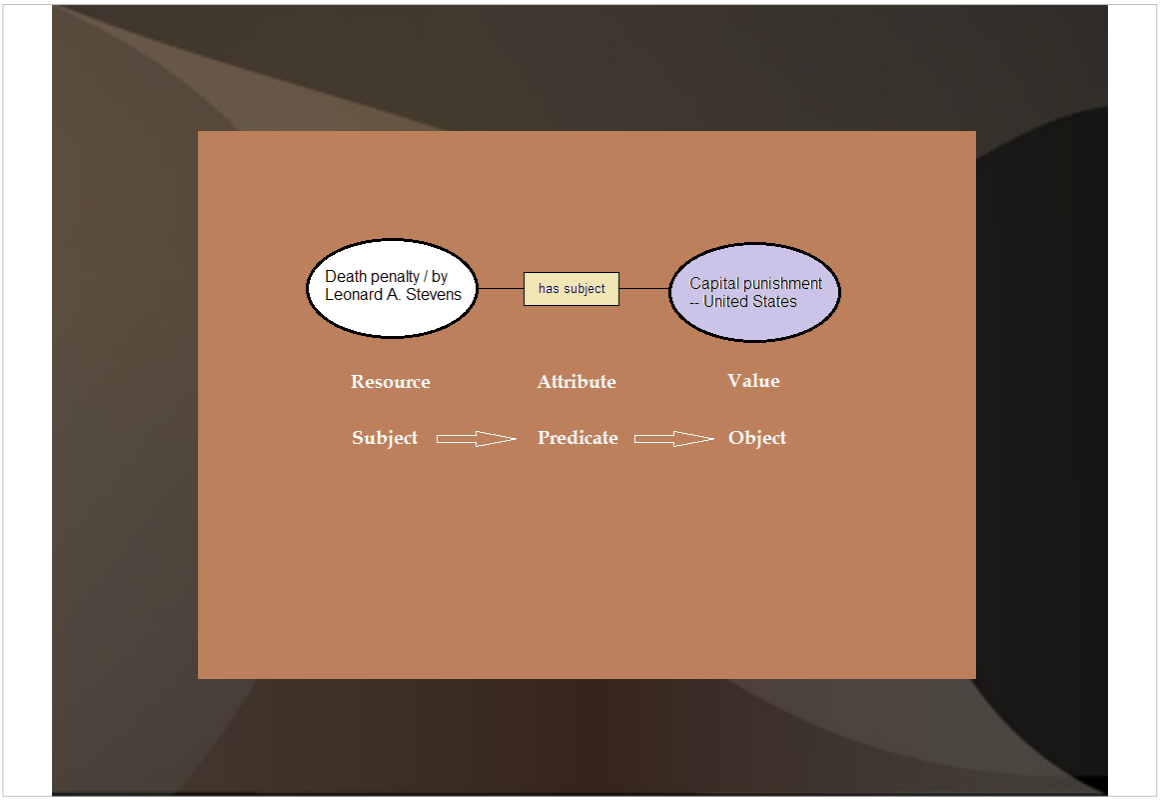


Here's how we currently represent that same data statement in our library catalogue. It's part of the information we provide in our MARC record ... the book with the title 'Death penalty' has the subject 'Capital punishment—United States'.

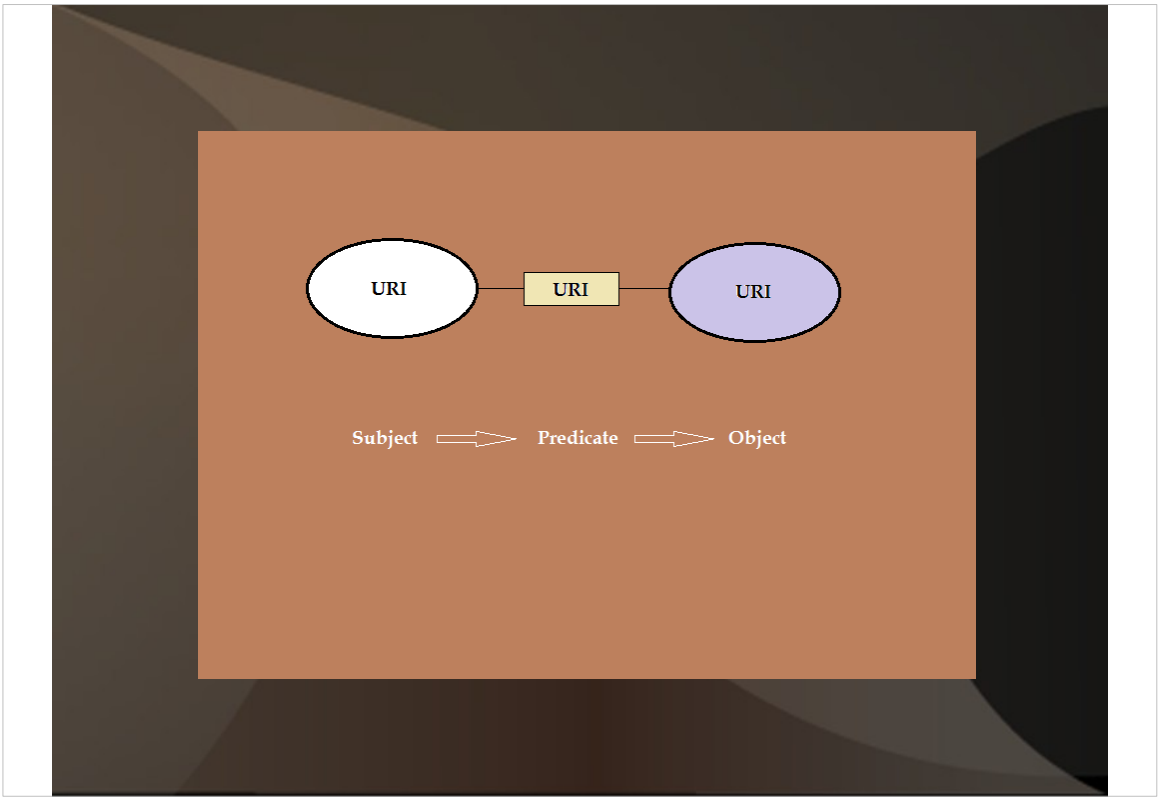
In our current environment we know that this subject relates to this title because both of these MARC fields appear in the same MARC record.



But that same information can just as easily be conveyed using an RDF triple statement like this one here.



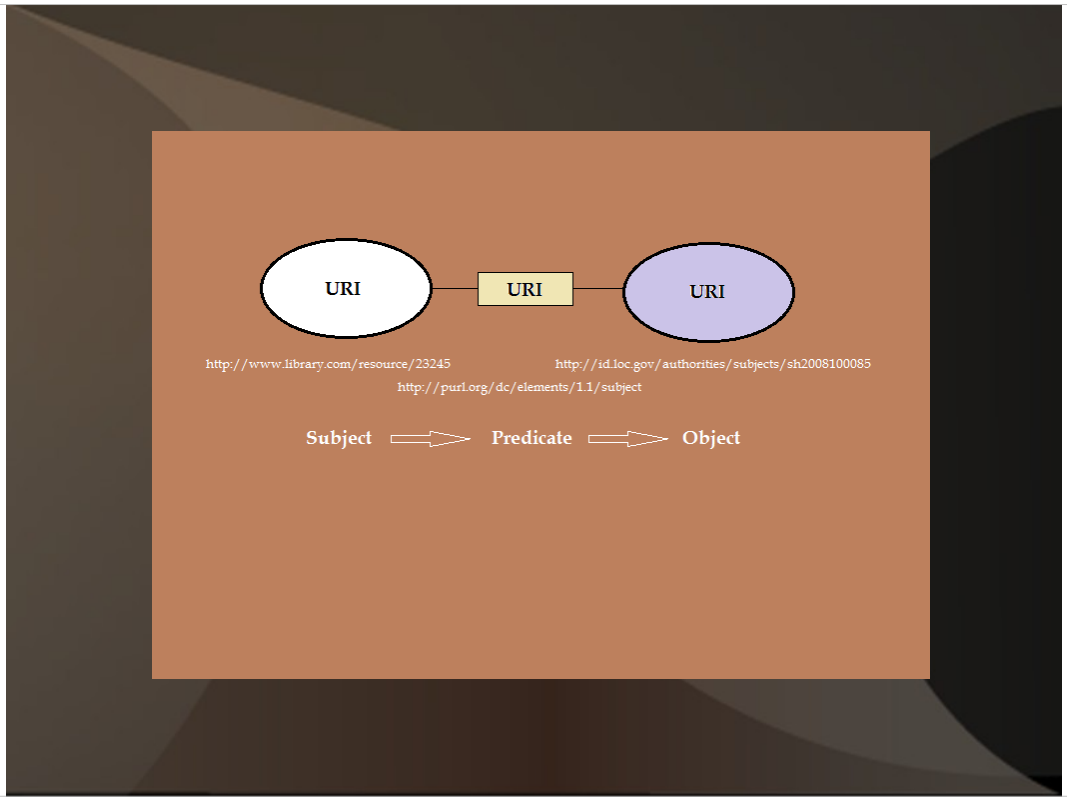
The three parts of an RDF triple are described this way: there is a 'subject' (in this case the title 'Death penalty'); there is a 'predicate' (hasSubject); and there is an 'object' (the value 'Capital punishment—United States.'



One thing the library community will have to get used to is the application of uniform resource identifiers, or URIs, instead of using a string of text to represent each of the parts in a triple.

Text strings are meant to be read by people. But a URI is something that a computer can interpret more easily which makes this data statement “machine actionable”. That means that the machines on the network will be able to make accurate connections to data between data sets.

It can be converted back into a text string when presented to our library users.



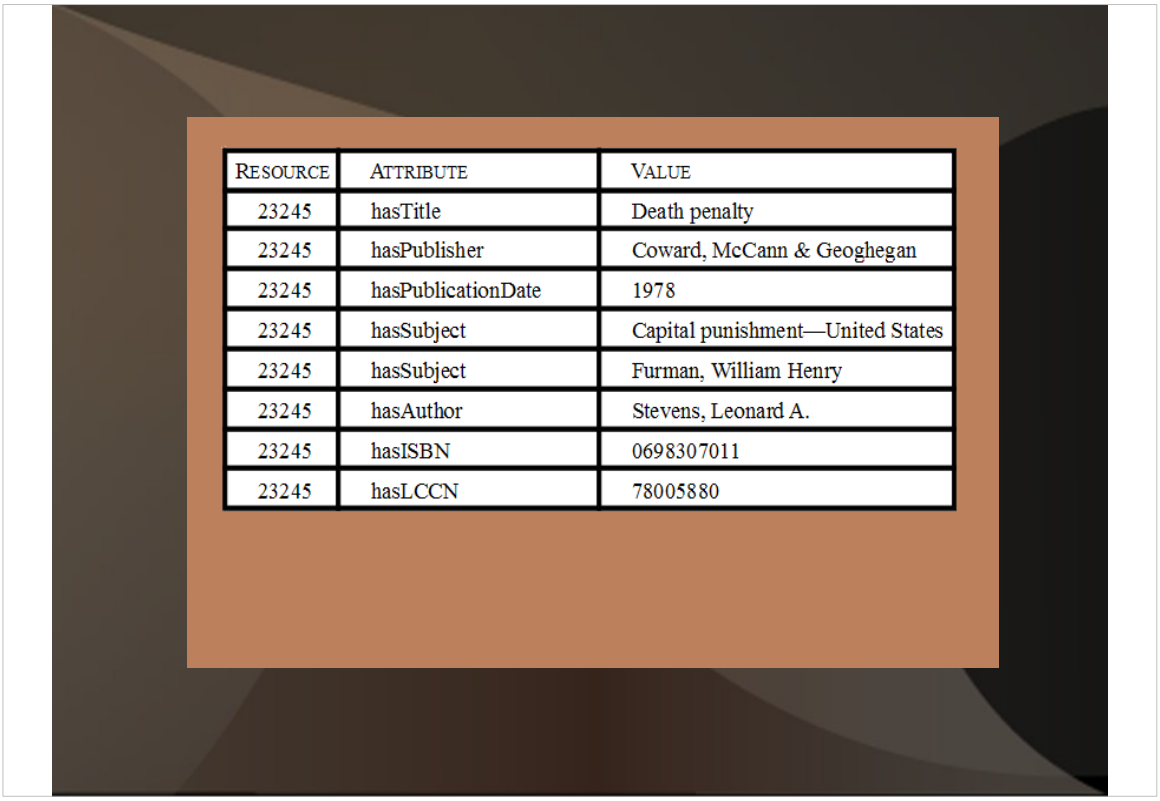
URIs for this triple might be something like this:

-there would be a URI for the title or an ID for the resource itself

-a URI for the predicate, in this case the Subject element from the Dublin Core element set

-and a URI for the actual subject, here a link to the LC authority record for Capital punishment—United Sates.

The Library of Congress recently announced that they have released their subject headings as linked data. So there will now be official URIs available for us to use.

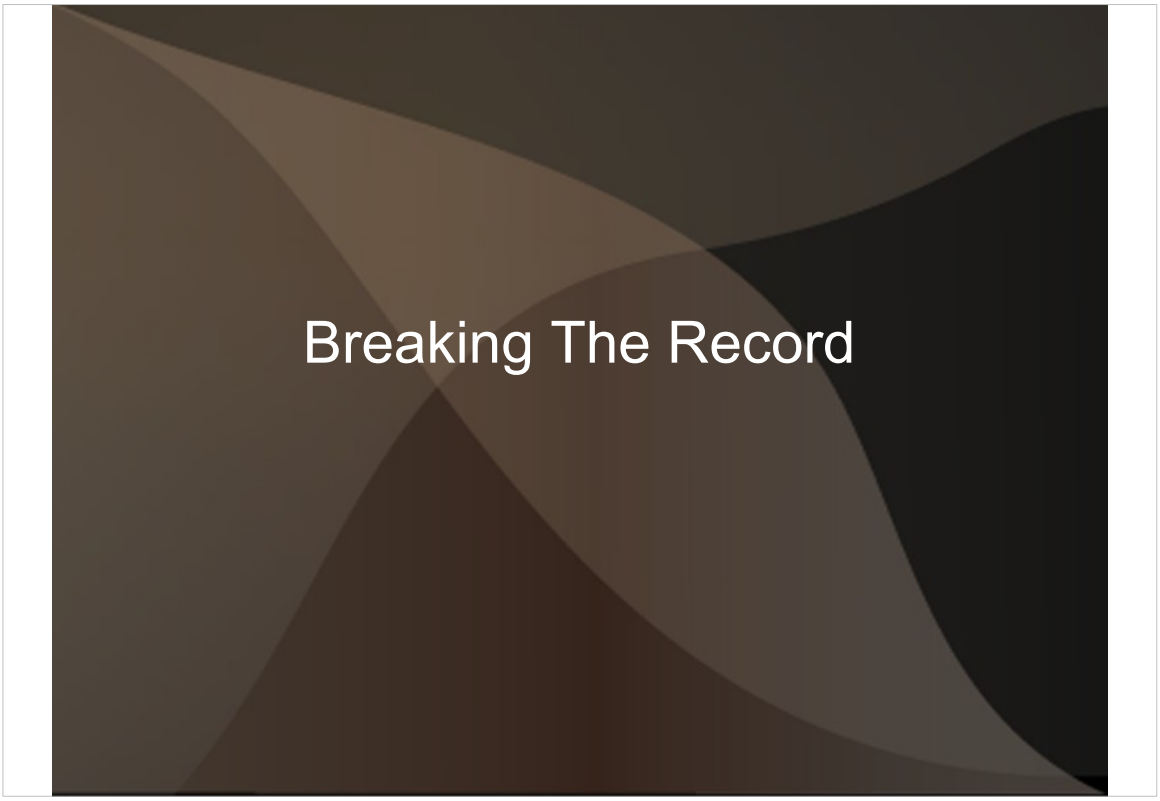


RESOURCE	ATTRIBUTE	VALUE
23245	hasTitle	Death penalty
23245	hasPublisher	Coward, McCann & Geoghegan
23245	hasPublicationDate	1978
23245	hasSubject	Capital punishment—United States
23245	hasSubject	Furman, William Henry
23245	hasAuthor	Stevens, Leonard A.
23245	hasISBN	0698307011
23245	hasLCCN	78005880

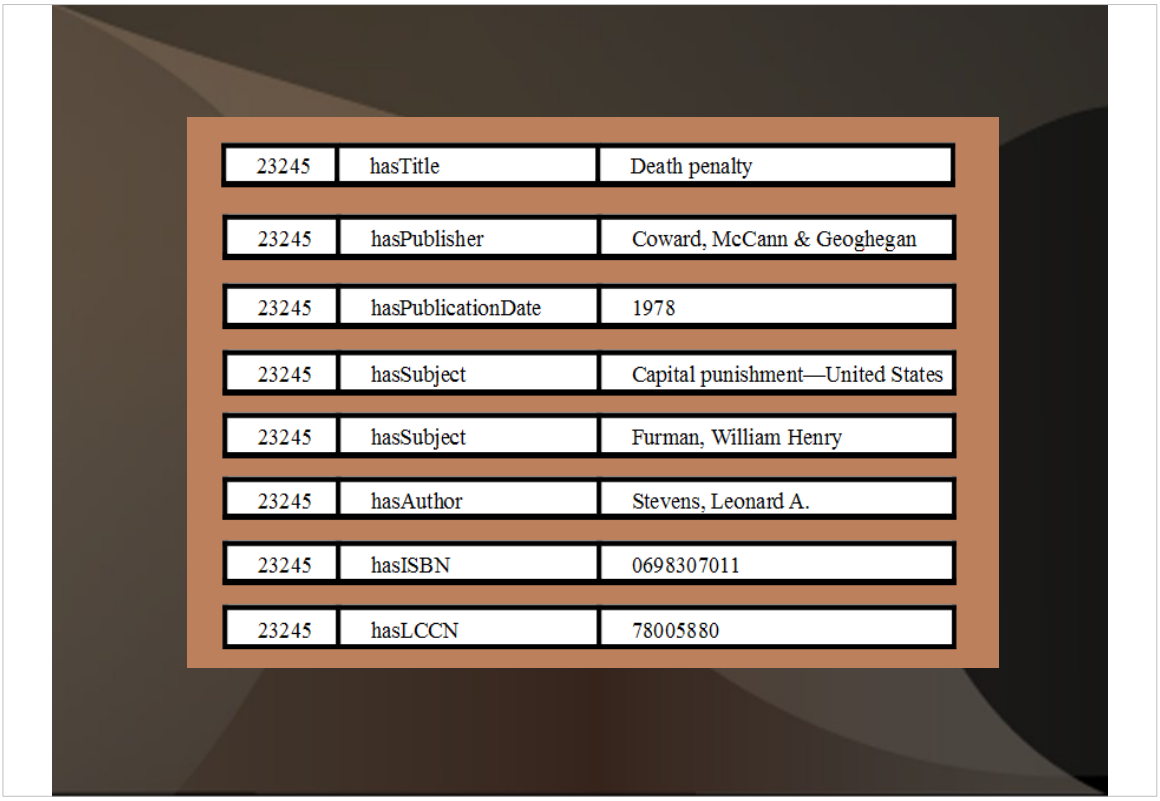
We could also express these relationships as a set in a table something like this.

Gordon Dunsire uses a similar table in his presentation, “Linked data and the implications for library cataloguing” delivered at the Canadian Library Association conference in 2011.

And this is very similar to the way our records are represented in the data tables found in our integrated library systems.

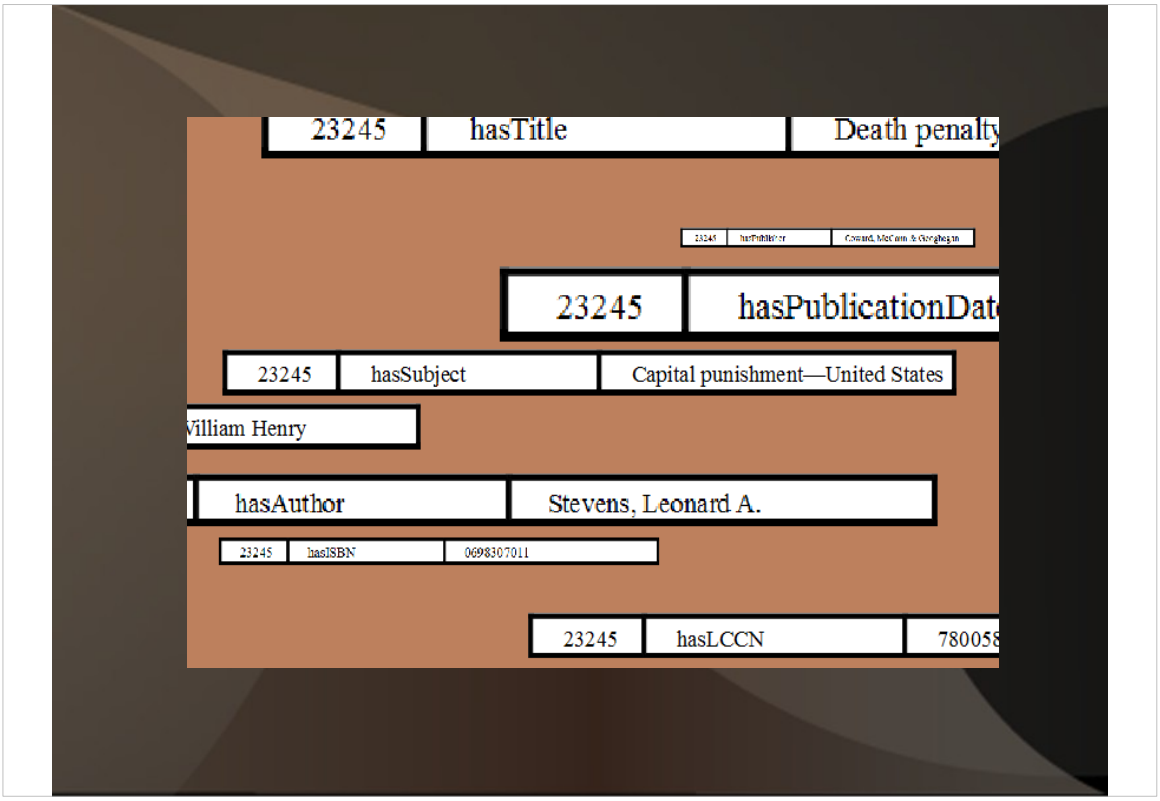


Once we've succeeded in doing that we can take the next step and ...

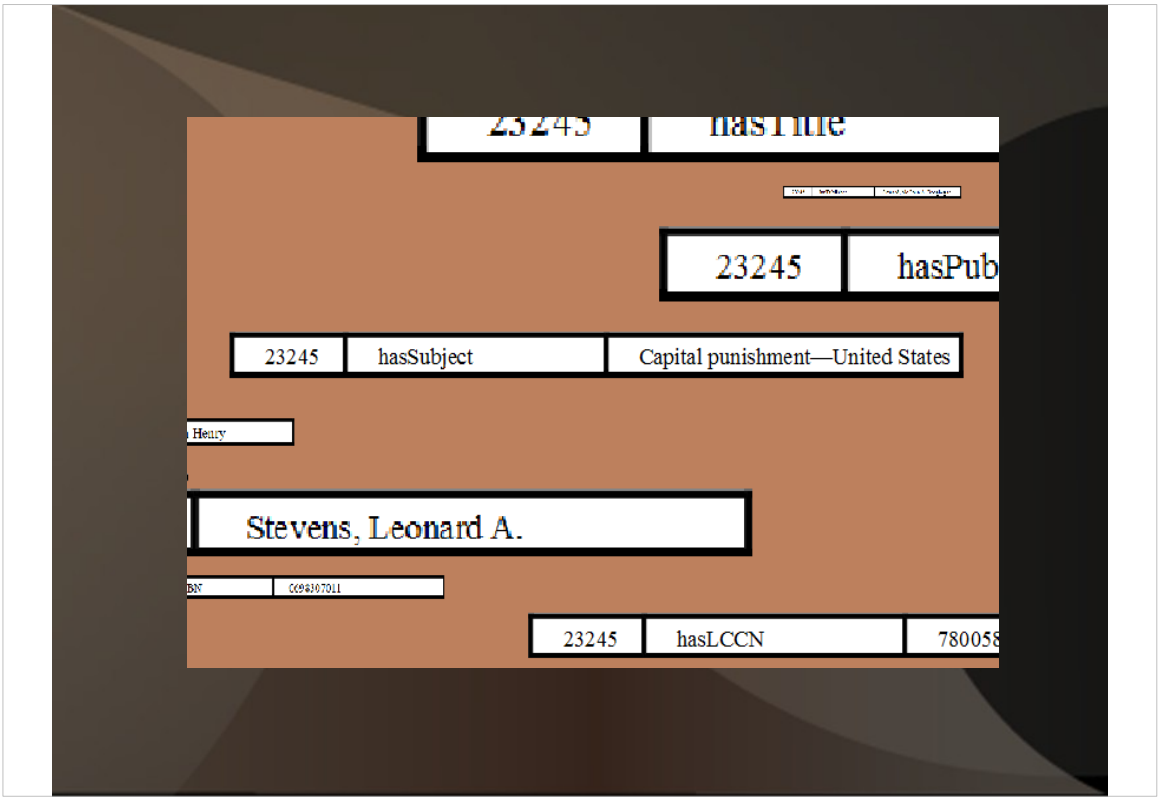


23245	hasTitle	Death penalty
23245	hasPublisher	Coward, McCann & Geoghegan
23245	hasPublicationDate	1978
23245	hasSubject	Capital punishment—United States
23245	hasSubject	Furman, William Henry
23245	hasAuthor	Stevens, Leonard A.
23245	hasISBN	0698307011
23245	hasLCCN	78005880

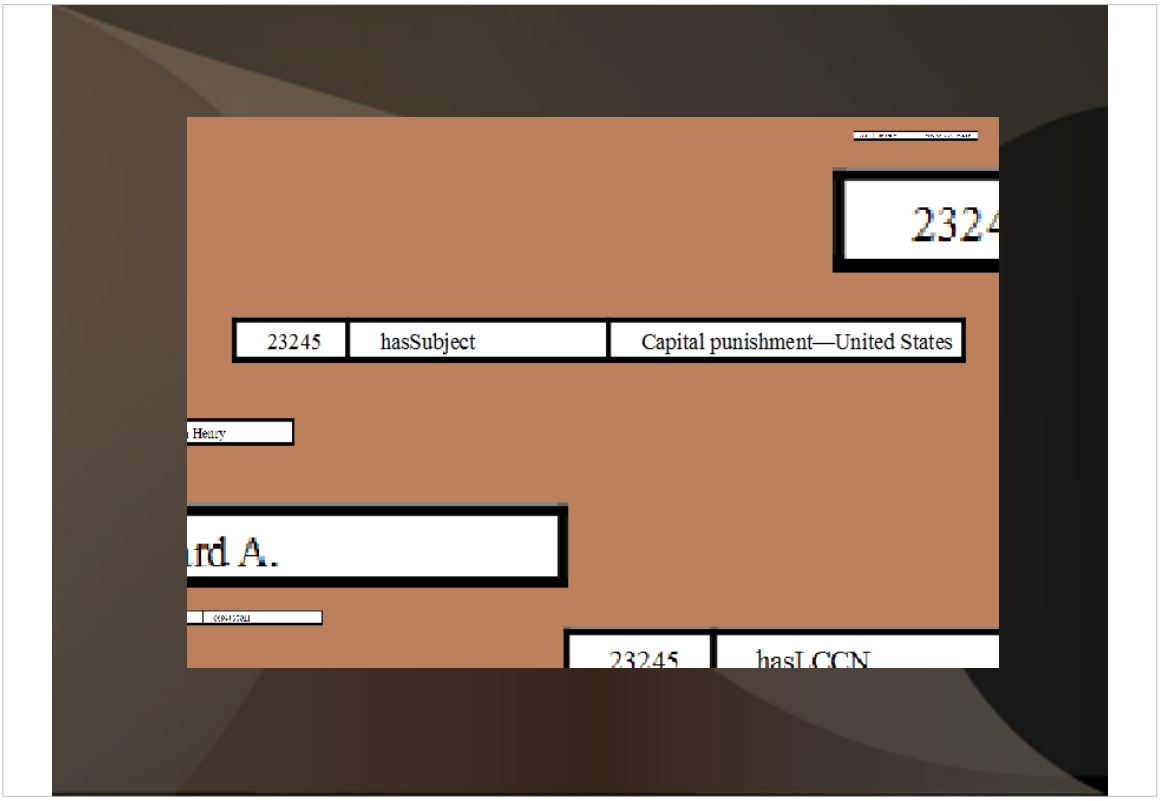
... we can break the record apart and free each of the data statements them from the data table.



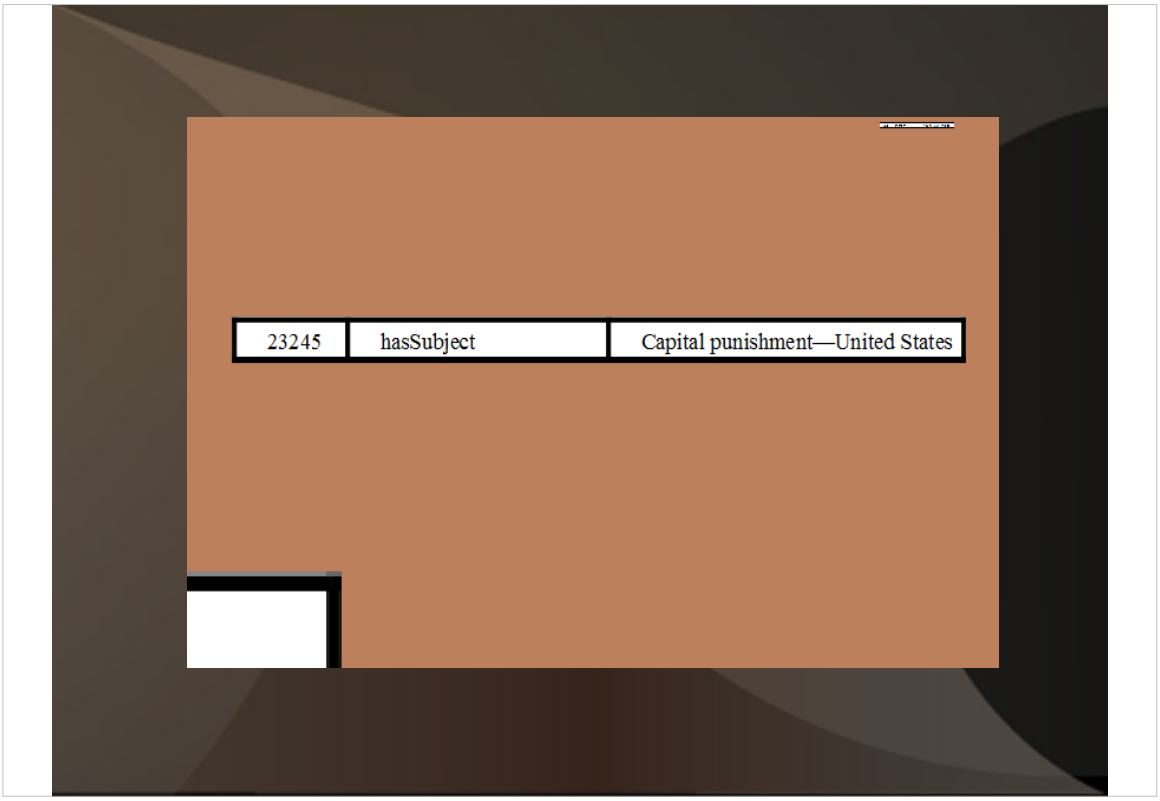
The triples representing the various relationships for resource "23245" are now free to float away ...



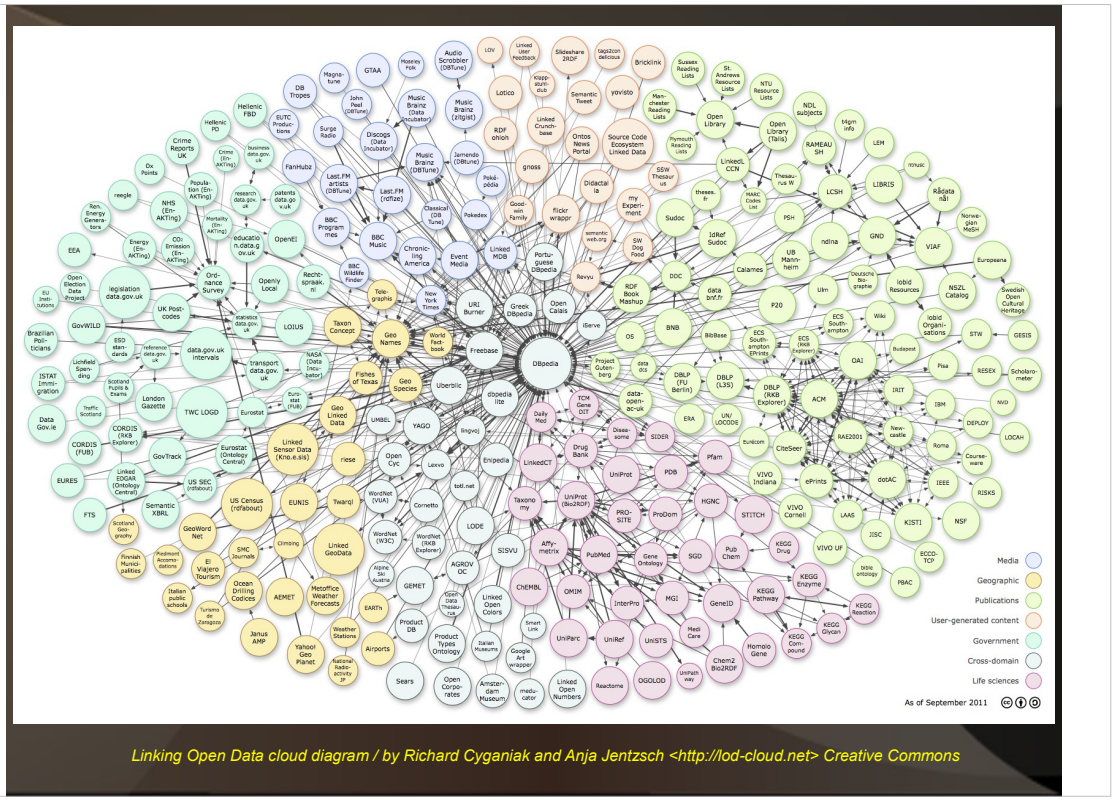
... into the semantic web ...



... and take their place ...

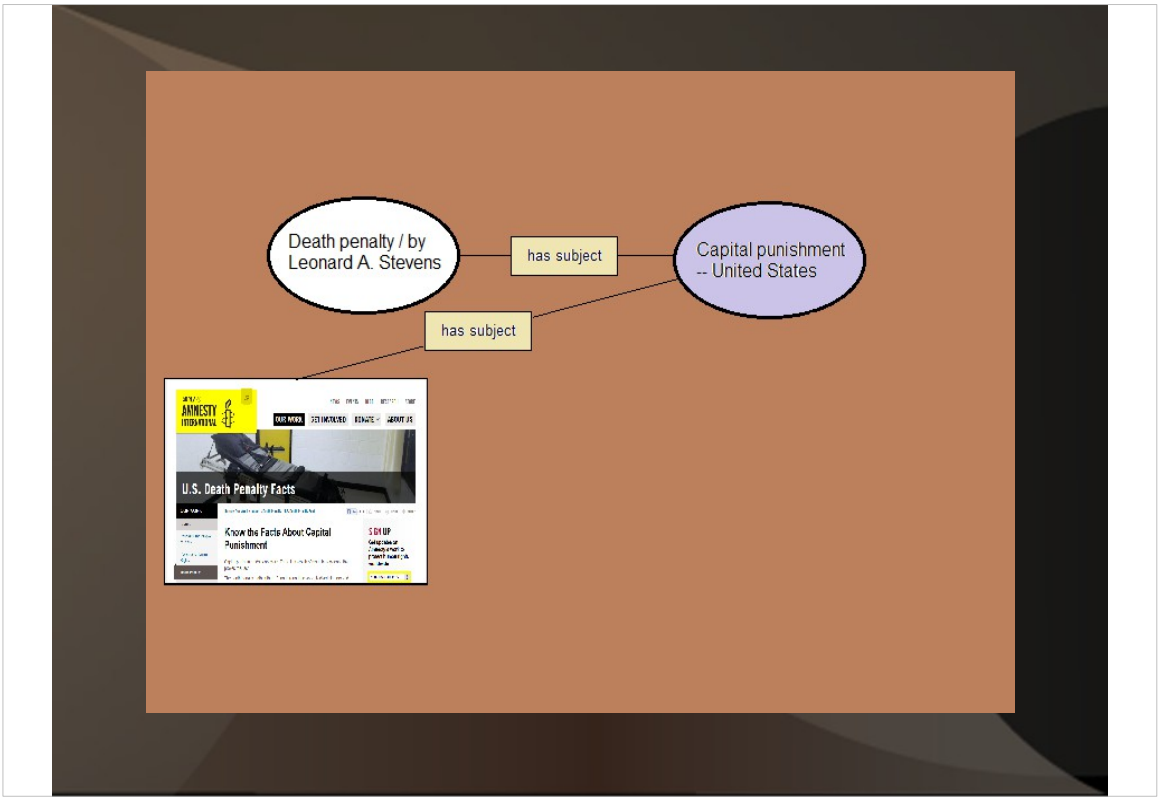


... with the other triples that form ...

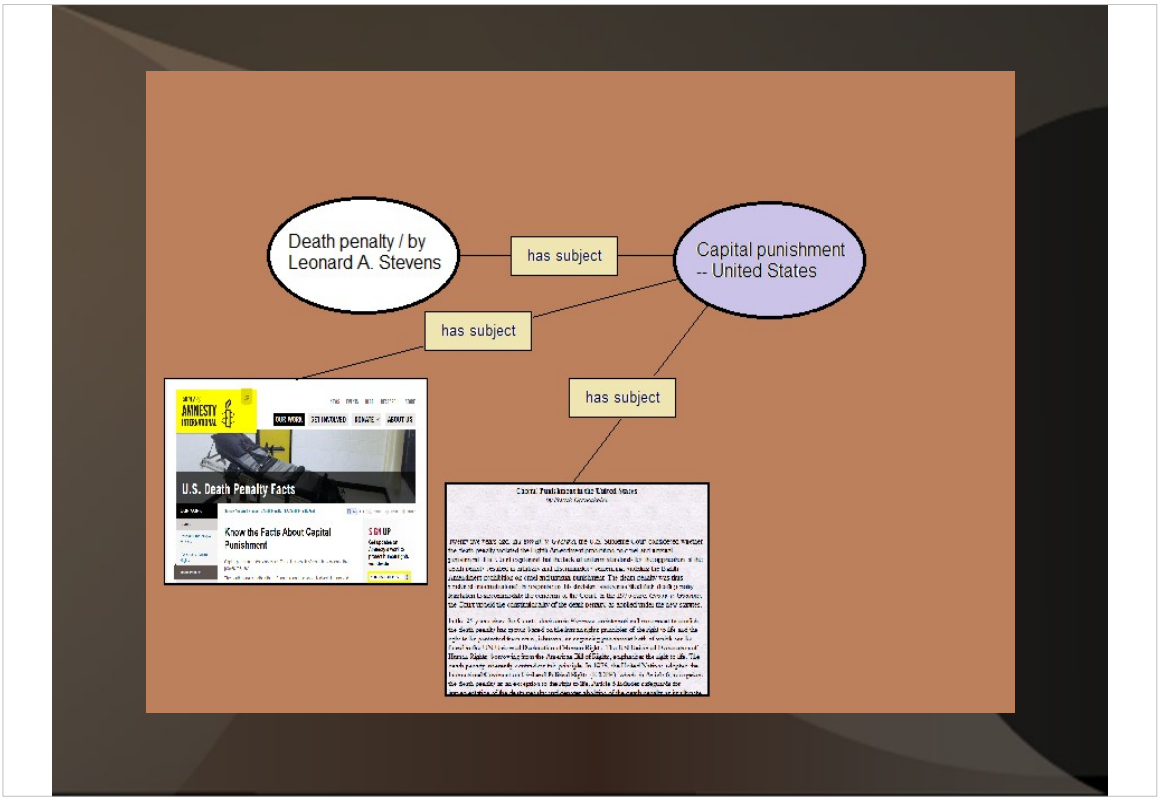


Linking Open Data cloud diagram / by Richard Cyganiak and Anja Jentzsch <http://lod-cloud.net> Creative Commons

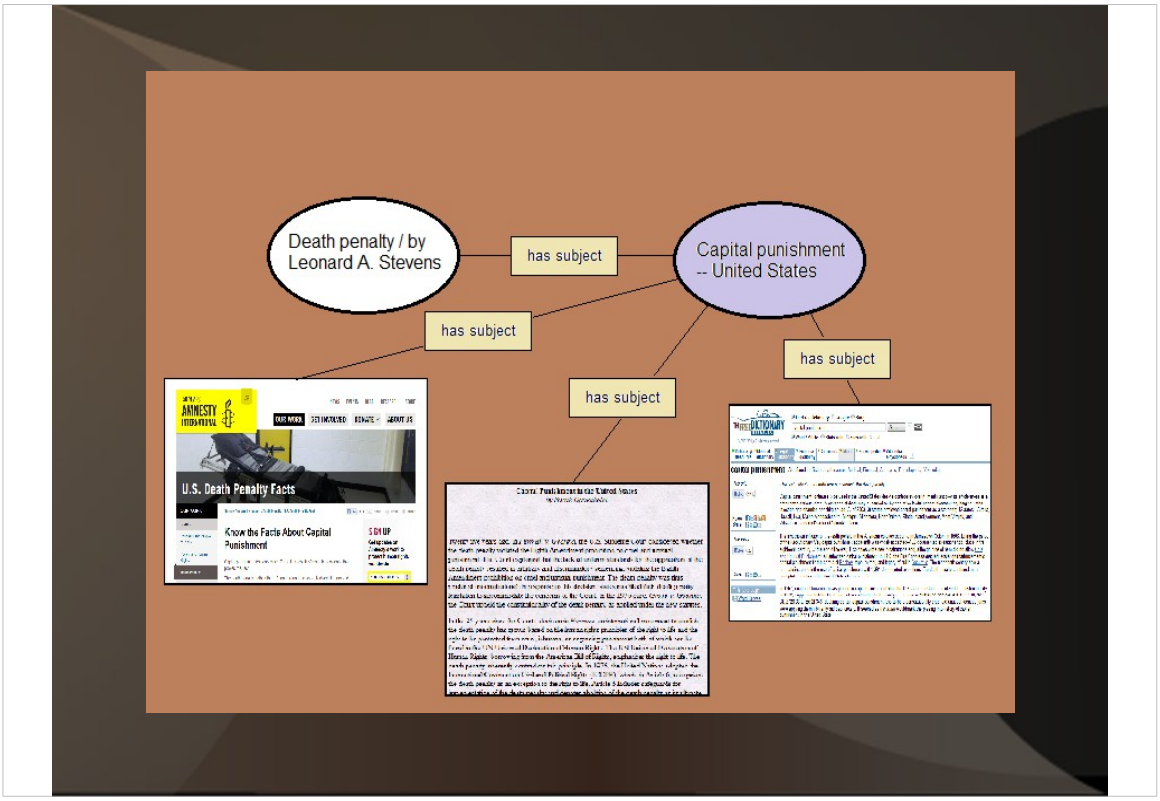
... the linked data cloud.



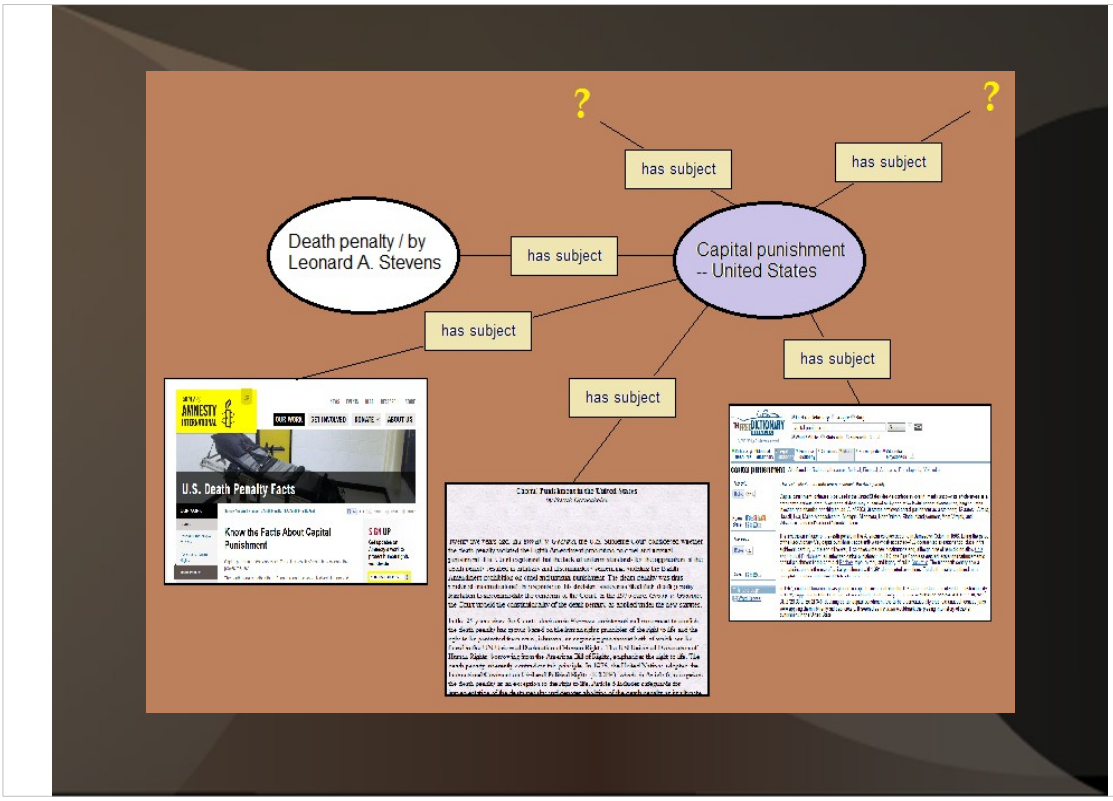
So now our bibliographic data might link up to this page about capital punishment in the U.S. by Amnesty International ....



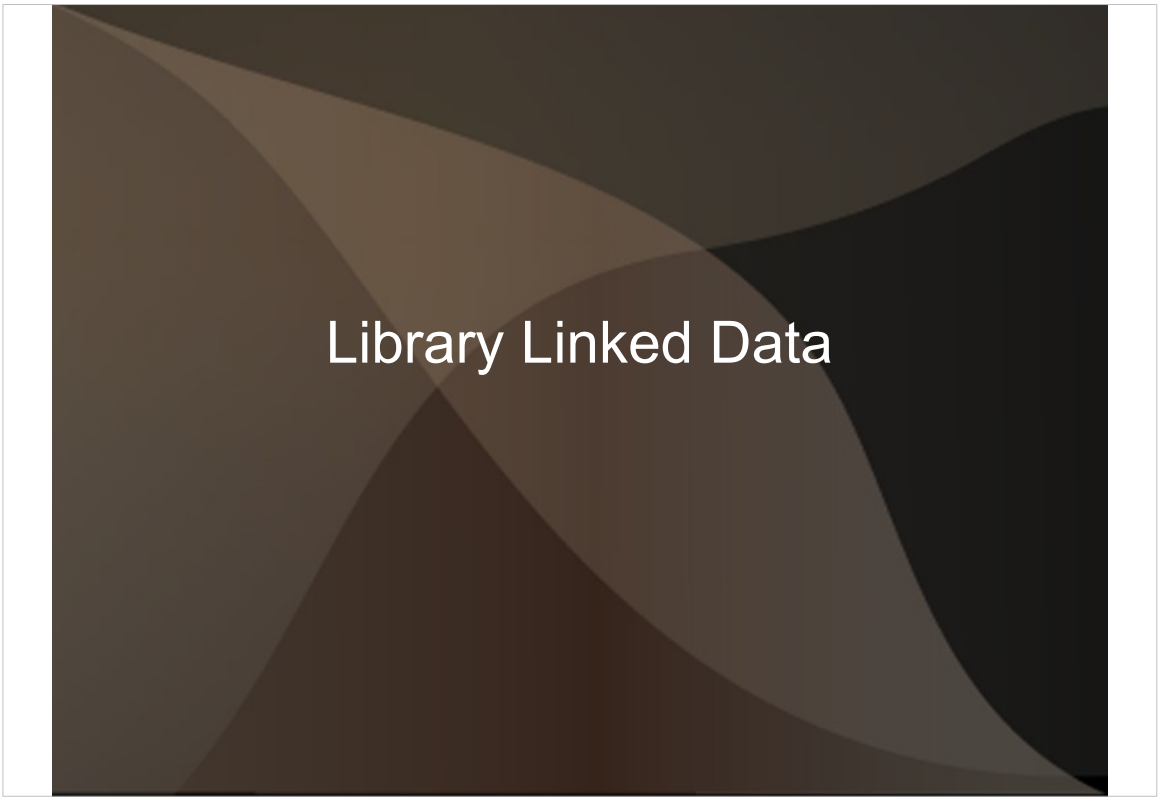
... or this article by Sarah Oppenheim, Chair of the Death Penalty Project at the National Lawyers Guild at American University's Washington College of Law ...



... or this entry on capital punishment from the Free Legal Dictionary by Farlex.

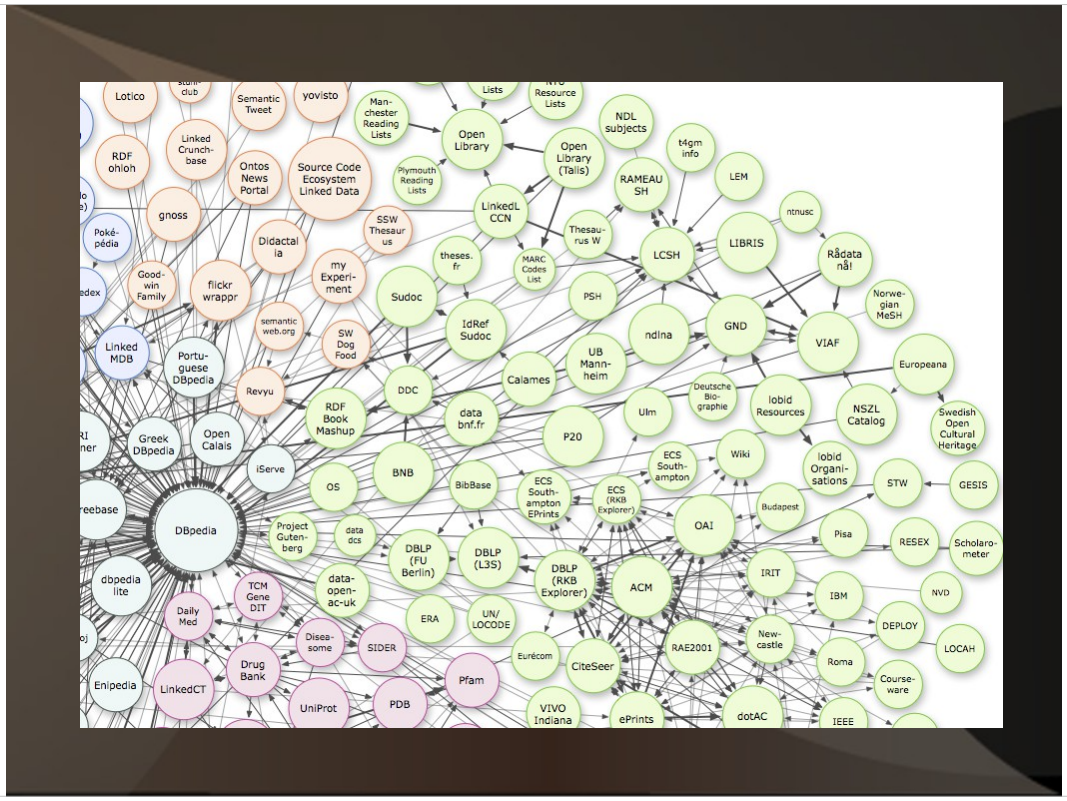


... or to any other resource that might have a relationship with this book on capital punishment in the United States.



There are many connections in the linked open data cloud including many triples from a number of library data sets.

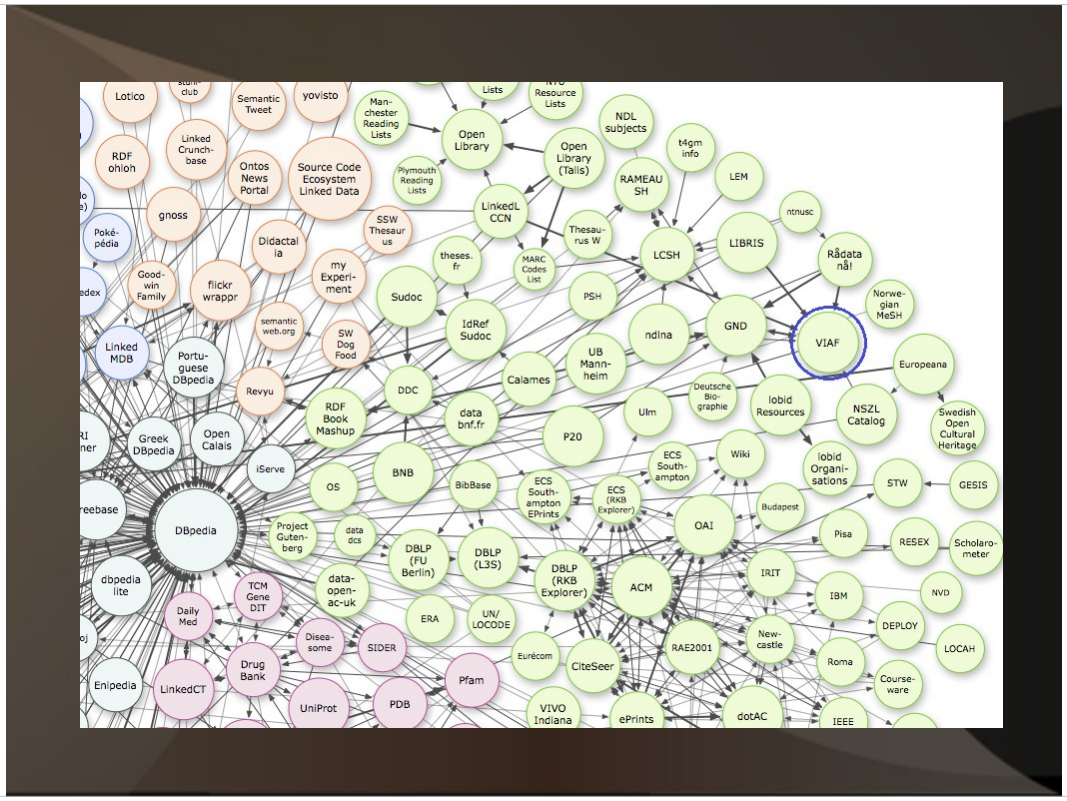
At the beginning of this presentation I said that this diagram represented about 295 data sets. That translates into about 31.6 **billion** triples.



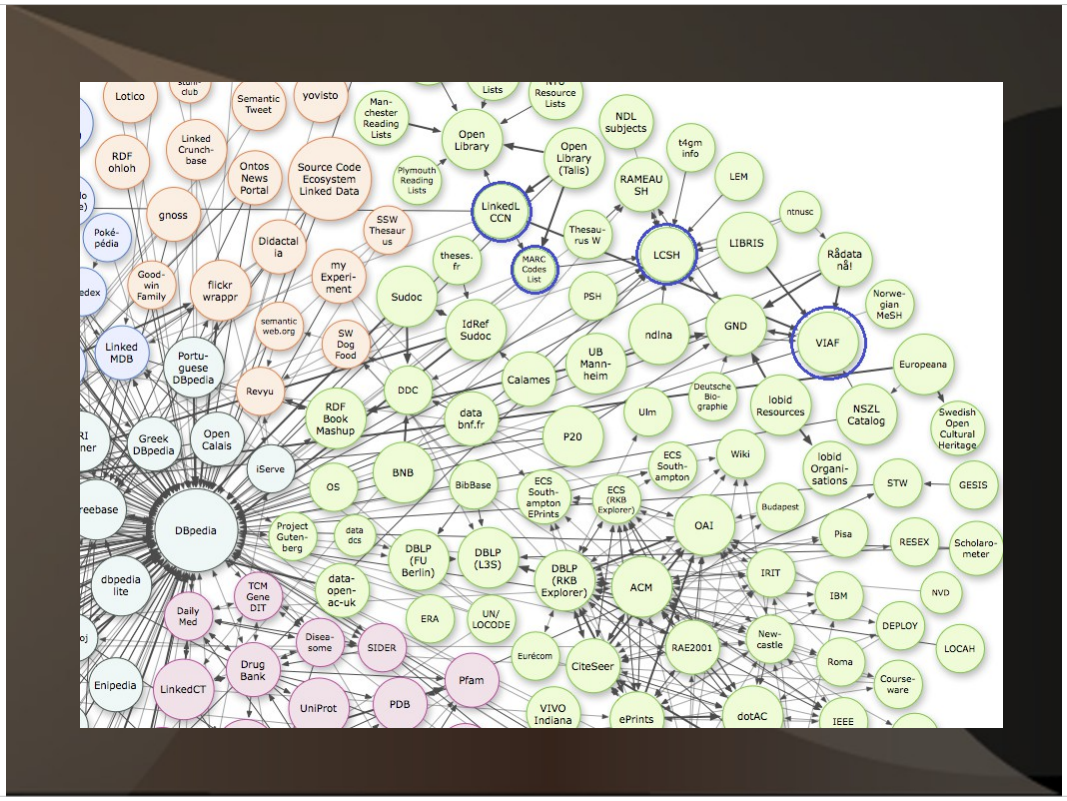
The library data sets are in the “Publications” section of the linked data cloud which can be seen here as the green circles.

This part of the cloud is what Gordon Dunsire has called the “library corner”.

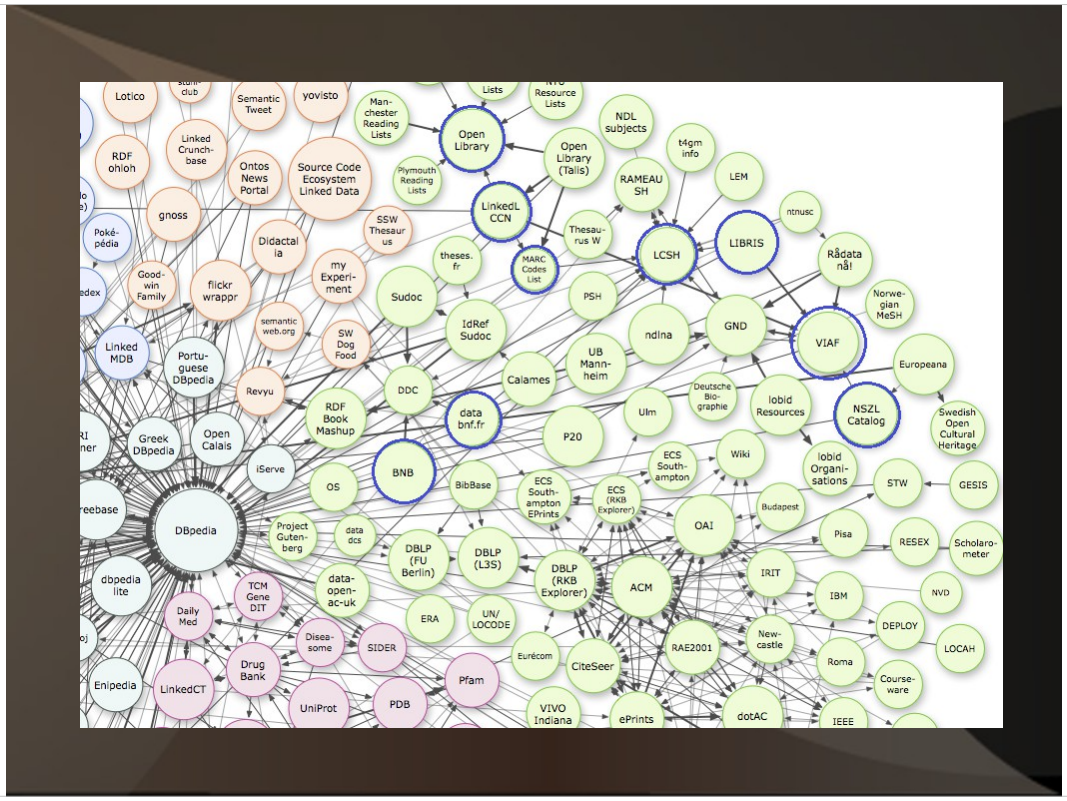
You can see some of the major library data sets in this part of the cloud.



There's the Virtual International Authority File (VIAF).



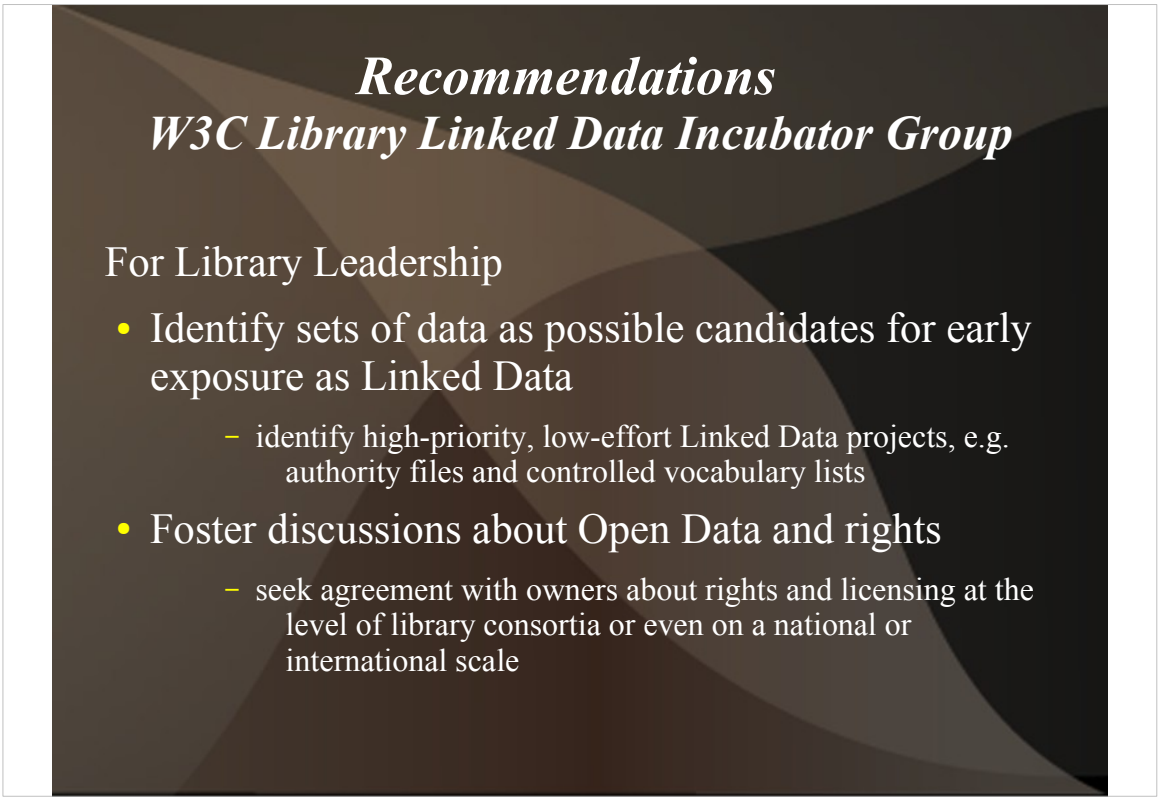
There is the Linked Library of Congress Control Numbers (LinkedLCCN), the MARC Code Lists and the Library of Congress Subject Headings.



There is also data from the British National Library (BNB) [85 million triples], the Bibliothèque nationale de France (data bnf.fr) [2 million], LIBRIS (data from the Swedish National Library) [50 million], NSZL Catalog (the Hungarian National Library) [19 million], and the Open Library [400 million].

Together these institutions have so far contributed about 765 million triples.

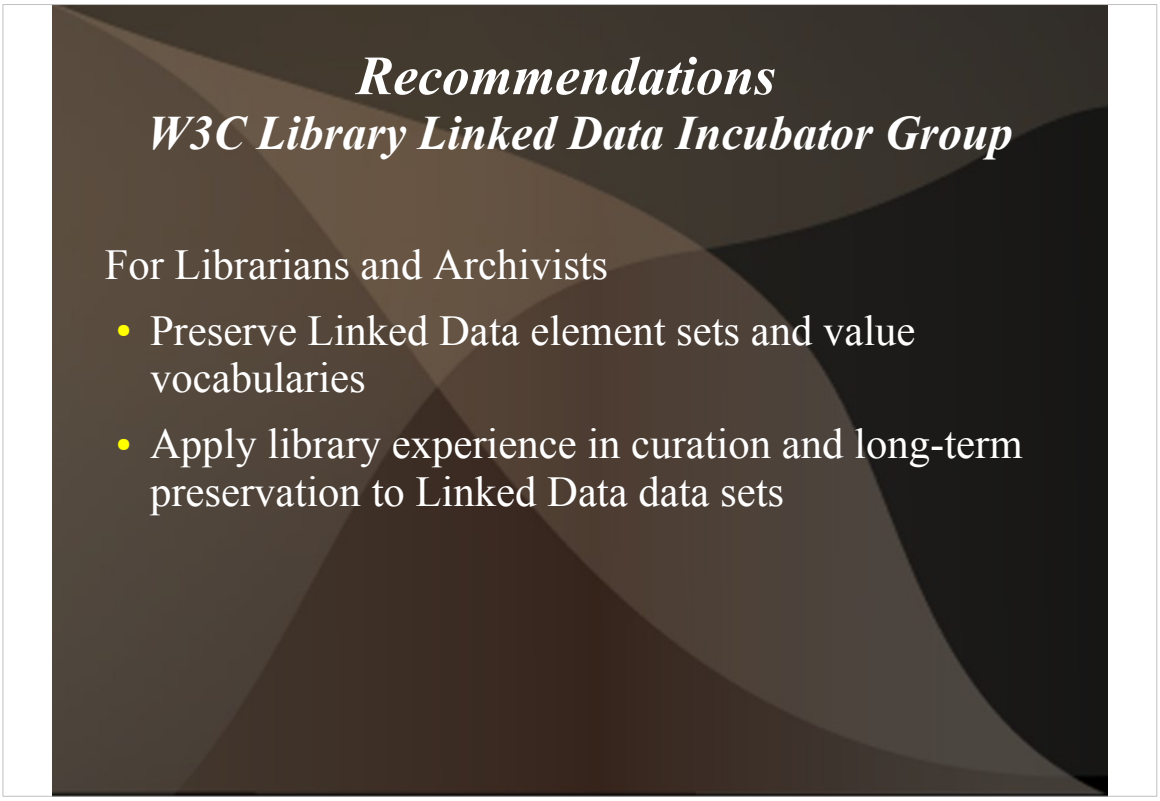
And there are many more bibliographic data sets already in the cloud and preparing to link into the cloud.



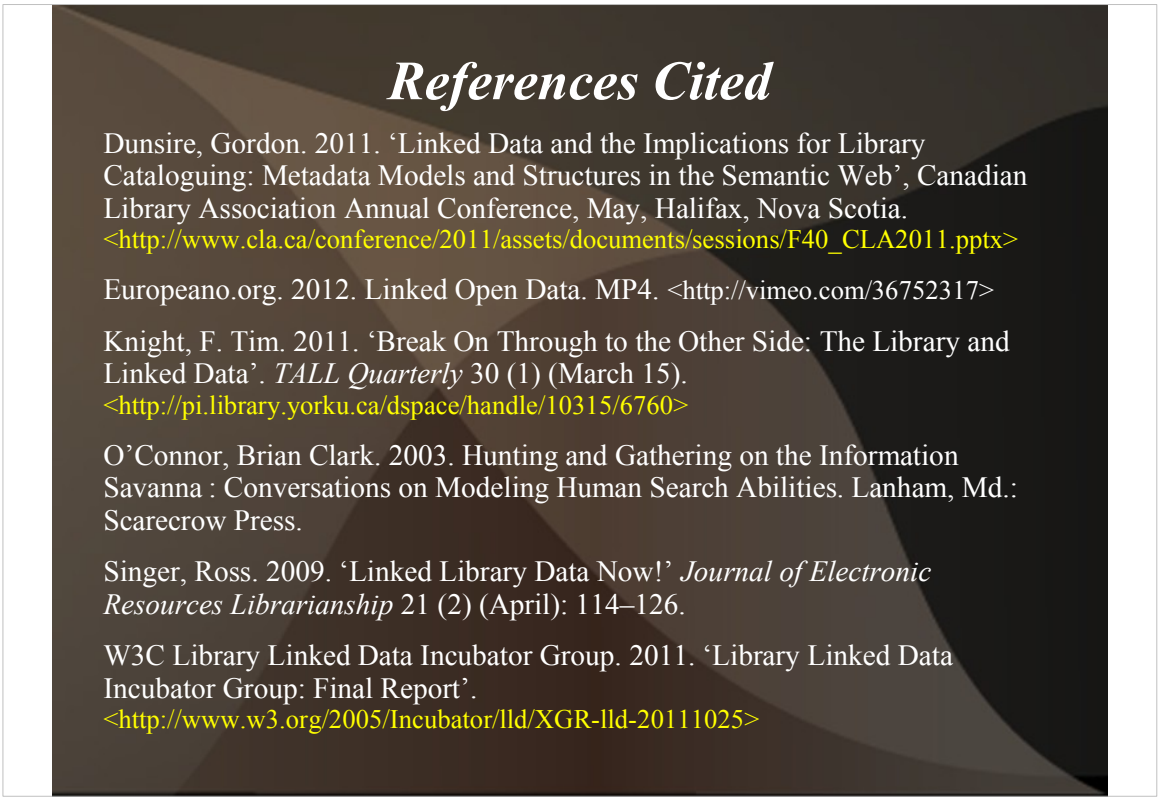
OK, so I'm just about done. I hope this has helped you to get a better sense of what linked data is and how libraries can play a role in this brave new world.

I want to leave you with some of the recommendations made in the Final Report of the W3C Library Linked Data Incubator Group which was released last October.

They recommend that for library leadership ... [read slide]



And for librarians and archivists they suggest that we [read slide] ...



Here is the information for the resources I cited during this presentation and I've also included a reference to an introductory article I wrote for the Toronto Association of Law Libraries last year.

Thanks for listening!

Thanks for listening!

tknight@osgoode.yorku.ca

Twitter: @freemoth

OSGOODE

OSGOODE HALL LAW SCHOOL
YORK UNIVERSITY