9-13-2013

Linked Data and Canadian Legal Resources

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

Follow this and additional works at: http://digitalcommons.osgoode.yorku.ca/librarians

Part of the Cataloging and Metadata Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Repository Citation
http://digitalcommons.osgoode.yorku.ca/librarians/24

This Conference Proceeding is brought to you for free and open access by the Law Library at Osgoode Digital Commons. It has been accepted for inclusion in Librarian Publications & Presentations by an authorized administrator of Osgoode Digital Commons.
First of all I’d like to thank CanLII for organizing this event, and Sarah Sutherland in particular, for inviting me to speak today ...

I'm here to talk about linked data and Canadian legal resources on the semantic web.
I'll begin with a quick review of the basic theory and building blocks that support linked data; then I'll talk about the growth of the so-called linked open data cloud; then we'll consider how this idea might be applied to legal resources and take a look at some of the challenges involved; and finally I'll touch on what organizations like CanLII can do to support these developments.
On the surface the basic theory of linked data is actually a pretty simple one.

It’s the practical and technological side of the equation that still holds many challenges and continues to make the implementation of linked data a rather complex exercise.

So for now I'll stick with the simple stuff ... :-(
So, first of all you've got a Subject ...
... and, an Object ...
... and in between these two things is, what's known in Semantic Web parlance as, the “Predicate.”

In grammatical terms the predicate tells us something about the subject, it’s a property or characteristic that the subject has.
I prefer to think of this as the Relationship: the thing, or property that connects the Subject to the Object.
When a Subject and an Object are connected by a Relationship we have a Triple.
Some common triples you might find associated with legal resources could be:

- a Case has a Citation
a Court has a Name
a Resource has a Format
etc.
If we look at a simple appellate case like this one, we can pull out some of the data elements and then express them as a collection of Triples.

For example, this entire set of data elements could be associated with the Subject of a bunch of Triples that we might call, “Case A.”
So, for example:

Case A has docket number C57108
Case A has plaintiff Royal Bank of Canada
Case A has defendant Leslie King
Case A was heard in the Court of Appeal for Ontario

And so on ...
If we then list these four triples you'd get a collection of triple statements like this. The Subject for these triples is ‘Case A’ each followed by a relationship or property of some sort and a value corresponding to that relationship or property.

For those of you familiar with relational databases …
… you might think this is starting to look very much like a table with a primary key set as ‘Case A.’

And traditionally that's how data like this has been represented: as a collection of data elements in a database record.

And really there's nothing wrong with that.

However, this data is confined to the relational database within which it was created.
The beauty of linked data is that it allows the record structure to break down ...
... with the resulting triples able to connect independently with information from any number of other data sources that might be out there.

The triples are available to float freely in the semantic web.
You could still bring all of these data elements together as a group with a query that sets out to find our original set of ‘Case A’ data elements.

And we could represent the result with a graph like this one where the circle in the centre is the Subject ‘Case A’ and the spokes around it are the various relationships and their associated objects.

But once our data is broken up into Triples other connections can be made automatically between data sources which can both enhance our data and enhance the data that others provide to us.
And suddenly, and I imagine quite quickly, once we get enough data to reach a critical mass, our data begins to link up with someone else's data ...
And before you know it there is a chain of resources linked together through their common data points.
Creating what I imagine as these crystalline structures that will be able to connect up in many different ways. Connections with the potential to reveal new relationships between the many Subjects, Objects and Relationships in existing data sets.
When I was in library school in the 90s we were introduced to the work of Vannevar Bush and in particular his article “As we may think” describing what he called the Memex system. Although this paper was published in 1945 I return to this periodically amazed by his forward looking ideas. Truly ahead of his time.

Well it seems that time is catching up to him and this passage citing what Bush calls 'associative indexing' is particularly relevant to today’s developments of linked data and how linked data works.

“... associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing.” — Vannevar Bush
So if I can take the liberty to make a minor adjustment and substitute a couple of contemporary terms it could be rewritten to read something like this:

“... associative indexing, the basic idea of which is a provision whereby any resource may be caused at will to select immediately and automatically another. This is the essential feature of linked data. The process of tying two items together is the important thing.”

And this process of tying two things together, the Triple, is really the kernel that leads to the potential of machine learning and machine actionable data.

OK, so let's take quick a look under the hood, so-to-speak.
Here’s an example of some RDF coded in XML. RDF stands for the Resource Description Framework and is one of the popular standards at work in the semantic web.

This particular example is based on the appellate case we looked at earlier.

I’ll just run through what this coding represents.
We start by declaring the namespaces to be used so that we know, and the machine knows, which terms and which standards we are drawing from.

The first few lines of code here respectively reference: the W3C RDF syntax document; the Dublin Core element set; and the last refers to a local namespace.

This last namespace is a fictitious one that I’ve called ‘lex’. Ideally this would lead to an actual metadata document containing the terms defined in that namespace.
Then we have the RDF description block that contains the code for the triples that represent the document, the appellate case we were looking at earlier.
The Subject for all of the triples expressed here, is what I previously referred to as ‘Case A’ in our example. In RDF the Subject is expressed using the RDF attribute ‘about’. So this is telling us that the RDF description is about ‘Case A’ only here I’m using a URI.

Below that line we’ve got the 4 Objects that we identified earlier and their Predicates.
The docket number.
The plaintiff.
The defendant.
And the name of the court.
I’ve added a couple of additional triples to illustrate the use of the Dublin Core namespace and how metadata schemes can interoperate with each other in RDF.

The Predicate is found in the information appearing before the equals sign here. So the Subject, defined by the URI for ‘Case A’ here, has the title ‘Royal Bank of Canada v. King.’
So when we serialize this code we’ll get the same 4 triples we had from the beginning …
http://canlii.ca/t/g0656 has docket number C57108
http://canlii.ca/t/g0656 has plaintiff Royal Bank of Canada
http://canlii.ca/t/g0656 has defendant Leslie King
http://canlii.ca/t/g0656 was heard in Court of Appeal for Ontario

… but ‘Case A’ is replaced by a URI here. This is the HTTP URI for the case as it’s found in the CanLII database …
... and also have the additional 2 triples used in the example referencing the Dublin Core. Ideally each of the parts of the triple will also be described using a URI.

OK, so that’s a general overview of linked data theory.
An important figure in all of this is of course Tim Berners-Lee: “inventor of the World Wide Web” and champion of the semantic web.

On a website that considers linked data design issues, Tim Berners-Lee provides the four basic principles of linked data:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things

So, a URI is a Uniform Resource Identifier and an example is the HTTP URI, which we all know as the familiar URL, like the one we just used in the RDF example.
SPARQL is a query language used to search across RDF data sets known as Triple Stores.
Many of you have likely seen this image of the linked open data cloud. It’s based on metadata collected by the Linked Open Data Cloud group on CKAN. CKAN is the Comprehensive Knowledge Archive Network which is described in Wikipedia as a “web-based system for the storage and distribution of data.”

This cloud represents the data sets that are openly available as linked data along with their links out to other linked open data sets on the Web. Although billed as the 'latest' this diagram was originally generated in September 2011 so it's actually a couple of years old now representing a “cloud” of 295 data sets.

The current number of linked open data sets recorded at CKAN via DataHub is currently listed as 337. This is a small fraction of the total number of data sets available in Datahub, which currently number 10,000. But these are proprietary and have restrictions on use. [http://datahub.io/group/lodcloud]
If you recall my earlier crystalline structure you might consider it as a very tiny piece of the linked open data cloud. And although there were only 295 linked open data sets in 2011 that still represented somewhere in the neighbourhood of 32 billion triples.
Which is remarkable considering the modest beginnings of the linked open data cloud which began with only 12 data sets in 2007.
As shown in this chart, the linked open data cloud has grown significantly over the last 4 or 5 years.
You can see that **DBPedia** figures prominently here, shown in the center of this cloud. DBPedia was created by Chris Bizer, one of the leading linked data developers working in the field today. He had the idea to extract the structured information already embedded in Wikipedia.

Since its creation many data set providers have linked their data sets to DBPedia which has made it the center of the linked open data hub.
This diagram, which looks very much like the early 2007 version of the Linked Open Data Cloud, is from a recent article called “Linked data in the legal domain.” It’s based on some recent work being done at the Institute of Legal Information Theory and Techniques of the Italian National Research Council.

This represents what the authors call the Legal Data Cloud and shows some relevant Italian and European Union legal data sets. I'd like to return to this at the end of the presentation and ask you what a similar diagram for Canadian legal resources might look like.

I'll turn now to a look at the nature and characteristics of legal resources.
I'm sure many of you have heard this phrase before:

“Law is chaos with an index” often attributed to Oliver Wendell Holmes. But Holmes apparently paraphrased this from a collection of essays by British lawyer, Sir Thomas Holland, which he had reviewed for the American Law Review.

Regardless of its origin this is an apt description of the legal domain.

In a short paper about linked legal data efforts in Finland the authors provide a very nice summary of this potential chaos in the opening paragraph of their report.
“Publishing and using juridical information is challenging in many ways. It is produced by different parties, such as governmental bureaus, ministries, different levels of courts, research organizations, and media. The content is heterogeneous and produced using differing tools, data formats, and practices. The links between documents are often informal and/or not made explicit. The law in general is a dynamic, changing entity: for example, it is important to be able to refer to different versions of a law at different points of time. These challenges can be addressed through the use of linked data techniques.”
Speaking about the need for semantic web technology in the legal domain Joost Breuker and others wrote in 2009,

“The traditional fields and practices of law are changing fast. Legal drafting, private contracting, judicial sentencing and administrative management have been enlarged with online dispute resolution initiatives and new forms of self-regulation and access to justice. Citizens, customers and consumers require a greater participation and faster and more effective ways of facing their legal activities.”

So along with increased user expectations for access to legal information over the past 5 years or so there has also been a change in how law is being practiced. In the same article, for example, the authors note that 2008 marked the point when the number of lawyers using free online services had, for the first time, overtaken those using for-fee services (89% vs. 83%).
As the authors suggest here, the heterogeneous reality of the legal domain and the increase in user expectations is something that might be successfully addressed in a linked data environment.

But it really depends on whether we can build a critical mass of data sources in a standard format like RDF/XML. Using RDF/XML helps to normalize the data making it more homogeneous and therefore easier for machines and humans alike to discover and make better connections between the various data sources that are available.

Once the data is up we can start to place our efforts on the development and consistent use of the metadata that describes and gets people to the documents relevant to assist with their legal issues.
One promising metadata project, still in the early stages of development, is URN:Lex. This proposal, submitted to the Internet Engineering Task Force in 2010, is a product of a number of groups and led by the Institute of Legal Information Theory and Techniques of the Italian National Research Council and also involving Cornell’s Legal Information Institute.

“The purpose of the "lex" namespace is to assign an unequivocal identifier, in standard format, to documents that are sources of law. The identifier is conceived so that its construction depends only on the characteristics of the document itself and is, therefore, independent from the document's on-line availability, its physical location, and access mode.”

So this is one area where efforts might be focused.
There are a number of other legal metadata initiatives reinforcing the notion that “standards are great because there are so many to choose from.” I’ve listed a few for you here that you can take a look at later.

And I draw your attention to Robert Richards list of resources dealing with “metadata specifically designed for legal information.”
And another area that disserves our attention is work on the development and use of ontologies for legal resources. It is the ontologies that define the metadata structures and how they are to be used.

Andre Valente in his chapter on Types and Roles of Legal Ontologies identifies five main uses or roles for ontologies:

(a) To organize and structure information;
(b) To provide reasoning and problem solving;
(c) To enable semantic indexing and search;
(d) To provide semantics integration and interoperation; and
(e) To understand the domain.

Types and Roles of Legal Ontologies

- Organize and structure information
- Reasoning and problem solving
- Semantic indexing and search
- Semantics integration and interoperation
- Understanding the domain

Types and Roles of Legal Ontologies / Andre Valente <http://link.springer.com/chapter/10.1007/978-3-340-32253-5_5>
There are also quite a few legal ontologies available. In the opening chapter of the book Law, Ontologies and the Semantic Web 23 separate legal ontologies are identified. I’ve highlighted a couple here and Valente also describes a dozen or so other legal ontologies in the book chapter cited earlier.
So, is there a role for CanLII in all of this?

I think there is an essential role for CanLII to play here. As a major aggregator of Canadian case law CanLII is already an important source for legal data. And contributing this already structured legal data to the linked data initiative is not only good for the linked open data community it supports CanLII’s mission and their goal to provide “Free and unrestricted access to legal information.”

And I applaud the recent release of the CanLII web API because it does expose this structured data that has already been collected. The API also enables new web-based applications to be built against the CanLII database which we should see examples of later this afternoon.
However as Chris Bizer pointed out in a presentation from way back in 2007, there are some disadvantages to relying solely on a web API for access to this data:

- The data remains enclosed in a proprietary database which isolates the data and perpetuates the data silo mentality
- Software developers must know you exist and where you are and then decide how they will use the available API parameters
- Links between different data sources must be explicitly created, i.e. hard coded rather than discovered as needed when a query is made
- Therefore the data is not part of the global database, i.e. the data is “on the web” but it’s not “of the web”.

Some disadvantages with a web API:

- Data is enclosed in a proprietary database perpetuating data silos
- Developers must know you exist and decide what to access based on available API parameters
- Links between different data sources must be explicitly created
- Data is not included in the global database i.e. “on the web” but not “of the web”
So the first step in my mind is to make the data available as linked open data.

AS Tim Berners-Lee said in his TED Talk in 2009:

“The important thing about data is, the more things you have to connect together, the more powerful it is.” — Tim Berners-Lee
If we want to support the linked open data initiative and be part of this emerging global database we should be working towards adding our structured legal data to the open cloud of linked data. Every contribution is useful and will lead to a critical mass of legal data.

Tim Berners-Lee had the audience chanting ‘Raw Data Now’, encouraging those people who are keepers of the data to put it up as linked data now …
There is a lot of work being done in Europe and the U.S. but activity in this country is rather quiet. But it seems like a logical next step for CanLII. CanLII is well positioned to take their place at the centre of the open legal data cloud in Canada. And I humbly encourage them to take this opportunity to do so.

I’m sure there are others in a similar position who might be interested and able to share their own data in this way.

I’d be interested to hear about any legally oriented or government organizations in Canada that might be considering participation in the development of a legal data cloud in Canada …
I’ve included a list of the references I consulted for this presentation here for your information.
Thank you for your attention!

tknight@osgoode.yorku.ca
Twitter: @freemoth