Smart Thesauri: Using Taxonomies with Linked Data

By Margie Hlava & Bob Kasenchak,
Access Innovations, Inc.

As interest in Linked Data (LD) continues to grow, many organizations—publishers, corporations, universities, libraries—are increasingly interested in strategies to jump-start LD initiatives. Any organization that has an existing taxonomy (or other controlled vocabulary) can expedite the move to LD by leveraging its existing semantic structures as a bridge to an advanced LD-based semantic strategy.

Why Linked Data?

There are three primary reasons motivating organizations to move towards LD:

1. To use resources on the Web to enhance internal knowledge environments.
Once a link to an external data source (e.g., DBpedia or Wikidata) is established, references to other content—Wikipedia articles, definitions, images, social media and news feeds, and other information—can be queried off and added to internal resources to enrich content or Web portals.

2. To add backlinks to internal resources and content to LD portals, thereby pointing to an organization as an authority on a topic or topics.
Adding reciprocal links from LD sources to content (or publically available web resources) enables other LD users to find and reference an organization's expertise on a given subject.

3. To add an organization's expertise to the growing Semantic Web of knowledge and information—i.e., to contribute to the sum of available human knowledge.

More idealistically, many LD enthusiasts are motivated by the goal of adding their curated and well-formed specialized knowledge to the expanding network of LD sources—in other words, attaching their datasets and knowledge organization systems (KOS) to the LD community.

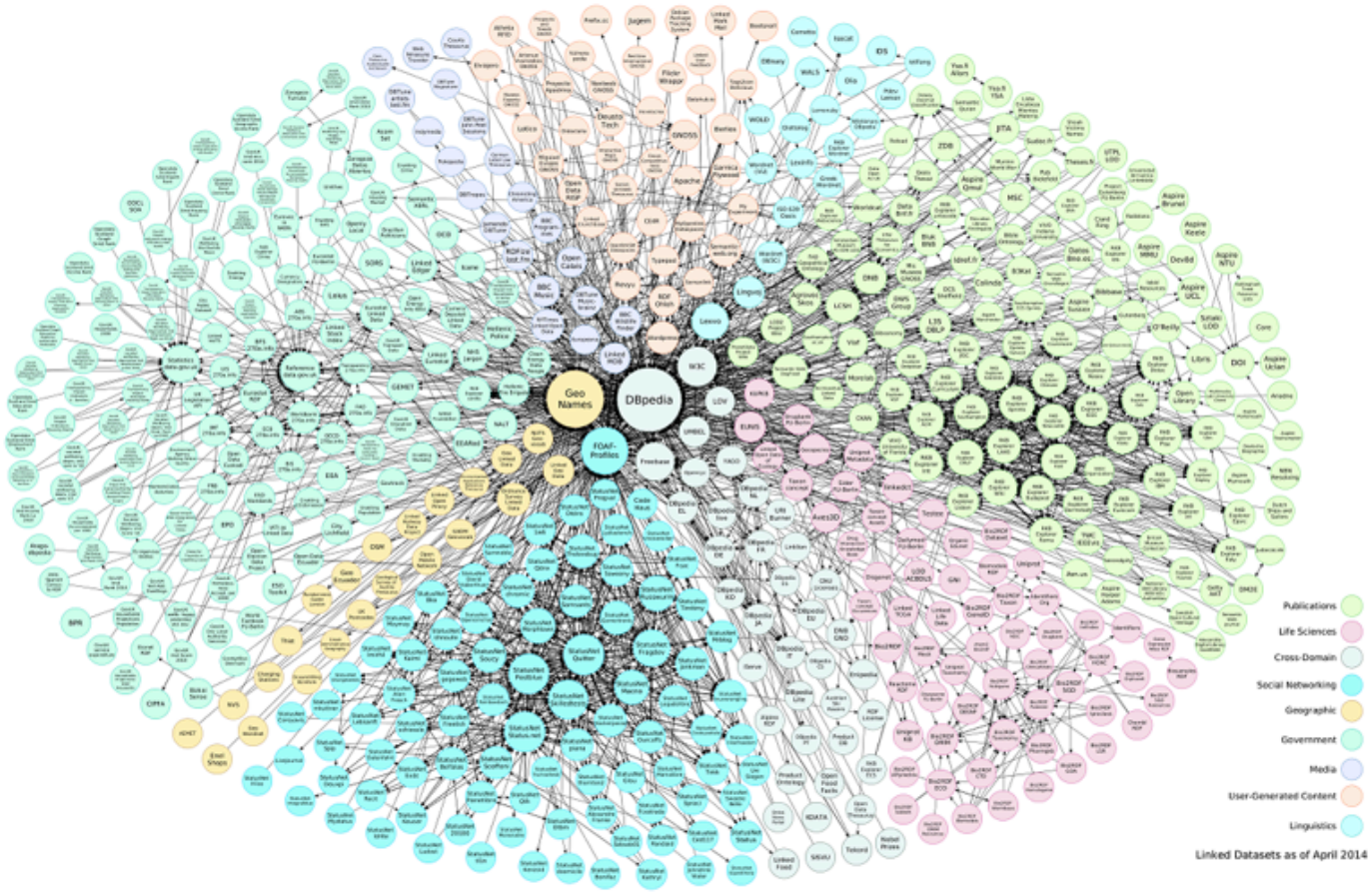

Figure 1: The Ubiquitous LD Graph Diagram. Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/. Reproduced here under the terms of the Creative Commons Attribution-ShareAlike 3.0 Unported license (CC BY-SA 3.0).

Note that in this diagram (as in most such graphics) DBpedia still occupies the central position, as it is the most robust and oft-linked universal (i.e., not subject-specific) LD source. Although it has its shortcomings, DBpedia is still the best choice for stable LD URIs and this is not likely to change anytime soon; accordingly, we will use DBpedia as the sample LD source in this article.

From Taxonomy to Linked Data

Although DBpedia Spotlight is good at recognizing existing LD concepts in a block of text, in most cases it doesn't have the robust synonymy and/or rule-based concept extraction used in most semantic platforms. Additionally, an organization with a well-formed taxonomy and rich semantic strategy will already have indexed content—possibly a very large volume of content—so "re-indexing" a large legacy dataset using DBpedia Spotlight is an unwieldy proposition.

Instead, by asserting a link between terms in an existing taxonomy and the corresponding concept in an LD source, it's not necessary to run legacy content through Spotlight (or a similar LD-matching service) since the link is asserted at the thesaurus level instead of the document level. This is both much more efficient and provides a simpler mechanism to curate and establish LD links in the future.

Sample Process

Imagine that you have a large set of content about physics—you could be a publisher, laboratory, or research organization—and that you have an existing, well-formed taxonomy of physics concepts and wish to pursue LD. Basically, you want to assert that each term (more on this later) in your taxonomy corresponds to a DBpedia URI on the same topic.

For example, you might have a term (or branch) in your vocabulary called "Optics". The first step is to add a field to the term record in your taxonomy management system (most commercial taxonomy applications have a mechanism for this) to hold a URI; this could be a dedicated 'live' URI field or just a simple text field. Next, in this field for your term "Optics", you add the corresponding DBpedia URI:

http://dbpedia.org/page/Optics

…and you're done.**

You have now achieved something like this:

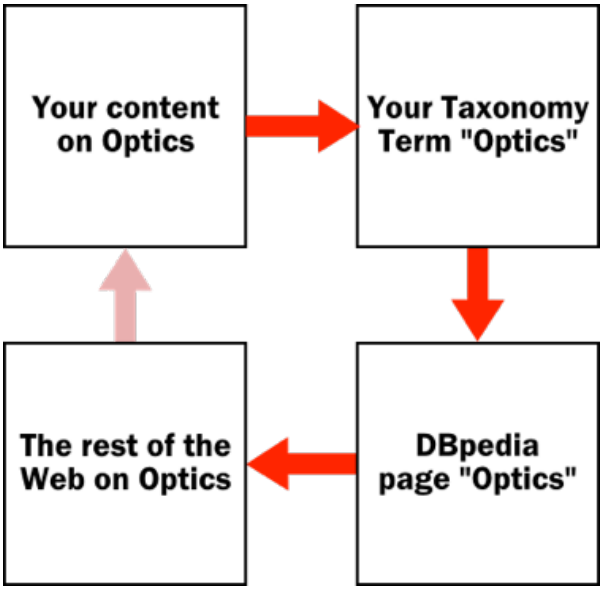

Figure 2: Connecting Resources on the Same Topic Using LD

Using a SPARQL endpoint, you can now query information from the DBpedia page or use it to reference or link to other pages from DBpedia. For example, if you have a topical web portal on Optics, you could automatically add definitions, images, news, social media feeds, or links to other publishers with Optics content.

Automation, Problems, and Quality Control

Can this process—matching terms from your taxonomy to DBpedia—be automated? Yes, but it requires careful quality control. We have found the most success using Spotlight as a starting point and validating the results by hand; this is faster than matching each term manually, and at the same time ensures accuracy.

As mentioned earlier, Spotlight is better at matching blocks of text (leveraging semantic proximity) than single concepts from a taxonomy, but it's accurate enough to decrease some of the effort of manual matching.

The primary sticking point, however, is that any specialized thesaurus will be more granular than DBpedia is. Your taxonomy on Physics, in the thought experiment above, is going to have far more specific terms than DBpedia in many, many places. The top two or three levels of your taxonomy will probably have corresponding LD pages, but the more specific topics will not.

** *A subsequent step might involve adding a backlink to your topical/library page on Optics to the DBpedia dbpedia-owl:wikiPageExternalLink field, if you want your content to be publicly available.*

For example, there's a robust DBpedia page on Optics, as well as one on Nonlinear optics; more specific topics within Nonlinear optics, however, are far less likely to have a corresponding page (e.g., "Photonic metamaterials" has no corresponding LD page we've found so far—and many scientific and technical vocabularies get even more granular than this).

Possible solutions to this problem are as follows:

- Forget it for the time being and check back later to see whether a corresponding page emerges.

This is the easiest option, but does not accomplish much.

- "Roll up" more granular topics to the next-nearest Broader Term in the taxonomy.

This procedure at least provides LD pages for every topic in the taxonomy, but leaves much to be desired; every term in a large branch might point to the same LD page, which is not particularly useful.

- Proactively add new DBpedia pages on not-yet-existing topics, and add the backlink to your content/vocabulary as the first link; the Web should come and fill in the blanks eventually.

Attractive, altruistic, and useful, though far more time consuming, this option is ideal from an information science perspective but may not be practical in the scope of your LD initiative.

REFERENCES

For more information, we recommend David Wood, Marsha Zaidman, Luke Ruth, and Michael Hausenblas *Linked Data: Structured data on the Web (Shelter Island, NY: Manning Publications 2013)*