Practical *P-P-P-P*roblems with Linked Data

by Mike Bergman - Monday, October 04, 2010

http://www.mkbergman.com/917/practical-p-p-p-problems-with-linked-data/

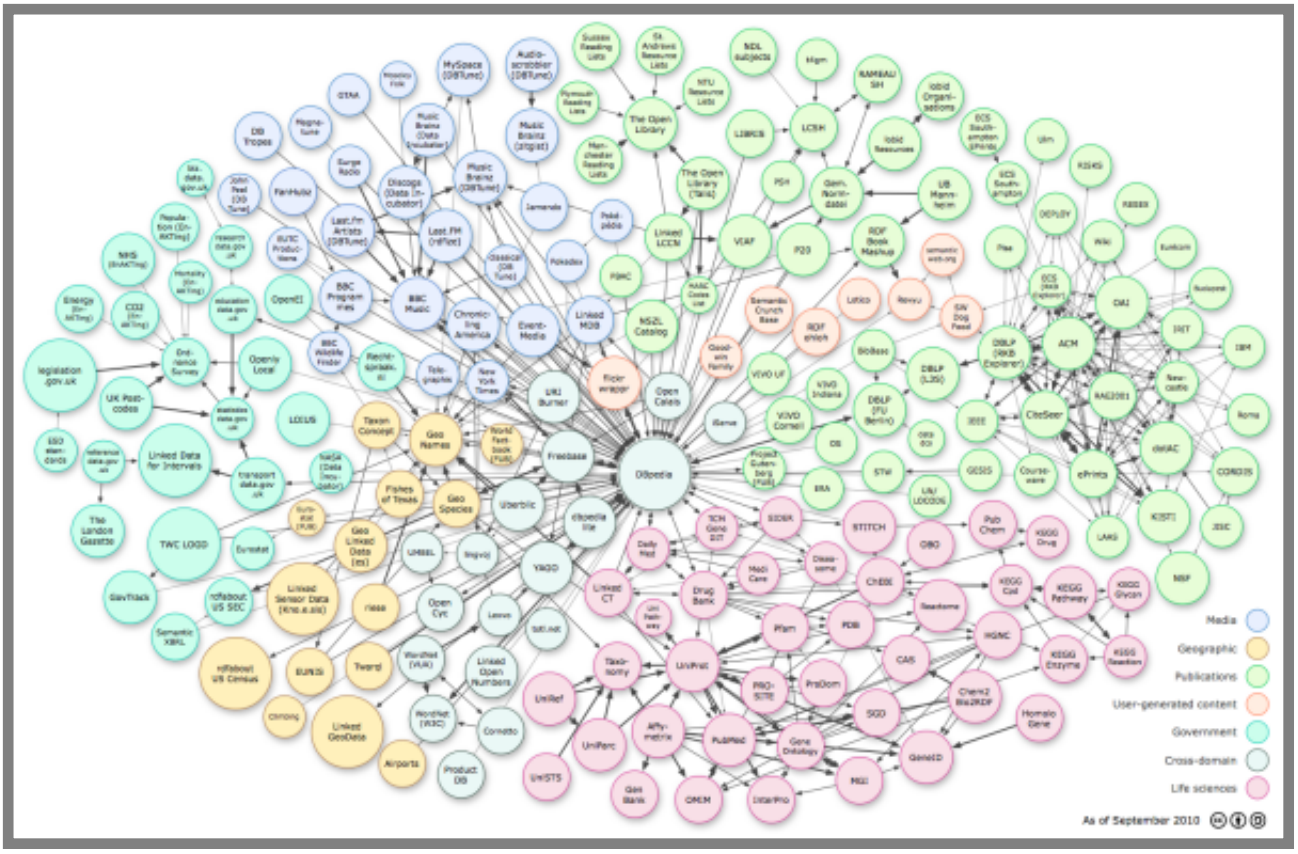
Dealing with the Four Ps to Broaden Actual Use

We have to again thank Richard Cyganiak and Anja Jentzsch -- as well as all of the authors and publishers of linked open datasets -- for the recent update to the linked data cloud diagram [1]. Not only have we seen admirable growth since the last update of the diagram one year ago, but the datasets themselves are now being registered and updated with standard metadata on the CKAN service. Our own UMBEL dataset of reference subject concepts is one of those listed.

Growth and the Linked Data Cloud

The linked open data (LOD) "cloud" diagram and its supporting statistics and archived versions are also being maintained on the http:lod-cloud.net site [1]. This resource, plus the CKAN site and the linked data site maintained by Tom Heath, provide really excellent starting points for those interested in learning more about linked open data. (Structured Dynamics also provides its own FAQ sheet with specific reference to linked data in the enterprise, including both open and proprietary data.)

As an approach deserving its own name, the practice of linked data is about three years old. The datasets now registered as contributing to this cloud are shown by this diagram, last updated about a week ago [1]:



(click to expand)

LOD was initially catalyzed by DBpedia and the formation of the Linked Open Data project by the W3C. In the LOD's first listing in February 2007, four datasets were included with about 40 million total triples. The first LOD cloud diagram was published three years ago (upper left figure below), with 25 datasets consisting of over two billion RDF triples and two million RDF links. By the time of last week's update, those figures had grown to 203 data sets (qualified from the 215 submitted) consisting of over 25 billion RDF triples and 395 million RDF links [2].

This growth in the LOD cloud over the past three years is shown by these archived diagrams from the LOD cloud site [1]:

2007-10-08 **2007-11-07** **2007-11-10** **2008-02-28** **2008-03-31**

2008-09-18 **2009-03-05** **2009-03-27** **2009-07-14** **2010-09-22**

(click on any to expand)

With growth has come more systematization and standard metadata. The CKAN (*comprehensive knowledge archive network*) is especially noteworthy by providing a central registry and descriptive metadata for the contributing datasets, under the lodcloud group name.

Still, Some Hard Questions

This growth and increase in visibility is also being backed by a growing advocacy community, which were initially academics but has broadened to also include open government advocates and some publishers like the NY Times and the BBC. But, with the exception of some notable sites, which I think also help us understand key success factors, there is a gnawing sense that linked data is not yet living up to its promise and advocacy. Let's look at this from two perspectives: growth and usage.

Growth

While I find the visible growth in the LOD cloud heartening, I do have some questions:

- Is the LOD cloud growing as quickly as its claimed potential would suggest? I suspect not. Though there has been about a tenfold growth in datasets and triples in three years, this is really from a small base. Upside potential remains absolutely huge
- Is linked data growing faster or slower than other forms of structured data? Notable comparatives here would include structure in internal Google results; XML; JSON; Facebook's Open Graph Protocol, others
- What is the growth in the use of linked data? Growth in publishing is one thing, but use is the ultimate measure. I suspect that, aside from specific curated communities, uptake has been quite slow (see next sub-section).

Perhaps one of these days I will spend some time researching these questions myself. If others have benchmarks or statistics, I'd love to see them.

Such data would be helpful to put linked data and its uptake in context. My general sense is that while linked data is gaining visible traction, it is still not anywhere close to living up to its promise.

Usage

I am much more troubled by the lack of actual use of linked data. To my knowledge, despite the publication of endpoints and the availability of central access points like Openlink Software's lod.openlinksw.com, there is no notable service with any traction that is using broad connections across the LOD cloud.

Rather, for anything beyond a single dataset (as is DBpedia), the services that do have usefulness and traction are those that are limited and curated, often with a community focus. Examples of these notable services include:

- The life sciences and biomedical community, which has a history of curation and consensual semantics and vocabularies
- FactForge from Ontotext, which is manually cleaned and uses hand-picked datasets and relationships, all under central control
- Freebase, which is a go-to source for much instance data, but is notorious for its lack of organization or structure
- Limited, focused services such as Paul Houle's Ookaboo (and, of course, many others), where there is much curation but still many issues with data quality (see below).

These observations lead to some questions:

- Other than a few publishers promoting their own data, are there any enterprises or businesses consuming linked data from multiple datasets?
- Why are there comparatively few numbers of links between datasets in the current LOD cloud?
- What factors are hindering the growth and use of linked data?

We're certainly not the first to note these questions about linked data. Some point to a need for more tools. Recently others have looked to more widespread use of RDFa (RDF embedded in Web pages) as possible enablers. While these may be helpful, I personally do not see either of these factors as the root cause of the problems.

The Four Ps

Readers of this blog well know that I have been beating the tom-toms for some time regarding what I see as key gaps in linked data practice [3]. The update of the LOD cloud diagram and my upcoming keynote at the Dublin Core (DCMI) DC-2010 conference in Pittsburgh have caused me to try to better organize my thoughts.

I see four challenges facing the linked data practice. These four problems -- the *four Ps* -- are *predicates*, *proximity*, *provision* and *provenance*. Let me explain each of these in turn.

Problem #1: Predicates

For some time, the quality and use of linking predicates with linked data has been simplistic and naïve. This problem is a classic expression of Maslow's hammer," if all you have is a hammer, everything looks like a nail." The most abused linking property (predicate) in this regard is `owl:sameAs`.

In order to make links or connections with other data, it is essential to understand what the nature is of the subject “thing” at hand. There is much confusion about actual “things” and the references to “things” and what is the nature of a “thing” within linked data [4]. Quite frequently, the use or reference or characterization of "things" between different datasets should not be asserted as exact, but as only approximate to some degree.

So, we might be referring to something that is ***about***, or ***similar to***, or ***approximate with*** or some other qualified linkage. Yet the actual semantics of the `owl:sameAs` predicate is quite exact and one with some of the strongest *entailments* (what do the semantics mean) defined. For `sameAs` to be applied correctly, every assertion about the linked object in one dataset must be believed to be true for every assertion about that linked object in the matching dataset; in other words, the two instances are being asserted as **identical** resources.

One of the most vocal advocates of linked data is Kingsley Idehen, and he perpetuates the misuse of this predicate in a recent mailing list thread. The question had been raised about a geographical location in one dataset that mistakenly put the target object into the middle of a lake. To address this problem, Kingsley recommended:

```
You have two data spaces: [AAA] and [BBB], you should make a third --
yours, which I think you have via [CCC].

Place the fixed (cleansed) data in your [CCC] data space, connect the
coreferenced entities using an "owl:sameAs" relation, scope queries
that are accuracy sensitive to your [CCC] data space. Use inference
rules for union expansion across [AAA] and [BBB] via "owl:sameAs",
when data quality requirements are low and data expanse requirements
high.

That's how you clean up the mess and potentially get compensated for
doing so, in the process.
```

The point here is not to pick on Kingsley, nor even to solely single out `owl:sameAs` as a source of this problem of linking predicates. After all, it is reasonable to want to relate two objects to one another that are mostly (and putatively) about the same thing. So we grab the best known predicate at hand.

The real and broader issue of linked data at present is firstly, actual linking predicates are often not used. And, then, secondly, when they are used, their semantics are too often wrong or misleading.

We do not, for example, have sufficient and authoritative linking predicates to deal with these "sort of" conditions. It is a key semantic gap in the linked data vocabulary at present. Just as SKOS was developed as a generalized vocabulary for modeling taxonomies and simple knowledge structures, a similar vocabulary is needed for predicates that reflect real-world usage for linking data objects and datasets with one another [5].

The idea, of course, with linked data resides in the term linked. And linkage means how we represent the relation between objects in different datasets. Done right, this is the beauty and power of linked data and offers us the prospect of federating information across disparate sources on the Web.

For this vision, then, to actually work, links need to be asserted and they need to be asserted correctly. If they are not, then all we are doing is shoveling triples over the fence.

Problem #2: Proximity (or, "is About")

Going back to our first efforts with UMBEL, a vocabulary of about 20,000 subject concepts based on the Cyc knowledge base [6], we have argued the importance of using well-defined reference concepts as a way to provide "aboutness" and reference hooks for related information on the Web. These reference points become like stars in constellations, helping to guide our navigation across the sea of human knowledge.

While we have put forward UMBEL as one means to provide these fixed references, the real point has been to have accepted references of any manner. These may use UMBEL, alternatives to UMBEL, or multiples thereof. Without some fixity, preferable of a coherent nature, it is difficult to know if we are sailing east or west. And, frankly, there can and should be multiple such reference structures, including specific ones for specific domains. Mappings can allow multiple such structures to be used in an overlapping manner depending on preference.

When one now looks at the LOD cloud and its constituent datasets, it should be clear that there are many more potential cross-dataset linkages resident in the data than the diagram shows. Reference concepts with appropriate linking predicates are the means by which the relationships and richness of these potential connections can be drawn out of the constituent data.

The use of reference vocabularies is rejected by many in the linked data community for what we believe to be misplaced ideological or philosophical grounds. Saying that something is "about" Topic A (or even Topics B and C in different reference vocabularies) does not limit freedom nor make some sort of "ontological commitment". There is also no reason why free-form tagging systems (folksonomies) can also not be mapped over time to one or many reference structures to help promote interoperability. Like any language, our data languages can benefit from one or more dictionaries of nouns upon which we can agree.

Linked data practitioners need to decide whether their end goal is actual data interoperability and use, or simply publishing triples to run up the score.

Problem #3: Provision of Useful Information

We somewhat controversially questioned the basis of how some linked data was being published in an article late last year, *When Linked Data Rules Fail* [4]. Amongst other issues raised in the article, one involved publishing large numbers of government datasets without any schema, definitions or even data labels for numerically IDed attributes. We stated in part:

. . . we have ABSOLUTELY NO INFORMATION ABOUT WHAT THE DATA CONTAINS OTHER THAN

A PROPERTY LABEL. There is much, much rich value here in data.gov, but all of it remains locked up and hidden.

The sad truth about this data release is that it provides absolutely no value in its current form. We lack the keys to unlock the value.

To be sure, early essential spade work has been done here to begin putting in place the conversion infrastructure for moving text files, spreadsheets and the like to an RDF form. This is yeoman work important to ultimate access. But, until a vocabulary is published that defines the attributes and their codes so we can unlock this value, it will remain hidden. And only when its further value (by connecting attributes and relations across datasets) through a schema of some nature is also published, the real value from connecting the dots will also remain hidden.

These datasets may meet the partial conditions of providing clickable URLs, but the crucial aspect of "providing useful information" as to what any of this data means is absent.

Some of these problems have now been fixed in the subject datasets, but in this circumstance and others we still see way too many instances within the linked data community of no definitions of terms, no human readable labels and the lack of other information by which a user of the data may gauge its meaning, interpretation or semantics. Shame on these publishers.

Really, in the end, the provision of useful information comes down to the need to answer a simple question: Link what?

The what is an essential component to staging linked data for actual use and interoperability. Without it, there is no link in linked data.

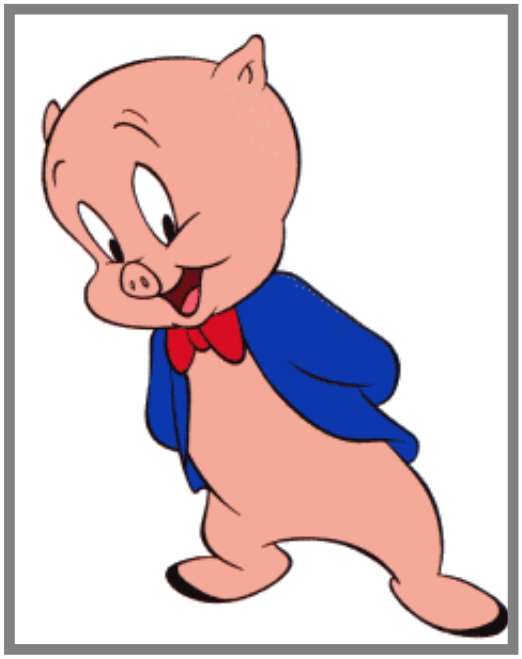
Problem #4: Provenance

There are two common threads in the earlier problems. One, semantics matter, because after all that is the arena in which linked data operates. And, second, some entities need to exert the quality control, completeness and consistency that actually enables this information to be dependable.

Both of these threads intersect in the idea of provenance.

Practice to date suggests that uncurated, linked datasets in the wild are unlikely to be useful nor used in combination with other datasets. Some agent or community will need to take charge -- that is, establish provenance -- to help ensure the consistency and quality upon which interoperability depends.

This assertion should not be surprising -- the standard Web needed some consistent attention with respect to directories and search engines. That linked data or the Web of data is no different, perhaps even more demanding, should be expected.



"That's Linked P-P-P-Problems, Folks!"

When we look to those efforts that are presently getting traction in the linked data arena (with some examples above), we note that all of them have quality control and provenance at their core. I think we can also say that only individual datasets that themselves adhere to quality and consistency will even be considered for inclusion in these curated efforts.

Where Will the Semantics Leadership Emerge?

The current circumstance of the semantic Web is that adequate languages and standards are now in place. We also see with linked data that techniques are now being worked out and understood for exposing usable data.

But what appears to be lacking are the semantics and reference metadata under which real use and interoperability take place. The W3C and its various projects have done an admirable job of putting the languages and standards in place and raising the awareness of the potential of linked data. We can now fortunately ask the question: What organizations have the authority to establish the actual vocabularies and semantics by which these standards can be used effectively?

When we look at the emerging and growing LOD cloud we see *potential* written with a capital ***P***. If the problem areas discussed in this article -- the contrasting *four Ps* -- are not addressed, there is a real risk that the hard-earned momentum of linked data to date will dissipate. We need to see real consumption and real use of linked data for real problems in order for the momentum to be sustained.

Of the *four Ps*, I believe three of them require some authoritative leadership. The community of linked data needs to:

- Find responsive *predicates*
- Publish reference concepts as *proximate* aids to orient and align data , and
- Do so with the *provenance* of an authoritative voice.

When we boil down all of the commentary above a single question remains: Where will the semantic leadership emerge?

---

[1] Linking Open Data cloud diagrams, by Richard Cyganiak and Anja Jentzsch, last updated in Sept. 2010. See http://lod-cloud.net/. Most of the diagrams are available in PNG, PDF and SVG formats, in colored (keyed) and uncolored versions. The site also contains many other useful links.

[2] The original W3C LOD project page, the SWEO Community Project, has continued to be maintained and updated even though the official project has now ended. This site is a useful source of archived data and news releases.

[3] Notable articles include [4] and M.K. Bergman, 2008. "A New Constellation in the Linking Open Data (LOD) Sky," *AI3:::Adaptive Information* blog, Oct. 5, 2008; see http://www.mkbergman.com/457/a-new-constellation-in-the-linking-open-data-lod-sky/; and M.K. Bergman, 2009. "Moving Beyond Linked Data," *AI3:::Adaptive Information* blog, Sept. 9, 2009; see http://www.mkbergman.com/802/moving-beyond-linked-data/.

[4] M.K Bergman and Fred Giasson, 2009. "When Linked Data Rules Fail," *AI3:::Adaptive Information* blog, Nov. 16, 2009. See http://www.mkbergman.com/846/when-linked-data-rules-fail/.

[5] A vocabulary of linking predicates would capture the variety and degrees to which individuals, instances, classes and concepts are similar or related to objects in other datasets. This purpose is different than, say, voiD (*Vocabulary of Interlinked Datasets*), which has as its purpose providing descriptive metadata about the nature of particular datasets.

[6] UMBEL (*Upper Mapping and Binding Exchange Layer*) is an ontology of about 20,000 subject concepts that acts as a reference structure for inter-relating disparate datasets. The reference concepts and their relationships are a direct sub-set extraction from the OpenCyc version of the Cyc knowledge base. UMBEL also has a second purpose of being a general vocabulary of classes and predicates designed for the creation of domain-specific ontologies.
