The Semantics of Context

by Mike Bergman - Tuesday, May 06, 2008

http://www.mkbergman.com/440/the-semantics-of-context/
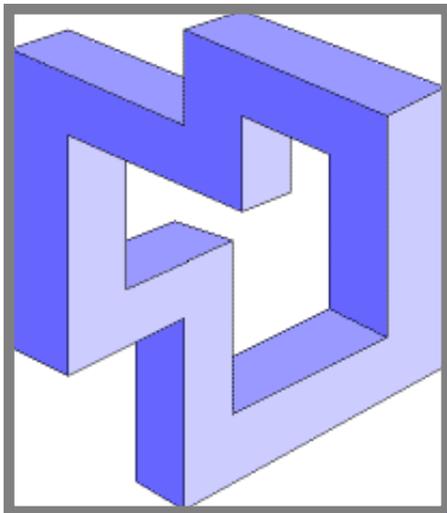
Many Kinds of RDF Links Can Provide Linked Data 'Glue'

In a recent blog post, Kingsley Idehen picked up on the UMBEL project's mantra of "context, Context, CONTEXT!" as contained in our recent slideshow. He likened *context* to the real estate phrase of "location, location, location". Just so. I like Kingsley's association because it reinforces the idea that *context* places concepts and things into some form of referential road map with respect to other things and concepts.

To me, *context* describes the relationships and environmental proximities of what UMBEL calls subject concepts and their instance sub-concepts and named entity members, the whole of which might be visualized as a graph of reference nodes in the firmament of a global knowledge space.

Indeed, it is this very 'cloud' of subject concept nodes that we tried to convey in an earlier piece on what UMBEL's backbone structure of 21,000 subject concepts might look like, shown at right. (Of course, this visualization results from the combination of UMBEL's OpenCyc contextual framework and specific modeling algorithms; the graph would vary considerably if based on other frameworks or models.)
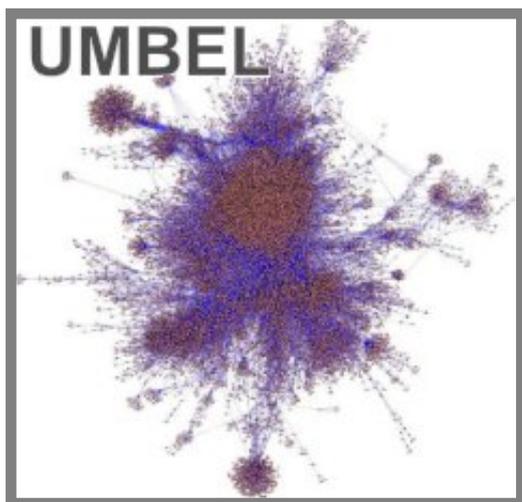
Yet in a comment to Kingsley's post, Giovanni Tummarello said, "If you believe in context so much then the linking open data idea goes bananas. Why? Because 'sameAs' is fundamentally wrong.. an entity on DBpedia IS NOT sameAs one on GeoNames because the context is different and bla bla... so it all crumbles." [1]

Well, hmmm. I must beg to differ.

I suspect as we now are seeing Linked Data actually enter into practice, new implications and understandings are coming to the fore. And, as we try new approaches, we also sometimes suffer from the sheer difficulty of explicating those new understandings in the context of the shaky semantics of the semantic Web.

Giovanni's comment raises two issues:

1. What the context or meaning of *context* is, and
2. The dominant RDF link 'glue' for Linked Data that has been used to date, the `owl:sameAs` predicate.

Therefore, since UMBEL is putting forth the argument for the importance of context in Linked Data, it is appropriate to be precise about the semantics of what is meant.

Context in Context

What is context? The tenth edition of Merriam-Websters Collegiate Dictionary (and the online version) defines it as:

context \?kän-?tekst\ *n*.;ME, weaving together of words, Latin *contextus* connection of words, coherence, from *contexere* to weave together, from *com- + texere* to weave (*ca*.1586)

1: the parts of a discourse that surround a word or passage and can throw light on its meanings

2: the interrelated conditions in which something exists or occurs: environment, setting <the historical

context of the war>.

Another online source I like for visualization purposes is <u>Visuwords</u>, which displays the accompanying graph relationships view based on <u>WordNet</u>.
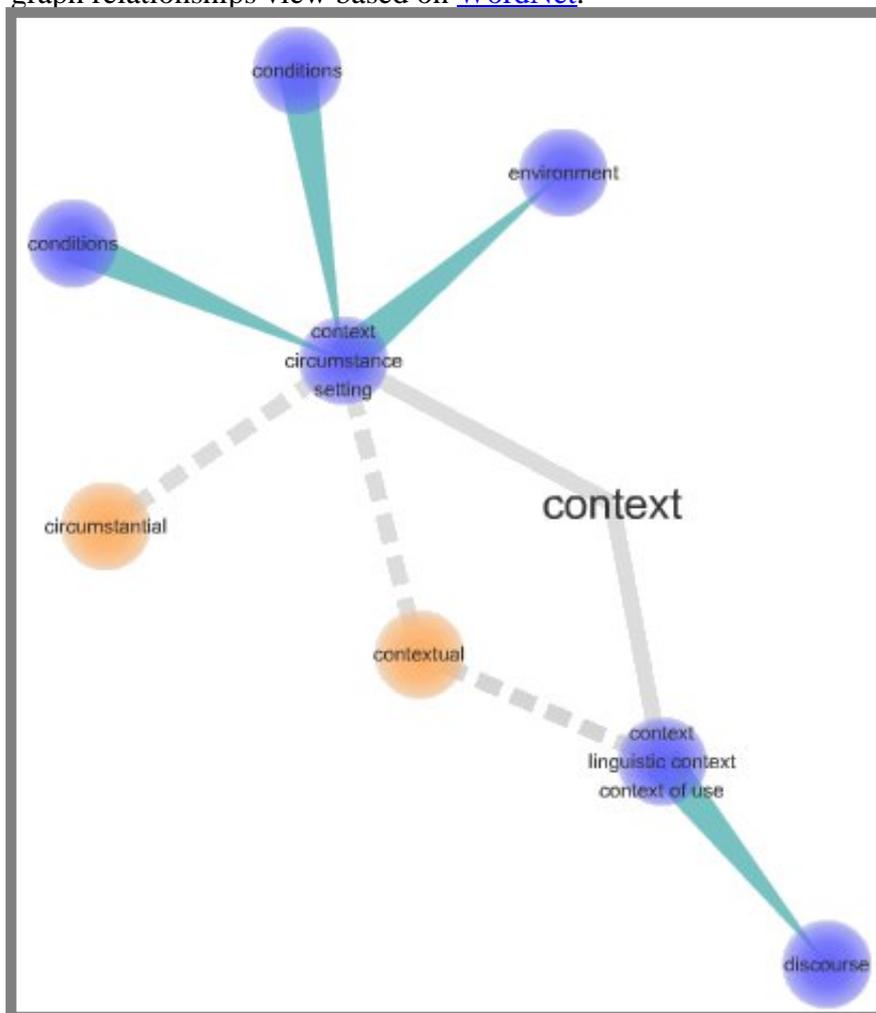
Both of these references, of course, base their perspective on language and language relationships. But, both also provide the useful perspective that context also conveys the senses of *environment*, *surroundings*, *interrelationships*, *connections* and *coherence*.

<u>*Context*</u> has itself been a focus of much research from linguistics to philosophy and computer science. Each field has its specific take on the concept, but I believe it fair to say that *context* is consensually used as a holistic reference structure that tries to put all worlds and views, including that of the observer and observed, into a consistent framework. Indeed, when that framework and its assertions fit and make sense, we give that a word, too: **coherent**.

Hu Yijun [2], for example, intersects the interplay of language, semantics and the behavior and circumstances of human actors to frame context. Yijun observes that an invariably-applied research principle is that **meaning** is determined by context. Context refers to environmental conditions surrounding a discourse and its parts which are related with it, and provides the framework to interpret that discourse. There are world views, relationships and interrelationships, and assertions by human actors

that combine to establish the context of those assertions and the means to interpret them.

In the concept of *context*, therefore, we see all of the components and building blocks of RDF itself. We have things or concepts (*subjects* or *objects*) that are related to one another (via properties or *predicates*) to form the basic assertions (*triples*). These are combined together and related in still more complex structures attempting to capture a world view or domain (*ontology*). These assertions have trust and credibility based on the actors (*provenance*) that make them.

In short, *context* is the essence of the semantic Web and Linked Data, not somehow in variance or conflict with it.

Without *context*, there is **_no_** *meaning*.

While one interpretation might be that the characteristics of one individual (say, *Quebec City*) might be oriented to latitude and longitude in a [GeoNames](#) source, while the characteristics of that individual may have a different context (say, population or municipal government) in the different [DBpedia](#) (Wikipedia) source, we need to be very careful of what is meant by *context* here. The identity of the individual (*Quebec City*) remains the same in both sources. The *context* does not change the individual nor its identity, only the nature of the characteristics used to provide different coherent information about it.

Not the Same Old sameAs

With the growth in Linked Data, we are starting to hear the rumblings around possible misuse and misapplication of the `sameAs` predicate [3]. Frankly, this is good, because I share the view there has been some confusion regarding the predicate and misapplications given its semantics.

The built-in OWL property `owl:sameAs` links an individual to an individual [4]. Such an owl:sameAs statement indicates that two URI references actually refer to the same thing: the individuals have the same "identity".

A link is a predicate is an assertion. It by nature ties ("glues") two resources to one another. Such an assertion can either: (1) be helpful and "correct"; (2) be made incorrectly; (3) assert the wrong or perhaps semantically poor relationship; or (4) be used maliciously or to deceive.

(Unlike email spam, #4 above has not occurred anywhere to my knowledge for Linked Data. Unfortunately, and most sadly, deceitful links will occur at some point, however. This inevitability is a contingency the community must be cognizant of as it moves forward.)

To date, almost all inter-source Linked Data links have occurred via `owl:sameAs`. If we liken this situation to early child language acquisition, it is like we only have one verb to describe the world. And because our vocabulary is relatively spare, we have tended to apply `sameAs` to situations and relations that, comparatively, have a bit of semblance to baby-talk.

So long as we have high confidence two disparate sources are referring to the ***same individual*** with the ***same identity***, `sameAs` is the semantically correct RDF link. In all other cases, the use of this predicate should be suspect.

Simple string or label matches are insufficient to make a `sameAs` assertion. If `sameAs` can not be confidently asserted, as might be the case where the relation of individual referents is perhaps likely but uncertain, we need to invoke new predicates or make no assertion at all. And, if the resources at hand are not individuals at all but classes, the need for new semantics increases still further.

As we increase the size of the [Linked Data 'cloud'](#) or show rapid growth in Linked Data, we should be aware that quality, not size, may be the most important metric powering acceptance. The community has made unbelievable progress in finally putting real data behind the semantic Web promise. The challenge now is to add to our vocabulary and ensure quality assertions for the linkages we publish.

Many Predicates Can Richen the RDF Link 'Glue'

One of UMBEL's purposes, for example, is to broaden our relations to the class level of subject concepts. As we move beyond the early days of FOAF and other early vocabularies, we will see further richening of our predicates. We also need predicates and predicate language that reflects the open-world nature [5] of public Linked Data and the semantic Web.

So, while `sameAs` helps us aggregate related information about the same identifiable individual, the predicates of class relations in context to other classes helps to put all information into context. And, if done right -- that is, if the semantics and assertions are relatively correct -- these desired contextual relations and interlinkages can blossom.

The new predicates forthcoming from the UMBEL project, to be published with technical documentation this month, and related to these purposes will include:

- `isAligned` -- the predicate for aligning external ontology classes to UMBEL subject concepts
- `isAbout` -- the predicate for relating individuals and instances to their contextual subject concepts, and
- `isLikely` -- the predicate for likely relations between the same identifiable individual, but where there is some ambiguity or uncertainty short of a `sameAs` assertion.

Assertions such as these that are open to ambiguity or uncertainty, while appropriate for much of the open-world nature of the semantic Web, may also be difficult predicates for the community to achieve consensus. Like our early experience with `sameAs`, these predicates -- or others that can just as easily arise in their stead -- will certainly prove subject to some growing pains. :)

Any Context is Better than No Context at All

Most people active in the semantic Web and Linked Data communities believe a decentralized Web environment leads to innovation and initiative. Open software, standards activities, and vigorous community participation affirm these beliefs daily.

The idea of context and global frames of reference, such as represented by UMBEL or perhaps any contextual ontology, could appear to be at odds with those ideals of decentralization. But one paradox is that without context, the basis for RDF linkages is made much poorer and therefore the potential for the

benefits (and thus adoption) of Linked Data lessen.

The object lesson should therefore not be a rejection of context. Indeed, any context is better than no context at all.

Of course, whether that context gets provided by UMBEL or by some other framework(s) remains to be seen. This is for the market to decide. But the ability of contextual frameworks to richen our semantics should be clear.

The past year with the growth and acceptance of Linked Data have affirmed that the mechanisms for linking and relating data are now largely in place. We have a simple, yet powerful and extensible data model in RDF. We have beginning vocabularies and constructs for conducting the data discourse. We have means for moving legacy data and information into this promising new environment.

Context and Linked Data are not in any way at odds, nor are context and `sameAs`. Indeed, context itself is an essential framework for how we can orient and grow our semantics. Human language required its referents in the real world in order to grow and blossom. Context is just as essential to derive and grow the semantics and meaning of the semantic Web.

The early innovators of the Linked Open Data community are the very individuals best placed to continue this innovation. Let's accept `sameAs` for what it is -- one kind of link in a growing menagerie of RDF link predicates -- and get on with the mission of putting our enterprise in context. I think we'll find our data has a lot more meaningfully to say -- and with more coherence.

[1] See the original post for the full comment; shown with some minor re-formatting.

[2] Hu Yijun, 2006. *On the Essence of Discourse: Context Coherence*, see http://www.paper.edu.cn/en/downloadpaper.php?serial_number=200606-221&type=1.

[3] For example, two papers presented at the the *Linked Data on the Web (LDOW2008) Workshop* at *WWW2008*, April 22, 2008, Beijing, China, two weeks ago highlight this issue. In the first, Bouquet et al. (Paolo Bouquet, Heiko Stoermer, Daniele Cordioli and Giovanni Tummarello, 2008. *An Entity Name System for Linking Semantic Web Data*, paper presented at *LDOW2008*, see http://events.linkeddata.org/ldow2008/papers/23-bouquet-stoermer-entity-name-system.pdf) state, "In fact it appears that the current use of `owl:sameAs` in the Linked Data community is not entirely in line with this definition and uses `owl:sameAs` more like a Semantic Web substitute forx [sp] a hyperlink instead of realizing the full logical consequences."

Also Jaffri et al. (Afraz Jaffri, Hugh Glaser and Ian Millard, 2008. *URI Disambiguation in the Context of Linked Data*, paper presented at *LDOW2008*, see http://events.linkeddata.org/ldow2008/papers/19-jaffri-glaser-uri-disambiguation.pdf) note that misuses of `sameAs` may result in a large percentage of entities being improperly conflated, entity references may be incorrect, and there is potential error propagation when mislabeled `sameAs` are then applied to new instances. As the authors state, "This will have a major impact on the Semantic Web when such repositories are used as data sources without any attempt to manage the inconsistencies or 'clean' the data." The authors also pose thoughtful mechanisms for addressing these issues built on the hard-earned co-referencing experience gained within the information extraction (IE) community.

These observations are not in any way meant to be critical or alarmist. They simply point to the need for quality control and accurate semantics when asserting relationships. These growing pains are a natural circumstance of rapid

growth.

[4] W3C, *OWL Web Ontology Language Reference*, W3C Recommendation, 10 February 2004. See
http://www.w3.org/TR/owl-ref/. Note that in OWL Full sameAs can also be applied to classes, but that is a special case, not applicable to how Linked Data has been practiced to date, and is not further discussed here.

[5] The "open world" assumption is defined in the SKOS ontology reference documentation (W3C, *SKOS Simple Knowledge Organization System Reference*, W3C Working Draft, 25 January 2008; see
http://www.w3.org/TR/2008/WD-skos-reference-20080125/#L881) as:

"RDF and OWL Full are designed for systems in which data may be widely distributed (*e.g.*, the Web). As such a system becomes larger, it becomes both impractical and virtually impossible to "know" where all of the data in the system is located. Therefore, one cannot generally assume that data obtained from such a system is "complete", *i.e*., if some data appears to be "missing", one has to assume, in general, that the data might exist somewhere else in the system. This assumption, roughly speaking, is known as the "open world" assumption."

_____

PDF generated by *AI3:::Adaptive Information* blog