When is Content Coherent?

by Mike Bergman - Friday, July 25, 2008

http://www.mkbergman.com/450/when-is-content-coherent/

Structure Demands Context; But, is that Enough?

Last week marked a red-letter day in my professional life with release of the UMBEL subject concept structure. UMBEL began as a gleam in the eye more than a year ago when I observed that semantic Web techniques, while powerful -- especially with regard to the RDF data model as a universal and simple (at its basics) means for representing any information and its structure -- still lacked *something*. It took me a while to recognize that the first something was ***context***.

Now, I have written and talked much about ***context*** before on this blog, with *The Semantics of Context* being the most salient article for the present discussion.

This is my mental image of Web content without context: Unconnected dust motes floating through a sun-lite space, moving slowly, randomly, and without connections, sort of like Brownian motion. Think of the sunlight on dust shown by the picture to the left.

By providing context, my vision saw we could freeze these moving dust motes and place them into a fixed structure, perhaps something like constellations in the summer sky. Or, at least, more stable, and floating less aimlessly and unconnected.

So, my natural response was to look for structural frameworks to provide that ***context***. And that was the quest I set forward at UMBEL's initiation.

At the time of UMBEL's genesis, the impact of Wikipedia and other sources of user-generated content (UGC) such as del.icio.us or Flickr or many, many others was becoming clear. The usefulness of tags, folksonomies, microformats and other forms of "bottom-up" structure was proven.

The evident -- and to me, exciting -- aspect of globally-provided UGC was that this was the ultimate democratic voice: the World has spoken, and the article about *this* or the tag about *that* had been vetted in the most interactive, exposed, participatory and open framework possible. Moreover, as the World changed and grew, these new realizations would also be fed back into the system in a self-correcting goodness. Final dot.

Through participation and collective wisdom, therefore, we could gain consensus and acceptance and avoid the fragility and arbitrariness of "wise man" or imposed from the "top-down" answers. The people have spoken. All voices have been heard. The give and take of competing views have found their natural resting point. Again, I thought, final dot.

Thus, when I first announced UMBEL, my stated desire (and hope) was that something like Wikipedia could or would provide that structural context. Here is a quote from the announcement of UMBEL, nearly one year ago to this day:

The selection of the actual subject proxies within the UMBEL core are to be based on consensus use. The subjects of existing and popular Web subject portals such as Wikipedia and the Open Directory Project (among others) will be intersected with other widely accepted subject reference systems such as WordNet and library classification systems (among others) in order to derive the candidate pool of UMBEL subject proxies.

Yet, that is not the basis of the structure announced last week for UMBEL. Why?

The Strengths of User-Generated Content

Before we probe the negative, let's rejoice the positive.

User-generated content (UGC) works, has rapidly proven itself in venues from authoritative subjects (Wikipedia), photos (Flickr), bookmarking and tagging (del.icio.us), blogs, video (YouTube) and every Web space imaginable. This is new, was not foreseen by most a few years ago, and has totally remade our perception of content and how it can be generated. Wow!

The nature of this user-generated content, of course, as is true for the Web itself, is that it has arisen from a million voices without coercion, coordination or a plan, spontaneously in relation to chosen platforms and portals. Yet, still, today, as to what makes one venue more successful than others, we are mostly clueless. My suspicion is that -- akin to financial markets -- when Web portals or properties are successful, they readily lend themselves to retrospective books and learned analysis explaining that success. But, just try to put down that "recipe" in advance, and you will most likely fail.

So, prognostication is risky business around these parts.

There is a reason why both the head and sub-head of this article are stated as questions: I don't know. For

the reasons stated above, I would still prefer to see user-generated structure (UGS) emerge in the same way that topic- and entity-specific content has on Wikipedia. However, what I can say is this: for the present, this structure has not yet emerged in a coherent way.

Might it? Actually, I hope so. But, I also think it will not arise from systems or environments exactly like Wikipedia and, if it does arise, it will take considerable time. I truly hope such new environments emerge, because user-mediated structure will also have legitimacy and wisdom that no "expert" approach may ever achieve.

But these are *what if*'s, and *nice to have*'s and *wouldn't it be nice*'s. For my purposes, and the clients my company serves, what is needed must be pragmatic and doable today -- all with acceptable risk, time to delivery and cost.

So, I think it safe to say that UGC works well today at the atomic level of the individual topic or data object, what might be called the nodes in global content, but not in the connections between those nodes, its structure. And, the key to the answer of why user-generated structure (UGS) has not emerged in a bottom-up way resides in that pivotal word above: ***coherence***.

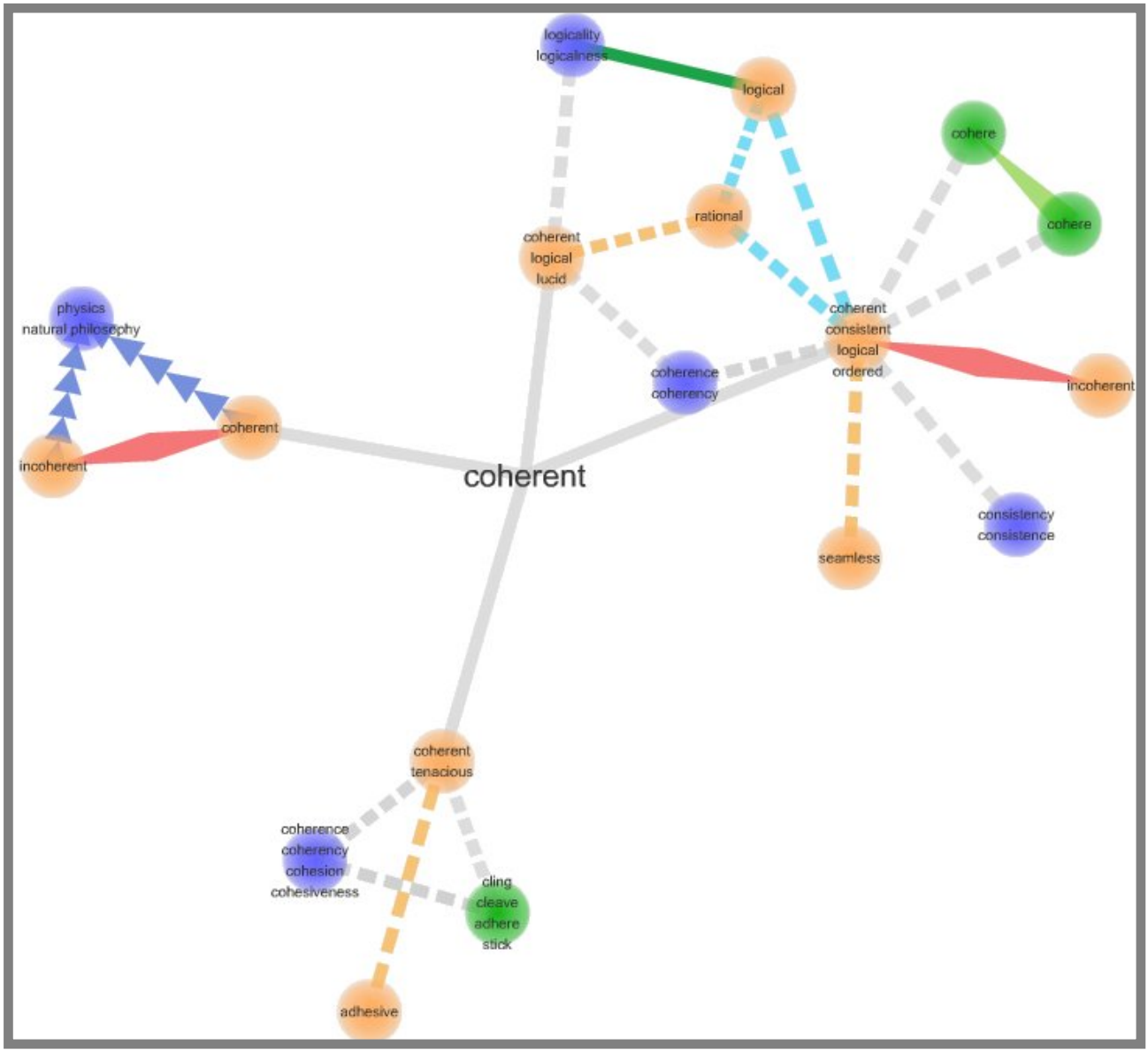
Coherence was the second *something* to accompany context as lacking missing pieces for the semantic Web.

Coherence in Context

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

coherent *k?-?hir-?nt\ adj.*;* Middle French or Latin; Middle French *cohérent*, from Latin *cohaerent-*, *cohaerens*, present participle of *cohaer?re* Date: (*ca*. 1555)**1: a:** logically or aesthetically ordered or integrated : consistent <coherent style> <a *coherent* argument> **b:** having clarity or intelligibility : understandable <a *coherent* person> <a *coherent* passage>
2: having the quality of cohering; especially : cohesive, coordinated <a *coherent* plan for action>
3: a: relating to or composed of waves having a constant difference in phase <*coherent* light> **b:** producing coherent light <a *coherent* source>.

Another online source I like for visualization purposes is Visuwords, which displays the accompanying graph relationships view based on WordNet.



Of course, coherent is just the adjectival property of having coherence. Again, the Merriam Webster dictionary defines *coherence* as ***1:*** *the quality or state of cohering: as* ***a:*** *systematic or logical connection or consistency* ***b:*** *integration of diverse elements, relationships, or values.*

Decomposing even further, we can see that coherence is itself the state of the verb, cohere. *Cohere*, as in its variants above, has as its etymology a derivation from the Latin *cohaer?re*, from *co-* + *haer?re* to stick, namely “to stick with”. Again, the Merriam Webster dictionary defines *cohere* as ***1: a:*** *to hold together firmly as parts of the same mass; broadly: stick, adhere* ***b:*** *to display cohesion of plant parts* ***2:*** *to hold together as a mass of parts that cohere* ***3: a:*** *to become united in principles, relationships, or interests* ***b:*** *to be logically or aesthetically consistent.*

These definitions capture the essence of coherence in that it is a state of logical, consistent connections, a logical framework for integrating diverse elements in an intelligent way. In the sense of a content graph, this means that the right connections (edges or predicates) have been drawn between the object nodes (or content) in the graph.

Bottom-up UGC: The Hip Bone is Connected to the Arm Bone

Structure without coherence is where connections are being drawn between object nodes, but those connections are incomplete or wrong (or, at least, inconsistent or unintelligible). The nature of the content graph lacks logic. The hip bone is not connected to the thigh bone, but perhaps to something wrong or silly, like the arm or cheek bone.

Ambiguity is one source for such error, as when, for example, the object "bank" is unclear as to whether it is a financial institution, billiard shot, or edge of a river. If we understand the object to be the wrong thing, then connections can get drawn that are in obvious error. This is why *disambiguation* is such a big deal in semantic systems.

However, ambiguity tends not to be a major source of error in user-generated content (UGC) systems because the humans making the connections can see the ***context*** and resolve the meanings. Context is thus a very important basis for resolving disambiguities.

A second source of possible incoherence is the *organizational structure* or *schema* of the actual concept relationships. This is the source that poses the most difficulty to UGC systems such as folksonomies or Wikipedia.

Remember in the definitions above that *logic*, *consistency* and *intelligibility* were some of the key criteria for a coherent system. Bottom-up UGS (user-generated structure) is prone to not meet the test in all three areas.

"In the context of an information organization framework, a structure is a cohesive whole or 'container' that establishes qualified, meaningful relationships among those activities, events, objects, concepts which, taken together, comprise the 'bounded space' of the universe of interest." -- J.T. Tennis and E.K. Jacob [1]

Logic and consistency almost by definition imply the application of a uniform perspective, a single world view. Multiple authors and contributors doing so without a common frame of reference or viewpoint are unable to bring this consistency of perspective. For example, how time might be treated with regard to famous people's birth dates in Wikipedia is very different than its discussion of time with respect to topics on geological eras, and Wikipedia contains no mechanisms for relating those time dimensions or making them consistent.

Logic and intelligibility suggest that the structure should be testable and internally consistent. Is the hip bone connected with the arm bone? No? and why not? In UGC systems, individual connections are made by consensus and at the object-to-object level. There are no mechanisms, at least in present systems, for resolving inconsistencies as these individual connections get aggregated. We can assign *dogs* as *mammals* and *dogs* as *pets*, but does that mean that all *pets* are *mammals*? The connections can get complicated fast and such higher-order relationships remain unstated or more often than not wrong.

Note as well that in UGC systems items may be connected ("assigned") to categories, but their "factual" relation is not being asserted. Again, without a consistency of how relations are treated and the ability to test assertions, the structures may not only be wrong in their topology, but totally lack any inference power. Is the hip bone connected with the cheek bone? UGC structures lack such fundamental logic underpinnings to test that, or any other, assertion.

From the first days of the Web, notably Yahoo! in its beginnings but many other portals as well, we have seen many taxonomies and organizational structures emerge. As simple heuristic devices for clustering large amounts of content, this is fine (though certainly there, too, there are some structures that are better at organizing along "natural" lines than others). Wikipedia itself, in its own structure, has useful organizational clustering.

But once a system is proposed, such as UMBEL, with the purpose of providing broad referenceability to virtually any Web content, the threshold condition changes. It is no longer sufficient to merely organize. The structure must now be more fully graphed, with intelligent, testable, consistent and defensible relations.

Full Circle to Cyc and UGC

Once the seemingly innocent objective of being a lightweight subject reference structure was established for UMBEL, the die was cast. Only a coherent structure would work, since anything else would be fragile and rapidly break in the attempt to connect disparate content. Relating content coherently itself demands a coherent framework.

As noted in the lead-in, this was not a starting premise. But, it became an unavoidable requirement once the UMBEL effort began in earnest.

I have spoken elsewhere about other potential candidates as possibly providing the coherent underlying structure demanded by UMBEL. We have also discussed why Cyc, while by no means perfect, was chosen as the best starting framework for contributing this coherent structure.

I anticipate we will see many alternative structures proposed to UMBEL based on other frameworks and premises. This is, of course, natural and the nature of competition and different needs and world views.

However, it will be most interesting to see if either *ad hoc* structures or those derived from bottom-up UGC systems like Wikipedia can be robust and coherent enough to support data interoperability at Web scale.

I strongly suspect not.

---

[1] Joseph T. Tennis and Elin K. Jacob, 2008. "Toward a Theory of Structure in Information Organization Frameworks," upcoming presentation at the *10th International Conference of the International Society for Knowledge Organization (ISKO 10)*, in Montréal, Canada, August 5th-8th, 2008. See http://www.ebsi.umontreal.ca/isko2008/documents/abstracts/tennis.pdf.
