Where are the Road Signs for the Structured Web?

by Mike Bergman - Tuesday, May 29, 2007

http://www.mkbergman.com/375/where-are-the-road-signs-for-the-structured-web/

The ***Why***, ***How*** and ***What*** of the Semantic Web are Becoming Clear -- Yet the Maps and Road Signs to Guide Our Way are Largely Missing

There has been much recent excitement surrounding RDF, its linked data, "RDFizers" and GRDDL to convert existing structured information to RDF. The belief is that 2007 is the breakout year for the semantic Web.

The *why* of the semantic Web is now clear. The *how* of the semantic Web is now appearing clear, built around RDF as the canonical data model and the availability and maturation of a growing set of tools. And the *what* is also becoming clear, with the massive new datastores of DBpedia, Wikipedia[3], the HCLS demo, Musicbrainz, Freebase, and the Encyclopedia of Life only being some of the most recent and visible exemplars.

Yet the *where* aspect seems to be largely missing.

By *where* I mean: *Where* do we look for our objects or data? If we have new objects or data, *where* do we plug into the system? Amongst all possible information and domains, *where* do we fit?

These questions seem simplistic or elemental in their basic nature. It is almost incomprehensible to wonder *where* all of this data now emerging in RDF relates to each other -- what the overall frame of reference is -- but my investigations seem to point to such a gap. In other words, a key piece of the emerging semantic Web infrastructure -- *where* is this stuff -- seems to be missing. This gap is across domains and across standard ontologies.

What are the specific components of this missing *where*, this missing infrastructure? I believe them to be:

- Lack of a central look-up point for *where* to find this RDF data in reference to desired subject

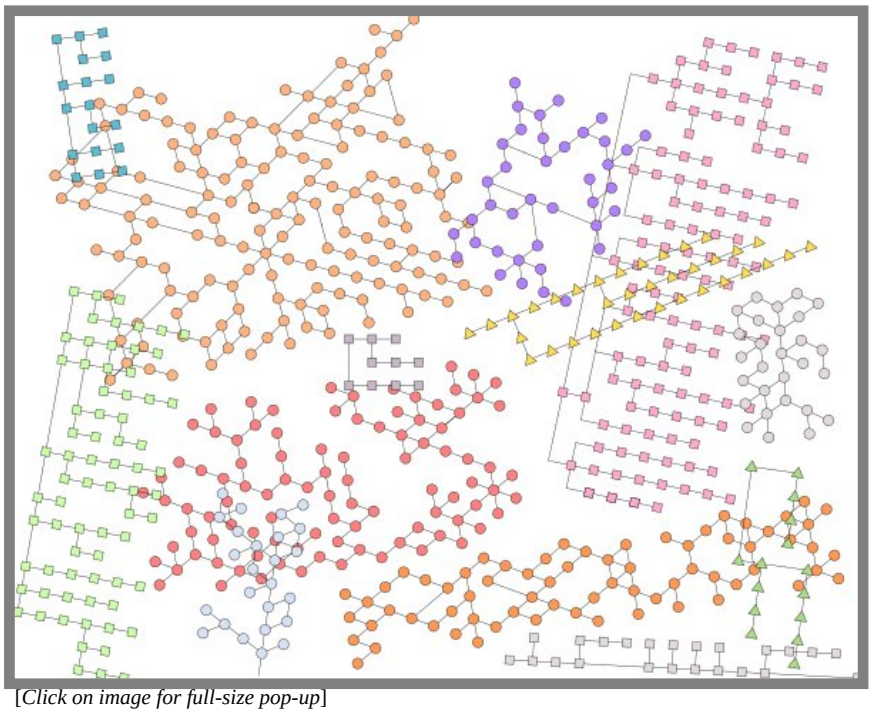
matter
- Lack of a reference subject context for *where* this relevant RDF data fits; *where* can we place this data in a contextual frame of reference -- animal, mineral, vegetable?
- Lack of a open means *where* any content structure -- from formal ontologies to "RDFized" documents to complete RDF data sets -- can "bind" or "map" to other data sets relevant to its subject domain
- Lack of a registration or publication mechanism for data sets that do become properly placed, the *where* of finding SPARQL or similar query endpoints, and
- In filling these gaps, the need for a broad community process to give these essential infrastructure components legitimacy.

I discuss these missing components in a bit more detail below, concluding with some preliminary thoughts on how the problem of this critical infrastructure can be redressed. The good news, I believe, is that these potholes on the road to the semantic Web can be relatively easily and quickly filled.

The Lack of Road Signs Causes Collisions and Missed Turns

I think it is fair to say that structure on the current Web is a jumbled mess.

As my recent *Intrepid Guide to Ontologies* pointed out, there are at least 40 different approaches (or types of *ontologies*, loosely defined) extant on the Web for organizing information. These approaches embrace every conceivable domain and subject. The individual data sets using these approaches span many, many orders of magnitude in size and range of scope. Diversity and chaos we have aplenty, as the illustrative diagram of this jumbled structural mess shows below.



[*Click on image for full-size pop-up*]

Mind you, we are not yet even talking about whether one dot is equivalent or can be related to another dot

and in what way (namely, connecting the dots via real *semantics*), but rather at a more fundamental level. Does one entire data set have a relation to any other data set?

Unfortunately, the essential precondition of getting data into the canonical RDF data model -- a challenge in its own right -- does little to guide us as to where these data sets exist or how they may relate. Even in RDF form, all of this wonderful RDF data exists as isolated and independent data sets, bouncing off of one another in some gross parody of Brownian motion.

What this means, of course, is that useful data that could be of benefit is overlooked or not known. As with problems of data silos everywhere, that blindness leads to unnecessary waste, incomplete analysis, inadequate understanding, and duplicated effort [1].

These gaps were easy to overlook when the focus of attention was on the *why*, *what* and *how* of the semantic Web. But, now that we are seeing RDF data sets emerge in meaningful numbers, the time is ripe to install the road signs and print up the maps. It is time to figure out *where* we want to go.

The Need for a Lightweight Subject Mapping Layer

As I discussed in an earlier posting, there's not yet enough backbone to the *structured Web*. I believe this structure should firstly be built around a lightweight subject- or topic-oriented **reference layer**.

> *An umbrella subject reference becomes the "super-structure" to which other specific ontologies can place themselves in an "info-spatial" context.*

Unlike traditional upper-level ontologies (see the Intrepid Guide), this backbone is not meant to be comprised of abstract concepts or a logical completeness of the "nature of knowledge". Rather, it is meant to be only the thinnest veneer of (mostly) hierarchically organized subjects and topic references (see more below).

This subject or topic vocabulary (at least for the backbone) is meant to be quite small, likely more than a few hundred reference subjects, but likely less than many thousands. (There may be considerably more terms in the overall controlled vocabulary to assist context and disambiguation.)

This "umbrella" subject structure could be thought of as the reference subject "super-structure" to which other specific ontologies could place themselves in a sort of locational or "info-spatial" context.

One way to think of these subject reference nodes is as the major destinations -- the key cities, locations or interchanges -- on the broader *structured Web* highway system. A properly constructed subject structure could also help disambiguate many common misplacements by virtue of the context of actual subject mappings.

For example, an ambiguous term such as "driver" becomes unambiguous once it is properly mapped to

one of its possible topics such as golf, printers, automobiles, screws, NASCAR, or whatever. In this manner, context is also provided for other terms in that contributing domain. (For example, we would now know how to disambiguate "cart" as a term for that domain.)

A high-level and lightweight subject mapping layer does not warrant difficult (and potentially contentious) specificity. The point is not to comprehensively define the scope of all knowledge, but to provide the fewest choices necessary for what subject or subjects a given domain ontology may appropriately reference. We want a listing of the major destinations, not every town and parish in existence.

(That is not to say that more specific subject references won't emerge or be appropriate for specific domains. Indeed, the hope is that an "umbrella" reference subject structure might be a tie-in point for such specific maps. The more salient issue addressed here is to create such an "umbrella" backbone in the first place.)

This subject reference "super-structure" would in no way impose any limits on what a specific community might do itself with respect to its own ontology scope, definition, format, schema or approach. Moreover, there would be no limit to a community mapping its ontology to multiple subject references (or "destinations", if you will).

The reason for this high-level subject structure, then, is simply to provide a reference map for where we might want to go -- no more, no less. Such a reference structure would greatly aid finding, viewing and querying actual content ontologies -- of whatever scope and approach -- wherever that content may exist on the Web.

This is not a new idea. About the year 2000 the topic map community was active with published subject indicators (PSIs) [2] and other attempts at topic or subject landmarks. For example, that report stated:

The goal of any application which aggregates information, be it a simple back-of-book index, a library classification system, a topic map or some other kind of application, is to achieve the "collocation objective;" that is, to provide binding points *from which everything that is known about a given subject can be reached. In topic maps, binding points take the form of topics; for a topic map application to fully achieve the collocation objective there must be an exact one-to-one correspondence between subjects and topics: Every topic must represent exactly one subject and every subject must be represented by exactly one topic.*

When aggregating information (for example, when merging topic maps), comparing ontologies, or matching vocabularies, it is crucially important to know when two topics represent the same subject, in order to be able to combine them into a single topic. To achieve this, the correspondence *between a topic and the subject that it represents needs to be made clear. This in turn requires subjects to be identified in a non-ambiguous manner.*

The identification of subjects is not only critical to individual topic map applications and to interoperability between topic map applications; it is also critical to interoperability between *topic map applications and other applications that make explicit use of abstract representations of subjects, such as RDF.*

From that earlier community, Bernard Vatant has subsequently spoken of the need and use of "hubjects" as organizing and binding points, as has Jack Park and Patrick Durusau using the related concept of "subject maps" [3]. An effort that has some overlap with a subject structure is also the Metadata Registry being maintained by the National Science Digital Library (NSDL).

However, while these efforts support the idea of subjects as partial binding or mapping targets, none of them actually proposed a reference subject structure. Actual subject structures may be a bit of a "third rail" in ontology topics due to the historical artifact of wanting to avoid the pitfalls of older library classification systems such as the Dewey Decimal Classification or the Library of Congress Subject Headings.

Be that as it may. I now think the timing is right for us to close this subject gap.

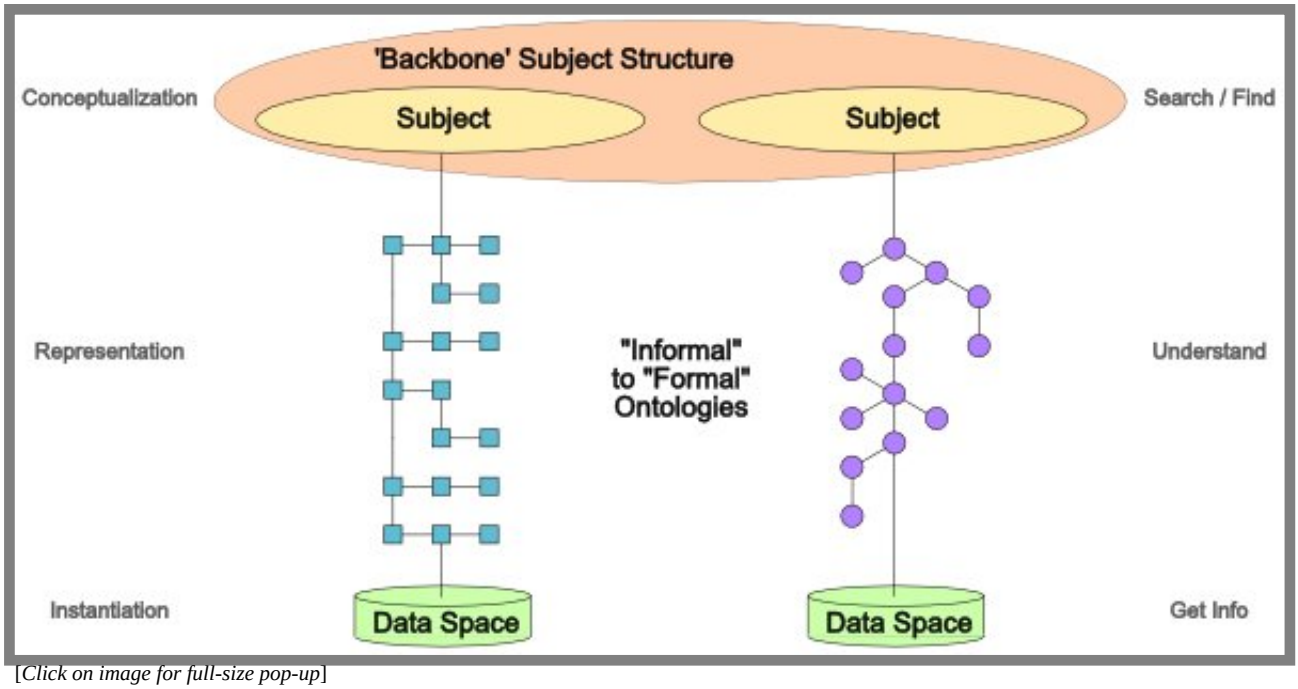
A General Conceptual Model

This mapping layer lends itself to a three-tiered general conceptual model. The first tier is the subject structure, the conceptualization embracing all possible subject content. This referential layer is the lookup point that provides guidance for where to search and find "stuff."

The second layer is the representation layer, made up of informal to "formal" ontologies. Depending on the formalism, the ontology provides more or less understanding about the subject matter it represents, but at minimum binds to the major subject concepts in the top subject mapping layer.

The third layer are the data sets and "data spaces" [4] that provide the actual content instantiations of these subjects and their ontology representations. This data space layer is the actual source for getting the target information.

Here is a diagram of this general conceptual model:



[*Click on image for full-size pop-up*]

The layers in this general conceptual model progress from the more abstract and conceptual at the upper level, useful for directing where traffic needs to go, to concrete information and data at the lower level, the real object of manipulation and analysis.

The data spaces and ontologies of various formalisms in the lower two tiers exist in part today. The upper mapping layer does not.

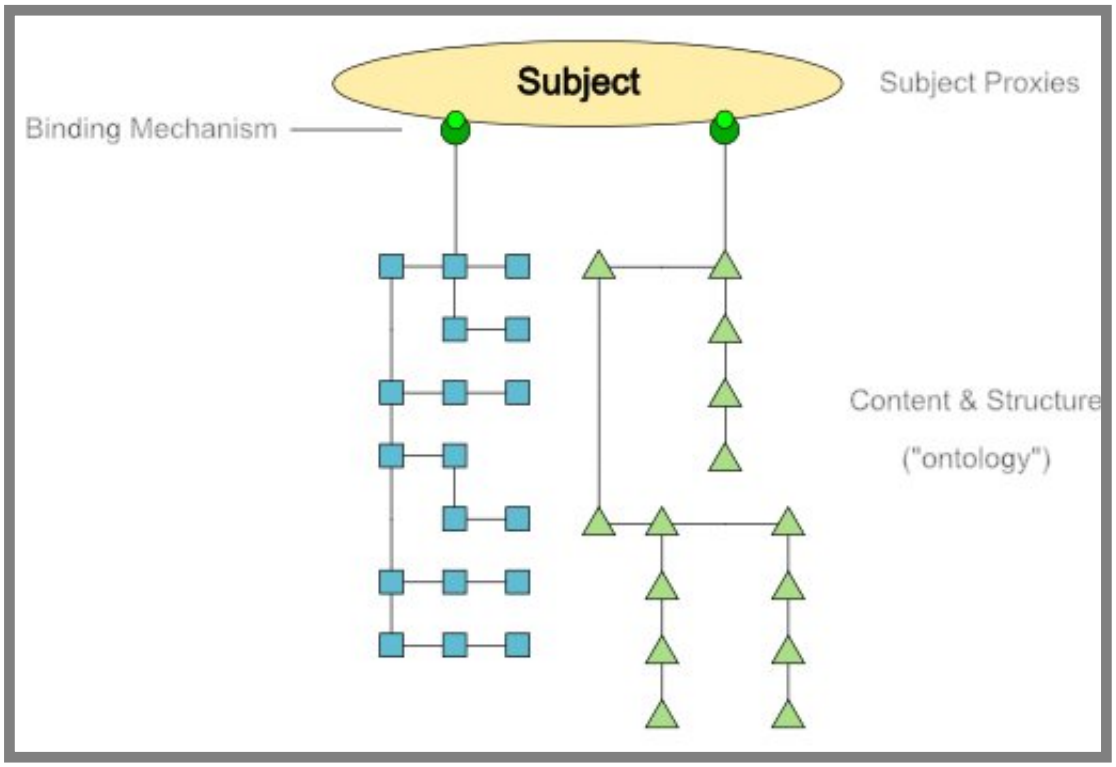
By its nature the general upper mapping layer, in its role as a universal reference "backbone," must be somewhat general in the scope of its reference subjects. As this mapping infrastructure begins to be fleshed out, it is also therefore likely that additional intermediate mapping layers will emerge for specific domains, which will have more specific scopes and terminology for more accurate and complete understanding with their contributing specific data spaces.

Six Principles for a Possible Lightweight Binding Mechanism

So, let's beg the question for a moment of what an actual reference subject structure might be. Let's assume one exists. What should we do with this structure? How should we bind to it?

- First, we need to assume that the various ontologies that might bind to this structure reside in the real world, and have a broad diversity of domains, topics and formality of structure. Therefore, we should: a) provide a binding mechanism responsive to the real-world range of formalisms (that is, make no grand assumptions or requirements of structure; each subject structure will be be provided as is); and b) thus place the responsibility to register or bound the subject mapping assignment(s) to the publisher of the contributing content ontology [5].
- Second, we can assume that the reference subject structure (light green below) and its binding ontology basis (dark green) are independent actors. As with many other RESTful services, this needs to work in a peer-to-peer (P2P) manner.
- Third, as the Intrepid Guide argues, RDF and its emergent referential schema provide the natural data model and characterization "middle ground" across all Web ontology formalisms. This observation leads to SKOS as the organizing schema for ontology integration, supplemented by the related RDF schema of DOAP, SIOC, FOAF and Geonames for the concepts of projects, communities, people and places, respectively. Other standard referents may emerge, which should also be able to be incorporated.
- Fourth, the actual binding point of "subjects" are themselves only that: binding points. In other words, this makes them representative "proxies" not to be confused with the actual subjects themselves. This implies no further semantics than a possible binding, and no assertion about the accuracy, relevance or completeness of the binding. How such negotiation may resolve itself needs to be outside of this scope of a simple mapping and binding reference layer (see again [3]).
- Fifth, the binding structure and its subject structure needs to have community acceptance; no "wise guys" here, just well-intentioned bozos.
- And, sixth, keep it simple. While not all publishers of Web sites need comply -- and the critical threshold is to get some of the major initial publishers to comply -- the truth like everything else on the Web is that the network effect makes thinks go. By keeping it simple, individual publishers and tool makers are more likely to use and contribute to the system.

How this might look in a representative subject binding to two candidate ontology data sets is shown below:

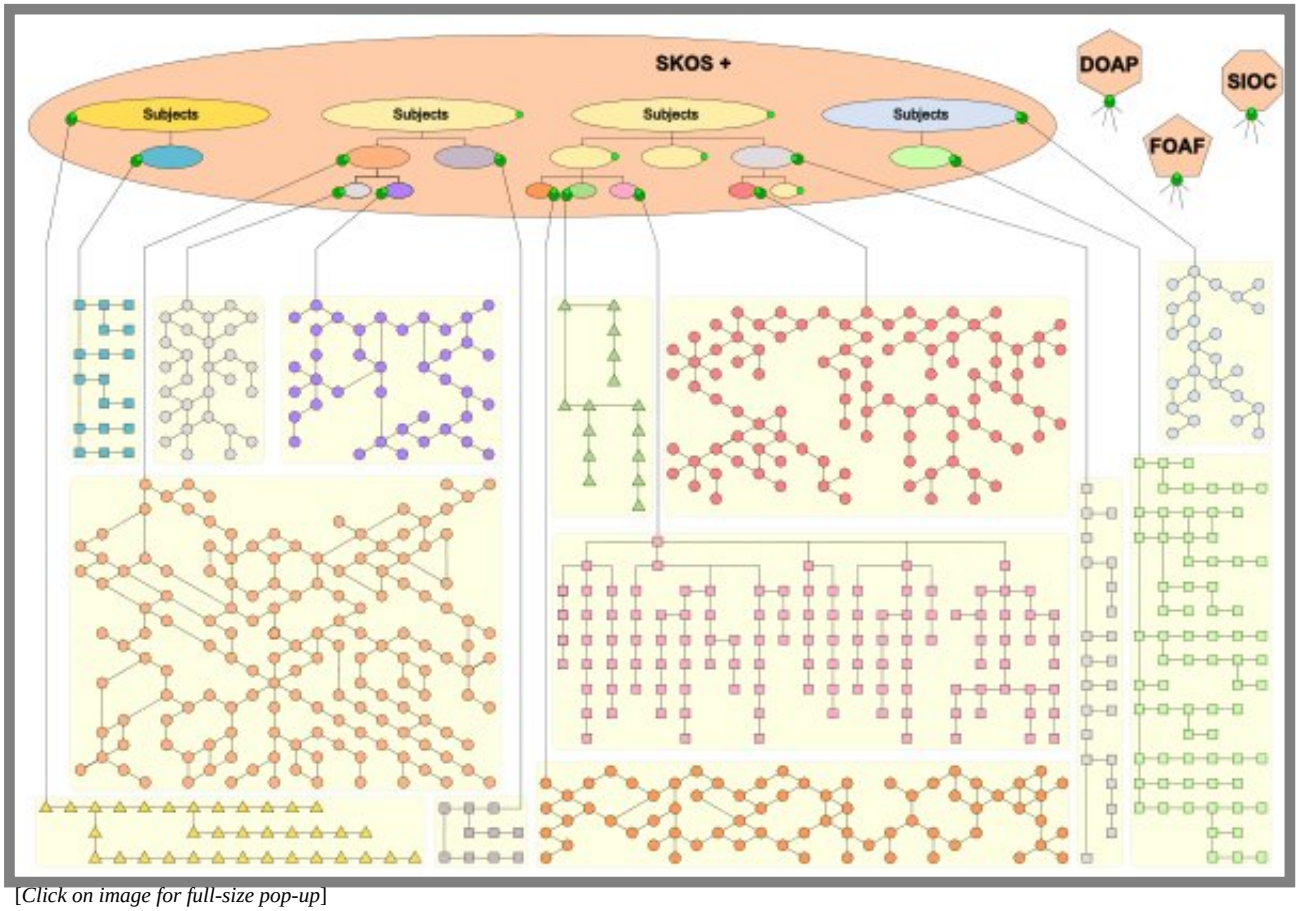


This diagram shows two contributing data sets and their respective ontologies (no matter how "formal") binding to a given "subject." This subject proxy may not be the only one bound by a given data set and its ontology. Also note the "subject" is completely arbitrary and, in fact, is only a proxy for the binding to the potential topic.

One Possible Road Map

Clearly, this approach does ***not*** provide a powerful inferential structure.

But, what it does provide is a quite powerful ***organizational*** structure. Access and the *where* for the sake of simplicity and adoption are given preference over inferential elegance. Thus, assuming now, again, a subject structure backbone, matched with these principles and the same jumbled structures as first noted above, we can now see an organizational order emerge from the chaos:



[*Click on image for full-size pop-up*]

The assumptions to get to this point are not heroic. Simple binding mechanisms matched with a high-level subject "backbone" are all that is required *mechanically* for such an approach to emerge. All we have done to achieve this road map is to follow the trails above.

Use Existing Consensus to Achieve Authority

So, there is nothing really revolutionary in any of the discussion to this point. Indeed, many have cited the importance of reference structures previously. Why hasn't such a subject structure yet been developed?

One explanation is that no one accepts any global "external authority" for such subject identifications and organization. The very nature of the Web is participatory and democratic (with a small "d"). Everyone's voice is equal, and any structure that suggests otherwise will not be accepted.

It is not unusual, therefore, that some of the most accepted venues on the Web are the ones where everyone has an equal chance to participate and contribute: Wikipedia, Flickr, eBay, Amazon, Facebook, YouTube, etc. Indeed, figuring out what generates such self-sustaining "magic" is the focus of many wannabe ventures.

While that is not our purpose here, our purpose is to set the preconditions for what would constitute a referential subject structure that can achieve ***broad acceptance*** on the Web. And in the paragraph above, we already have insight into the answer: Build a subject structure *from already accepted sources* rich in

subject content.

A suitable subject structure must be adaptable and self-defining. These criteria reflect expressions of actual social usage and practice, which of course changes over time as knowledge increases and technologies evolve.

One obvious foundation to building a subject structure is thus Wikipedia. That is because the starting basis of Wikipedia information has been built entirely from the bottom up -- namely, what is a deserving topic. This has served Wikipedia and the world extremely well, with now nearly 1.8 million articles online in English alone (versions exist for about 100 different languages) [6]. There is also a wealth of internal structure within Wikipedia's "infobox" templates, structure that has been utilized by DBpedia (among others) to actually transform Wikipedia into an RDF database (as I described in an earlier article). As socially-driven and -evolving, I foresee Wikipedia to continue to be the substantive core at the center of a knowledge organizational framework for some time to come.

But Wikipedia was never designed with an organizing, high-level subject structure in mind. For my arguments herein, creating such an organizing (yes, in part, hierarchical) structure is pivotal.

One innovative approach to provide a more hierarchical structural underpinning to Wikipedia has been YAGO ("yet another great ontology"), an effort from the Max-Planck-Institute SaarbrÃ¼cken [7]. YAGO matches key nouns between Wikipedia and WordNet, and then uses WordNet's well-defined taxonomy of synsets to superimpose the hierarchical class structure. The match is more than 95% accurate; YAGO is also designed for extensibility with other quality data sets.

I believe YAGO or similar efforts show how the foundational basis of Wikipedia can be supplemented with other accepted lexicons to derive a suitable subject structure with appropriate high-level "binding" attributes. In any case, however constructed, I believe that a high-level reference subject structure must evolve from the global community of practice, as has Wikipedia and WordNet.

I have previously described this formula as **W + W + S + ?** (for Wikipedia + WordNet + SKOS + other?). There indeed may need to be "other" contributing sources to construct this high-level reference subject structure. Other such potential data sets could be analyzed for subject hierarchies and relationships using fairly well accepted ontology learning methods. Additional techniques will also be necessary for multiple language versions. Those are important details to be discussed and worked out.

The real point, however, is that existing and accepted information systems already exist on the Web that can inform and guide the construction of a high-level subject map. As the contributing sources evolve over time, so could periodic updates and new versions of this subject structure be generated.

Though the choice of the contributing data sets from which this subject structure could be built will never be unanimous, using sources that have already been largely selected through survival of the "fittest" by large portions of the Web-using public will go a long ways to establishing authoritativeness. Moreover, since the subject structure is only intended as a ***lightweight reference structure*** -- and not a complete ***closed-world definition*** -- we are also setting realistic thresholds for acceptance.

Conclusion and Next Steps

The specific topic of this piece has been on a **subject** reference mapping and binding layer that is lightweight, extensible, and reflects current societal practice (broadly defined). In the discussion, there has been recognition of existing schema in such areas as people (FOAF), projects (DOAP), communities (SIOC) and geographical places (Geonames) that might also contribute to the overall binding structure. There very well may need to be some additional expansions in other dimensions such as time and events, organizations, products or whatever. I hope that a consensus view on appropriate high-level dimensions emerges soon.

There are a number of individuals presently working on a draft proposal for an open process to create this subject structure. What we are working quickly to draft and share with the broader community is a proposal related to:

1. A reference umbrella subject binding ontology, with its own high-level subject structure
2. Lightweight mechanisms for binding subject-specific community ontologies to this structure
3. Identification of existing data sets for high-level subject extraction
4. Codification of high-level subject structure extraction techniques
5. Identification and collation of tools to work with this subject structure, and
6. A public Web site for related information, collaboration and project coordination.

We believe this to be an exciting and worthwhile endeavor. Prior to the unveiling of our public Web site and project, I encourage any of you with interest in helping to further this cause to contact me directly at *mike at mkbergman dot com* [8].

---

[1] I attempted to quantify this problem in a white paper from about two years ago, Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents, BrightPlanet Corporation, 42 pp., July 20, 2005. Some reasons for how such waste occurs were documented in a four-part series on this **AI3** blog, Why Are $800 Billion in Document Assets Wasted Annually?, beginning in October 2005 through parts two, three and four concluding in November 2005.

[2] See Published Subjects: Introduction and Basic Requirements (OASIS Published Subjects Technical Committee Recommendation, 2003-06-24.

[3] See especially Park and Durusau, Avoiding Hobson's Choice In Choosing An Ontology and Towards Subject-centric Merging of Ontologies.

[4] The concept of "data spaces" has been well-articulated by Kingsley Idehen of OpenLink Software and Frédérick Giasson of Zitgist LLC. A "data space" can be personal, collective or topical, and is a virtual "container" for related information irrespective of storage location, schema or structure.

[5] If the publisher gets it wrong, and users through the reference structure don't access their desired content, there will be sufficient motivation to correct the mapping.

[6] See Wikipedia's statistics sections.

[7] Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum, "Yago - A Core of Semantic Knowledge" (also in bib or ppt). Presented at the 16th International World Wide Web Conference (WWW 2007) in Banff, Alberta, on May 8-12, 2007. YAGO contains over 900,000 entities (like persons, organizations, cities, etc.) and 6 million facts about these entities, organized under a hierarchical schema. YAGO is available for download (400Mb) and converters are available for XML, RDFS, MySQL, Oracle and Postgres. The YAGO data set may also be queried directly online.

[8] I'd especially like to thank Frédérick Giasson and Bernard Vatant of Mondeca for their reviews of a draft of this posting. Fred was also instrumental in suggesting the ideas behind the figure on the general conceptual model.
