What is a Reference Concept?

by Mike Bergman - Monday, December 06, 2010

http://www.mkbergman.com/938/what-is-a-reference-concept/

And, Seven Guidelines for this Second of Two Semantic 'Gaps'

I have been writing and speaking of late about next priorities to promote the interoperability of linked data and the semantic Web. In a talk a few weeks back to the Dublin Core (DCMI) annual conference, I summarized these priorities as the need to address two aspects of the semantic "gap":

1. One aspect is the need for vetted reference sources that provide the entities and concepts for aligning disparate content sources on the Web, and
2. A second aspect is the need for accurate mapping predicates that can represent the often approximate matches and overlaps of this heterogeneous content.

I discussed the second aspect in an earlier post [1]. In today's installment, we now focus on the "gap" relating to reference concepts.

The Web *Increases* the Need for Organization

Interoperability comes down to the nature of things and how we describe those things or quite similar things from different sources. Given the robust nature of semantic heterogeneities in diverse sources and datasets on the Web (or anywhere else, for that matter!) [2], how do we bring similar or related things into alignment? And, then, how can we describe the nature or basis of that alignment?

Of course, classifiers since Aristotle and librarians for time immemorial have been putting forward various classification schemes, controlled vocabularies and subject headings. When one wants to find related books, it is convenient to go to a central location where books about the same or related topics are clustered. And, if the book can be categorized in more than one way -- as all are -- then something like a

card catalog is helpful to find additional cross-references. Every domain of human endeavor makes similar attempts to categorize things.

On the Web we have none of the limitations of physical books and physical libraries; locations are virtual and copies can be replicated or split apart endlessly because of the essentially zero cost of another electron. But, we still need to find things and we still want to gather related things together. According to Svenonius, "Organizing information if it means nothing else means bringing all the same information together" [3]. This sentiment and need remains unchanged whether we are talking about books, Web documents, chemical elements or linked data on the Web.

Like words or terms in human language that help us communicate about things, how we organize things on the Web needs to have an understood and definable meaning, hopefully bounded with some degree of precision, that enables us to have some confidence we are really communicating about the same something with one another. However, when applied to the Web and machine communications, the demands for how these definitions and precisions apply need to change. This makes the notion of a Web basis for organization both easier and harder than traditional approaches to classification.

It is easier because everything is virtual: we can apply multiple classification schema and can change those schema at will. We are not locked into historical anomalies like huge subject areas reserved for arcane or now historically less important topics, such as the Boer Wars or phrenology. We need not move physical books around on shelves in order to accommodate new or expanded classification schemes. We can add new branches to our classification of, say, nanotechnology as rapidly as the science advances.

Yet it is harder because we can no longer rely on the understanding of human language as a basis for naming and classifying things. Actually, of course, language has always been ambiguous, but it is manifestly more so when put through the grinder of machine processing and understanding. Machine processing of related information adds the new hurdles of no longer being able to rely on text labels ("names") alone as the identifier of things and requires we be more explicit about our concept relationships and connections. Fortunately, here, too, much has been done in helping to organize human language through such lexical frameworks as WordNet and similar.

The Idea and Role of Reference Concepts

Many groups and individuals have been grappling with these questions of how to organize and describe information to aid interoperability in an Internet context. Among many, let me simply mention two because of the diversity their approaches show.

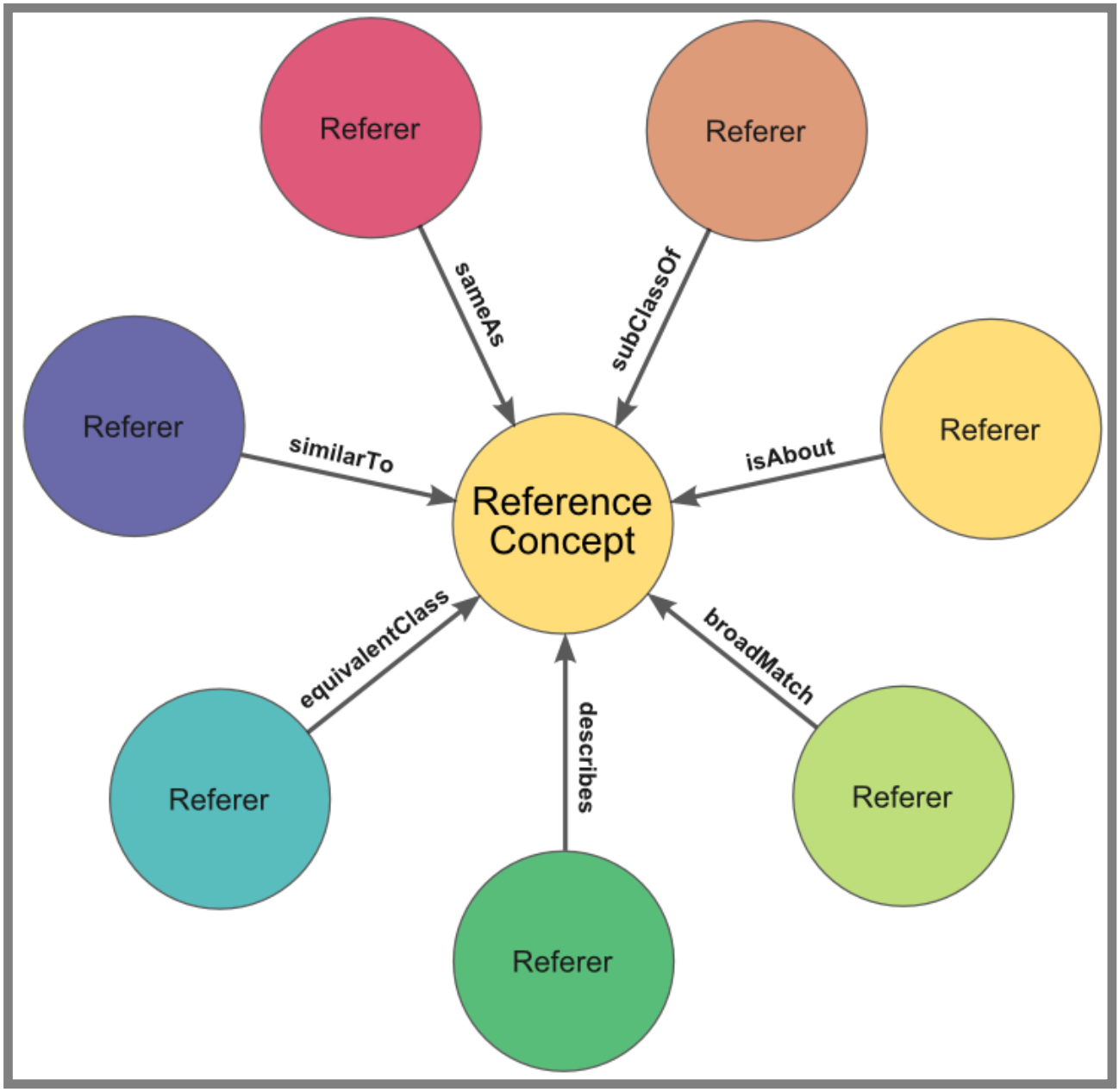
Bernard Vatant, for one, has with his colleagues been an advocate for some time for the need for what he calls "hubjects." With an intellectual legacy from the Topic Maps community, the idea of "hubjects" is to have a flat space of reference subjects to which related information can link and refer. Each hubject is the hub of a spoked wheel of representations by which the same subject matter from different contexts may be linked. The idea of the flat space or neutrality in the system is to place the "hubject" identifier (referent) outside of other systems that attempt to organize and provide "meta-frameworks" of knowledge organization. In other words, there are no inherent suggested relationships in the reference "hubject" structure: just a large bin of defined subjects to which external systems may link.

A different and more formalized approach has been put forward by the FRSAD working group [4], dealing with subject authority data. Subject authority data is the type of classificatory information that deals with the subjects of various works, such as their concepts, objects, events, or places. As the group stated, the scope of this effort pertains to the "aboutness" of various conceptual works. The framework for this effort, as with the broader FRBR effort, are new standards and approaches appropriate to classifying electronic bibliographic records.

Besides one of the better summaries and introductions to the general problems of subject classification in general, the FRSAD approach makes its main contribution in clearly distinguishing the *idea of something* (which it calls a *thema*, or entity used as the subject of a work) from the name or label of something (which it calls *nomen*). For many in the logic community, steeped in the Peirce triad of *sign-object-interpretant* [5], this distinction seems rather obvious and straightforward. But, in library science, labels have been used interchangeably as identifiers, and making this distinction clean is a real contribution. The FRSAD effort does not itself really address how the thema are actually found or organized.

The notion of a *reference concept* used herein combines elements from both of these approaches. A *reference concept* is *the idea of something*, or a *thema* in the FRSAD sense. It is also a reference hub of sorts, similar to the idea of a "hubject". But it is also much more and more fully defined.

So, let's first begin by representing a reference concept in relation to its referers and possible linking predicates as follows:



A referer needs to link appropriately to its reference concept, with some illustrative examples shown on the arrows in the diagram. These links are the predicates, ranging from the exact to the approximate, discussed in the first semantic "gap" posting. (**Note:** see that earlier post for a longer listing of existing, candidate linking predicates. No further comment is made in this present article as to whether those in that earlier posting or the example ones above are "correct" or not; see the first post for that discussion.)

If properly constructed and used, a reference concept thus becomes a fixed point in an information space. As one or more external sources link to these fixed points, it is then possible to gather similar content together and to begin to organize the information space, in the sense of Svenonius. Further, and this is a key difference from the "hubject" approach, if the reference concept is itself part of a coherent structure, then additional value can be derived from these assignments, such as inference, consistence testing, and alignments. (More on this latter point is discussed below.)

Seven Guidelines for a Reference Concept

If the right factors are present, it should be possible to relate and interoperate multiple datasets and knowledge representations. If present, these factors can result in a series of fixed reference points to which external information can be linked. In turn, these reference nodes can form constellations to guide the traversal to desired information destinations on the Web.

Let's look at the seven factors as to what constitutes guidelines for best practices.

Guideline #1: Persistent URI

By definition, a Web-based reference concept should adhere to linked data principles and should have a URI as its address and identifier. Also, by definition as a "reference", the vocabulary or ontology in which the concept is a member should be given a permanent and persistent address. Steps should be taken to ensure 24x7 access to the reference concept's URI, since external sources will be depending on it.

As a general rule, the concepts should also be stated as single nouns and use CamelCase notation (that is, class names should start with a capital letter and not contain any spaces, such as `MyNewConcept`).

Guideline #2: Preferred Label

Provide a preferred label annotation property that is used for human readable purposes and in user interfaces. For this purpose, a construct such as the SKOS property of `skos:prefLabel` works well. Note, this label is ***not*** the basis for deciding and making linkages, but it is essential for mouseovers, tooltips, interface labels, and other human use factors.

Guideline #3: Definition

Give all concepts and properties a definition. The matching and alignment of things is done on the basis of concepts (not simply labels), which means each concept must be defined [6]. Providing clear definitions (along with the coherency of its structure) gives an ontology its semantics. Remember not to confuse the label for a concept with its meaning. For this purpose, a property such as `skos:definition` works well, though others such as `rdfs:comment` or `dc:description` are also commonly used.

The definition is the most critical guideline for setting the concept's meaning. Adequate text and content also aid semantic alignment or matching tasks.

Guideline #4: Tagset

Include explicit consideration for the idea of a "*semset*" or "tagset", which means a series of alternate labels and terms to describe the concept. These alternatives include true synonyms, but may also be more expansive and include jargon, slang, acronyms or alternative terms that usage suggests refers to the same concept. The *semset* construct is similar to the "synsets" in Wordnet, but with a broader use understanding. Included in the *semset* construct is the single (per language) preferred (human-readable)

label for the concept, the `prefLabel`, an embracing listing of alternative phrase and terms for the concept (including acronyms, synonyms, and matching jargon), the `altLabels`, and a listing of prominent or common misspellings for the concept or its alternatives, the `hiddenLabels`.

This tagset is an essential basis for tagging unstructured text documents with reference concepts, and for search not limited to keywords. The tagset, in combination with the definition, is also the basis for feeding many NLP-driven methods for concept or ontology alignment.

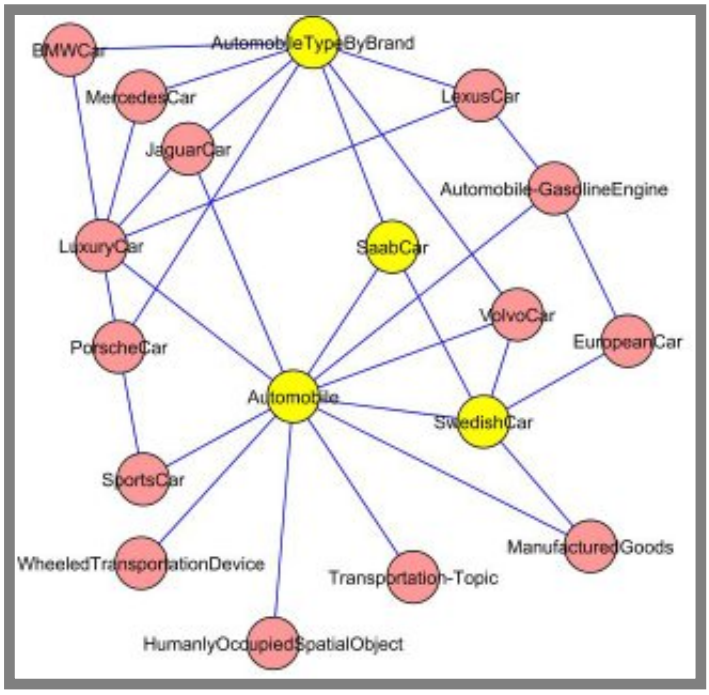
Guideline #5: Language Independent

The practice of using an identifier separate from label, and language qualified entries for definition, preferred label and tagset (alternative labels) means that multi-lingual versions can be prepared for each concept. Though this is a somewhat complicated best practice in its own right (for example, being attentive to the `xml:lang="en"` tag for English), adhering to this practice provides language independence for reference concepts.

Sources such as Wikipedia, with its richness of concepts and multiple language versions, can then be a basis for creation of alternative language versions.

Guideline #6: Range and Domain

Use of domains and ranges assists testing, helps in disambiguation, and helps in external concept alignments. Domains apply to the subject (the left hand side of a triple); ranges to the object (the right hand side of the triple). Domains and ranges should not be understood as actual constraints, but as axioms to be used by reasoners. In general, domain for a property is the range for its inverse and the range for a property is the domain of its inverse.



Guideline #7: Part of Coherent Structure

When reference concepts, properly constructed as above, are also themselves part of a coherent structure, further benefits may be gained. These benefits include inferencing, consistency testing, discovery and navigation. For example, the sample at right shows that a retrieval for Saab cars can also inform that these are automobiles, a brand of automobile, and a Swedish kind of car.

To gain these advantages, the coherent structure need not be complicated. RDFS and SKOS-based lightweight vocabularies can meet this test. Properly constructed OWL ontologies can also provide these benefits.

When best practices are combined with being part of a coherent structure, we can refer to these structures as *reference ontologies* or *domain ontologies*.

The State of Reference Concepts

In part, these best practices are met to a greater or lesser extent by many current vocabularies. But few provide complete coverage, and across a broad swath of domain needs, major gaps remain. This unfortunate observation applies to upper-level ontologies, reference vocabularies, and domain ontologies alike.

Upper-level ontologies include the Suggested Upper Merged Ontology (SUMO), the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), PROTON, Cyc and BFO (Basic Formal Ontology). While these have a coherency of construction, they are most often incomplete with respect to reference concept construction. With the exception of SUMO and Cyc, domain coverage is also very general.

Our own UMBEL reference ontology [7] is closest to meeting all criteria. The reference concepts are constructed to standard. But coverage is fairly general, and not directly applicable to most domains (though it can help to orient specific vocabularies).

Wikipedia, as accessed via the DBpedia expression, has good persistent URIs, labels, altLabels and proxy definitions (via the first sentences abstract). As a repository of reference concepts, it is extremely rich. But the organizational structure is weak and provides very few of the benefits for coherent structures noted above.

Going back to 1997, DCMI has been involved in putting forward possible vocabularies that may act as "qualifiers" to `dc:subject` [8]. Such reference vocabularies can extend from the global or comprehensive, such as the Universal Decimal Classification or Library of Congress Subject Headings, to the domain specific such as MeSH in medicine or Agrovoc in agriculture [9]. One or more concepts in such reference vocabularies can be the object of a `dc:subject` assertion, for example. While these vocabularies are also a rich source of reference concepts, they are not constructed to standards and at most provide hierarchical structures.

In the area of domain vocabularies, we are seeing some good pockets of practice, especially in the biomedical and life sciences arena [10]. Promising initiatives are also underway in library applications [11] and perhaps other areas unknown to the author.

In summary, I take the state of the art to be quite promising. We know what to do, and it is being done in some pockets. What is needed now is to more broadly operationalize these practices and to extend them across more domains. If we can bring attention to and publicize exemplar vocabularies, we can start to realize the benefits of actual data interoperability on the Web.

---

[1] See M. K. Bergman, 2010. "The Nature of Connectedness on the Web," *AI3:::Adaptive Information* blog, November 22, 2010. See http://www.mkbergman.com/935/the-nature-of-connectedness-on-the-web/.

[2] See M. K. Bergman, 2006. "Sources and Classification of Semantic Heterogeneities," *AI3:::Adaptive Information* blog, June 6, 2006. See http://www.mkbergman.com/232/sources-and-classification-of-semantic-heterogeneities/.

[3] From a quote on page 10 by Elaine Svenonius, 2000. *The Intellectual Foundation of Information Organization*, MIT Press, 2000, 255pp. I'd like to thank Karen Coyle for recently posting this quote on the Linked Library Data (LLD) mailing list.

[4] Marcia Lei Zeng, Maja Žumer, Athena Salaba, eds., 2010. *Functional Requirements for Subject Authority Data (FRSAD): A Conceptual Model*, prepared by the IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR), June 2010, 75 pp. See http://www.ifla.org/files/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf. This effort is part of the broader and well-known FRBR (Functional Requirements of Bibliographic Records) initiative.

[5] C.S. Peirce's sign relations are covered under the discussion about Semiotic Elements under the Sign section on Peirce in Wikipedia. In the the context of this discussion, the *sign* corresponds to any of the labels or identifiers associated with the (reference concept) *object*, the meaning of which is provided by its *interpretant* definition and useful language labels. See also John Sowa, 2000. "Ontology, Metadata, and Semiotics," presented at ICCS'2000 in Darmstadt, Germany, on August 14, 2000; see http://www.jfsowa.com/ontology/ontometa.htm.

[6] As another commentary on the importance of definitions, see http://ontologyblog.blogspot.com/2010/09/physician-decries-lack-of-definitions.html.

[7] 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. It is also a general vocabulary of classes and predicates designed for the creation of domain-specific ontologies.

[8] Rebecca Guenther, 1997. *Dublin Core Qualifiers/Substructure*, October 15, 1997. See http://www.loc.gov/marc/dcqualif.html.

[9] Two, among many, metadata listings of potential reference vocabularies are http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/controlling-your-language-links-to-metadata-vocabularies/ and http://hilt.cdlr.strath.ac.uk/hilt2web/Sources/thesauri.html.

[10] For example, see the Open Biological and Biomedical Ontologies (OBO) initiative and the W3C's Semantic Web Health Care and Life Sciences Interest Group.

[11] See the W3C's Linked Library Data initiative, with particular attention to topics and use cases.

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