In Search of 'Gold Standards' for the Semantic Web

by Mike Bergman - Monday, February 28, 2011

http://www.mkbergman.com/947/in-search-of-gold-standards-for-the-semantic-web/

Wikipedia + UMBEL + Friends May Offer One Approach

In the first part of this series we argued for the importance of *reference structures* to provide the structures and vocabularies to guide interoperability on the semantic Web. The argument was made that these reference structures are akin to human languages, requiring a sufficient richness and terminology to enable nuanced and meaningful communications of data across the Web and within the context of their applicable domains.

While the idea of such reference structures is great -- and perhaps even intuitive when likened to human languages -- the question is begged as to what is the basis for such structures? Just as in human languages we have dictionaries, thesauri, grammar and style books or encyclopedia, what are the analogous reference sources for the semantic Web?

In this piece, we tackle these questions from the perspective of the entire Web. Similar challenges and approaches occur, of course, for virtually every domain and specific community. But, by focusing on the entirety of the Web, perhaps we can discern the grain of sand at the center of the pearl.

Bootstrapping the Semantic Web

The idea of bootstrapping is common in computers, compilers or programming. Every computer action needs to start from a basic set of instructions from which further instructions or actions are derived. Even starting up a computer ("booting up") reflects this bootstrapping basis. Bootstrapping is the answer to the classic chicken-or-egg dilemma by embedding a starting set of instructions that provides the premise at start up [1]. The embedded operand for simple addition, for example, is the basis for building up more complete mathematical operations.

So, what is the grain of sand at the core of the semantic Web that enables it to bootstrap meaning? We start with the basic semantics and "instructions" in the core RDF, RDFS and OWL languages. These are very much akin to the basic BIOS instructions for computer boot up or the instruction sets leveraged by compilers. But, where do we go from there? What is the analog to the compiler or the operating system that gives us more than these simple start up instructions? In a semantics sense, what are the vocabularies or languages that enable us to understand more things, connect more things, relate more things?

To date, the semantic Web has given us perhaps a few dozen commonly used vocabularies, most of which are quite limited and simple pidgin languages such as DC, FOAF, SKOS, SIOC, BIBO, etc. We also have an emerging catalog of "things" and concepts from Wikipedia (via DBpedia) and similar. (Recall, in this piece, we are trying to look Web-wide, so the many fine building blocks for domain purposes such as found in biology, medicine, finance, astronomy, etc., are excluded.) The purposes and scope of these vocabularies widely differ and attack quite different slices of the information space. SKOS, for example, deals with describing simple knowledge structures like taxonomies or thesauri; SIOC is for describing social media.

By virtue of adoption, each of these core languages has proved its usefulness and role. But, as skew lines in space, how do these vocabularies relate to one another? And, how can all of the specific domain vocabularies also relate to those and one another where there are points of intersection or overlap? In short, after we get beyond the starting instructions for the semantic Web, what is our language and vocabulary? How do we complete the bootstrap process?

Clearly, like human languages, we need rich enough vocabularies to describe the things in our world and a structure of the relationships amongst those things to give our communications meaning and coherence. That is precisely the role provided by reference structures.

The Use and Role of 'Gold Standards'

To prevent reference structures from being rubber rulers, some fixity or *grounding* needs to establish the common understanding for its referents. Such fixed references are often called 'gold standards'. In money, of course, this used to be a fixed weight of gold, until that basis was abandoned in the 1970s. In the metric system, there are a variety of fixed weights and measures that are employed. In the English language, the Oxford English Dictionary (OED) is the accepted basis for the lexicon. And so on.

Yet, as these examples show, none of these gold standards is absolute. Money now floats; multiple systems of measurement compete; a variety of dictionaries are used for English; most languages have their own reference sets; etc. The key point in all gold standards, however, is that there is wide acceptance for a defined reference for determining alignments and arbitrating differences.

Gold standards or reference standards play the role of referees or arbiters. What is the meaning of this? What is the definition of that? How can we tell the difference between this and that? What is the common way to refer to some thing?

Let's provide one example in a semantic Web context. Let's say we have a dataset and its schema A that we are aligning with another dataset with schema B. If I say two concepts align exactly across these datasets and you say differently, how do we resolve this difference? On one extreme, each of us can say our own interpretation is correct, and to heck with the other. On the other extreme, we can say both interpretations are correct, in which case both assertions are meaningless. Perhaps papering over these extremes is OK when only two competing views are in play, but what happens when real problems with many actors are at stake? Shall we propose majority rule, chaos, or the strongest prevails?

These same types of questions have governed human interaction from time immemorial. One of the reasons to liken the problem of operability on the semantic Web to human languages, as argued in Part I, is to seek lessons and guidance for how our languages have evolved. The importance of finding common *ground* in our syntax and vocabularies -- and, also, critically, in how we accept changes to those -- is the basis for communication. Each of these understandings needs to be codified and documented so that they can be referenced, and so that we can have some confidence of what the heck it is we are trying to convey.

For reference structures to play their role in plugging this gap -- that is, to be much more than rubber rulers -- they need to have such grounding. Naturally, these groundings may themselves change with new information or learning inherent to the process of human understanding, but they still should retain their character as references. Grounded references for these things -- 'gold standards' -- are key to this consensual process of communicating (interoperating).

Some 'Gold Standards' for the Semantic Web

The need for gold standards for the semantic Web is particularly acute. First, by definition, the scope of the semantic Web is all things and all concepts and all entities. Second, because it embraces human knowledge, it also embraces all human languages with the nuances and varieties thereof. There is an immense gulf in referenceability from the starting languages of the semantic Web in RDF, RDFS and OWL to this full scope. This gulf is chiefly one of vocabulary (or lack thereof). We know how to construct our grammars, but we have few words with understood relationships between them to put in the slots.

The types of gold standards useful to the semantic Web are similar to those useful to our analogy of human languages. We need guidance on structure (syntax and grammar), plus reference vocabularies that encompass the scope of the semantic Web (that is, everything). Like human languages, the vocabulary references should have analogs to dictionaries, thesauri and encyclopedias. We want our references to deal with the specific demands of the semantic Web in capturing the lexical basis of human languages and the connectedness (or not) of things. We also want bases by which all of this information can be related to different human languages.

To capture these criteria, then, I submit we should consider a basic starting set of gold standards:

- RDF/RDFS/OWL -- the data model and basic building blocks for the languages
- Wikipedia -- the standard reference vocabulary of things, concepts and entities, plus other structural guidances
- WordNet -- lexical language references as an aid to natural language processing, and
- UMBEL -- the structural reference for the connectedness of things for basic coherence and inference, plus a vocabulary for mapping amongst reference structures and things.

Each of these potential gold standards is next discussed in turn. The majority of discussion centers on

Wikipedia and UMBEL.

RDF/RDFS/OWL: The Language

Naturally, the first suggested gold standard for the semantic Web are the RDF/RDFS/OWL language components. Other writings have covered their uses and roles [2]. In relation to their use as a gold standard, two documents, one on RDF semantics [3] and the other an OWL [4] primer, are two great starting points. Since these languages are now in place and are accepted bases of the semantic Web, we will concentrate on the remaining members of the standard reference set.

Wikipedia: The Vocabulary (and More)

The second suggested gold standard for the semantic Web is Wikipedia, principally as a sort of canonical vocabulary base or lexicon, but also for some structural aspects. Wikipedia now contains about 3.5 million English articles, by far larger than any other knowledge base, and has more than 250 language versions. Each Wikipedia article acts as more or less a reference for the thing it represents. In addition, the size, scope and structure of Wikipedia make it an unprecedented resource for researchers engaged in natural language processing (NLP), information extraction (IE) and semantic Web-related tasks.

For some time I have been maintaining a listing called SWEETpedia of academic and research articles focused on the use of Wikipedia for these tasks. The latest version tracks some 250 articles [5], which I guess to be about one half or more of all such research extant. This research shows a broad variety of potential roles and contributions from Wikipedia as a gold standard for the semantic Web, some of which is detailed in the tables below.

An excellent report by Olena Medelyan *et al.* from the University of Waikato in New Zealand, Mining Meaning from Wikipedia, organized this research up through 2008 and provided detailed commentary and analysis of the role of Wikipedia [6]. They noted, for example, that Wikipedia has potential use as an encyclopedia (its intended use), a corpus for testing and modeling NLP tasks, as a thesaurus, a database, an ontology or a network structure. The Intelligent Wikipedia project from the University of Washington has also done much innovative work on "automatically learned systems [that] can render much of Wikipedia into high-quality semantic data, which provides a solid base to bootstrap toward the general Web" [7].

However, as we proceed through the next discussions, we'll see that the weakest aspect of Wikipedia is its category structure. Thus, while Wikipedia is unparalleled as the gold standard for a reference vocabulary for the Web, and has other structural uses as well, we will need to look elsewhere for how that content is organized.

Major Wikipedia Initiatives

Many groups have recognized these advantages for Wikipedia, and have built knowledge bases around it. Also, many of these groups have also recognized the category (schema) weaknesses in Wikipedia and have proposed alternatives. Some of these major initiatives, which also collectively represent a large number of the research articles in SWEETpedia, include:

Project	Schema Basis	Comments
DBpedia	Wikipedia Infoboxes	excellent source for URI identifiers; structure extraction basis used by many other projects
Freebase	User Generated	schema are for domains based on types and properties; at one time had a key dependence on Wikipedia; has since grown much from user-generated data and structure; now owned by Google
Intelligent Wikipedia	Wikipedia Infoboxes	a broad program and a general set of extractors for obtaining structure and relationships from Wikipedia; was formerly known as KOG; from Univ of Washington
SIGWP	Wikipedia Ontology	the Special Interest Group of Wikipedia (Research or Mining); a general group doing research on Wikipedia structure and mining; schema basis is mostly from a thesaurus; group has not published in two years
UMBEL	UMBEL Reference Concepts	RefConcepts based on the Cyc knowledge base; provides a tested, coherent concept schema, but one with gaps regarding Wikipedia content; has 28,000 concepts mapped to Wikipedia
WikiNet	Extracted Wikipedia Ontology	part of a long-standing structure extraction effort from Wikipedia leading to an ontology; formerly known as WikiRelate; from the Heidelberg Institute for Theoretical Studies (HITS)
Wikipedia Miner	N/A	generalized structure extractor; part of a wider basis of Wikipedia research at the Univ of Waikato in New Zealand
Wikitology	Wikipedia Ontology	general RDF and ontology-oriented project utilizing Wikipedia; effort now concluded;

		from the Ebiquity Group at the Univ of Maryland
YAGO	WordNet	maps Wordnet to Wikipedia, with structured extraction of relations for characterizing entities

It is interesting to note that none of the efforts above uses the Wikipedia category structure "as is" for its schema.

Structural Sources within Wikipedia

The surface view of Wikipedia is topic articles placed into one or more categories. Some of these pages also include structured data tables (or templates) for the kind of thing the article is; these are called *infoboxes*. An infobox is a fixed-format table placed at the top right of articles to consistently present a summary of some unifying aspect that the articles share. For example, see the listing for my home town, Iowa City, which has a *city* infobox.

However, this cursory look at Wikipedia in fact masks much additional and valuable structure. Some early researchers noted this [8]. The recognition of structure has also been a key driver for the interest in Wikipedia as a knowledge base (in addition to its global content scope). The following table is a fairly complete listing of structure possibilities within Wikipedia (see Endnotes for any notes):

Wikipedia Structure	Potential Applications	Note
Corpus		
Entire Corpus	knowledge base; graph structure; corpus for n-grams, other constructions	[9]
Categories		
Category	category suggestion; semantic relatedness; query expansion; potential parent category	
Contained Articles	semantically-related terms (siblings)	
Hierarchy	hyponymic and meronymic relations between terms	
Listing Pages/Categories	semantically-related terms (siblings)	
Patterned Categories	functional metadata	[9]
Infobox Templates		

Attributes	synonyms; key-value pairs	
Values	units of measure; fact extraction	[9]
Items	category suggestion; entity suggestion	
Geolocational	coordinates; places; geolocational; (may also appear in full article text)	
Issue Templates		
Multiple Types	exclusion candidates; other structural analysis; examples include Stub, Message Boxes, Multiple Issues	[9]
Category Templates		[13]
Category Name	disambiguation; relatedness	
Category Links	semantic relatedness	
Articles		
First Paragraph	definition; abstract	
Full Text	complete discussion; related terms; context; translations; NLP analysis basis; relationships; sentiment	
Redirects	synonymy; spelling variations, misspellings; abbreviations; query expansion	
Title	named entities; domain specific terms or senses	
Subject	category suggestion (phrase marked in bold in first paragraph)	
Section Heading(s)	category suggestion; semantic relatedness	[9]
See Also	related concepts; query expansion	[9]
Further Reading	related concepts	[9,10]
External Links	related concepts; external harvest points	
Article Links		
Context	related terms; co-occurrences	

Label	synonyms; spelling variations; related terms; query expansion	
Target	link graph; related terms	
LinksTo	category suggestion; functional metadata	
LinkedFrom	category suggestion; functional metadata	
References		
Citations	external harvest points	[9,10]
Media		
Images	thumbnails; image recognition for disambiguation; controversy (edit/upload frequency)	[11]
Captions	related concepts; related terms; functional metadata	[9]
Disambiguation Pages		
Article Links	sense inventory	
Discussion Pages		
Discussion Content	controversy	
Redux for Article Structure	see Articles for uses	
History Pages		
Edit Frequency	topicalness; controversy (diversity of editors, reversions)	
Edit Basis	lexical errors	[9]
Lists		
Hyponyms	instances; named entity candidates	
Alternate Language Versions		
Redux for All Structures	see all items above; translation; multilingual alignment; entity disambiguation	[12]

The potential for Wikipedia to provide structural understandings is evident from this table. However, it should be noted that, aside from some stray research initiatives, most effort to date has focused on the major initiatives noted earlier or from analyzing linking and infoboxes. There is much additional research that could be powered by the Wikipedia structure as it presently exists.

From the standpoint of the broader semantic Web, the potential of Wikipedia in the areas of metadata enhancement and mapping to multiple human languages [12] are particularly strong. We are only now at the very beginning phases of tapping this potential.

Structural Weaknesses

The three main weaknesses with Wikipedia are its category structure [14], inconsistencies and incompleteness. The first weakness means Wikipedia is not a suitable organizational basis for the semantic Web; the next two weaknesses, due to the nature of Wikipedia's user-generated content, are constantly improving.

Our recent effort to map between UMBEL and Wikipedia, undertaken as part of the recent UMBEL *v* 1.00 release, spent considerable time analyzing the Wikipedia category structure [15]. Of the roughly half million categories in Wikipedia, only about 85,000 were found to be suitable candidates to participate in an actual schema structure. Further breakdowns are shown by this table resulting from our analysis:

Wikipedia Category Breakdowns		
Removals		**20.7%**
Administrative	15.7%	
Misc Cleaning	5.0%	
Functional (not schema)	**61.8%**	
Fn Dates	10.1%	
Fn Nationalities	9.6%	
Fn Listings, related	0.8%	
Fn Occupations	1.0%	
Fn Prepositions	40.4%	
Candidates	**17.4%**	
SuperTypes	1.7%	

General Structure	15.7%
TOTAL	**100.0%**

Fully 1/5 of the categories are administrative or internal in nature. The large majority of categories are, in fact, not structural at all, but what we term *functional categories*, which means the category contains faceting information (such as subclassifying *musicians* into *British musicians*) [16]. Functional categories can be a rich source of supplementary metadata for its assigned articles -- though, no one has yet processed Wikipedia in this manner -- but are not a useful basis for structural conceptual relationships or inferencing.

This weakness in the Wikipedia category system has been known for some time [17], but researchers and others still attempt to do mappings on mostly uncleaned categories. Though most researchers recognize and remove internal or administrative categories in their efforts, using the indiscriminate remainder of categories still leads to poor precision in resulting mappings. In fact, in comparison to one of the more rigorous assessments to date [18], our analysis still showed a 6.8% error rate in hand inspected categories.

Other notable category problems include circular references, skipped intermediate categories, misassigned categories and incomplete assignments.

Nonetheless, Wikipedia categories do have a valuable use in the analysis of local relationships (one degree of relatedness) and for finding missing category candidates. And, as noted, the functional categories are also a rich and untapped source of additional article metadata.

Like any knowledge base, Wikipedia also has inconsistent and incomplete coverage of topics [19]. However, as more communities accept Wikipedia as a central resource deserving completeness, we should see these gaps continue to get filled.

The DBpedia Implementation

One of the first database versions of Wikipedia built for semantic Web purposes is DBpedia. DBpedia has an incipient ontology useful for some classification purposes. Its major structural organization is built around the Wikipedia infoboxes, which are applied to about a third of Wikipedia articles. DBpedia also has multiple language versions.

DBpedia is a core hub of Linked Open Data (LOD), which now has about 300 linked datasets; has canonical URIs used by many other applications; has extracted versions and tools very useful for further processing; and has recently moved to incorporate live updates from the source Wikipedia [20]. For these reasons, the DBpedia version of Wikipedia is the suggested implementation version.

WordNet: Language Relationships

The third suggested gold standard for the Semantic Web is WordNet, a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. There are over 50 languages covered by wordnet approaches, most mapped to this English WordNet [21].

Though it has been used in many ontologies [22], WordNet is most often mapped for its natural language purposes and not used as a structure of conceptual relationships *per se*. This is because it is designed for words and not concepts. It contains hundreds of basic semantic inconsistencies and also lacks much domain applicability. Entities, of course, are also lacking. In those cases where WordNet has been embraced as a schema basis, much work is generally expended to transform it into an ontology suitable for knowledge representation.

Nonetheless, for word sense disambiguation and other natural language processing tasks, as well as for aiding multi-lingual mappings, WordNet and its various other language variants is a language reference gold standard.

UMBEL: A Coherent Structure

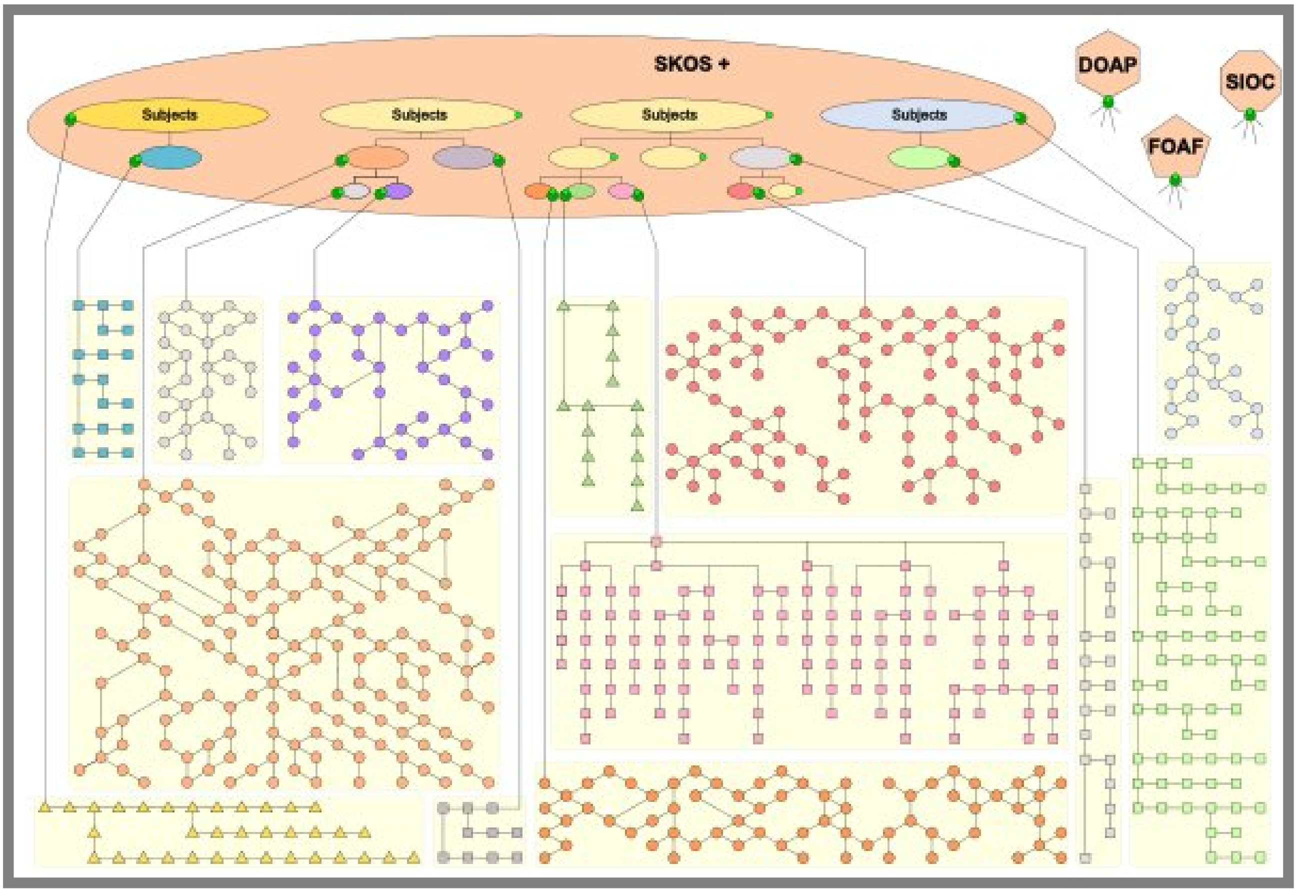
So, with these prior gold standards we gain a basic language and grammar; a base (canonical) vocabulary and some structure guidance; and a reference means for processing and extracting information from input text. Yet two needed standards remain.

One needed standard is a conceptual organizing structure (or schema) by which the canonical vocabulary of concepts and instances can be related. This core structure should be constructed in a coherent [23] manner and expressly designed to support inferencing and (some) reasoning. This core structure should be sufficiently large to embrace the scope of the semantic Web, but not so detailed as to make it computationally inefficient. Thus, the core structure should be a framework that allows more focused and purposeful vocabularies to be "plugged in", depending on the domain and task at hand. Unfortunately, the candidate category structures from our other gold standards in Wikipedia and WordNet do not meet these criteria.

A second needed standard is a bit of additional vocabulary "glue" specifically designed for the purposes of the semantic Web and ontology and domain incorporation. We have multiple and disparate world views and contexts, as well as the things described by them [24]. To get them to interoperate -- and to acknowledge differences in alignment or context -- we need a set of relational predicates (vocabulary) that can capture a range of mappings from the exact to the approximate [25]. Unlike other reference vocabularies that attempt to capture canonical definitions within defined domains, this vocabulary is expressly required by the semantic Web and its goal to federate different data and schema.

UMBEL has been expressly designed to address both of these two main needs [26]. UMBEL is a coherent categorization structure for the semantic Web and a mapping vocabulary designed for dataset and conceptual interoperability. UMBEL's 28,000 reference concepts (RefConcepts) are based on the Cyc knowledge base [27], which itself is expressly designed as a common sense representation of the world with express variations in context supported via its 1000 or so microtheories. Cyc, and UMBEL upon which it is based, are by no means the "correct" or "only" representations of the world, but they are coherent ones and thus internally consistent.

UMBEL's role to allow datasets to be "plugged in" and related through some fixed referents was expressed by this early diagram [28]:



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

The idea -- which is still central to this kind of reference structure -- is that a set of reference concepts can be used by multiple datasets to connect and then inter-relate. These are shown by the nested subjects (concepts) in the umbrella structure.

UMBEL, of course, is not the only coherent structure for such interoperability purposes. Other major vocabularies (such as LCSH; see below) or upper-level ontologies (such as SUMO, DOLCE, BFO or PROTON, etc.) can fulfill portions of these roles, as well. In fact, the ultimate desire is for multiple reference structures to emerge that are mapped to one another, similar to how human languages can inter-relate. Yet, even in that desired vision, there is still a need for a bootstrapped grounding. UMBEL is the first such structure expressly designed for the two needed standards.

Mappings to the Other Standards

UMBEL is already based on the central semantic Web languages of RDF, RDFS, SKOS, and OWL 2.

The recent version 1.00 now maps 60% of UMBEL to Wikipedia, with efforts for the remaining in process. UMBEL provides mappings to WordNet, via its Cyc relationships. More of this is in process and will be exposed. And the mappings between UMBEL and GeoNames [29] for locational purposes is also nearly complete.

The Gold Resides in Combining These Standards

Each of these reference structures -- RDF/OWL, Wikipedia, WordNet, UMBEL -- is itself coherent and recognized or used by multiple parties for potential reference purposes on the semantic Web. The advocacy of them as standards is hardly radical.

However, the gold lies in the combination of these components. It is in this combination that we can see a grounded knowledge base emerge that is sufficient for bootstrapping the semantic Web.

The challenge in creating this reference knowledge base is in the mapping between the components. Fortunately, all of the components are already available in RDF/OWL. WordNet already has significant mappings to Wikipedia and UMBEL. And 60% of UMBEL is already mapped to Wikipedia. The remaining steps for completing these mappings are very near at hand. Other vocabularies, such as GeoNames [29], would also beneficially contribute to such a reference base.

Yet to truly achieve a role as a gold standard, these mappings should be fully vetted and accurate. Automated techniques that embed errors are unacceptable. Gold standards should not themselves be a source for propagation of errors. Like dictionaries or thesauri, we need reference structures that are quality and deserving of reference. We need canonical structures and canonical vocabularies.

But, once done, these gold standards themselves become reference sources that can aid automatic and semi-automatic mappings of other vocabularies and structures. Thus, the real payoff is not that these gold standards themselves get actually embedded in specific domain uses or whatever, but that they can act as reference referees for helping align and *ground* other structures.

Like the bootstrap condition, more and more reference structures may be brought into this system. A reference structure does not mean reliance; it need not even have more than minimal use. As new structures and vocabularies are brought into the mix, appropriate to specific domains or purposes, reference to other grounding structures will enable the structures and vocabularies to continue to expand. So, not only are reference concepts necessary for grounding the semantic Web, but we also need to pick good mapping predicates for properly linking these structures together.

In this manner, many alternative vocabularies can be bootstrapped and mapped and then used as the dominant vocabularies for specific purposes. For example, at the level of general knowledge categorization, vocabularies such as LCSH, the Dewey Decimal Classification, UDC, etc., can be preferentially chosen. Other specific vocabularies are at the ready, with many already used for domain purposes. Once grounded, these various vocabularies can also interoperate.

Grounding in gold standards enables the freedom to switch vocabularies at will. Establishing fixed reference points via such gold standards will power a virtuous circle of more vocabularies, more mappings, and, ultimately, functional interoperability no matter the need, domain or world view.

This is the last of a two-part series on the importance and choice of *reference structures* (Part I) and *gold standards* (Part II) on the semantic Web.

[1] For example, according to the Wikipedia entry on Machine code, "A machine code instruction set may have all instructions of the same length, or it may have variable-length instructions. How the patterns are organized varies strongly with the particular architecture and often also with the type of instruction. Most instructions have one or more opcode fields which specifies the basic instruction type (such as arithmetic, logical, jump, etc) and the actual operation (such as add or compare) and other fields that may give the type of the operand(s), the addressing mode(s), the addressing offset(s) or index, or the actual value itself."

[2] See, for example, M.K. Bergman, 2009. "Advantages and Myths of RDF," *AI3:::Adaptive Information* blog, April 8, 2009; see http://www.mkbergman.com/483/advantages-and-myths-of-rdf/ and M.K. Bergman, 2010. "Ontology Tutorial Series," *AI3:::Adaptive Information* blog, September 27, 2010; see http://www.mkbergman.com/916/ontology-tutorial-series/.

[3] Patrick Hayes, ed., 2004. *RDF Semantics*, W3C Recommendation 10 February 2004. See http://www.w3.org/TR/rdf-mt/.

[4] Pascal Hitzler et al., eds., 2009. *OWL 2 Web Ontology Language Primer*, a W3C Recommendation, 27 October 2009; see http://www.w3.org/TR/owl2-primer/.

[5] See SWEETpedia from the *AI3:::Adaptive Information* blog, which currently lists about 250 articles and citations.

[6] Olena Medelyan, Catherine Legg, David Milne and Ian H. Witten, 2008. Mining Meaning from Wikipedia, Working Paper Series ISSN 1177-777X, Department of Computer Science, The University of Waikato (New Zealand), September 2008, 82 pp. See http://arxiv.org/ftp/arxiv/papers/0809/0809.4530.pdf. This paper and its findings is discussed more in M.K. Bergman, 2008. "Research Shows Natural Fit between Wikipedia and Semantic Web," *AI3:::Adaptive Information* blog, October 15, 2008; see http://www.mkbergman.com/460/research-shows-natural-fit-between-wikipedia-and-semantic-web/.

[7] For a comprehensive treatment, see Fei Wu, 2010. *Machine Reading: from Wikipedia to the Web*, a doctoral thesis to the Department of Computer Science, University of Washington, 154 pp; see http://ai.cs.washington.edu/www/media/papers/Wu-thesis-2010.pdf. To my knowledge, this paper also was the first to use the "bootstrapping" metaphor.

[8] Quite a few research papers have characterized various aspects of the Wikipedia structure. One of the first and most comprehensive was Torsten Zesch, Iryna Gurevych, Max Mühlhäuser, 2007b. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource, and the longer technical report. See http://www.ukp.tu-darmstadt.de/software/JWPL. Also, 2008. In *Proceedings of the Biannual Conference of the Society for Computational Linguistics and Language Technology*, pp. 213221. Also, for another early discussion, see Linyun Fu, Haofen Wang, Haiping Zhu, Huajie Zhang, Yang Wang and Yong Yu, 2007. Making More Wikipedians: Facilitating Semantics Reuse for Wikipedia Authoring. See http://data.semanticweb.org/pdfs/iswc-aswc/2007/ISWC2007_RT_Fu.pdf

[9] This structural basis in Wikipedia is largely untapped.

[10] Citations and references appear to be highly selective (biased) in Wikipedia; nonetheless, those available are useful seeding points for more suitable harvests.

[11] Images have been used a thumbnails and linked references to the articles they are hosted in, but have not been analyzed much for semantics or file names.

[12] There are a variety of efforts underway to use Wikipedia as a multi-language cross-reference based on its 250 language versions; search, for example, on "multiple language" in SWEETpedia. Both named entity and concept matches can be used to correlate in multiple languages. This is greatly aided by inter-language links.

[13] When present, these appear at the bottom of an article and have many related categories; see this one for the semantic Web.

[14] See further http://en.wikipedia.org/wiki/Wikipedia:Category and http://en.wikipedia.org/wiki/Wikipedia:Categorization_FAQ for a discussion of use and guidelines for Wikipedia categories.

[15] For the release notice, see http://umbel.org/content/finally-umbel-v-100. *Annex H* to the *UMBEL Specifications* provides a description of the mapping methodologies and results.

[16] *Functional categories* combine two or more facets in order to split or provide more structured characterization of a category. For example, Category:English cricketers of 1890 to 1918, has as its core concept the idea of a cricketer, a sports person. But, this is also further characterized by nationality and time period. Functional categories tend to have a A x B x C construct, with prepositions denoting the facets. From a proper characterization standpoint, the items in this category should be classified as a Person --> Sports Person --> Cricketer, with additional facets (metadata) of being English and having the period 1890 to 1981 assigned.

[17] See, for example, Massimo Poesio et al., 2008. *ELERFED: Final Report*, see http://www.cl.uni-heidelberg.de/~ponzetto/pubs/poesio07.pdf, wherein they state, "We discovered that in the meantime information about categories in Wikipedia had grown so much and become so unwieldy as to limit its usefulness." Additional criticisms of the category structure may be found in S. Chernov, T. Iofciu, W. Nejdl and X. Zhou, 2006. "Extracting Semantic Relationships between Wikipedia Categories," in *Proceedings of the 1st International Workshop: SemWiki'06—From Wiki to Semantics*., co-located with the 3rd Annual European Semantic Web Conference ESWC'06 in Budva, Montenegro, June 12, 2006; and L Muchnik, R. Itzhack, S. Solomon and Y. Louzon, 2007. "Self-emergence of Knowledge Trees: Extraction of the Wikipedia Hierarchies," in *Physical Review E* 76(1). Also, this blog post from Bob Bater at *KOnnect*, "Wikipedia's Approach to Categorization," September 22, 2008, provides useful comments on category issues; see http://iskouk.wordpress.com/2008/09/22/wikipedias-approach-to-categorization/.

[18] Olena Medelyan and Cathy Legg, 2008. Integrating Cyc and Wikipedia: Folksonomy Meets Rigorously Defined Common-Sense, in *Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI08 Conference*, Chicago, US. See http://www.cs.waikato.ac.nz/~olena/publications/Medelyan_Legg_Wikiai08.pdf.

[19] As two references among many, see A. Halavais and D. Lackaff, 2008. "An Analysis of Topical Coverage of Wikipedia," in *Journal of Computer-Mediated Communication* 13 (2): 429–440; and A. Kittur, E. H. Chi and B. Suh, 2009. "What's in Wikipedia? Mapping Topics and Conflict using Socially Annotated Category Structure," in *Proceedings of the 27th Annual CHI Conference on Human Factors in Computing Systems*, pp 4–9.

[20] See DBpedia.org, especially DBpedia reference.

[21] See http://www.globalwordnet.org/gwa/wordnet_table.htm for a listing of known wordnets by language.

[22] For example, see this listing in Wikipedia.

[23] M.K. Bergman, 2008. "When is Content Coherent?," *AI3:::Adaptive Information* blog, July 25, 2008; see http://www.mkbergman.com/450/when-is-content-coherent/.

[24] For a couple of useful references on this topic, first see this discussion regarding contexts (and the possible relation to Cyc microtheories): Ramanathan V. Guha, Rob McCool, and Richard Fikes, 2004. "Contexts for the Semantic Web," in Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, eds., *International Semantic Web Conference*, volume 3298 of Lecture Notes in Computer Science, pp. 32-46. Springer, 2004. See http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.58.2368&rep=rep1&type=pdf. For another discussion about local differences and contexts and the difficulty of reliance on "common" understandings, see: Krzysztof Janowicz, 2010. "The Role of Space and Time for Knowledge Organization on the Semantic Web," in *Semantic Web* 1: 25–32; see http://iospress.metapress.com/content/636610536x307213/fulltext.pdf.

[25] OWL already provides the exact predicates; see further M.K. Bergman, 2010. "The Nature of Connectedness on the Web," *AI3:::Adaptive Information* blog, November 22, 2010, 2008; see http://www.mkbergman.com/935/the-nature-of-connectedness-on-the-web/ and the UMBEL mapping predicates in this vocabulary listing.

[26] UMBEL is a reference of 28,000 concepts (classes and relationships) derived from the Cyc knowledge base. The reference concepts of UMBEL are mapped to Wikipedia, DBpedia ontology classes, GeoNames and PROTON. UMBEL is designed to facilitate the organization, linkage and presentation of heterogeneous datasets and information. It is meant to lower the time, effort and complexity of developing, maintaining and using ontologies, and aligning them to other content. See further the UMBEL Specifications (including Annexes A – H), Vocabulary and RefConcepts.

[27] Cyc is an artificial intelligence project that has assembled a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal to provide human-like reasoning. The OpenCyc version 3.0 contains nearly 200,000 terms and millions of relationship assertions. Started in 1984, by 2010 an estimated 1000 person years had been invested in its development.

[28] This image and more related to the general question of interoperability in relation to a reference structure is provided in M.K. Bergman, 2007, "Where are the Road Signs for the Structured Web?," *AI3:::Adaptive Information* blog, May 29, 2007; see http://www.mkbergman.com/375/where-are-the-road-signs-for-the-structured-web/.

[29] GeoNames is a geographical database available for free download under a Creative Commons Attribution license. It contains over 10 million geographical names and consists of 7.5 million unique features, of which 2.8 million are populated places. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes. Given the importance of locational information, GeoNames is a natural complement to the gold standards mentioned herein. See further its Web site, which also showcases a nifty browser of mappings to Wikipedia.