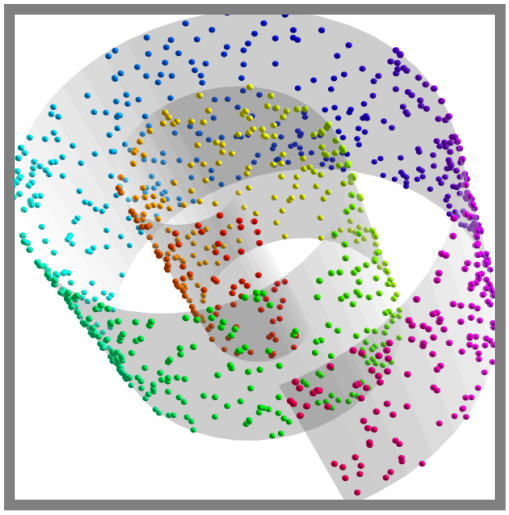
Shaping Wikipedia into a Computable Knowledge Base

by Mike Bergman - Tuesday, March 31, 2015

http://www.mkbergman.com/1847/shaping-wikipedia-into-a-computable-knowledge-base/



Part of a Vision for Information Interoperability on the Web

Wikipedia is arguably the most important information source yet invented for natural language processing (NLP) and artificial intelligence, in addition to its role as humanity's largest encyclopedia. Wikipedia is the principal information source for such prominent services as IBM's Watson [1], Freebase [2], the Google Knowledge Graph [3], Apple's Siri [4], YAGO [5], and DBpedia [6], the core reference structure for linked open data [7]. Wikipedia information has assumed a prominent role in NLP applications in word sense disambiguation, named entity recognition, co-reference resolution, and multi-lingual alignments; in information retrieval in query expansion, multi-lingual retrieval, question answering, entity ranking, text categorization, and topic indexing; and in semantic applications in topic extraction, relation extraction, entity extraction, entity typing, semantic relatedness, and ontology building [8].

The massive size of Wikipedia -- with more than 26 million articles across 250 different language versions [9,10] -- makes it a rich resource for reference entities and concepts. Structural features of Wikipedia that help to inform the understanding of relationships and connections include articles (and their embedded entities and concepts), article abstracts, article titles, infoboxes, redirects, internal and external links, editing histories, categories (in part), discussion pages, disambiguation pages, images and associated metadata, templates, tables, and pages of lists, not to mention Wikipedia as a whole being used as a corpus or graph [11]. It is no wonder that Wikipedia is referenced in about 665,000 academic articles and books [12]. And all of this occurs in a phenomenon that is not yet 15 years old!

Wikipedia is unparalleled as a resource for mining these resources of structure, concepts and entities. But, and here is the challenge, Wikipedia is never itself used as a computable knowledge base. It is a resource for other knowledge systems, but not a coherent knowledge base unto itself. Wikipedia feeds other useful knowledge bases, but does not play those roles alone. Why this is and how it can be remedied is the subject of this article.

Three Basic Problems

Wikipedia has been cited for three weaknesses relevant to its role as a knowledge base. The first is that its coverage is imbalanced. Various studies have evaluated the scope of Wikipedia [13, 14, 15, among many] and have found areas of popular culture such as games, movies, music and actors to be over-represented, while areas of philosophy, technology, academics and history, to be under-represented. While still perhaps true in terms of absolute numbers of articles, the actual domain coverage has been improving in recent years.

The second Wikipedia problem is incompleteness. Wikipedia tends to be spotty in terms of providing complete and equal representation in populating certain categories (or classes) with articles (instances). It also tends to be incomplete in how well embedded or structured various articles may be. An example of the representation problem is in economy or commerce and the coverage of companies or products. The notability criterion [16] is a tricky one here; some companies or products with seemingly equivalent notability get listed, while others do not. Another example is the kingdom of life where some life forms are extremely well represented, while others are not. The incompleteness of structure relates to which articles or entire categories have infoboxes or ones that are well populated, as well as how category assignments are incomplete or inconsistent. The existence of "stub" articles is one evidence for such incompleteness. As Wikipedia has gotten more structured and complicated, the number of active editors has declined. The growing use of bots, however, is often compensating for this and in some cases bringing better consistency and equivalent treatment [17,18].

But the biggest problem of Wikipedia has been its category structure. Categories were not part of the original design, but were added to Wikipedia in 2004. Various reviewers have likened Wikipedia more to a thesaurus than a classification scheme [19], others that it is different than classical knowledge organization systems in that it has no specified root or hierarchy [20]. This improved a wee bit from 2006 to 2010, when the main Wikipedia topics were organized according to top-level and main topics [21]. Still, typical commentaries point to the fact that Wikipedia's category structure is "noisy, ill-formed, and difficult to make sense of" [22]. Its crowdsourced nature has led to various direct and indirect cycles in portions of the category structure [23]. All of these problems lead to the inability to do traditional reasoning or inference over the Wikipedia category graph [24].

Besides these lacks of computability, the Wikipedia graph is bloated with "artificial" categories (see further below) that just add noise to trying to understand or navigate the Wikipedia category structure. In short, while Wikipedia is a goldmine of resources and partial structure, its organization is incoherent at a global level, and it is unable to support reasoning and other tasks that might be expected from a truly functional knowledge base.

The real shame -- but also the real opportunity -- is that this lack of coherency makes it more difficult to validate and improve the information already in Wikipedia. So, there are both external reasons of linkage and internal reasons of improved authority for which it is desirable to shape Wikipedia into a true knowledge base.

Efforts to Recast Wikipedia

These faults are not unrecognized and the prospect of better leverage from Wikipedia has stimulated many efforts. Gazing inward, it is not uncommon to find efforts that attempt to clean up the existing Wikipedia structure [25], or various attempts to use the content of Wikipedia article categories [26] to re-constitute new taxonomies [27] or concept networks [28]. Clean up appears essential, and is a relative constant in other attempts to recast Wikipedia [29].

The choice of Wikipedia's founders to make its full content available electronically for free and without restriction was a masterstroke. This has stimulated many to grab the Wikipedia content and to recast it in other ways. One of the first, and most successful, was DBpedia, with an emphasis on making (much of) Wikipedia available in RDF and linked data. DBpedia emphasized the structured content of Wikipedia's infoboxes and eventually derived a typology of entities and their properties expressed as the DBpedia ontology [30]. It is not hyperbole to state that DBpedia nucleated the entire linked data phenomenon [7].

The key insight of YAGO [5] was the recognition that the resource richness of Wikipedia lacked a unifying structure, with WordNet chosen as the replacement organizing framework. Also, by patterned analysis of Wikipedia's article titles structure, YAGO was able to infer and select many attribute relationships between entities. This enabled YAGO to posit what, in essence, was a much-expanded category structure for Wikipedia expressed as predicates. Many other efforts have also chosen WordNet as their organizing framework for Wikipedia [31,32].

Freebase [2], itself another attempt to use crowdsourcing with explicit attention to structured data, struggled in its early years until it embraced and incorporated Wikipedia. That marked the take-off point for Freebase, which was later acquired by Google to form the backbone of its knowledge graph. Freebase is now being shut down with its assets being transferred to Wikidata.

Wikidata [33] is itself an interesting case of how the Wikipedia model is being expanded. Wikidata, a sister project to Wikipedia under the Wikimedia banner, takes as its starting point the structured data about entities evident in Wikipedia infoboxes. Rather than extracting and cleaning that entity information as DBpedia does, the role of Wikidata is to be the multilingual source for all entities feeding the Wikimedia network, including Wikipedia. The approach leads to more uniformity and consistency, and provides a central Wikimedia access point for structured data. However, somewhat akin to Wikipedia, Wikidata also has struggled to find an appropriate typology (or ontology) for its millions of entities [34].

Other approaches to the Wikipedia classification challenge have been to map -- or "express" -- Wikipedia articles in relation to established external vocabularies or structures, such as the Library of Congress Classification [35], Library of Congress Subject Headings [23, 36], Universal Decimal classification (UDC) [37], Cyc [38] or UMBEL [39], among others. The idea here is that accepted organizational schemes provide more coherence than the Wikipedia category structure, with sometimes additional benefits as well.

Though not complete topical recastings, certain aspects of Wikipedia have also proven their usefulness for general knowledge acquisition. Using article (concept or entity) content can inform topical tagging using explicit semantic analysis (ESA) [40], automatic topic identification [41], information extraction [42] or a myriad of others.

Making a Natural Wikipedia Category Scheme

Whether "cleaned" or recasted, taking the existing Wikipedia structure in its existing form is problematic. Though some category cleaning sometimes takes place with some of these uses of Wikipedia, that is not uniformly nor universally so. The cleaning that does take place is often limited to administrative categories (relating to internal Wikipedia conventions or management). However, other Wikipedia conventions (such as lists) and the proliferation of user-generated "artificial" categories actually represent the bulk of the total number of categories.

Charles S. Peirce was the first, by my reading, who looked at the question of "natural classes," which are now sometimes contraposed against what are called "artificial classes" (we tend to use the term "compound" classes instead). A "natural class" is a set with members that share the same set of attributes, though with different values (such as differences in age or hair color for humans, for example). Some of those attributes are also more essential to define the "type" of that class (such as humans being warm-blooded with live births and hair and use of symbolic languages). Artificial classes tend to only add one or a few shared attributes, and do not reflect the essence of the type [43].

"Compound" (or artificial) categories (such as `Films directed by Pedro Almodóvar` or `Ambassadors of the United States to Mexico`) are not "natural" categories, and including them in a logical evaluation only acts to confuse attributes from classification. To be sure, such existing categories should be decomposed into their attribute and concept components, but should not be included in constructing a schema of the domain.

"Artificial" categories may be identified in the Wikipedia category structure by both syntactical and heuristic signals. One syntactical rule is to look for the head of a title; one heuristic signal is to select out any category with prepositions. Across all rules, "compound" categories actually account for most of what is removed in order to produce "cleaned" categories.

We can combine these thoughts to show what a "cleaned" version of the Wikipedia category structure might look like. The 12/15/10 column in the table below reflects the approach used for determining the candidates for SuperTypes in the UMBEL ontology, last analyzed in 2010 [44]. The second column is from a current effort mapping Wikipedia to Cyc [45]:

	12/15/10	**3/1/15**
Total Categories	100%	100%
Administrative Categories	14%	15%
Orphaned Categories	10%	20%
Working Categories	76%	66%
"Artificial" Categories	44%	34%
Single Head	33%	23%
Plural Head		24%

"Clean" Categories	33%	46%

Two implications can be drawn from this table. First, without cleaning, there is considerable "noise" in the Wikipedia category structure, equivalent to about half to two-thirds of all categories. Without cleaning these categories, any analysis or classification that ensues is fighting unnecessary noise and has likely introduced substantial assignment errors. Second, approaches, assumptions and how filters get sequenced differ between "cleaning" attempts, which both makes comparability a challenge but also represents areas for discussion and testing to derive best practices. This lack of comparability due to differences in staging Wikipedia for analysis makes it difficult to draw comparisons between different studies and approaches. One study is not necessarily relatable to other studies.

Today, in chaotic and uncoordinated ways, we see Wikipedia feeding much analysis through partial aspects of its structure and supplying many reference concepts and entities. But each analysis is done for different purposes using different bases; they are thus incompatible. Coherency, usability and insight suffer. Any prior efforts to map to or use Wikipedia categories that do not remove these artificial categories only introduce noise and are therefore likely to be in substantial error.

Benefits of a Reference Knowledge Base

If we could overcome these shortcomings by taking the steps to make Wikipedia a true reference knowledge base, what might the benefits be? Or, said another way, *why* should we care?

One benefit is that reference structures of any kind provide a focus, by definition, of common or canonical referents. This commonality leads to better defined, better understood and more widely used referents. Common referents become a kind of common vocabulary for the space, upon which other vocabularies and datasets can reference. A common language, of sorts, can begin to emerge.

Reference structures also provide a grounding, a spoke-and-hub design [46], that leads to an efficient basis for external vocabularies and datasets to refer to one another. Of course, any direct mapping can provide a means to relate this information, but such pairwise mappings are not scalable nor efficient. In a spoke-and-hub design, the number of mappings required goes down significantly with the number of datasets or items requiring mapping. The spoke-and-hub design, for example, is at the heart of such disciplines as master data management.

Another benefit of common reference structures is that they provide a common target for the development of tools and best practices. These kinds of "network effects" lead to still further tooling and practices. Thus, while we see literally tens of thousands of academic papers and approaches leveraging Wikipedia in one way or another, we see little of a practice or a community that has been built around it as a knowledge base. It is as if we are still looking a bit at the shadow of Wikipedia and its possible role, a chimera for its potential as a true knowledge base.

But the ultimate benefit of Wikipedia as a reference knowledge base will reside in its computability. When we can reason over Wikipedia's content, use it for testing and analyzing assertions or new facts, when its coherent organization can be applied to such tasks as informing how to map and interoperate

data together or remaking whole legacy applications such as enterprise information integration or MDM, all of which in cross-lingual ways, we will finally see the realization of Wikipedia's inherent potential. And, as these latent capabilities get exploited, we will see supporting knowledge sources such as Wikidata also get pulled into the ecosystem.

Seven Requirements for a Computable Knowledge Base

So, if we buy into the benefits of a computable Wikipedia -- or any other useful knowledge source for that matter -- what are the guideposts for doing so? How do we assess the gaps and then fill them?

The importance of working with a "clean" version of the Wikipedia structure is obvious, yet ultimately there are higher-order requirements for what it takes, in our view, to become a "true" reference knowledge base. By our definition, such KBs have these aspects:

- ***Coherent*** -- does it hold together conceptually, logically, does it make sense? Either internally via consistency tests and such, or externally via testing against known facts and knowledge, the structure of the knowledge base should be defensible and meet the "common sense" test
- ***Comprehensive*** -- does the knowledge base have the scope of domains to which it is likely to interact? For a Web reference, the KB need not be global, but be relevant to an important domain of discourse. The biomedical domain, and its constituent and biological sub-domains, is an example. Something like Wikipedia represents a more "global" domain, and is thus central to the idea
- ***Referencable*** -- is the knowledge source authoritative? does it use URIs for referencing its objects?
- ***Open Standards*** -- which also implies, does it meet open standards? Open standards, by virtue of their decision processes, represent well-reasoned bases. Open standards are also easier to interoperate with and have more tooling available
- ***Computable*** -- the combination of the above can lead to a KB structure that supports reasoning, inference, set selection, relations, attributes, datatypes, and filtering and retrieval. These aspects make the KB "computable" [47], the threshold qualifier for a "true" knowledge base
- ***Testable*** -- but now, once the KBs are computable, they are also testable. That means the entire KB structure may be tested, verified, validated, scored, and evaluated
- ***Multi-lingual*** -- if not already multi-lingual, does it have a structure (such as ID v label-based) that supports multiple languages? Is there attention paid to encoding and transfer standards so as to promote consumption and use of the KB data? Multi-linguality may sound like icing on the cake, but it represents the next phase of bringing structure to the question of how to better identify, discern, and disambiguate information.

Wikipedia, and other publicly available knowledge sources [48], already fulfill many of these requirements. With focused attention, any current reference source should be able to be lifted to meet these seven major requirements.

Outlines of a General Staging Pipeline

OK, then: what might such a KB processing (or "lifting") approach look like?

Well, the first point is that it should be a pipeline. It is important to be able to swap in and out various options at multiple points from input to desired output. Then, because there are disparate sources and different formats to accommodate, it is also important to use canonical syntaxes and standards for expressing the products and specifications at the various steps along that pipeline.

The very notion of pipeline implies workflows, which are the actual drivers for how the pipeline should be designed. Key workflow steps include:

- Clean the input sources
- Express the sources in a canonical form [49]
- Identify and extract concepts
- Map the structure to KB concepts
- Identify and extract entities
- Identify and extract relations
- Type the entities, concepts, and relations
- Extract attributes and values for identified entities
- Test these against the existing KB
- Update reference structures, including placement of the new assertions, as appropriate
- Characterize and log to files
- Commit to the KB
- Rinse, repeat.

Much information gets processed in these pipelines, and the underlying sources update frequently. Thus, the pipelines themselves need to be performant and based on solid code. Automation, within the demanding bounds of quality, is also an essential condition to be scalable. Improving on that is a process, not a state.

Time to Make Some Sausage

Most of these observations are really not new or innovative [39,50]. Possibly what is new is to articulate the situation for major reference sources on the Web, and to then analyze and propose how to process them in the service of information interoperability.

Because, you see, we're still at the very, very earliest phases of how the Internet is changing the abilities to gather, understand, and represent the information in our world. We're about ready to embark on the next stage in that journey.

---

[1] *IBM Journal of Research and Development* 56(3/4), Special Issue on “This is Watson”, 2012

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM, 2008.

[3] A. Singhal: Introducing the Knowledge Graph: Things, not Strings. *Google Blog*, May 16, 2012

[4] Gruber, T. "Siri: a virtual personal assistant." In keynote presentation at *Semantic Technologies conference*

(SemTech09), 2009.

[5] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." In *Proceedings of the 16th international conference on World Wide Web*, pp. 697-706. ACM, 2007.

[6] Sören Auer, Chris Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak and Zachary Ives, 2007. DBpedia: A nucleus for a web of open data, in *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, Busan, South Korea, volume 4825 of LNCS, pages 715728, November 2007. See http://iswc2007.semanticweb.org/papers/ISWC2007_IU_Auer.pdf.

[7] Heath, Tom, and Christian Bizer. "Linked data: Evolving the web into a global data space." *Synthesis lectures on the semantic web: theory and technology* 1, no. 1 (2011): 1-136.

[8] 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.

[9] Mesgari, Mostafa, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. "The sum of all human knowledge: A systematic review of scholarly research on the content of Wikipedia," *Journal of the Association for Information Science and Technology* 66, no. 2 (2015): 219-245.

[10] However, only 1/10th of the different language Wikipedias have more than 100,000 articles; see https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics.

[11] See the discussion of 'structural sources' in M.K. Bergman, 2011. "In Search of 'Gold Standards' for the Semantic Web," in *AI3:::Adaptive Informatio*n blog, February 28, 2011.

[12] This count is from Google Scholar absent references in citations and patents with the query, http://scholar.google.com/scholar?as_vis=1&q=wikipedia&hl=en&as_sdt=1,16. Also, see the SWEETpedia listing 250 articles relating to this topic on this *AI3:::Adaptive Information* blog; I ceased updating the list about five years ago because it was growing too large to manage.

[13] Halavais, Alexander, and Derek Lackaff. "An analysis of topical coverage of Wikipedia." *Journal of Computer Mediated Communication* 13, no. 2 (2008): 429-440.

[14] Holloway, Todd, Miran Bozicevic, and Katy Börner. "Analyzing and visualizing the semantic coverage of Wikipedia and its authors." *Complexity* 12, no. 3 (2007): 30-40.

[15] Samoilenko, Anna, and Taha Yasseri. "The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics." *EPJ Data Science* 3, no. 1 (2014): 1-11.

[16] See http://en.wikipedia.org/wiki/Wikipedia:Notability

[17] Halfaker, Aaron, and John Riedl. "Bots and cyborgs: Wikipedia's immune system."*Computer* 3 (2012): 79-82.

[18] See https://en.wikipedia.org/wiki/Category:Wikipedia_bots_by_purpose

[19] Voss, J. Collaborative Thesaurus Tagging the Wikipedia Way, (2006)

[20] Suchecki, Krzysztof, Alkim Almila Akdag Salah, Cheng Gao, and Andrea Scharnhorst. "Evolution of Wikipedia's Category Structure." *Advances in Complex Systems* 15, no. supp01 (2012).

[21] For Wikipedia's main topics, see http://en.wikipedia.org/wiki/Category:Main_topic_classificationsReference; for Wikipedia's top-level categories, see http://en.wikipedia.org/wiki/Category:Fundamental_categories.

[22] Kittur, A., Chi, E. H. and Suh, B., 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 (CHI'2009)*, New York, USA, 2009, pp. 1509–1512.

[23] Joorabchi, Arash, and Abdulhussain E. Mahdi. "Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts." Journal of Information Science 40, no. 2 (2014): 211-221.

[24] Paulheim, Heiko, and Christian Bizer. "Type inference on noisy rdf data," In *The Semantic Web–ISWC 2013*, pp. 510-525. Springer Berlin Heidelberg, 2013.

[25] Maciej Janik and Krys Kochut, 2007. Wikipedia in Action: Ontological Knowledge in Text Categorization, *University of Georgia, Computer Science Department Technical Report no. UGA-CS-TR-07-001*. See http://lsdis.cs.uga.edu/~mjanik/UGA-CS-TR-07-001.pdf. Also, see Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Mohamed Tmar, and Abdelmajid Ben Hamadou. "Wikipedia Category Graph and New Intrinsic Information Content Metric for Word Semantic Relatedness Measuring." *Computing* 10, no. 13 (2012): 35-37.

[26] Vivi Nastase and Michael Strube, 2008. Decoding Wikipedia Categories for Knowledge Acquisition, in *Proceedings of the AAAI08 Conference*, Chicago, US, , pp.1219-1224.

[27] Simone Paolo Ponzetto and Michael Strube, 2007a. Deriving a Large Scale Taxonomy from Wikipedia, in Association for the Advancement of Artificial Intelligence (AAAI2007).

[28] Andrew Gregorowicz and Mark A. Kramer, 2006. Mining a Large-Scale Term-Concept Network from Wikipedia, *Mitre Technical Report*, October 2006.

[29] Wu, Fei, and Daniel S. Weld. "Autonomously semantifying wikipedia." In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 41-50. ACM, 2007.

[30] Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. "DBpedia-A crystallization point for the Web of Data." *Web Semantics: science, services and agents on the world wide web* 7, no. 3 (2009): 154-165.

[31] Marius Pasca, 2009. Outclassing Wikipedia in Open-Domain Information Extraction: Weakly-Supervised Acquisition of Attributes over Conceptual Hierarchies, in *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 639–647, Athens, Greece, 30 March – 3 April 2009. See http://www.aclweb.org/anthology/E/E09/E09-1073.pdf.

[32] Fei Wu and Daniel S. Weld, 2008. Automatically Refining the Wikipedia Infobox Ontology, presented at the *17th International World Wide Web Conference (WWW 2008)*

[33] Vrande?i?, Denny, and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase." *Communications of the ACM* 57, no. 10 (2014): 78-85.

[34] From scratch, in a bit over three years, Wikidata has grown to cover about 19 million entities according to February 2015 statistics. However, there has yet to emerge an overarching typology or ontology for these entities, with the typing system that does exist growing from the bottom up. For some background, see https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Migrating_away_from_GND_main_type

[35] There is an alternate entry point to Wikipedia provided by http://en.wikipedia.org/wiki/Library_of_Congress_Classification

[36] Kiyota, Yoji, Hiroshi Nakagawa, Satoshi Sakai, Tatsuya Mori, and Hidetaka Masuda. "Exploitation of the wikipedia category system for enhancing the value of LCSH." In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pp. 411-412. ACM, 2009.

[37] Salah, Almila Akdag, Cheng Gao, Krzysztof Suchecki, and Andrea Scharnhorst. "Need to categorize: A comparative look at the categories of universal decimal classification system and Wikipedia," *Leonardo* 45, no. 1 (2012): 84-85.

[38] Pohl, Aleksander. "Classifying the Wikipedia articles into the OpenCyc taxonomy." In *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, vol. 5, p. 16. 2012.

[39] Upper Mapping and Binding Exchange Layer (UMBEL) Specification, *UMBEL.org*, retrieved February 16, 2015.

[40] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, January 2007.

[41] Hassan, Mostafa. "Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge." PhD dissertation, *University of Waterloo*, 2013.

[42] Fei Wu, Raphael Hoffmann and Daniel S. Weld, 2008b. Information Extraction from Wikipedia: Moving Down the Long Tail, in *Proceedings of the 14th ACM SigKDD International Conference on Knowledge Discovery and Data Mining (KDD-08)*, Las Vegas, NV, August 24-27, 2008, pp. 635-644. See http://www.cs.washington.edu/homes/wufei/papers/kdd08.pdf.

[43] Menno Hulswit, 1997. "Peirce's teleological approach to natural classes," in *Transactions of the Charles S. Peirce Society* (1997): 722-772. See http://repository.ubn.ru.nl/bitstream/handle/2066/29577/29577___.PDF?sequence=1

[44] Upper Mapping and Binding Exchange Layer (UMBEL) Specification, Annex G: UMBEL SuperTypes Documentation, *UMBEL.org*, retrieved February 16, 2015.

[45] Aleksander Smywinski-Pohl, Krzysztof Wróbel, Michael K. Bergman and Bartosz Zió?ko, 2015. "cycloped.io: An Interoperable Framework for Web Knowledge Bases," manuscript in preparation.

[46] The main advantage of a grounding reference is that it allows a spoke-and-hub design for data mapping, which is tremendously more efficient than pairwise mappings. In a spoke-and-hub design, where the reference ontology is the common node at the hub, only $n - 1$ routes are necessary to connect all sources, meaning that it scales linearly with the number of sources and attributes. Without a grounding reference, these same mapping capabilities would require routes in a pairwise (point-to-point) approach, that also scales poorly as a quadratic function. A system of ten datasets would require 9 composite mappings in the reference grounding case, but 45 in a pairwise approach. And, of course,

datasets themselves contain tens to thousands of attributes, compounding the map scaling problem further.

[47] For example, WordNet is a coherent lexical ontology, but is not computable.

[48] See the knowledge bases section of M.K. Bergman, 2014. "Knowledge-based Artificial Intelligence," in *AI3:::Adaptive Information* blog, November 14, 2014.

[49] Galárraga, Luis, Geremy Heitz, Kevin Murphy, and Fabian M. Suchanek. "Canonicalizing open knowledge bases." In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1679-1688. ACM, 2014.

[50] See, for example, Suchanek, Fabian M., and Gerhard Weikum. "Knowledge Bases in the Age of Big Data Analytics." *Proceedings of the VLDB Endowment* 7, no. 13 (2014): 1713-1714.
