# Semantic Wikipedia – Checking the Premises

Rainer Hammwöhner

Institut für Medien-, Informations- und Kulturwissenschaft
Universität Regensburg
Universitätsstraße
93040 Regensburg
rainer.hammwoehner@sprachlit.uni-regensburg.de

**Abstract:** Enhancing Wikipedia by means of semantic representations seems to be a promising issue. From a formal or technical point of view there are no major obstacles in the way. Nevertheless, a close look at Wikipedia, its structure and contents reveals that some questions have to be answered in advance. This paper will deal with these questions and present some first results based on empirical findings.

## 1 Introduction

Up to now Wikipedia has accumulated an enormous wealth of information by the effort of an open community of volunteers. This information however is semi-structured at best and therefore imposes restrictions on automatic processing. Automatic processing of Wikipedia contents is desirable for a couple of reasons. Enhanced information services can improve the utility of Wikipedia itself. Implicit knowledge scattered over separated parts of the corpus can be brought together and made explicit. Consistency of the corpus can be enforced by autonomous agents operating on semantic representations. Information extracted from Wikipedia can be used in other contexts.

There are several approaches to this task, but two very general types may be distinguished. The first approach employs information extraction from Wikipedia based on the interpretation of existing explicitly defined structures [AL07]. The main sources of information are templates embedded within Wikipedia's articles. The resulting knowledge is represented in terms of a formal language and may be subject to viewing and querying via the OntoWiki software [ADR06]. The second approach requires a modification of the markup language in order to allow for link typing and attribute assignment [Vö06]. A process of information extraction and representation will again lead to formal representations that may be employed by inference processes. A necessary prerequisite of this approach is an extension to the MediaWiki software that is the technical core of Wikipedia [KVV06].

According to [Vö06] the following key elements are necessary to achieve the intended semantic annotation of Wikipedia's articles: *categories* classify articles according to their content, *types* express the meaning of links connecting Wikipedia's articles and *attributes* capture atomic properties related to the contents of an article. Categories are the only of these devices being already in use and ready for evaluation. Thus the notion of categorizing Wikipedia's articles will play a crucial role within the theoretical and practical considerations of this paper.

## 2 The Premises

Introducing at least one of the approaches mentioned above will be of major consequence to the users of Wikipedia. New information services will be available on the one hand and the authoring process will be more demanding on the other hand. The success of this project is bound to some central premises that should be made explicit and checked before the effort of large scale implementation is to be taken.

***P1*** Technical feasibility: Prototypes for both of the approaches have been implemented.

***P2*** Formal soundness: The proposed semantic representations are based on rigidly defined structures. However, there is some lack of clarity about the further use of typed links. As far as no terminological reasoning is intended, no problems should arise.

***P3*** Reliability of results: Recent studies have attested Wikipedia's convenient average quality [Ha07a, Ha07b, Wi07]. However, Wikipedia articles of abysmal quality can be found easily. The user of Wikipedia needs the competence to distinguish reliable from erroneous information. Semantic operations on Wikipedia should not accumulate errors and must not blur the user's view by hiding the sources of errors. It is not quite clear, whether this criterion is met by the proposed approaches.

***P4*** Reliability of the authoring process: The first approach does not impose additional tasks on the author. No new problems should arise here. The second approach relies heavily on the proper assignment of link types and categories by the user. The author can decide which and how many link types or categories to use. He can select from predefined denominators or enter new link types and categories at his will. Obviously, problems can arise out of the inconsistent and ambiguous use of type and category identifiers. [Vö06, section 4.1] infer from the seemingly unproblematic use of the category system that a consistent use of a link type system is to be expected too. This conclusion is problematic simply, because there is no empirical evidence of a proper use of the category system at all. It is the major objective of this paper to present some observations which are relevant to this issue.

***P5*** Multi-lingual system: Approaches to realizing a Semantic Wikipedia should consider that Wikipedia is a multi-lingual information base. At least an interlingual mapping mechanism for link types and attributes corresponding to interlingual category mapping should be developed.

***P7*** Usability: All efforts in enhancing Wikipedia by innovative information services will be futile unless they are integrated within an environment devoted to strict usability criteria. This applies for the authors and information seekers as well.

The list introduced above may not be complete. But the relevance of the mentioned premises does not seem to be questionable. ***P4*** occupies a key position since a fundamental question is involved here. Usable interfaces may be revamped, formal systems can be redesigned, but the competence of a large user community can be adjusted only in the long run. Thus ***P4*** may be the decisive criterion in the choice between more or less demanding approaches to a Semantic Wikipedia.

## 3 Is Wikipedia's category system a sound thesaurus?

The category system of Wikipedia is intended to provide an additional navigation structure on the set of articles [Wi07a]. It is not used as a device of query support primarily. The proper assignment of categories is defined by a set of rules of thumb [Wi07a]. The question, whether this category system is a thesaurus, was firstly brought up by [Vo06]. In his comprehensive overview on the category system of Wikipedia Voss examines the statistical distribution of category features and compares this category system to other means of knowledge organization - thesauri (MeSH: Medical Subject Headings), hierarchical classifications (Dewey Decimal Classification) and folksonomies (del.icio.us). This comparison is of major importance to Semantic Wikipedia, because formal properties of the category system may be inferred from the result. Voss arrives at the conclusion, that Wikipedia's category system is a thesaurus, since the requirements of ISO 2788 [ISO86] are met. The *equivalence relation* connecting synonymous terms may be represented using redirects. The *hierarchical relation* between broader and narrower terms is expressed by the category ⇒ subcategory relation. *Associations* between related terms are represented by hyperlinks. Obviously the mark-up language of Wikipedia is capable of expressing thesaurus structures. The question, however, is, whether the existing category systems *are* thesauri. [Vo06] further elaborates his conclusions by comparing excerpts from the MeSH thesaurus and from the English Wikipedia. The presented structures are reasonably similar. But counter examples may be found easily at least within the English Wikipedia (as observed at. 0.6.07.2007):

> *categories* ⇒ *fundamental* ⇒ *thought* ⇒ ***knowledge*** ⇒ *academia* ⇒ *academic institutions* ⇒ *school counseling* ⇒ *personal development* ⇒ *personal finance* ⇒ *microeconomics* ⇒ *information, knowledge and uncertainty* ⇒ *information* ⇒ ***knowledge*** ⇒ *nature* ⇒ *life* ⇒ *death* ⇒ *extinction* ⇒ *fossils* ⇒ *dinosaurs*

This illustrative example demonstrates the existence of cycles (*knowledge*) within the category ⇒ subcategory relation. Cyclic structures conform to Wikipedia's rule set [Wi07a], but not to ISO 2788 since the resulting structure is no hierarchy. The category ⇒ subcategory relation does not lead generally from broader to narrower terms, but in many cases to related terms. Thus, the category ⇒ subcategory relation may not be considered as a transitive relation representing terminological subordination.

As a consequence there is no support of terminological reasoning by the English category system. Even retrieval support, e.g. by spreading activation, may lead to unwanted results, if the terminology is as weakly structured as the example suggests. The same criticism is valid for the French Wikipedia as well. The category systems of the Italian and German Wikipedia are quite different in structure. They contain a few cycles only, their hierarchy has a considerably lower depth (s. table 1). This applies to the maximal descriptor level (first value) and the longest observed path within the hierarchy (value in brackets) as well. A substantial difference between both of the depth values indicates a lack of balance within the category system.

| | articles | basic categories | all categories | max depth | superord. per cat. (median) | cycles |
|---|---|---|---|---|---|---|
| de (en) | 152 | 366 | 1740 | 10 (15) | 2 | 4 |
| de (fr,it) | 169 | 394 | 1816 | 10 (15) | 2 | 4 |
| en | 152 | 581 | 6274 | 14 (156) | 2 | 493 |
| it | 167 | 321 | 1091 | 12 (15) | 2 | 7 |
| fr | 134 | 360 | 3116 | 14 (83) | 2 | 424 |

Table 1: Basic features of category systems

The data presented above are derived from the following samples: two bilingual samples of de-en (size 152) and de-it (size 169) were chosen at random using interlingua links. A sample of 134 French articles was added to the latter one, once more using interlingua links. The basic categories describing these articles were sampled as well as all of their superordinate categories. The example suggests, that sample size has some influence on the number of basic categories, less influence on the total number of categories and no impact on the depth of hierarchy and number of cycles. It can be assumed, that deep category systems are error prone. Authors will have difficulties to get an overview on the overall structure since the number of paths to the top category shows exponential growths behaviour. An additional example will illustrate the pitfalls of big category hierarchies in Wikipedia. It shows the first 99 categories of the longest path within the category ⇒ subcategory multi-hierarchy as found in the sample of the English Wikipedia:

*digital revolution ⇒ cryptography ⇒ application of cryptography ⇒ authentication methods ⇒ personal identification ⇒ biometrics ⇒ physical anthropology ⇒ human evolution ⇒ evolutionary psychology ⇒ memetics → anticipatory thinking ⇒ strategic management ⇒ product management ⇒ product development ⇒ design ⇒ built environment ⇒ architecture ⇒ architecture and engineering occupations ⇒ building engineering ⇒ building materials ⇒ metals ⇒ alloys ⇒ copper alloys ⇒ bronze ⇒ bronze age ⇒ ancient near east → ancient near eastern religions → ancient semitic religions ⇒ Abrahamic religions ⇒ Judaism → messianism ⇒ Jesus ⇒ doctrines and teachings of Jesus → nonviolence → peace ⇒ peace churches ⇒ anabaptism ⇒ amish ⇒ simple living ⇒ environmentalism ⇒ environmental ethics ⇒ extinction ⇒ extinct species*

*⇒ extinct animals ⇒ prehistoric animals ⇒ mesozoic animals ⇒ cynodonts ⇒ mammals⇒ primates ⇒ apes ⇒ humans ⇒ anthropology ⇒ prehistory ⇒ archaeology ⇒ periods and stages in archaeology ⇒ ancient history ⇒ ancient mysteries ⇒ astrology → ~~astrological factors~~ → classical elements → earth⇒ earth sciences ⇒ environmental science ⇒ environment ⇒ urban studies and planning ⇒ transportation ⇒ travel ⇒ tourism ⇒ cultural heritage ⇒ cultural history ⇒ cultural movements ⇒ art genres ⇒ graphic design ⇒ printing ⇒ books → fiction ⇒ fictional ⇒ fictional abilities ⇒ superhuman powers ⇒ psychic powers → prediction ⇒ futurology ⇒ population ⇒ demography ⇒ ethnicity ⇒ ethnicity in politics ⇒ anti-national sentiment ⇒ prejudices ⇒ bias ⇒ appearance ⇒ aesthetics ⇒ arts ⇒ visual arts → communication design ⇒ mass media ⇒ media by format ⇒ digital media ⇒ software ⇒ software engineering ⇒ ...*

This example was extracted from the English Wikipedia at 15th of June and verified at the 7th of August 2007. In the meantime one category and 10 category ⇒ subcategory links have been deleted (→), a super-category has been added to *digital revolution* again. Some of these deletions lead to a simplification of the overall structure; some others were caused by the insertion of additional hierarchy levels. It is an open question, which effects will result from the volatility of the category system as observed in this example. These findings, however, have to be confirmed using bigger samples or the complete data set. It would be desirable to develop diagnostic tools which could identify problematic category inclusions. One promising approach is the comparison of category systems from various Wikipedias. If a category ⇒ subcategory inclusion is present in more than one Wikipedia, it is likely to be valid. If it occurs in one Wikipedia only, it can be invalid or culture specific as well.

## 4 What does this mean to Semantic Wikipedia

This small study, based more on illustrative examples than on statistical evidence, suggests that Wikipedia's category system is not obviously a sound base for the development of a more demanding semantic system. The proliferation of the category system indicates what may happen to a link type system that may freely be extended by the user. This aspect is of crucial importance since evaluation of link typing had controversial results even in more controlled settings [Ma91]. As a consequence more empirical studies on category assignment are needed in order to understand the unfolding of the rather different category systems within the German and Italian Wikipedia on one side and the French and English Wikipedia on the other. Various settings – for instance with open and closed link type systems – should be considered before modifications at the existing encyclopaedia are brought into effect. Nevertheless, the introduction of more semantic features into Wikipedia has lots of promising aspects, too. The category system can be relieved from alien tasks like fact representation. The problem of redundant assignment of categories and subcategories [Wi07b] to Wikipedia articles can be solved by simple inference processes in combination with appropriate presentation tools. These are just examples of the positive effects that can be achieved by Web 2.0 techniques. Furthermore, the technical soundness and good performance of the existing prototypes promises that experiments may be carried out with reasonable effort.

## References


[AL07] Auer, S.; Lehmann, J.: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. accepted at ESWC 2007. http://www.informatik.uni-leipzig.de/~auer/publication/ExtractingSemantics.pdf, cited 20.05.2007

[ADR06] Auer, S.; Dietzold, S.; Riechert, T.: OntoWiki - A Tool for Social, Semantic Collaboration. In I. Cruz et al. (Eds.): Proceedings of 5th International Semantic Web Conference, Nov 5th-9th, Athens, GA, USA, LNCS 4273, pp. 736-749, 2006. Springer-Verlag Berlin Heidelberg 2006. http://www.informatik.uni-leipzig.de/~auer/publication/ontowiki.pdf, cited 20.05.2007

[Ha07a] Hammwöhner, R. et.al.: Qualität der Wikipedia. Eine vergleichende Studie. In Oßwald, A.; Stempfhuber, M; Wolff, C. (eds.) Open Innovation. Proc. 10th Int. Symposium on Information Science in Cologne. UVK, 2007, pp. 77-90.

[Ha07b] Hammwöhner, R.: Qualitätsaspekte der Wikipedia. In: Stegbauer, C.; Schmidt, J.; Schönberger, K. (eds): Wikis: Diskurse, Theorien und Anwendungen, Sonderausgabe von kommunikation @ gesellschaft, Jg. 8, 2007, Online-Publication: http://www.soz.uni-frankfurt.de/K.G/B3_2007_Hammwoehner.pdf

[ISO86] ISO 2788: 1986: Guidelines for the establishment and development of monolingual thesauri.

[KVV06] Krötzsch, M.; Vrandecic, D.; Völkel, M.: Semantic Mediawiki. In Proc. 5th Int. Semantic Web Conf. (ISWC06). http://korrekt.org/papers/KroetzschVrandecicVoelkel_ISWC2006.pdf, cited 20.05.2007

[Ma91] Marshall, C.C. et.al.: Aquanet: a hypertext tool to hold your knowledge in place. In *Proc. Hypertext' 91, San Antonio,* S. 261-275, New York, 1991. ACM.

[Vö06] Völkel, M. et.al.: Semantic Wikipedia. In Proc. 15th Int. Conf. on World Wide Web, WWW 2006, Edinburgh, Scotland, May 23-26, 2006. http://www.aifb.uni-karlsruhe.de/WBS/hha/papers/SemanticWikipedia.pdf, cited 20.05.2007

[Vo06] Voss, J.: Collaborative thesaurus tagging the Wikipedia way. (v2; 2006-04-27; http://arxiv.org/abs/cs.IR/0604036) – Wikimetrics research papers, volume 1, issue 1 (cited 0.6.07.2007).

[Wi07] Wiegand, D.: Entdeckungsreise, Digitale Enzyklopädien erklären die Welt, c't, Magazin für Computer und Technik, Nr. 6, 2007, S. 136-145.

[Wi07a] Wikipedia:Categorization. In Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Wikipedia:Category, cited 27.05.2007.

[Wi07b] Wikipedia: Categorization and subcategories. In Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Wikipedia:Categorization_and_subcategories, cited 27.05.2007.

[Wi07c] Wikipedia Diskussion: Kategorien. In Wikipedia, The Free Encyclopedia. http://de.wikipedia.org/w/index.php?title=Wikipedia_Diskussion:Kategorien/Archiv1, …Archiv2, …Archiv3, cited 27.05.2007