# Categories in Wikipedia from a View of Information Science

## (text to presentation)

January 2011

**Lucie Sakastrová**
Institute of Information Studies and Librarianship
Faculty of Arts and Philosophy, Charles University in Prague

Text to presentation from Wikiconference (http://www.wikikonference.cz/) focuses on the most important characteristics and possibilities of usage of encyclopedia **Wikipedia**'s (http://www.wikipedia.org) **categorial system** (folksonomy). The first section of the text provides basic similarities and differences between Wikipedia's categorial system (http://en.wikipedia.org/wiki/Category:Contents) and Wikipedia as a whole (see slide 2).

Wikipedia and its categorial system differ mainly that Wikipedia is an **encyclopedia**, while the categorial system is a **folksonomy**, which means it is the system of collective organizing or sorting of information content [Folksonomy, 2010]. Both are created **collectively** and work on the **wiki** principle. This means that anyone can participate in the creation of their content and anyone can edit it [Wiki, 2010]. These contents are also **open**. This means that they can be, under the condition of obeying their licenses, free copied, used and distributed [*Wikipedie*, 2011].

**Multilingualism** and **multiculturalism** of Wikipedia and its categorial system is doubtful, because Wikipedia is not so much a multilingual encyclopedia, but rather consists of more than 250 monolingual encyclopedias, which are interconnected via the so-called interlanguage links (http://en.wikipedia.org/wiki/Help:Interlanguage_links). However, their contents are not identical. There are versions that have as few tens of articles, while the English version has more than 3.5 million articles [Wikimedia Foundation, 2011]. It means that the results of each language version are very individual and determined both by the used language and its "power" and culture or cultures associated with it. The same characteristics apply for the categorial system. It means that also the categorial system is not multilingual, but consist of tens of monolingual categorial systems, which may be interconnected via interlanguage links, but their results are in terms of scope, structure, classified contents, etc. not the same (compare, for example, the slide 6 and 7).

**Quality management tools** are the same by the categorial system and Wikipedia as a whole. This includes preventive tools, which include system of help, documentation or categorization of Wikipedia users according to the allocated rights to existing user groups (http://en.wikipedia.org/wiki/Special:ListGroupRights). The subsequent tools include e.g. blocking mechanism (http://en.wikipedia.org/wiki/Wikipedia:Blocking_policy), deletion mechanism, etc. (http://en.wikipedia.org/wiki/Wikipedia:Maintenance). And the hybrid tools include some special services such as service called "Recent changes" (http://en.wikipedia.org/wiki/Help:Recent_changes) or service called "Watchlist" (available only after login), etc.

The following part provides a list of important similarities and differences between Wikipedia's categorial system and systems of classification of information created by experts (see slide 3).

The most important difference between Wikipedia's categorial system and systems of classification of information created by experts is the risk of **vandalism**, which threatens probably only Wikipedia. Other characteristics and risks, such as **error rate**, **redundancy** or **inconsistence** threatened Wikipedia's categorial system and the systems of classification of information in general. It can be only assumed that by the systems of classification of information created by experts will be the quantity of errors, redundancy and inconsistence lower due to higher probability of homogeneity of this group (knowledge of the field, knowledge of the ways of classification of information, binding norms and manuals, etc.). Furthermore, it can be assumed that the redundancy, which will appear in the systems of classification of information created by experts, will have at least a presumption of purpose and meaning, while this don't have to always be in the case of Wikipedia.

Other important characteristic of systems of classification of information is generally also a **level of detail** (the depth of the hierarchy). Generally speaking, the level of detail of the classification systems of information created by experts is usually lower (smaller depth of the hierarchy), otherwise the system becomes uncoordinated and unsustainable. In the case of Wikipedia's categorial system is the level of detail unlimited, but in different locations of it's tree of decomposition also very individual. There are even categories dedicated to individuals (see, for example, the category "Michael Jackson" on the slides 6 and 7), which are in the case of systems of classification of information created by experts, whose purpose is purely scientific, practically unthinkable (like, for example, more detailed classification of type "homo sapiens sapiens" to concrete people in the case of biological taxonomy).

Of course, Wikipedia's categorial system or another system of classification of information, are not objective, because objectivity does not exist. Which exists, is only expression the degree of **intersubjectivity** consensus among any number of interested individuals. In the case of Wikipedia, there are more interested individuals than in the case of systems of classification of information created by experts. Interested individuals of Wikipedia are also much more diverse. In the case of Wikipedia, intersubjectivity is given by the Wikipedia principles of collective work, in the case of systems created by experts the intersubjectivity is given by the degree of compliance among professionals in areas such as knowledge of the field and views on methods of classification of information, etc.

Because of the much higher number of Wikipedia users in comparison with number of experts created professional systems, the **greater update rate** can be assumed by Wikipedia, while the **more speed of agreement** among experts in the case of professional systems. This is related to the risk of **instability** of Wikipedia, and the risk of narrow-mindedness and elderliness of systems of classification of information created by experts, which are not able to accent changes in the scientific and other kind of knowledge quickly enough. With the number of users also bears the risk of bias and it can be assumed that this risk will be lower in the case of Wikipedia because of a significantly greater number and diversity of its users. But it can not be said with certainty.

The question of degree of **scientism** of systems is similar to the question of objectivity. No system of classification of information is 100% scientific (e.g. due to new discoveries and new views on these findings, which requires revision of the existing status, etc.). However, it can be assumed that in the case of systems of classification of information created by experts the degree of scientism is higher, but it also can not be excluded that some parts of the Wikipedia's categorial system are managed by professionals and reflect some degree of scientism, too.

The following parts of the presentation include concrete examples of problematic issues of categorization of content in the Czech and English versions of Wikipedia (see slides 4-8).

These examples are concerned with problematic of used **terminology** in Czech version (see slide 4), presentation of **controversial information** both in Czech and English version (e.g. age, rase, information about collaboration with political police, etc.) (see slides 4,6 and 7), problems of apparent **illogicality** of some classification criteria (see e.g. slides 4,5 and 8) or **cohesion** of some topics **with culture** or **religion** (see e.g. slide 5 or compare slides 6 and 7).

The final part of the presentation deals with the differences in the usage of Wikipedia's categorial system and systems of classification of information created by experts with focus on the importance of Wikipedia's categorial system for the purpose of building the semantic web (see slides 9-10).

The usage of Wikipedia's categorial system is more suitable for the **purposes of interests**, **hobbies or entertainment**, while the systems of classification of information created by experts for **professional purposes** (see slide 9). **Accuracy** and **completeness** of the information found in Wikipedia is not as important as accuracy and completeness of the information found in specialized databases, such as in the case of some patent database, where the loss of only one document by searching can cause substantial financial loss in eventual legal dispute. While systems of classification of information created by experts are generally designed for both **browsing** and **searching** (separate search fields, rotated index, etc.), categorial system of Wikipedia is more suitable for browsing, even though it can be searchable, too (alphabetic index of categories, advanced search limited to the type of content of categories, etc.).

The usage of systems of classification of information created by experts are suitable for **information resources that are created and managed by professionals**, while the usage of Wikipedia's categorial systems is suitable for **open information resources, which are created collectively**, and also for external systems and services working with Wikipedia's content, such as **internet search engines**, **linking and reference services** or **services of the semantic web** (which are the services of the new evolutionary level of the existing web, in which information is structured and stored according to standardized rules, which makes them easier to find and to process [Semantický web, 2010]).

While in the case of databases and search engines is usually necessary to formulate search queries through a more or less complex **query language**, in the case of semantic web services queries can be formulated also in **natural language** sentences (see slide 10). The relevant service is able to translate the query formulated in natural language into the query formulated in the query language so that the relevant **robot** could work with it and put the relevant output results.

So the conclusion of the presentation and this text, too, is that, even though some of the methods and results of the categorization of content in Wikipedia appear from the view of information science to be random, unnecessary, controversial, inappropriate or pointless, these results can be apllied in the **semantic web** and similar advanced services, which means that the **importance** of this results will continue in its upward trend, not only in the field of **information science**.

## Literature


- Folksonomy. In *Wikipedia : the free encyclopedia* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2001- , last modified on 16 December 2010 at 12:23 [cit. 2011-01-08]. Anglické rozhraní. Dostupné z WWW: <http://en.wikipedia.org/wiki/Folksonomy>.
- Sémantický web. In *Wikipedie : otevřená encyklopedie* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2002- , naposledy editována 26. 11. 2010 v 18:56 [cit. 2011-01-09]. České rozhraní. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Sémantický_web>.
- Wiki. In REITZ, Joan M. *ODLIS : Online Dictionary of Library and Information Science* [online]. Westport (CT) : Libraries Unlimited, 2004-2010, last updated March 9, 2010 [cit. 2011-01-08]. Vyšel i v tištěné formě. Dostupné z WWW: <http://lu.com/odlis/>.
- Wikimedia Foundation. 2011. *Wikimedia : meta-wiki* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2001-, last modified on 4 January 2011, at 19:09 [cit. 2011-01-08]. List of Wikipedias. Anglické rozhraní. Dostupné z WWW: <http://meta.wikimedia.org/wiki/List_of_Wikipedias>.
- *Wikipedie : otevřená encyklopedie* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2002- , naposledy editována 3. 1. 2011 v 13:16 [cit. 2011-01-08]. Wikipedie. České rozhraní. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Wikipedie>.