Anita Nuopponen
Assistant professor in Applied Linguistics
University of Vaasa

SUBJECT CLASSIFICATION AS A TERMINOLOGY RETRIEVAL TOOL IN AN ONLINE ENCYCLOPEDIA

There are several means to assist an information seeker in online information services e.g. in online glossaries, terminological databases, online encyclopedias, bibliographies, and link collections. A widely utilised method in free online glossaries from different subject fields is a simple search function where the information seeker types a word in a box. This kind of search, however, requires that the user knows one or more search terms to start with. Often only a single article with a term and a definition, or an equivalent in another language is shown to the user. This search method does not support browsing in cases in which the information seeker does not have an exact term in mind. In some glossaries, the word search leads the user to the wanted term entry in an alphabetical list of entries, thus showing all the term entries around it. In another type of typical online glossary, the material can be browsed by the alphabet. Sometimes the word search and the letter search are combined, see Figure 1.

INTRODUCTION

Remarks concerning ontological and epistemological issues

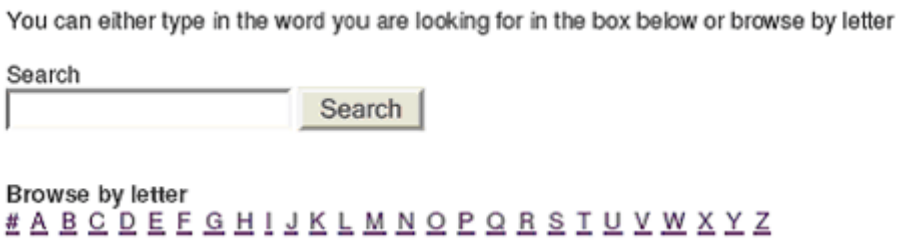

Figure 1. A typical interface of an online glossary1

Browsable subject (field) categories have been successfully used to assist information seekers in different kinds of online information resources such as link collections, e.g. Yahoo Directories2 . They are also used in online encyclopedias, e.g. Wikipedia (categories), Encarta (categories3), and Encyclopaedia Britannica Online (subjects4). The purpose of Wikipedia category classification is defined as follows:

> "Categories (along with other features like cross-references, lists, and infoboxes) help users find information, **even if they don't know that it exists or what it's called**."[5]

Browsable subject (field) classifications offer a solution when the information seeker does not have an exact search term in mind. This is also why I am interested in testing the function of the classification of Wikipedia. Traditional terminological databases utilise subject field classification, but not normally for browsing purposes. Instead, the classification assists in delimiting the field of the search term (see Figure 2), or to give information on the domain of the concept retrieved. Traditional terminological databases, such as Eurodicautom, are geared for translators and support searches where the user has a term to start the search with.

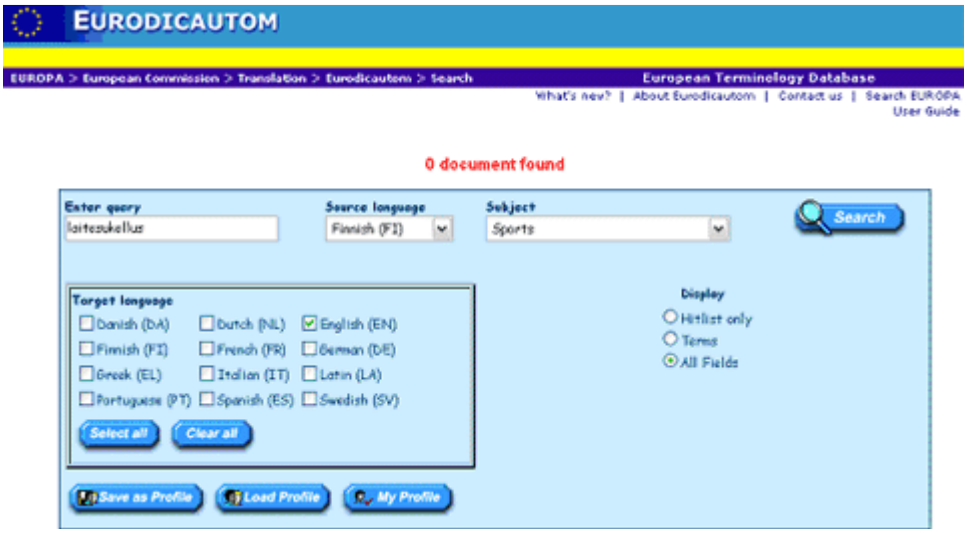

Figure 2. The interface of Eurodicautom terminology database.

The purpose of this paper is to discuss the use of browsable categories as a retrieval tool for terminological information. Instead of looking at existing or planned terminological vocabularies or data bases, this paper takes a look at an Internet based encyclopedia, Wikipedia. The question to be asked here is How to find desired terminological information by using the category classification provided by the encyclopedia? I set two information seeking tasks for myself in order to test the classification:

> 1) Find a definition of the concept Japanese tea ceremony.
> 2) Find English equivalents for the Finnish terms 'laitesukellus' (direct translation: "apparatus diving"), 'avovesisukellus' ("open water diving") and 'parisukellus' ("pair diving") in a case in which the information seeker is familiar with the concepts but is not sure about the English equivalents and cannot find them in normal dictionaries.

Before discussing this pilot study and the implications of the findings for a terminological hypertext system, the general ideas behind encyclopedias, Wikipedia and subject classification systems are scrutinized.

1 ENCYCLOPEDIA

Traditionally, encyclopedias are printed books or sets of books. They contain authoritative information about a variety of fields in the form of factual articles, normally subject to editorial approval. Articles vary in length, but are longer than the ones in glossaries and dictionaries. There are also vocabularies or glossaries, which are called 'encyclopedias' and encyclopedias, which are called 'terminologies', 'glossaries', or 'dictionaries'. Encyclopedias may be general in scope, such as the Encyclopedia Britannica or subject-oriented, such as The Encyclopedia of Philosophy. Articles in most printed general encyclopedias are organised in alphabetical order and the user must know what she/he is looking for. Indexes covering terms, which appear in articles and references in the text help in this. Articles in encyclopedias typically also contain references to authoritative books and articles on the subject.

The tasks of encyclopedias are, on the one hand, "to document and maintain authentic knowledge, ensuring and testifying its preservation over time and space", and on the other hand, "to provide adequate categorization and systematization of knowledge, providing easy access to knowledge for any interested person" (Lehner et al.: 3). For terminologists and other seekers of terminological information, encyclopedias have always been important reference sources. They systematize and define concepts and present the information in the form of a compact text. For an encyclopedia article, each subject or field is researched and presented as a whole. This makes it easy to extract terms, concepts, and definitions and to draw a graphical presentation of the concept system as well as to compile a glossary. Because of the systematic nature of encyclopedia articles, they are also used as sources for several ongoing projects in which terminology is extracted and knowledge or terminology bases are built automatically using natural language technologies (see e.g. Sui, Cui et al.: 2005).

Today, several printed encyclopedias have migrated into the Internet, e.g. The Encyclopædia Britannica Online6 and The Columbia Encyclopedia7. A further development are encyclopedias, which exist only on the www, e.g. Encyclopedia.com and Wikipedia. Many online encyclopedias are copies of the print version, or imitate one.

2 WIKIPEDIA

Wikipedia was selected as the object of study because it differs from other online encyclopedias in several ways, and because its popularity as a source of terminological information is growing all the time. Its content is dynamic: it is written collaboratively by volunteers, and anyone can add or change an article. Wikipedia's content is free whereas the commercial online encyclopedias may show only the beginning of the articles to those who have not registered (see e.g. Encyclopædia Britannica). Wikipedia was started in 2001 and is now operated by the non-profit Wikimedia Foundation. In November 2006, it had over 5 million pages, including more than 1,504,000 (compared to 961,000 in February 2006) in the English-language version, 88,000 (48,000) in Finnish, 195,600 (135,000) in Swedish, and over 7,500 (4,000) pages in Latin, just to mention a few of the numerous languages which have their own independent Wikipedia versions.8 The English version and its category classification will be used here.

Wiktionary, a sister project, is a similar collaborative project with the aim "to produce a free multilingual dictionary in every language, with definitions, etymologies, pronunciations, quotations, synonyms, antonyms and translations".9 Wiktionary also contains glossaries of different subject fields. However, they are not yet extensive enough and do not yet utilise the possibilities available on the web. Wikipedia itself is still more interesting for terminological purposes with its huge network of knowledge on any subject in various languages.

3 ORDERING AND CLASSIFICATION SYSTEMS

Different types of classification systems have been created for ordering and searching for books and information. Library and other classifications have found new life as knowledge organisation systems for networked knowledge. Originally, library classifications were meant for ordering "the fields of knowledge in a systematic way", bringing "related items together in the most helpful sequence", providing "orderly access to the shelves", and providing "an exact location for an item on the shelf." (Suman & Karmakar 2002.) These needs still exist, but on the World Wide Web, bookshelves are replaced by hypertext pages or databases. Library classifications have been designed to cover everything under the sun and above and must therefore remain on a high level of abstraction. As an example could be mentioned The Universal Decimal Classification (UDC), see Table 1.

Table 1. Outline of the UDC.

0	GENERALITIES
1	PHILOSOPHY. PSYCHOLOGY
2	RELIGION. THEOLOGY
3	SOCIAL SCIENCES
4	VACANT
5	NATURAL SCIENCES
6	TECHNOLOGY
7	THE ARTS
8	LANGUAGE. LINGUISTICS. LITERATURE
9	GEOGRAPHY. BIOGRAPHY. HISTORY

Like so many other information services, Wikipedia has its own subject field classification, Categories. In February and July 2006, the classification consisted of 9 main categories, which all have a large number of subcategories (see Table 2).

Table 2. Wikipedia's categories (12.07.2006)

Art and Culture	Arts and crafts · Cultural movements · Entertainment · Films · Food and drink · Games · Languages · Literature · Mass media · Museums · Music · Mythology · Parties · Performing arts · Pets · Popular culture · Radio · Sports · Television · Traditions · Tourism · Toys
Geography	Africa · Antarctica · Asia · Australia · Europe · North America · Oceania · South America · Cities · Climate · Countries · Landforms · Maps · Parks · Subterranea · Towns · Villages
History	Africa · Asia · Australia · Eurasia · Europe · North America · Oceania · South America · By Period · By Region · By Country · By Topic · Colonialism · Historiography · Timelines
Mathematics	Algebra · Analysis · Arithmetic · Economics · Education · Equations · Geometry · Logic · Measurement · Numbers · Proofs · Theorems · Trigonometry · Statistics
Natural sciences	Applied sciences · Astronomy · Biology · Chemistry · Earth sciences · Ecology · Heuristics · Health sciences · History of science · Information science · Medicine · Scientific method · Physics · Protoscience · Scientists · Space · Systems theory
Philosophy and Religion	Lists · Branches · Movements · Schools and traditions · Theories · Arguments · Philosophers · Literature · History · By era · By region · Aesthetics · Epistemology · Ethics · Logic · Metaphysics · Buddhism · Christianity · Confucianism · Cults · Deism · Judaism · Islam · Religious Movements · Satanism · Taoism
Social sciences	Anthropology · Archaeology · Cultural studies · Demographics · Economics · International relations · Linguistics · Psychology · Media studies · Political science · Social scientists · Sociology · Sexology
Society and People	Biographies · Business · Communication · Education · Ethnic groups · Family · Finance · Gender · Government · Health · Home · Industries · Labor · Law · Mass media · Organizations · Politics · War
Technology	Agriculture · Architecture · Automation · Automobiles · Big Science · Biotechnology · Chemical processes · Computing · Electronics · Energy · Engineering · History of technology · Information technology · Internet · Manufacturing · Nanotechnology · Nuclear technology · Sound · Telecommunications · Technology forecasting · Tools · Transportation · Vehicles

The classification in Table 2 will be used in this study, because when I made the analyses of the paths in January/February 2006 and in July 2006, the classification had not been changed. In November 2006, when checking some information, however, I found that the classification of categories had been extended to cover 2 new main categories. Also, the names of the main categories had been changed: Reference (new), Geography and places, Health and fitness (new), History and events, Mathematics and abstractions, Natural sciences and nature, People and self, Philosophy and thinking, Religion and belief systems, Social sciences and society, and Technology and applied sciences.10

The subcategories in the Wikipedia classification are linked to pages, which often contain a short introduction to the subject. On these introductory pages, there is a list of further subcategories as well as links to articles directly placed under the category in question. Every subcategory has these elements on its own page, too. Thus, some articles can be found after a couple of clicks, but for others, more clicking will be needed as we see below where the paths to the example information are described.

4 PATHS LEADING TO THE INFORMATION

In what follows, the paths leading to the desired information will be discussed in four stages: selecting the entry points (4.1), finding the right articles (4.2), tracing backwards to find alternative paths (4.3), and finally revisiting Wikipedia four months later (4.4).

4.1 Finding the entry point

Even though the Wikipedia category classification is comprehensive, it proved to be difficult to find the topmost categories for both tasks to start with. For finding the article "Japanese tea ceremony" the first guess, *Arts and Culture*, was not successful. *Ikebana* and *origami* were found directly under *Arts and Culture*'s subcategory *Arts and Craft*, but not Japanese tea ceremony which is very often associated with them as a hobby or an art form originating from Japan. The second guess was the path Geography: Asia: Japan, but it did not bring any results during the first two visits to the site in February or July.[11]

As to the second task, the target was an article or articles on the type of diving, in which breathing apparatus is utilized (fi *laitesukellus*) and a couple of its subtypes. The first guess was the main category Society and People, since the first supposition was that it would include Sports (see Figure 4). This proved to be a dead end, too.

In both cases, I had to return to the higher-level categories once more and read them through more carefully than I had initially done. The main category *Arts and Cultures* eventually proved to be a fruitful starting point for both tasks: for the tea ceremony its subcategory *Food and drink* and for diving *Sports*.

After I found the category Sports under *Art and Culture* – which I did not at first think of as the most obvious choice – the task of finding diving related information became easier: Sports had a subcategory *Water sports*, and *Diving* could be found under it.

4.2 Detours and ambiguities

Even though in both cases the entry points were found, there were still some problems before the desired information was discovered. The category classifications did not lead the information seeker directly to the actual articles but category articles were needed as intermediaries.

In order to find the path to tea ceremony, the subcategory *Food and drink* was tried. It had a subcategory *Ceremonial food and drink* which did not bring any results, but a parallel subcategory *Beverages* had a linked note on its introductory page saying: "*The main article for this category is* ***Beverages***". This category article was on a page of its own and it listed links to articles on different types of beverages including "Tea". In the article on tea, the Japanese tea ceremony was briefly mentioned , but without linking the term to the article "Japanese tea ceremony". A link to the article was finally found on the same page under the "See also" heading.

The first task was to find a definition for Japanese tea ceremony. In the introductory part of the article found, the Japanese tea ceremony is defined in the following way:

> "The **Japanese tea ceremony** (**chadō**, or **sadō**) is a traditional ritual influenced by Zen Buddhism in which powdered green tea, or matcha, is ceremonially prepared by a skilled practitioner and served to a small group of guests in a tranquil setting".

Some slight confusion is caused by the use of the terms as equivalents for the term 'Japanese tea ceremony' (see above) on the one hand, and for "the study or doctrine of the tea ceremony" on the other. This ambiguity is, however, something, which appears in other sources, too, and the hoped solution for it was not found here either even thought the right article was found.

Ambiguity or polysemy was also encountered in the case of the diving article, but on another level. In the main article "Diving", the term 'diving' did not, after all, refer to the concept, which I was looking for, not even to its superordinate concept. Actually, it was a co-ordinated concept, a fact clarified by a note on the page:

> "This article refers to the sport of jumping into water, often acrobatically. For swimming below the surface of the water, see underwater diving."[12]

When following this link, the "Underwater diving" article told me that

"**Underwater diving** refers to the practice of going underwater with or without breathing apparatus. [...] There are several types of underwater diving. [...] Scuba diving and surface supplied diving: swimming or walking underwater with breathing apparatus."[13]

The information provided in the article confirmed that the term 'scuba diving' was the equivalent of the Finnish term 'laitesukellus'. Following the link provided from the term 'scuba diving', an article was found which concentrated on this type of diving. In the article, however, no links or mention of the two further concepts I was looking for were found. Returning to the category Diving proved to be fruitful. There was a link to the article "Open-water diving", which confirmed that the Finnish term 'avovesisukellus' ("open water diving") can actually be translated with 'open-water diving' - it is not always certain that such a word for word equivalent is correct. As to the Finnish 'parisukellus' ("pair diving"), the task was slightly more difficult, since no direct translation of the term was to be found in the category Diving. It had, however links to the articles "Buddy check" and "Buddy system", both of which referred to the practices in diving in pairs. The nearest equivalent to the Finnish term proved to be 'buddy system', or 'buddy diving":

"When using the buddy system, the group dives together and co-operate with each other, so that they can help or rescue each other in the event of an emergency. […] With buddy diving, each of the divers is presumed to have a responsibility to the other. The "buddies" are expected to monitor each other […]" 14 In the article, the term 'buddy diving' is also used alternately with 'buddy system'. Thus all the tasks were performed and the desired information was found despite many dead ends.

4.3 Multiple paths

Once the relevant articles had been found, other possible pathways could be traced by following upwards the links provided at the bottom of every page. They lead to the superordinate categories. For both tasks, several alternative paths were found. Tracing upwards showed that there is a category called Tea, which has a subcategory Tea ceremony, the whole path being: Art and Culture – Food and drink – Beverages – Non-alcoholic beverages – Tea – Tea ceremony – Japanese tea ceremony (see Figure 3). It reflects the actual generic concept systems ("Beverages" and "Tea ceremony"), but for an information seeker it may be a bit too long. At least I did not even expect to find it and tried rather more associative paths at first. Another alternative and much shorter path was found when tracing upwards from the linked category Rituals via Culture to the uppermost category Society, which was not found in the beginning.

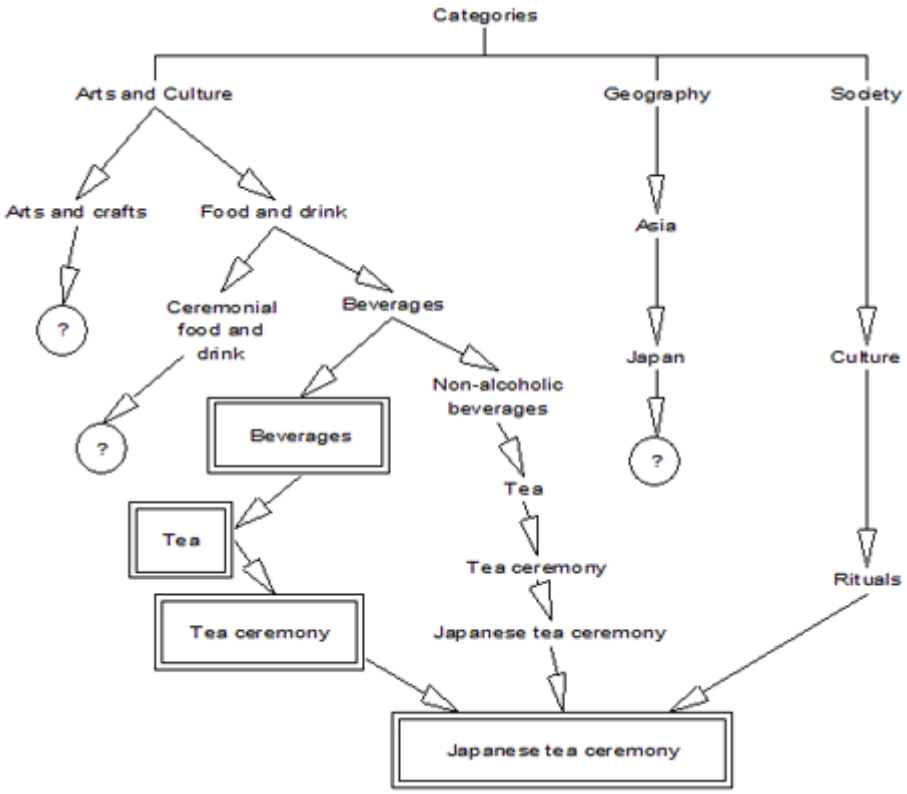

Figure 3. Paths to Japanese tea ceremony15

Tracking upwards showed that Diving could have been found through many different paths, the most obvious ones of which are included in Figure 4 together with the paths, which were found earlier.

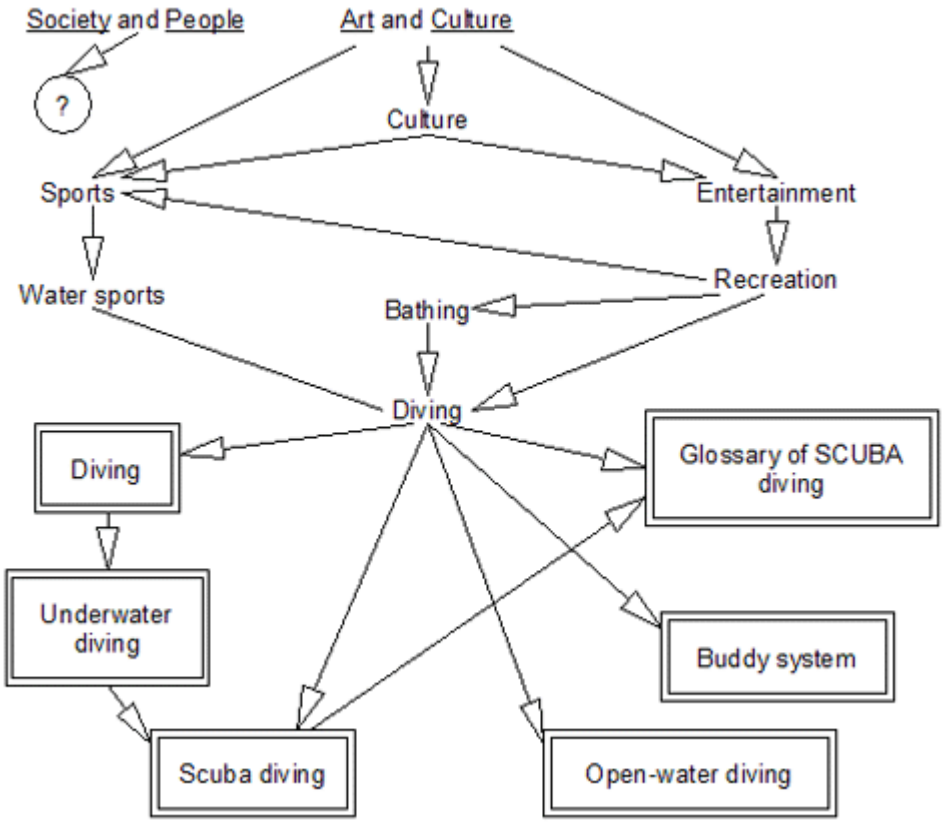

Figure 4. The paths to scuba diving

4.4 Opening new paths

As mentioned above, Wikipedia is dynamic in nature since new articles are added all the time and old ones updated and completed. Not only had the main categories been updated, but subcategories had also been changed somewhat after the visits in January/February and July. In my first visit, I was looking in vain for the article on tea ceremony via the path Geography: Asia: Japan. In November however, under the category Japan, there was a subcategory called Japanese Culture, and according to my earlier expectations, this path now leads to the subcategory Japanese tea ceremony. The classification of the higher categories in Wikipedia had undergone major reformulation, however, and the path had been lengthened (Geography and places: Countries: Countries by continent: Asian countries).

Some changes had also taken place in the article on scuba diving. Earlier it had no direct links to subtypes of scuba diving, but now on the third visit, it had an introductory section and direct links to some subtypes of scuba diving (e.g. recreational diving and technical diving) had been added. However, as before, the links to buddy system and open water diving were still missing.16

5 DISCUSSION

Collaborative online encyclopedias such as Wikipedia are revolutionising the genre of encyclopedia, and present an interesting object of study to terminologists, too. How to represent, organise, and find information in terminological glossaries and databases have been important topics for terminologists since the very beginning of the theory of terminology.

People often tend to think that all of knowledge can be organised in a single hierarchy, a comprehensive classification of all the phenomena in the universe, which makes up a summary of total human knowledge. At first sight, it may seem that Wikipedia categories follow strict hierarchic concept categories (see Table 2). Only the topmost categories are more or less mutually exclusive. Nevertheless, when we proceed to the categories at lower levels of abstraction, crossing paths will be found (see e.g. Figure 4). The same categories and articles may be found under several different categories. Instead of a single hierarchy, several overlapping ordering systems and concept systems interact". When browsing the categories, I even noticed cases in which a superordinate category became subordinate to its own subordinate category. However, Wikipedia instructions do not encourage this kind of looping. 17

Multiple entry points and overlapping hierarchies offer a smooth way to find the desired information. Different people have different points of view and start their search for the same item from different points of departure. In terminological analysis, we have also seen that the same concepts may belong to several different concept system types, e.g. generic, partitive, or functional concept systems. Instead of a hierarchy, the structure of a macro concept system rather forms a network of concepts (Nuopponen 1994). Despite the fact that the Wikipedia paths of categories were sometimes quite logical ones, including generic concept relations (e.g. Food and drink – Beverages – Non-alcoholic beverages – Tea – Tea ceremony – Japanese tea ceremony), they did not necessarily attract the information seeker as much as a supposedly shorter one with associative concept relations (e.g. here origination relation: Japan – Japanese culture – Japanese tea ceremony).

In addition to the hierarchical category classification, the English Wikipedia has an alphabetical listing of all the main and sub-categories, in which the categories Japanese tea ceremony and Diving could be found directly. It was also easy to find the relevant articles by means of the search function, e.g. by using the terms diving or diving apparatus, but the purpose was to navigate through the classification provided – after all, the classification exists for the purpose of assisting the information seeker. It must be added, however, that in both tasks I lost my way during the first efforts, and instead of finding the path along the categories to the target, I was able to navigate forwards by means of the articles and links in them. Thus the category classification system, together with the articles attached to them, formed a functioning retrieval tool: when categories did fail, articles helped me on and vice versa. In both cases, paths along categories were discovered afterwards when tracing the links to the upper categories from the articles. In a terminological hypertext system, all these four retrieval tools (categories, alphabetical lists, search, and links) would be needed to complement each other. In addition to these, Wikipedia has

other tools for information retrieval: overviews, featured content, lists of lists, glossaries, portals, and timelines. It also tries to apply other classification systems, e.g. the Library of Congress classification.18

When browsing Wikipedia's categories and looking for terminological information in the articles, an idea for further discussions and projects came up: There would certainly be a need for a collaborative terminological hypertext system similar to Wikipedia, where the lenghty encyclopedia articles would be replaced by entries with definitions and other terminological data compiled by means of terminological methods. The entries would be connected to each other via a subject category classification and links. Instead of building a hypertext system for terminology from scratch, would it be possible to publish terminological glossaries as part of Wikipedia's Wiktionary? Wikipedia and its sister projects as well as wiki applications offer versatile tools for collaboration and for publishing terminological glossaries online. Are these tools suitable for terminology work? Resources and copyrights are certainly an issue, but we could start with students' terminological projects and theses. These questions and ideas will be, however, left for future discussions and projects.

[1]Nowodiver's Glossary of scuba diving terms, http://www.nowodiver.net/en_ glossary.php
[2]http://dir.yahoo.com/
[3]http://encarta.msn.com/artcenter_0/Encyclopedia_Articles.html
[4]http://www.britannica.com/eb/subject
[5]Wikipedia; http://en.wikipedia.org/wiki/Wikipedia:Categorization (12.7.2006)
[6]http://www.britannica.com
[7]http://www.bartleby.com/65/
[8]http://meta.wikimedia.org/wiki/List_of_Wikipedias
[9]http://en.wiktionary.org/wiki/Main_Page
[10]http://en.wikipedia.org/wiki/Wikipedia:Categorical_index (28.11.2006)
[11]See Figure 4; dead ends are marked with a question mark.
[12]Wikipedia; http://en.wikipedia.org/wiki/Diving (7.2.2006)
[13]Wikipedia; http://en.wikipedia.org/wiki/Underwater_diving (7.2.2006)
[14]Wikipedia; http://en.wikipedia.org/wiki/Buddy_system (7.2.2006)
[15]Dead ends are marked with question marks and articles with double frames.
[16]Wikipedia; http://en.wikipedia.org/wiki/Scuba_diving (27.11.2006)
[17]Wikipedia; http://wikipedia/WikipediaCategorisation_FAQ.htm (12.7.2006),
http:// en.wikipedia.org/wiki/Wikipedia:Categorization (28.11.2006)
[18]Wikipedia; http://en.wikipedia.org/wiki/Wikipedia:Contents (11.12.2006)

REFERENCES

Lechner, Ulrike, Beat Schmid, Salome Schmid-Isler, Katarina Stanoevska-Slabeva (1999). Structuring and Systemizing Knowledge on the Internet - Realizing the Encyclopedia Concept as a Knowledge Medium. In Proc. Int. Conf. on Information Resource Management (IRMA'99). Available 28.11.2006: http://www. alexandria.unisg.ch/EXPORT/DL/10280.pdf

Nuopponen, Anita (1994). Begreppssystem för terminologisk analys. [Concept systems for terminological analysis] Acta Wasaensia. University of Vaasa, Vaasa.

Sui Zhifang, Gaoying Cui, Ding Wansong, Zhang Qinlong (2005). Domain Knowledge Engineering Based on Encyclopedias and the Web Text. The 5th Workshop on Asian Language Resources. Available 3.12.2006: http://www.cl.cs.titech.ac.jp/ALR/WS/5th/5thALR.pdf

Suman S., Debanshu Karmakar (2002). The Role of Library Classification in Organizing the Web. Available 28.11.2006: http://drtc.isibang.ac.in/~saiful/ greenstone/collect/Pdf/archives/HASH0193.dir/doc.pdf