

Knowledge organization and multilingual vocabularies

Jörn Sieglerschmidt, Bibliotheksservice-Zentrum Baden-Württemberg (BSZ)

Habet natura scientiarum thesauros innumerabiles,
qui nullis aetatibus exauriri possunt (Jean Bodin).¹

1. Introduction

Knowledge organization got its name in the context of enhancing the processing of information throughout an institution. In a networked environment knowledge organization should provide for the optimal allocation of information resources to the right person(s), at the right time(s) and places(s), in an expectable and understandable format. In the past knowledge organization was occupied with the classical topics of controlled vocabularies: classifications, thesauri, their theory, development, and usage. The topics have not changed dramatically, but the growing impact of the internet has shifted the focus somewhat to such topics as metadata standards, ontologies, semantic web etc. The politics and ethics of knowledge acquisition and distribution was and is a main topic, too. Here knowledge organization shows strong ties to the social sciences.

The expanding space of internet services has brought together the different language communities, but in most of the cases by neglecting the vernacular language of the internet user. English is most common and seldom perfectly spoken or understood by non-native speakers: so there are modern variants (beside American English e. g. continental English) and dialects (e. g. conference pidgin). Automatic translation produces funny results normally; it works only in very specific environments with a basic vocabulary of around 500 words (e. g. most sciences and applied sciences). So there is a strong need of processing information on multilingual platforms for the cultural heritage (ch) domain. There are many endeavors to meet these needs, but none is convincing until now. Therefore several possible solutions will be discussed below.

Classification is the basis of every naming activity, because giving a thing (nomina), a property (adiectiva), or an action (verba) a name makes it necessary to define the name in relation to the semantic and syntactic environment. Classification is not a basic human need, as one might be inclined to anthropologize this form of intellectual behaviour, but if men or animals use language, they must discriminate the meanings of the signs processed. Thus classification is a transcendental prerequisite of sign processing rather than an innate ability,² although famous scholars like e. g. Noam Chomsky have pleaded for mentalism.

Controlled vocabulary of any sort reflects the intellectual horizon of a certain time, a certain place, certain individuals. So the problem of controlled vocabulary is on the one hand the permanent change of things/thinking and the language expressing it, on the other hand the description of sometimes very specialized collections. This pro-

¹ The nature of knowledge encompasses innumerable thesauri which can at no time be exhausted. Bodinus, Ioannes: *Methodus ad facilem historiarum cognitionem: accurate denuo recusus*. Lyon: Ioannes Mareschallus 1583, 310; this reference relates to the medieval and early modern debates about cosmology and the limited duration of the world: cf. Weichenhan, Michael: 'Ergo perit coelum...'. *Die Supernova des Jahres 1572 und die Überwindung der aristotelischen Kosmologie*. (Boethius 49) Stuttgart: Steiner 2004, here pp. 77-130. The French theologian and philosopher Jean Bodin developed the theory of sovereignty of the state.

² Batley, Sue: *Classification in theory and practice*. Oxford/New Hampshire: Chandos 2005, 1.

blem multiplies, of course, if multilingual vocabulary is at stake. Controlled vocabulary must be evaluated therefore according to its ability to be flexible, to accommodate new subjects, and to meet specialized and/ or local needs.³ Furthermore: different user communities can have very different needs, especially regarding the depth of hierarchy and detail. And it should be borne in mind, that vocabulary can lose its strength of organizing knowledge, if the professional background – the artistry of concept formation - lacks.

2. Automatic indexing vs. controlled vocabulary

2.1. Automatic indexing

Big amounts of information like e. g. union catalogues, internet sites, databases are prepared for purposes of retrieval by an indexing machine. These machines read the whole text material using the usual separators to produce a list of single words. Certain words like articles and the like are removed. Almost all indexing machines can display stemming devices, i. e. linguistic procedures to reduce inflected forms of words to their stem form respectively, so that different inflected forms of the same word are not treated as different words. Nevertheless multilingual linguistic features are not available by now.

Most of the indexing machines use ranking procedures, in order to make a difference between more and less important parts of the information. The ranking procedures normally use not very sophisticated statistical measures. Nevertheless the ranking procedures produce an useful relevance index.

The big advantage of automatic indexing is the simple fact, that it can be done fast, without any intellectual endeavour (leaving aside the fact that the programming of the procedure is – hopefully - an intellectual endeavour), and for an almost unlimited amount of text. The recall of a search is therefore very high, but the precision very low. Proper names, homonyms, synonyms and the like cannot be identified. The precision can be enhanced by using and decoding metadata standards, in order to mark up certain parts of the information as more important than others. But this makes a difference only for the ranking of the results. The inclusion of controlled vocabulary during the retrieval process increases the recall by including e. g. synonyms of the searched term, but doesn't raise the precision.

2.2. Controlled vocabulary

Thesauri and classifications have a long historical tradition, if one includes encyclopedias, dictionaries and the like as similar forms of knowledge organization.⁴ The following remarks will concentrate on some actual examples of thesauri and classifications with multilingual extensions.

The big advantage of controlled vocabulary: it solves the problems of synonymy, homonymy and conceptual hierarchy. Furthermore a unique identifier provides for a precise reference. So the precision of the search results is very high. The disadvantages of such controlled vocabulary are:

- high amount of intellectual work must be done,
- big institutions (e. g. national libraries) must guarantee the maintenance and development of the vocabulary,

³ *ibid.*, 94f., 107f.

⁴ See the newest account on alphabetical ordering by Küster, Marc Wilhelm: *Geordnetes Weltbild. Die Tradition des alphabetischen Sortierens von der Keilschrift bis zur EDV. Eine Kulturgeschichte.* Tübingen: Niemeyer 2006.

- less flexibility regarding new branches of knowledge,
- (normally) the payment of license fees.

The necessity of the institutional background must be stressed. The example of the Universal Decimal Classification (UDC) shows, that the lack or a weakening of institutional support endangers the whole endeavour. Since the Deutsches Institut für Normung (DIN), the German institution maintaining and developing standards, has removed from the development of the German UDC, the systematic development of the German UDC stopped.⁵

Most discussions on controlled vocabulary are devoted to subject matters (what, how) leaving aside other facets of knowledge: proper names of natural persons, legal bodies (who), places (where), and times (when). In the big thesauri of national libraries or some classifications those are included. But the big vocabularies in the different areas are far more elaborated.

Place names are available as authority files in different forms, e. g. geographic or administrative, and different amount of completeness. The only authority file, that tries to combine all relevant information, is the Thesaurus of Geographic Names (TGN). The TGN is supported by the Getty Foundation. All place names included in the TGN cover administrative (in Germany administrative units down to the Landkreise, comparable to counties), geographic (at least the coordinates), in many cases historical information. The vernacular name (beside the American) and the synonyms of other languages (e. g. Firenze, Florence, Florenz) are indicated. So it seems reasonable to concentrate international efforts on the further development of the TGN, especially since some regions of the world are not represented in much detail.

There are many national authority files comprising names of persons and/ or legal bodies. But there exists no database combining the informations from different countries. Even the special area of artist's names, e. g. Union List of Artist Names (Getty Foundation) or Allgemeines Künstlerlexikon, doesn't know such efforts. The national libraries hold big authority files of persons. They exchange title information and with it informations about creators, publishers etc. These institutions should cooperate, in order to produce a common authority file.

Time specifications are very important for the search on ch sites. Nevertheless the use of clear data formats is not very common in the museum community. Only if data input deliver precise data about times and periods (of production, of an event etc.) functionalities of time search, e. g. visual helps like a time bar, could be displayed.

2.2.1. Thesauri

Thesauri are in general very flexible compared with classifications regarding the integration of new knowledge or of whole domains of knowledge. They are very rich in detail and very apt for the ch domain.

2.1.1.1. Subject headings of the national libraries

The endeavours to coordinate the subject headings of the British Library with those of the Bibliothèque National de France and of the German National Library are pending. The MACS project has ended with a fragment of the planned result. The Criss Cross project concentrates on the integration of the DDC and the authority files of the national libraries respectively. The planned multilingual tool for the description of ch objects is not in sight.

⁵ The last, abridged edition based on the *Classification décimale universelle* of the years 1927 until 1933 was published 1967: *Dezimalklassifikation. DK-Handausgabe. Internationale mittlere Ausgabe der universellen Dezimalklassifikation.* 2 vol. Berlin/ Köln: Beuth 1967.

2.1.1.2. UNESCO thesaurus

The UNESCO thesaurus covers all branches of knowledge, but with about 7500 descriptors on a very abstract level. The parts of the thesaurus relevant for ch will be translated to many European languages as one of the deliverables of the MICHAEL^{plus} project. As a rough tool of orientation the UNESCO thesaurus is very useful. It should therefore be considered as one source in all efforts of constructing a multilingual vocabulary.

2.1.1.3. Eurovoc

The eurovoc thesaurus has been developed by the European Union and is available in the 21 official languages of the EU. It comprises nearly 7000 descriptors and in some languages more than 10000 non-descriptors. The topics covered are: politics, international relations, European Communities, law, economics, trade, finance, social questions, education and communications, science, business and competition, employment and working conditions, transport, environment, agriculture, forestry and fisheries, agri-foodstuffs, production, technology and research, energy, industry, geography, international organizations. As expectable the focus of the thesaurus lies on economic, technological and administrative affairs. Ch topics are much less represented than in the UNESCO thesaurus and listed under the top term *social affairs*. Nevertheless the thesaurus should be consulted for the structuring of multilingual vocabulary.

2.2.2. Classifications

There are many multilingual and monolingual classifications covering many special branches of knowledge. Here it must be sufficient to discuss two universal classifications claiming to represent all branches of knowledge.

2.2.2.1. Dewey Decimal Classification (DDC)

The Dewey Decimal Classification is the oldest classification scheme, first published in 1876, and is hosted now by Online Computer Library Center (OCLC). The DDC has been translated in many languages and is used by many library institutions worldwide, among them many national libraries like the German. Beside the knowledge organization the main aim of the DDC is an aid for shelving books. The DDC shows very strongly the cultural restrictions of its origin. Fundamental changes of the classification scheme had to be avoided because of the many users, who expected a continuity of the scheme. The DDC allows classification on a more generic level and is a bit inflexible regarding the combination of descriptors. The DDC should be consulted in every multilingual vocabulary project, but it is not very appropriate for purposes of specialized ch collections.

2.2.2.2. Universal Decimal Classification (UDC)

A younger offspring of the DDC, first published 1905-1907 and in a much extended version 1927-1933 as *Classification Décimale universelle*, is the UDC. The UDC was not conceived as shelving aid in first line, but as documentation aid. It comprises features of faceted classifications.⁶ The flexibility is possible by using linking signs (+, /, :) and auxiliary tables that meet the needs of the ch domain by combining the main classes and subclasses with extensions giving information about material, persons, places, nationality, form, language, time.⁷ The UDC is available in many languages, but mostly not complete. The new *master reference file* (MRF) is based on the English middle – abridged – edition, comprises some 66000 numbers, and is

⁶ Batley 2005 (see fn. 2), 81f.

⁷ *ibid.*, 92-101.

available in English, although English and French terms are available for parts of the MRF. Revisions of the structure of the UDC are discussed, but not yet realized for the same reasons as noted above for the DDC.⁸ The UDC Consortium maintains and develops the UDC.

3. Ontologies and topic maps

3.1. Ontologies

Ontologies are discussed more frequently, since the semantic web claims to develop automatic devices for the linking of services in the internet. Beside topic maps ontologies are very apt to make the structure of knowledge transparent. It is not necessary to write more about the CRM here,⁹ but it should be clear, that efforts to develop multilingual vocabularies should use the CRM structure as a basis. That all types of controlled vocabulary should conform to the scheme of the simple knowledge organization system (SKOS) is clear likewise.

The screenshot shows the 'Associative Search in LEWI' interface in Mozilla Firefox. The search term 'genetic engineering' is entered in the search box. On the left, a topic map (Aquadrowser) is displayed with 'genetic engineering' at the center, connected to various related terms like 'bioengineered', 'herbicide', 'application', 'engineering', 'food', 'engeneering', 'cloning', 'gentechnologie', 'issue', 'genetik', 'animal', 'science', 'plant', 'biotechnology', 'Genetik', 'génétiqne', 'risk', 'resistance', and 'genetisch'. The search results on the right show a list of books with details such as title, author, year, found terms, shelf code, and location. The interface also includes navigation links like 'Home', 'New', 'Contact', 'Topics', and 'DEUTSCH'.

Fig. 1: Associative search with Aquabrowser

⁸ cf. McIlwaine, I.C.: UDC in the twenty-first century. In: Marcella, Rita/ Maltby, Arthur (eds.): The future of classification. Aldershot: Ashgate 2000, 93-104.

⁹ See my contribution to the Gothenburg meeting last year:

3.2. Topic maps

Topic maps are a very interesting device of combining possibly the intellectual work of ontologies, of controlled vocabulary, its common development – or that of folksonomies -, and the techniques of indexing. Topic maps are important as ontologies for the further development of the semantic web, because they define the structure of knowledge processed on the internet. Ontologies and topic maps will help to enhance the retrieval of relevant resources. A practical example is the search function of a German library specialized on issues of ethics in the sciences (with emphasis on life sciences):¹⁰. On the right side one can find the facets offered, on the left side there are a topic map that can be changed dynamically by the user. After such changes a new list of literature – and a new semantic net - will appear.

4. WordNet

A wordnet serves word sense disambiguation.¹¹ It is derived from text corpora and dictionaries. A wordnet consists of synsets, that include one concept and its semantic space defined by the following relations: synonymy (different graphemes, same concept), homonymy (same term, different concepts), antonymy (opposite concepts), hyponymie (subterm), hyperonymie (generic term), meronymy (whole-part-relation), implication (follow-up-relation), causation (causal relationship), association (nearness of concepts). GermaNet comprises about 53000 synsets, among them 38000 nouns, 9000 verbs, and 5500 adjectives.

EuroWordNet has developed a model of the most important and frequent concepts of eight European languages (English, Spanish, Netherlands, Italian, French, German, Czech, Estonian). The interlingual index (ILI) serves as language independent component linking and coordinating the different monolingual wordnets and their synsets respectively.

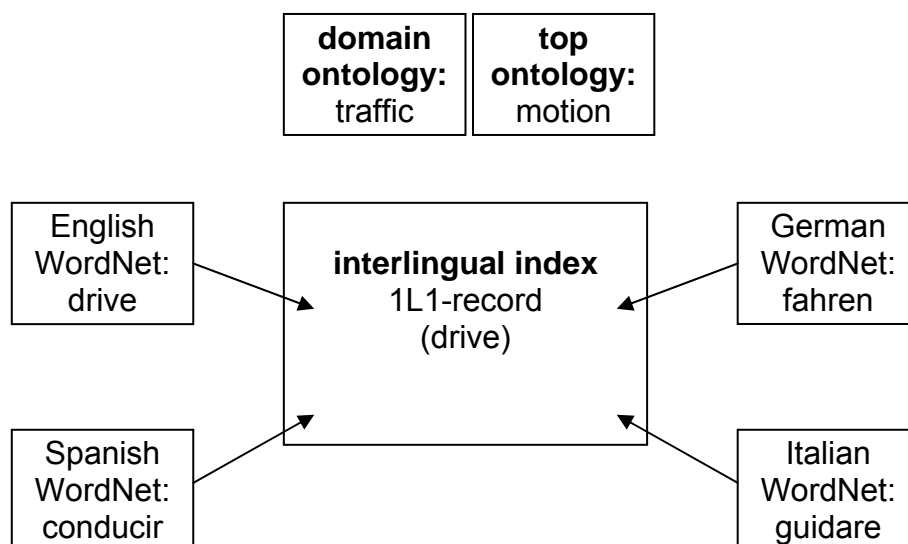


Fig. 2 EuroWordNet-architecture¹

¹⁰ <http://www.izew.uni-tuebingen.de/lewi/assoziativ_en.html>


¹¹ Kunze, Claudia: Lexikalisch-semantische Wortnetze in Sprachwissenschaft und Sprachtechnologie. In: Information Wissenschaft Praxis 57 (2006), 309-314.

Wordnets seem to be very apt to model and structure multilingual vocabulary. There are no application for the ch domain by now.

5. Folksonomies

There are many projects and efforts to build up folksonomies in the last years, since the Wikipedia has begun to develop procedures of the common and cooperative development of vocabularies. Whereas the community of documentation professionals looked down on these lay endeavors in the beginnings, the attitude has changed since, because the community of contributors to the Wikipedia has shown a remarkable skill. Furthermore the Wikipedia offers the most multilingual vocabulary platform available by now. Although the refinement of articles is very different in the different languages, every language variant offers a translation of a descriptor, perhaps much more information about the concepts linked with the descriptor, partly an elaborated taxonomy.

The following image shows the article about furniture in the Wikipedia. The German version categorizes the concept of furniture in four areas: style, material, function, construction. The newest German thesaurus of furniture is cited. The article is available in 23 languages: Afrikaans, العربية, Български, Brezhoneg, Català, Český, Dansk, Deutsch, English, Esperanto, Español, Suomi, Français, Gàidhlig, Ido, Íslenska, Italiano, Македонски, Nederlands, Polski, Português, Română, Русский, Simple English, Slovenščina, Svenska, Türkçe, Українська.



WIKIPEDIA
Die freie Enzyklopädie

Navigation

- [Hauptseite](#)
- [Über Wikipedia](#)
- [Themenportale](#)
- [Von A bis Z](#)
- [Zufälliger Artikel](#)

Mitmachen

- [Hilfe](#)
- [Autorenportal](#)
- [Letzte Änderungen](#)
- [Spenden](#)

Suche

Werkzeuge

- [Links auf diese Seite](#)
- [Änderungen an verlinkten Seiten](#)
- [Hochladen](#)
- [Spezialseiten](#)
- [Druckversion](#)
- [Permanenlink](#)
- [Artikel zitieren](#)

Andere Sprachen

- [Afrikaans](#)
- العربية
- Български
- Brezhoneg
- Català
- Český
- Dansk
- ?????
- English
- Esperanto
- Español
- Suomi
- Français


Möbel

Dieser Artikel oder Abschnitt weist folgende inhaltlich problematische Lücken auf: *Geschichtliches, bisher nur Kurzdefinition mit Listen*
 Hilf Wikipedia, indem du die fehlenden Informationen [recherchierst](#) und [einfügst!](#)


Der Begriff **Möbel** bzw. Mobiliar (sächlich) (von lat. *mobilis* = beweglich; im Gegensatz zu unbeweglichen Dingen = *Immobilien*) ist der Oberbegriff für Einrichtungsgegenstände in *Wohnungen*, Geschäften oder *Bürräumen*, aber auch im Außenbereich. Ein Möbelstück ist zweckgebunden und dient der Aufnahme von Gegenständen, dem Verrichten von Tätigkeiten, dem Sitzen oder Liegen. Die Einteilung in bestimmte Möbelgruppen ist nicht immer eindeutig und kann nach verschiedenen Kriterien erfolgen.

So können sie nach folgenden Kriterien katalogisiert werden:


- nach der **Stilrichtung**:
 - [Postmoderne](#)
 - [Bauhaus](#)
 - [Jugendstil](#)
 - [Historismus](#)
 - [Gründerzeit](#)
 - [Biedermeier](#)
 - [Barock](#)
 - [Renaissance](#)
 - [Gotik](#)
 - [Romanik](#)
- nach dem verwendeten **Material**:
 - [Holzmöbel](#)
 - [Plastikmöbel](#)
 - [Pappmöbel](#)
 - [Metallmöbel](#)
 - [Korbmöbel](#)
 - [Polstermöbel](#)
 - [Weichmöbel](#)
- nach ihrer **Funktion**:
 - [Behältnismöbel](#)
 - [Schrank](#)
 - [Wandschrank](#)
 - [Sitzmöbel](#)
 - [Arbeitsfläche \(Tisch\)](#)
 - [Kleinformöbel](#)
 - [Liegemöbel](#)
 - [Ambient-Möbel](#)
- nach ihrer **Konstruktion**:



Behältnismöbel Schrank
(Gesellenstück 2006, Hamburg)



Sitzmöbel Bank (Gesellenstück 2005, Hamburg)



Typische Papphocker; Deutscher Evangelischer Kirchentag Hannover 2005

It might be reasonable to cooperate with the Wikipedia in order to use the vocabulary work of so many people for the multilingual access to ch resources, because the furniture is not the only relevant example.

6. The problem of globally unique identifiers

If vocabulary will be merged, in order to build up common authority files, each term must be identified by a unique alphanumerical string. Most big controlled vocabularies have such identifiers for each term, so a matching between two e. g. place name authority files should produce a table paralleling the identifiers of the two files in two columns. If common vocabularies are an aim, there is a need for common registries providing the unambiguity of data, holding and administering it.

7. Summary

As has been shown there are many ways to get to the same destination. Some are very long, but convincing with regard to the needs of ch domain. None of the offers is so complete, that it could be brought into practical action at this moment. So many work has to be done. Even the biggest European project in the ch domain – *EDLnet* – has found no easy way to multilingual vocabularies. Hopefully the EU funds projects that will bring practical results – as soon as possible. It might be expected that European projects like *MultiMatch* and *STITCH* are able to deliver such results.