

# An ontology based semantic library catalogue

Dariusz Daćko, Joanna Józefowska, Agnieszka Ławrynowicz

Poznań University of Technology  
ul. Piotrowo 2, 60-965 Poznań  
tarest@poczta.fm, jjozefowska,alawrynowicz@cs.put.poznan.pl

## Abstract

In this paper we present an ontology created for the controlled vocabulary used in library catalogues. The vocabulary as well as the rules underlying the vocabulary creation are shortly described. The resulting ontology is presented. The vocabulary is used to describe the books in traditional and electronic subject catalogues that enable search for books on a given subject. Application of the developed ontology makes it possible to create a semantic search engine with high quality of search. Some ideas for future applications are provided. The research is based on the controlled vocabulary KABA used in Polish scientific libraries, although it seems that similar results may be obtained for any vocabulary complying with LCSH.

## 1. Introduction

Each book in a library catalogue is described by one or several subject headings that determine the book subject. For example, the book entitled “No lights, no sirens”, which is a biography of a New York policeman is described by three following headings:

Cea, Robert (1961- ) -- Diaries  
Police corruption -- United States  
Police brutality -- United States

since it describes corruption and violence in police forces. The first heading occurs because Robert Cea is a well recognized person. All subject headings are indexed in the subject access vocabulary. Therefore finding all books on a given subject, e.g. police corruption is possible. It is sufficient to find a card in the subject catalogue with the desired heading and see the cards of all books following the subject heading. However, there is a problem with using such catalogue. How can we find all books about policemen? If we just look for the appropriate heading (“Policemen”) we find a list of books concerning policemen in general. We do not find subjects like training of policemen, women in police forces, police professional ethics including violence and corruption in police forces, because these subjects have separate headings. On the list of books under the heading “Policemen” we do not find the famous biography of Robert Cea, either. In order to find it we need to search all subject headings which are narrower than “Policemen” and search the lists of books under each heading. This may be time consuming. The first problem that we encounter is determining all subject headings narrower than the heading under consideration. We start with all headings that are alphabetically close to the considered heading. Some of them may be narrower, like e.g. “Policemen -- Training of” or “Policemen -- Professional ethics”. If the narrower headings are distant in the alphabet from the considered heading then its hyponyms are explicitly provided, i.e. they are listed in the subject heading card. Another problem is to browse the lists of book titles corresponding to the narrower headings since they may often contain duplicates. The procedure has to be repeated if any heading has narrower headings itself and thus it is usually time consuming. It would be less annoying if the search was performed by a computer. This is possible if the subject access vocabulary is organized in an ontology. The ontology can be used to

find (recursively) all headings narrower than the heading interesting for the user. Next, the lists of books described by the original heading as well as the narrower headings are combined into a single list without duplicates. The final list is presented to the user. Moreover, the list may be ordered according to the decreasing frequency of being borrowed in the past. The user automatically obtains the requested list of books starting with the most popular ones, which are most likely also the most interesting ones for the user.

Developing an ontology for the subject access vocabulary is not straightforward. As we have mentioned above, some hypernyms and hyponyms are explicitly given in the subject heading card (record). Others have to be found automatically. Subject headings are created according to some rules following from the syntax of the subject heading language (SHL). Most of the hypernyms and hyponyms can be found with high level of certainty by exploring these rules.

In the following sections we present the subject heading language and derive rules that can be used to create implicit hierarchical relations. Next, we propose the ontology with an API and discuss its possible applications. Conclusions from the performed research close the paper.

It should be noted that currently available subject catalogues do not differ from traditional card catalogues. They use different data carrier but the format and the way of using the data remain unchanged. The new carrier caused some terminology changes: in traditional catalogues subject headings and bibliographic descriptions of books are stored on catalogue cards, in computer databases they are stored in records. Further on we use the terminology relevant to Online Public Access Catalogue (OPAC).

The presented research is based on the subject access vocabulary KABA [1] which is commonly used in Polish libraries. The structure of the vocabulary is similar as for other vocabularies of this type used all over the world which means that it is consistent with the Library of Congress Subject Headings (LCSH). Consistency means that the results of the presented research may be useful also in other national languages.

Although further in this paper, the examples are provided in English we still analyse the KABA vocabulary.

## 2. Subject heading language

Subject headings belong to controlled vocabulary and their structure is consistent with the subject heading language [2]. Controlled vocabulary schemes mandate the usage of predefined, authorised terms that have been preselected by the designer of the controlled vocabulary as opposed to natural language vocabularies where there is no restriction on the vocabulary that can be used. Controlled vocabularies ensure that each concept is described by only one authorised term and each authorised term in the controlled vocabulary describes only one concept. If a word has more than one meaning in the natural language it should be uniquely defined by its domain added in parenthesis. For example in LCSH vocabulary we have the heading “Cancer” for a disease, “Cancer (Crustacea)” for a genus of marine crabs and “Cancer (Astrology)” for the constellation.

If a heading can be represented by other terms as well, these terms are given as optional (so-called rejected terms) which are also written in the heading record. For example, an option of the heading “Cancer” is “Malignant tumors”.

The heading record contains also the following information:

- heading type – topical names (common nouns), place names (geographical names), corporate names, personal names, event names and publication titles (topical and place names constitute the majority),
- heading description in a natural language,
- hypernyms and hyponyms which cannot be inferred from the syntax of the subject headings,
- semantic relations to other headings (other than hypernymies),
- synonym heading in LCSH.

The subject heading language sets the rules of constructing the subject headings, both accepted and rejected ones. The subject heading consists of the subject and, optionally, a subdivision specifying the subject meaning. General (topical names), geographical or chronological subdivisions are distinguished. Subjects as well as subdivisions may have additional comments explaining the domain of the subject, if it is not obvious. An additional comment may be a topical or a geographical name. Below we present an example of a subject heading with two subdivisions (geographical and chronological ones).

Evidence (Criminal law) -- Poland -- 19 century

The heading relates to the evidence used in crime investigations in Poland in the 19<sup>th</sup> century.

Another example is a subject heading with one general subdivision:

Policemen -- Training of

The heading relates to the training of police forces.

The subjects and subdivisions used in the headings should be accepted terms, while additional comments can also be rejected terms. Several general subdivisions, while at most one geographical<sup>1</sup> and one chronological subdivision may appear in a heading.

<sup>1</sup> For the purpose of the research the KABA language is simplified. Originally, geographical subdivisions are recorded

The subject itself may sometimes represent a relation between two other subjects. For example a heading “Police and press” represents relations between the police and the press. Dependency headings consist of two headings linked by conjunction “and”, e.g. “Police” and “Press”.

Subjects and subdivisions may consist of multiple words. Often they contain a noun followed by one or more attributes, e.g. “Authors, Polish” (“Polish authors”), “Police, Military” (“Military police”), “Science, Political” (“Political science”). The order of terms is important and assumes that the first term is described by the following ones. This assumption has implications on the alphabetic catalogue as well as on the way of searching for implicit hypernyms.

## 3. Sources of hypernyms which are not explicitly given

As we have mentioned in the Introduction not all hyponyms and hypernyms appear directly in the heading record. Usually relations that can be inferred from the subject heading language are not explicitly recorded. The following relations follow from the SHL:

- a) Heading consisting of a subject and subdivisions is a hyponym of any heading obtained as its subsequence. For example the heading “Evidence (Criminal law) -- Poland” has two hypernyms: “Evidence (Criminal law)” and “Poland”. The heading “Jews -- Bavaria (Germany) -- History” also has two direct hypernyms: “Jews -- History” and “Bavaria (Germany) -- History”, and three following from the transitivity: “Jews”, “Bavaria (Germany)” and “History”. “Jews -- Bavaria (Germany)” is not a hypernym, because such heading does not appear in the vocabulary.
- b) A simple heading (subject without subdivisions) containing additional comments is a hyponym of a heading consisting of the additional comment only. For example the heading “Evidence (Criminal law)” has a hypernym “Criminal law”.
- c) Dependency heading is a hyponym of its components. For example “Police and press” has two hypernyms: “Police” and “Press”.
- d) Simple headings starting with another heading are *almost always* its hyponyms. For example the heading “Authors” has hyponyms: “Authors, Polish”, “Authors, English” etc. The heading “Aerodynamics” has two hyponyms: “Aerodynamics, Supersonic”, “Aerodynamics, Transonic” and also by transitivity: “Aerodynamics, Hypersonic”.

The last rule is a heuristic one, but taking into account the formal nature of the SHL it is almost always correct. The above rules may be applied to accepted as well as

---

as two geographical entries. The first entry is the name of the state and the second one is the name of the geographical object. Two entries are useful for alphabetical ordering of headings. In the semantic browser alphabetic grouping should be replaced by semantic grouping. Therefore the record can be simplified and instead of writing „Churches -- Poland -- Poznań” we can write „Churches -- Poznań (Poland)”. It is convenient, since the heading „Poznań (Poland)” appears in the vocabulary, while „Poland -- Poznań” does not.

rejected terms, although the latter concept has not been exhaustively tested, yet.

### **Additional comments**

Additional comments may belong to accepted as well as rejected terms. They may also appear in different grammatical number than the original heading. Therefore, in order to recognize an additional comment it should be searched in both grammatical numbers of the heading.

The change of the grammatical number is performed using the vocabulary method. First, every word from combined Polish language vocabularies *ispell* and *aspell* is recognized by a morphological analyzer SAM [4, 5]. Next, nouns in plural nominative form are selected. With each noun SAM associates its singular nominative form. In this way we obtain a mapping of the set of plural forms onto the set of singular forms of the considered nouns, which enables the appropriate change of the grammatical number of ordinary nouns. Additional comment being rare nouns or noun expressions are not recognized by the system.

Finding hypernyms of the additional comments is particularly important for headings which are instances of concepts, for example for Indian tribes such as Apache or Sioux. In such case hypernymies are not given explicitly, because there are too many of them.

### **Errors and ambiguities**

During the process of importing the KABA vocabulary and constructing the hypernymies numerous errors in the records were found. These errors are currently being corrected by NUKAT – the management center for KABA.

For numerous reasons hypernyms of some headings could not be determined or (more rarely) were determined incorrectly. The first reason is that NUKAT sometimes accepts additional comments which are not elements of the controlled vocabulary, but more arbitrary descriptions. In order to increase the ontology potential these comments can be substituted by proper headings. The second reason of ambiguities is that comments are sometimes ambiguous since they point to headings appearing in the vocabulary in both grammatical numbers. In such situation it is not clear with which form the comment should be associated. The third reason is that some headings are not completely consistent with the considered SHL.

Errors and ambiguities follow sometimes also from the shortcomings of the algorithms used to create implicit hypernymies. The algorithms can be improved for example by implementing declination of ordinary noun expressions (using the vocabulary method) and rare nouns (using the non-vocabulary method).

### **Possible improvements of the algorithms**

In addition to the improvement mentioned above the following improvements of the proposed approach may be applied:

a) We analyse only accepted terms. Hypernymies may be created also on the basis of rejected ones.

b) An element from a complex heading can be substituted by its hypernym, instead of being rejected. For example the heading “Jews -- Bavaria (Germany)” has also the hypernym “Jews -- Germany” and, by transitivity, “Jews -- Europe, Central”.

c) Associating a simple heading (without subdivisions) with headings that starts with this heading may lead to a semantic error. For example the heading “Architecture” is incorrectly classified as hypernym of the heading “Architecture, Computer”. Such errors may be detected by the users and reported to the vocabulary management centre. Since the set of the controlled vocabulary is relatively small such approach is quite reasonable.

Some hypernyms may only be found using an additional ontology. For example information about hyponyms of the heading “Metals” which are particular metals (e.g. copper or iron) is stored in a record as the following text: “see also names of particular metals”. There are about 4000 such cases in the considered vocabulary. Nonetheless, such hypernymies could be added manually to the vocabulary as explicit ones or as additional comments for headings of particular metals.

## **4. The ontology**

The subject heading vocabulary contains about 112 thousand of headings, including about 52 thousand topical names (together with names expanded by a subdivisions), 19 thousand geographical names, 10 thousand corporate names, 27 thousand personal names, 500 event names and 2 thousand titles of the most important publications.

The headings are related by about 70 thousand hypernymies. Using the presented algorithms additional 70 thousand hypernymies, which do not appear explicitly in the vocabulary were found. A part of them follows probably from the transitivity and they should not be added to the vocabulary. Removing superfluous relations has not been implemented yet. Using additional sources of hierarchy will result in new hypernymies. It should be noted that some sources of hypernymies have been used only for topical and geographical names.

A heading is created only if there exists a book on particular subject. This means that the vocabulary represents the current state of knowledge. Every subject heading is important enough that at least a part of a book can be written on the subject.

The headings describe general subjects, e.g. “Social sciences” as well as detailed ones, e.g. “Lie detectors”. Detailed headings should be connected by hierarchical relations via middle headings to general headings. In Figure 1 we present about 50 headings related to the heading “Police” and its hypernyms with links up to the general heading “Social sciences”.

## **5. Proposed applications of the ontology**

### **Ontology browser**

As we showed in the Introduction, determining, location and browsing the content of all hyponyms of a heading is very time consuming. It seems that presentation of all hyponyms (explicit and implicit ones) as hyperlinks would make the search more convenient. Numerous

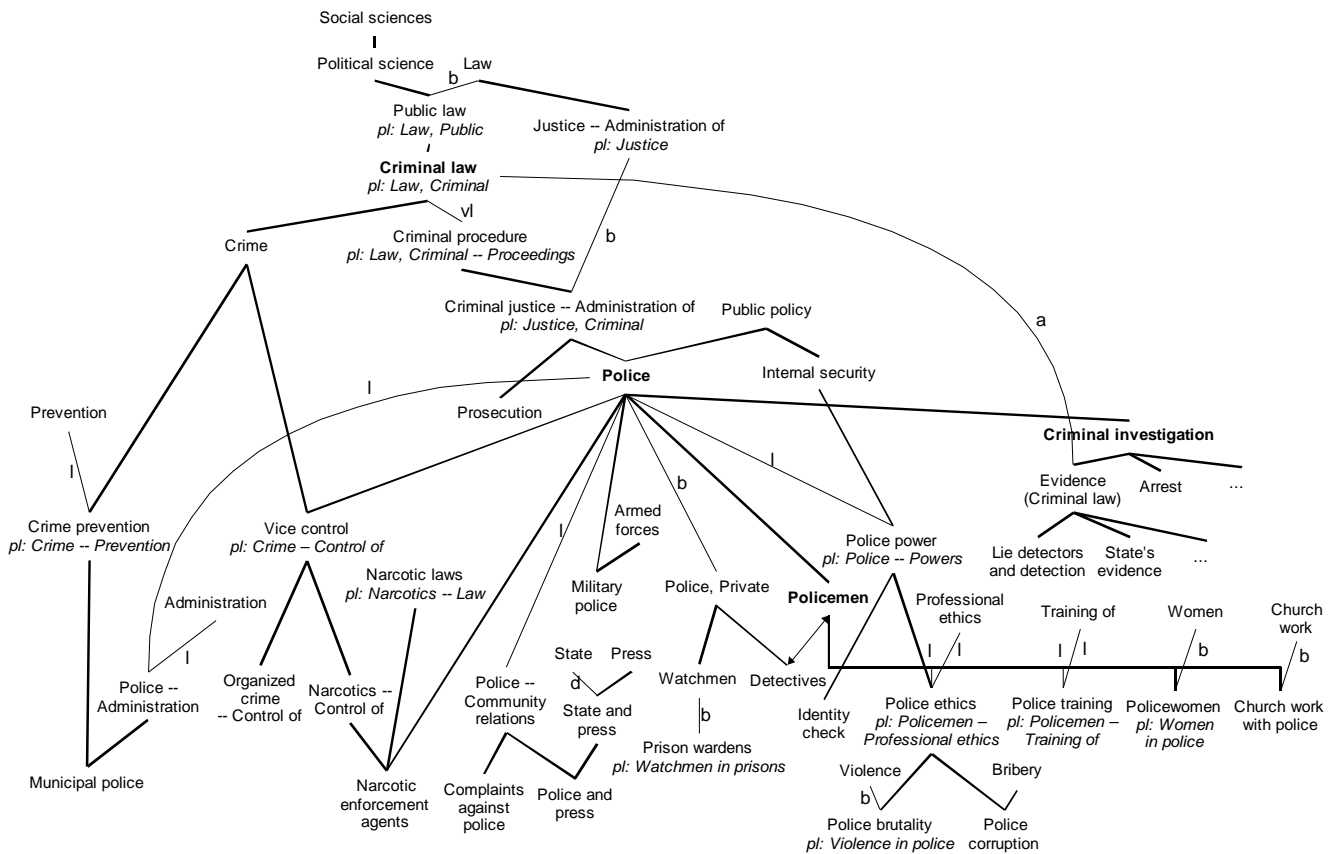


Figure 1. A fragment of the KABA vocabulary with about 50 headings related to the concept “Police”. All broader terms of the heading “Police” are shown.

The headings which are main concepts are shown in boldface. The heading names are given in LCSH. Their structure differs sometimes from the original one, so we also provide the Polish structure of the heading (in italic). This format should be used by the reader while analysing implicit relations between headings. Dots denote that more headings were available, however they are not important for the presented example.

Hypernymies given explicitly in the heading records are shown in boldface. An example of a semantic relation between the headings “Policemen” and “Detectives” is also presented. Semantic relations are not yet used in the system. Implicit hypernymies are marked with thin labelled lines. The label denotes the source of the relation: l – based on lexemes (subjects or subdivisions), a – based on additional comments, d – based on dependency names, b – based on the beginnings of the headings, vx – based on a rejected term (e.g. vx – based on rejected term lexemes).

Almost all relations following from the transitivity have been removed. It is not clear if it increases or decreases the clarity of the ontology browser. Only one relation following from the additional comment is presented in order to illustrate this source of relations.

hyponyms of a given heading could be grouped semantically, for example according to the type of the heading or to the heading domain (in case of headings possessing another, different hypernym).

In addition to hierarchical relations also other semantic relations should be displayed as hyperlinks.

### Book search engine

Applying an ontology in a book search engine has been described in the Introduction. This Subsection presents the search engine in more detail.

In order to speed up the search process we determine and store in a database all hypernymies – also those which can be obtained by transitivity. This step ensures a fast conclusion that a book described by the heading “Police corruption” addresses also the problems of “Bribery”, “Police ethics”, “Professional ethics”, “Police power”, “Policemen”, “Police”, ..., “Law” and “Social science”. Connection of such a book with the most general subjects is weaker, although it still exists.

Such structure may be read also reversely, i.e. all hyponyms (direct and indirect) of a given heading can be easily found.

Now, if a user looks for the list of all books on a given subject, he should select a heading representing the subject. It may be done by using the hierarchy browser, the alphabetical index or both. Next, a list of headings is created including the heading selected by the user and all its hyponyms (direct and indirect). For each heading a list of all books described by this heading is created. Finally, all lists are merged and duplicated titles are removed. The final list is sorted according to the decreasing frequency with which the book was borrowed in the past. The above mentioned procedure can be effectively performed using a simple database query. Notice that books on most general subjects appear at the top of the list because it is most likely that they are borrowed most often.

The books on the list can also be grouped, for example, according to direct hyponyms of the selected heading. Grouping could hide the information which books are the most interesting ones. Therefore we suggest a solution

where the original order of books is preserved and the relevant group is marked graphically, e.g. with an icon or font colour.

The list of books presented to the user can be seen as a bibliography on a given subject. It should be noted that the information about popularity (and so usefulness for the reader) of the book is not achievable in a regular OPAC.

If a user is searching for books dealing with multiple subjects and cannot find any appropriate heading (extended heading like "Policemen -- Training of", dependency name like "Police and press" nor attribute heading like "Authors, Polish") he can choose several headings. An intersection of sets containing the search results performed for each heading separately is then presented to the user. For example, one could search for books dealing with frogs living in Wielkopolska by selecting two headings: "Frogs" and "Wielkopolska (Poland)". If no book described by these two subjects exists, the search is extended to hypernyms of any of these two headings meaning that either the books on all amphibians in Wielkopolska or on frogs in Poland are searched next. The extension can be performed automatically. Results of such extended search should be presented also in the case where more general books are much more popular than books on the specific subject.

As we have shown there are numerous possible applications of using an ontology in the OPAC search engine.

### Statistical analysis of book subjects

It is also an interesting question which subjects appear in the biggest number of books or, in other words, which subjects are most popular among the books in a given library. For each heading the number of books described by a given heading or its hyponyms can be easily calculated. The subjects with the largest representation are then presented to the user. The subject headings may be selected from the entire vocabulary or just from the highest level of hierarchy.

Such analysis can be performed not only for the entire library resources but also for a given subject heading. For example, we may check which sub-heading of the heading "Mammals" is represented by the biggest number of books.

This task may be extended to several keywords and determine, for example, which sub-headings of the headings "Poland" and "Russia" are the most popular ones. In this case the hyponyms of both headings are considered.

Another possible extension is to search for the subjects that appear in *borrowed* books most often. Thus, instead of counting the books in the catalogue, the number of times the book was borrowed in the past is to be counted.

Popularity of direct hyponyms may be used as the rule of sorting the hyponyms in the browser.

### Non-library applications

It is also possible to use the proposed ontology as a general Polish language ontology. Although creating the vocabulary with a focus on library cataloguing applications caused that the headings denote domain

subjects rather than natural language concepts, the ontology has the following advantages:

- it enables to determine actual popularity of headings;
- it enables the translation of headings to other languages;
- it enables to obtain a bibliography on any subject heading.

## 6. Programming code

The program library implemented imports a vocabulary in the MARC Exchange format. Then, all entries (such as subject, subdivisions, additional comments, etc.) of the headings are recognized and their correctness is checked as well as the correctness of the entire vocabulary. The transformed vocabulary is available through Java API. A separate module uses API to construct implicit hypernymies in the vocabulary.

## 7. Summary

It is easy to notice the analogy between exploring the library resources and exploring the Internet. However, the library resources have some features that make their exploration easier compared to the Internet. Contrary to the web documents, all library documents are described using a controlled vocabulary. In consequence, it is easy to use an ontology to analyse the vocabulary. The syntax of the SHL makes the morphological analysis of the subject headings possible. Semantic analysis, based on the proposed ontology, may be also possible in the future. Moreover, subject heading languages have a very similar syntax in various countries, which is not typical for the web pages. Also the number of documents stored in a library is significantly smaller than on the web. Many of the Internet documents are available only in a short period of time, while in the library their availability is much more stable.

The above features decide that it is easier to analyse and develop a search system for the library resources. Algorithms for counting the popularity of particular documents are also more straightforward, since it is directly related to the number of times the book was borrowed from the library.

In library browsers, like in web browsers users may be personalized in order to position a document. Due to the above features a "Semantic library" may become available much sooner than the "Semantic Web".

## References

1. NUKAT - Polish central subject catalogue using KABA subject access vocabulary - [www.nukat.edu.pl](http://www.nukat.edu.pl).
2. *Język haseł przedmiotowych KABA: zasady tworzenia słownictwa*, red. Głowacka T., Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, Warszawa 2000.
3. LCSH subject access vocabulary of the Library of Congress subject catalogue - <http://authorities.loc.gov/>
4. Szafran K.: *Analizator morfologiczny SAM-95. Opis użytkowy*, Instytut Informatyki UW, 1996. Available online at: [www.mimuw.edu.pl/~kszafran/SAM-dists/tr226.ps](http://www.mimuw.edu.pl/~kszafran/SAM-dists/tr226.ps).
5. Tokarski J.: *Schematyczny indeks a tergo polskich form wyrazowych*, PWN, Warszawa 1993.