Thesauri and Ontologies for Digital Libraries

© Pavel Smrž   Anna Sinopalnikova   Martin Povolný

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {smrz,anna,xpovolny}@fi.muni.cz

Abstract

The paper discusses the importance of thesauri and ontologies for digital libraries. The role of multilingual ontologies as well as other lexical resources is tackled. We briefly survey and compare different forms of assistance used to make information stored in digital libraries more accessible for humans and more understandable for computers. A special attention is payed to the role of thesauri and ontologies in structuring, indexing and searching digital document collections. We also present a new software tool that enables to store and efficiently retrieve all the mentioned data.

1 Introduction

The size and the complexity of information collections stored in digital libraries grow rapidly. Metadata becomes fundamental for many projects ranging from corporate digital libraries to freely accessible web-based collections. It holds especially in the context of the envisioned Semantic Web.

The Semantic Web is a vision of data defined and linked in a way, that it can be processed and understood by machines, not just by humans. The notion of the Semantic Web as promoted by Tim Berners-Lee and Jim Hendler represents a transformation of the current World Wide Web so that the information and services are understandable and usable by computers as well as humans. However, metadata — a key idea of the Semantic Web — first saw the light long ago and yet only a negligible portion of today's web pages employ it. Thus, the evolution of the web itself shows that the vision of the Semantic Web cannot realise its potential unless intelligent automatic procedures help to analyse and to transform present-day natural language knowledge into the Semantic Web representation.

Many researchers understand the motivation behind the Semantic Web now but the number of real-world applications of the ideas is much more limited. One of the reasons lies in the fact that the Semantic Web vision, as well as the future of general digital libraries, depends on the definition of thesauri and ontologies. The emerging information systems need the definition of common understanding for their application domains. No future real-world knowledge-based system can achieve satisfactory results without an explicit formal specification of the concepts and relations between them. It is clear that to be able to enter the new era successfully, the problems of building and integration of ontologies must be solved.

Metadata are currently defined and used for a restricted set of applications, e.g. the Dublin Core set of bibliographical metadata such as title, author, etc. However, our research suggests that the aims of future digital libraries cannot be accomplished without an employment of a careful design of thesauri and ontologies. The emerging standards of the XML family, especially RDF Schema [1] and OWL[12], can help in this respect. These standards will surely play a crucial role in the definition of thesauri and ontologies.

The design of a broad-coverage general-purpose thesaurus or an ontology is extremely labour-intensive when prepared from scratch. The most promising approach in the development of large standard ontologies is therefore the effort to clean-up, refine and merge the existing resources, e.g. WordNet (`http://www.cogsci.princeton.edu/wn/`), HowNet (`http://www.keenage.com/zhiwang/ezhiwang.html`), CoreLex (`http://www.cs.brandies.edu/~paulb/CoreLex/overview.html`), the available part of Cyc (`http://www.cyc.com/`), etc. These databases are known under different names — semantic networks, lexical knowledge bases, lexical ontologies, ..., and their primary objective was often very different from providing a standard ontology (modelling the human mental lexicon in the case of WordNet, regular polysemy in CoreLex).

A promissing resource that has been applied for ontology preparation only recently [10] is the word association thesaurus. It arises from word-association tests and comprises thousands word stimulus-response pairs resulting from experiments with hundreds of persons. The refinement of the existing ontologies as well as the use of word association thesauri are discussed in Section 2.

There are several systems that enable storing and retrieving data and corresponding metadata. How-

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

ever, only a few of these tools provide support for an efficient integration of thesauri and ontologies and manipulation with them under the same "roof" as all other data. This fact led us to the decision to implement a system for the efficient storage and retrieval of arbitrary document collections in XML. The system is called DEB and is currently developed at the Faculty of Informatics, Masaryk University in Brno, Czech Republic.

The system takes full advantage of the client/server architecture. The server part manages data storage and retrieval, clients mediate the communication with users, query definition and a presentation of results. DEB also incorporates mechanisms for transforming selected parts of existing ontologies into OWL. All these functions are described in detail in Section 3.

The last section brings conclusions and states the future directions of our research.

The considerations presented in the paper are based on the experience gained from our participation in the EuroWordNet project (parallel wordnets for eight European languages — `http://www.hum.uva.nl/~ewn/`), the recently started Balkanet project (an addition of five other languages of Balkan countries — `http://www.ceid.upatras.gr/Balkanet/`) and RussNet (semantic network for Russian linking lexical semantics with derivational morphology — `http://www.phil.pu.ru/depts/12/RN/Database.html`).

2 The Role of Thesauri and Ontologies in Digital Libraries

The Semantic Web, discussed in the previous section as a motivation, is still labeled as the *future* of WWW. Thus, the crucial role ontologies play in this vision could someone lead to the false impression of a minor importance of the discussed lexical resources for the *present* digital libraries. It is therefore essential to emphasize the value they can add to the quality of the current systems.

Thesauri, ontologies and other lexical resources have been widely used in the digital library field. They are taking part in librarian-oriented as well as user-oriented tasks. The most obvious example is their inclusion in the process of structuring and classification of digital data. The automatic conceptual document indexing can supplement (or even replace in some cases) the standard bibliographic classifications.

On the other hand, the benefit of the more elaborated resources for information retrieval has been chalenged in many papers These works that consider the evaluation in term of standard precision/recall measures [5, 9]. We belive that the contradictory results are due to the nature of the tests and that the future research in this area should focus more on user aspects of navigation through documents rather than on the improvement of precision and recall.

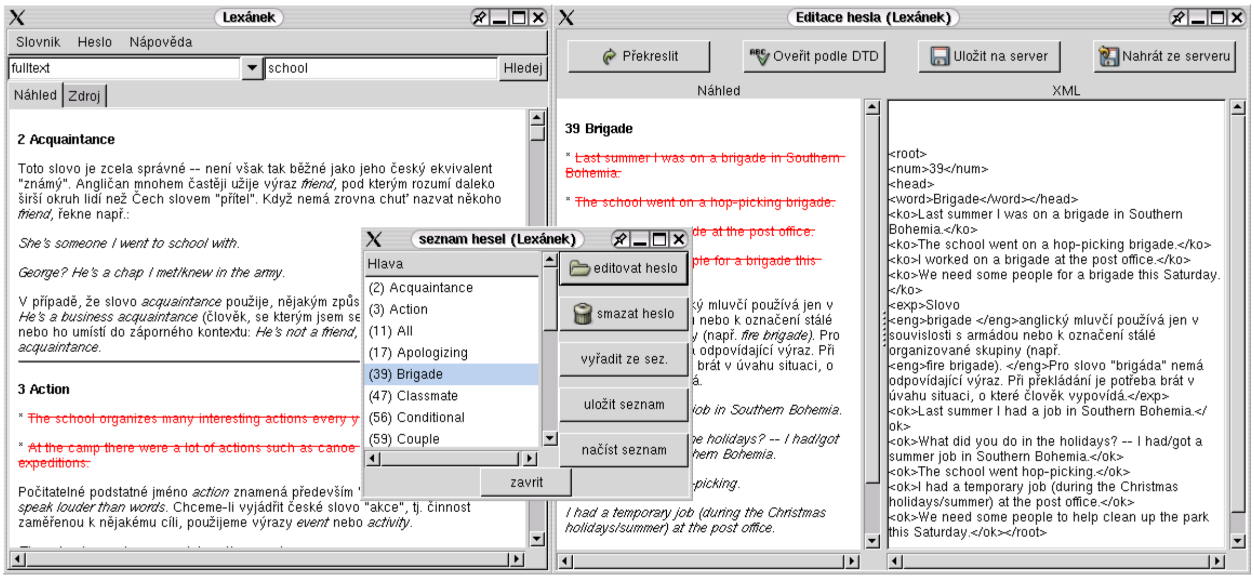
One type of information that is an essential part of different thesauri is the relation of synonymy. Query expansion by synonyms has often been studied and it has been shown that considerable improvements can be obtained at least in the configuration where the user can validate the expansion. Our understanding of synonymy here is rather broad, including all the style, register, regional or orthographic variants. The last is particularly important for dealing with proper names. The difficulty of this task can be illustrated by the spelling normalisation of of the name Muammar Quaddafi (the Libyan leader): the variants presented in the Library of Congress document collections vary from Moamar al-Kaddafi to Gheddafi Mu Ammar, their number exceeding 45 [13].

Hierarchical structure of thesauri and ontologies is usually defined by hyponymy and meronymy relations. The most obvious use of this information is also in the query expansion. A more elaborated application of hyponymy relation is e.g. in the named entity recognition (see e.g. [3]). For example, knowing that *lake* has hypernym *body of water* it can be deduced that *Lake Pleshcheyevo* is also a *body of water*.

A lot of papers in the IR and DL fields describe experiments with synonymy and hyponymy relations. However, there is an enormous number of other types of relations that (although not general enough) can help to a considerable extent. Terms interconnected by "related-to" or "see-also" links naturally define document topics and thus allow concept search as opposed to a context search. A question arises where to take these relations from as they are usually neglected in wordnet-like linguistic resources while the quality of information automatically extracted from corpora is rather low (limited). The application of Word Association Thesauri (WAT) showed to be very promising here.

WAT is a linguistic resource presenting the results of large-scale psycholinguistic experiments (free association test), carried out as follows: a list of stimulus words is given to subjects, who respond with the first word that is evoked. The list of stimuli, lists of responses and their absolute frequencies constitute the body of WATs. Increasing the number of stimulus words (e.g. 170,000) and subjects involved (e.g. 1,500), and varying the individual parameters of subjects (age, sex, profession, etc.) one can guarantee the reliability of data. The simplicity of WAT techniques allows to acquire various information about the language and knowledge of the world, since all the relations between lexical items (words, idioms, proper names etc.) relevant to the particular language system can be explicated. Nowadays, large WAT (monolingual as well as bilingual) are available for such languages as English, German, Russian, Czech, etc., and it seems to be unreasonable not to use them as a source of linguistic information, which cannot be acquired from other resources.

Cross-lingual information retrieval, i.e. finding a document written in a language different from the one used in the query, is another task where thesauri and ontologies find their application. Of course, we consider multilingual resources here, e.g. the results of EuroWordNet and BalkaNet projects in the form



2 Acquaintance

Toto slovo je zcela správné -- není však tak běžné jako jeho český ekvivalent "známý". Angličan mnohem častěji užije výraz *friend*, pod kterým rozumí daleko širší okruh lidí než Čech slovem "přítel". Když nemá zrovna chuť nazvat někoho *friend*, řekne např.:

She's someone I went to school with.

George? He's a chap I met/knew in the army.

V případě, že slovo *acquaintance* použije, nějakým způs
He's a business acquaintance (člověk, se kterým jsem se
nebo ho umístí do záporného kontextu: *He's not a friend,*
acquaintance.

3 Action

* ~~The school organizes many interesting actions every y~~

* ~~At the camp there were a lot of actions such as canoe expeditions.~~

Počitatelné podstatné jméno *action* znamená především '
speak louder than words. Chceme-li vyjádřit české slovo "akce", tj. činnost zaměřenou k nějakému cíli, použijeme výrazy *event* nebo *activity*.

(2) Acquaintance
(3) Action
(11) All
(17) Apologizing
(39) Brigade
(47) Classmate
(56) Conditional
(59) Couple

39 Brigade

* ~~Last summer I was on a brigade in Southern Bohemia.~~

* ~~The school went on a hop-picking brigade.~~

~~de at the post office.~~

~~ple for a brigade this~~

ký mluvčí používá jen v
nebo k označení stálé
y (např. *fire brigade*). Pro
odpovídající výraz. Při
brát v úvahu situaci, o
á.

iob in Southern Bohemia.

e holidays? -- I had/got
hem Bohemia.

-picking.

I had a temporary job (during the Christmas holidays/summer) at the post office.

```
<root>
<num>39</num>
<head>
<word>Brigade</word></head>
<ko>Last summer I was on a brigade in Southern
Bohemia.</ko>
<ko>The school went on a hop-picking brigade.</ko>
<ko>I worked on a brigade at the post office.</ko>
<ko>We need some people for a brigade this Saturday.
</ko>
<exp>Slovo
<eng>brigade </eng>anglický mluvčí používá jen v
souvislosti s armádou nebo k označení stálé
organizované skupiny (např.
<eng>fire brigade). </eng>Pro slovo "brigáda" nemá
odpovídající výraz. Při překládání je potřeba brát v
úvahu situaci, o které člověk vypovídá.</exp>
<ok>Last summer I had a job in Southern Bohemia.</
ok>
<ok>What did you do in the holidays? -- I had/got a
summer job in Southern Bohemia.</ok>
<ok>The school went hop-picking.</ok>
<ok>I had a temporary job (during the Christmas
holidays/summer) at the post office.</ok>
<ok>We need some people to help clean up the park
this Saturday.</ok></root>
```

Figure 1: Lexánek — Example of a DEB client

of wordnets for several languages linked by ILI. The basic schema of cross-lingual IR can be further extended by the cross-lingual information extraction and the translation of answer back to the language of user query.

Information contained in thesauri and ontologies can be employed not only in the process of information retrieval per se but also in the visualisation of information retrieved [3]. Terms referring to hierarchically organized concepts can be highlighted in many ways, e.g. all kinds of animals can be presented in red, geographical regions in blue etc. in the result of query "animals in Afrika". concepts

3 XML Document Management System

A considerable research effort today concentrates on developping tools for efficient storing and retrieving of data in XML. The biggest players on the RDBMS field extent their systems to be able to manipulate data in XML. Other companies offer specialized software tools that use XML as the native format of stored data.

On the other hand, the support for an efficient management of XML data is much less spread in the area of open-source publically-available systems. The designed and implemented system DEB is developed with the aim to fill in this gap in the GNU PL community.

DEB is based on the client/server architecture. The server side originated as a practical outcome of two Master theses at the Faculty of Informatics, Masaryk University in Brno, Czech Republic [6], [7].

The DEB server is responsible for the storage and retrieval of data. It is based on a specialised module for data storage, on FINLIB text indexing library for the conversion of data into a binary format and FININDEX for efficient retrieving.

DEB clients use XSLT for transforming data into HTML, which is then presented to the user with the help of a HTML widget. Figure 1 one such client called Lexánek. It not only implements the viewing of lexical databases but it also fully supports the editation using the DEB server.

Clients benefit from the client-side caching of parsed entries in DOM [4] and from the use of XPath [2] for extraction of important parts of the entries.

Users can modify the data view by supplying their own XSLT sheet. It gives DEB clients an additional level of flexibility. Last, but not least, most of the described features can be included in a "thin" client application accessible by standard web browsers.

A special extension of the standard XSLT processor allows nested queries that provide the efficiency of processing. XSLT sheets can request data from the server based on the entry being processed. The dataflow schema which has been described in detail in [11] is shown in Figure 2.

The creation of such an extension is possible even without breaking the rules of the XSLT processor (by schema we understand the part of URI before the double slash, such as `http`, `ftp`, `file`, etc.). The schema creates a virtual space of XML documents which are results of the queries. From the XSLT processor point of view, accessing the server data is the same as accessing any other external resources.

4 Conclusions and Future Directions

The work described in this paper focuses on the problem of building a reliable and practically applicable thesauri and ontologies that will be used in the area of digital libraries. The presented refinements and extensions of the lexical semantic database are applied to improve the quality of the Czech part of the multilingual lexical resource developed under the current Balkanet project and will be employed in the process of the building of RussNet. The described XML tool is able to store and retrieve the defined data but is also applicable to a whole range of XML data manipulation.

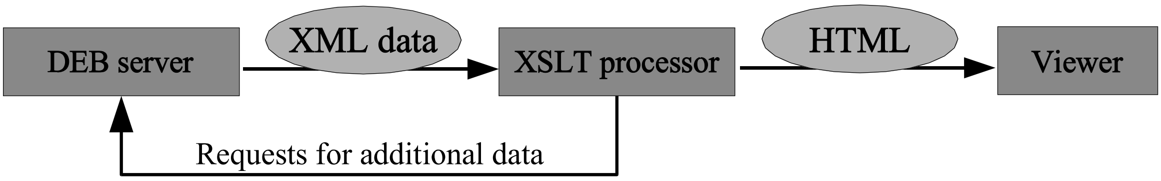


Figure 2: XSLT processor running queries nested in the XSLT script

There are still many open research problems related to the conceptual design of lexical resources. One of them concerns the attempts to integrate generative concepts to the structure of the knowledge base. Such integration usually calls for dynamic entities in the knowledge structure that can be implemented in the form of generative rules. The coexistence of these dynamic issues together with the much more static information in the standard knowledge base gives one of the directions for our research.

In our future work on the software tool we would like to implement XPath evaluation inside the server to make its interface closer to W3C standards. We will also consider the implementation of some data versioning features in the server.

We would also like to separate the core functionality from existing DEB clients and make it a separate layer in order to get a three-level architecture which would further ease the development of specialized applications for thin clients such as WWW browsers.

Current work on standardization of language resources [8] focuses on creation of unified format for representation of multilingual lexical data. We strongly believe that DEB could be used as a standard tool for browsing and editing of data in the proposed RDF-based formats.

References

[1] Dan Brickley and R.V. Guha. Rdf vocabulary description language 1.0: Rdf schema, 2003. `http://www.w3.org/TR/2003/WD-rdf-schema-20030123/`.

[2] James Clark and Steve DeRose. XML Path Language (XPath) Version 1.0, 1999. `http://www.w3.org/TR/xpath`.

[3] Claude de Loupy, Vanessa Combet, and Eric Crestan. Linguistic resources for information retrieval. In *ENABLER/ELSNET Workshop on International Roadmap for Language Resources*, 2003.

[4] Philippe Le Hégaret. Document Object Model (DOM) Bindings, 2002. `http://www.w3.org/DOM/Bindings`.

[5] Eduard Hovy. Building semantic/ontological knowledge by text mining. In *Proceedings of SemaNet'02: Workshop on Building and Using Semantic Networks*. Taipei, Taiwan, 2002.

[6] Luboš Karásek. System for creation and presentation of multilingual and one language dictionaries. Master's thesis, Faculty of Informatics, Masaryk University, Brno, 2000.

[7] Josef Kořeněk. System for maintainance and questioning of large XML dictionaries. Master's thesis, Faculty of Informatics, Masaryk University, Brno, 2002.

[8] Alessandro Lenci and Nancy Ide. The MILE Lexical Model, linguistic and formal architecture, 2002. ISLE Workshop.

[9] Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[10] Anna Sinopalnikova. How to use word association thesauri in linguistic studies. In *Proceedings of the Thirtieth Conference of Postgraduate Students and Lecturers*. St.-Petersburg State University, 2001.

[11] Pavel Smrz and Martin Povolny. DEB — Dictionary Editing and Browsing. In Nancy Ide, Laurent Romary, and Graham Wilcock, editors, *Proceedings of the 3rd Workshop NLPXML 2003, EACL 2003 Workshop*, pages 49–55, Budapest, Hungary, 2003.

[12] Frank van Harmelen, Jim Hendler Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL web ontology language reference, 2003. `http://www.w3.org/TR/owl-ref/`.

[13] Ian H. Witten and David Bainbridge, editors. *How to Build a Digital Library*. Morgan Kaufmann, 2003.