From Web Pages to Dictionary: a Languge-Independent Dictionary Writing System

Karel Pala and Aleš Horák
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pala@fi.muni.cz – hales@fi.muni.cz

Abstract

DEB II platform is a new language independent development platform for lexicographic tools based on the client/server architecture. It uses web Mozilla Development Platform for building special interface modules allowing to build the resulting lexicographic application such as a dictionary writing system from several components. The DEB II consists of a dictionary browser and editor for more XML dictionaries, an integrated Czech morphological analyzer and can cooperate with other applications such as corpus manager Manatee/Bonito and Word Sketch Engine or a geographical information system GRASS (used for building a Czech onomastic dictionary).

The properties and assets of the DEB II platform will be presented as one of its current applications, the tool named PRALED, which is designed for preparation of the new Czech Lexical Database in the Institute of Czech Language in Prague.

1 Introduction

There is a number of software systems that are able to store and process dictionary-like data, many of them using XML as the core element. We can mention, for example, the well known systems like Papillon which is able to manage multilingual dictionaries and various lexical databases and operates primarily with data in XML, or TshwaneLex that is another software tool for compilation of the dictionaries, consisting of the online and electronic dictionary modules (see http://www.tshwanedje.com/tshwanelex/). We also have to mention the Linguists Shoebox which is able to integrate various kinds of text data: lexical, cultural and grammatical (http://www.sil.org/computing/shoebox/).

However, these and similar tools may not be able to efficiently retrieve data needed for recently emerging ontologies or semantic networks and their browsing or editing. The Princeton WordNet, which is one of the most popular lexical resources in the NLP field (see nlp.fi.muni.cz/projects/visdic) has to be quoted as an example. In the EuroWordNet framework the specialized software tools for browsing and editing wordnets have been designed and implemented, the Polaris (and Periscope) in particular. In Balkanet project VisDic tool has been developed and implemented (see http://nlp.fi.muni.cz/projects/visdic/). It can serve also for editing standard alphabetically ordered bilingual or multilingual dictionaries. Ontologies and networks related to the Semantic Web can be browsed with that the tool called Visual Browser developed in the NLP Lab at FI MU enables one to convert WordNet-like databases into the RDF notation and to visualize quite complicated semantic relations between the objects which ontologies consist of.

The main motivation for building DEB2 platform has been a need to prepare new software tools that are going to be used during the creation of the new Czech lexical database at the Institute of Czech Language, Czech Academy of Science in Prague. In this paper we are going present the structure of the DEB2 platform and show its functionality that is being developed in close cooperation with lexicographers from the mentioned Institute.

2 The Structure of the DEB II Platform

The acronym DEB2 (Dictionary Editor and Browser) denotes a platform for building dictionary writing applications. It is based on the architecture client/server, thus the application falls into two parts (see the schema on Fig. 1). The server includes the majority of the required functions, the client part, on the other hand, serves as a user graphical interface which transfers user's requirements to the server, which returns the demanded data.

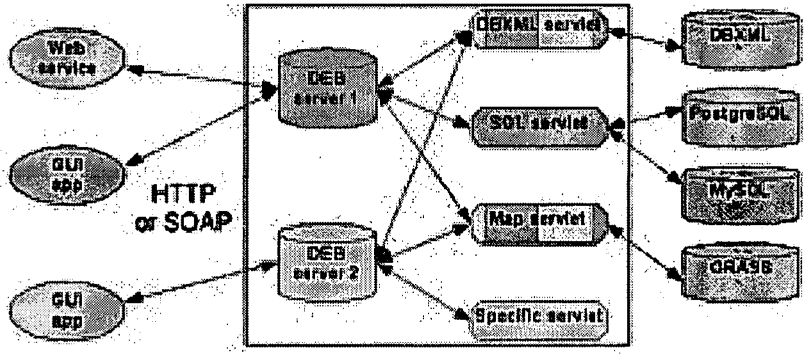


Figure 1 The schema of the DEB II platform

2.1 Server

Server part consists of the smaller components called servlets. They can be shared between different servers and in this way exploit repeatedly already existing functions. All the data are stored in the Berkeley DB XML database using format XML. Presently we use the following servlets:

- general servlet for processing the documents – it reads and stores data from and to the XML database, allows to ask various queries and searches the required data. It supports the XSLT transformations which enable to change appropriately the appearance of the output data.
- SQL servlet – yields the interface for the communication with the relational database such as PostgreSQL,
- GRASS servlet – works with the geographical information system GRASS, which is used for generating maps in the course of creating Czech onomastic dictionary,
- other servlets suitable for individual applications.

2.2 Client

The clients communicate with servlets using HTTP requests in a manner similar to recently popular concept in web development called AJAX (Asynchronous JavaScript and XML) with the usage of the SOAP protocol. The data are transported (using plain HTTP) in RDF, generic XML or plain-text formats or they are marshaled using SOAP. We have decided to exploit the Mozilla Development Platform for its flexibility and openness. It offers the complete kit of tools for the development of graphical applications. The main role here is played by the web browser Firefox or e-mail client Thunderbird.

Mozilla Platform belongs to a open source software and it works under all main operation systems, i.e. under Windows, Linux and Mac OS X (also under others).

Server and client communicate through the protocol HTTP which is used as web standard. More complex requirements can be transferred by means of the SOAP protocol. In our case, the client works as a web page. This is a new and important feature that makes the whole tool extremely flexible and appropriate for team cooperation ensuring the unified structure of the data. Server also allows to manipulate with the access rights to the data according to the needs of the particular project.

3 Current Tools within the DEB II Platform

3.1 DebDict Dictionary Browser

One of the clients on the DEB2 platform is a dictionary and lexical databases named DebDict designed for all the users who need to work with the various versions of Machine Readable Dictionaries to obtain the necessary lexical information. DebDict allows to work with any number of the electronic dictionaries without modifying them. The queries can be quite complex.

3.2 DEBVisDic – Browser and Editor of WordNet-like Lexical Databases

DEBVisDic is a new version of the successful wordnet semantic network editor and browser VisDic mentioned above. It can capture all the relevant semantic relations such as synonymy, near synonymy, hypero/hyponymy, antonymy, meronymy, holonymy as well as others if there is a need. With it the projection between the individual wordnets is easily possible giving as a result multilingual wordnet database exploiting an Interlingual Index (ILI, as it was done for 8 EuroWordNet and 6 Balkanet languages).

4 The PRALED Lexicographic Station

It is designed for the development of the Czech Lexical Database (CLD) and it serves as a main tool in preparation of the new comprehensive and exhaustive database of lexicographic information for Czech language. The user's part of the PRALED tool is presently under the development in the Institute of Czech Language (ICL), Czech Academy of Sciences, Prague. The presented demo will concentrate on showing PRALED‘s present functionality mostly designed by the lexicographers from ICL.

4.1 The PRALED Tool Functionality

PRALED offers users the following functions:

• queries to several XML dictionaries (of different underlying structures), particularly to all relevant Czech dictionaries, i.e. SSJC, SSC, SCS, SCFI, DIDEROT (see the References),

• editing existing or writing new dictionary entries. A lexicographer can use a forms which defines the structure of the entry and fill in all relevant fields (see Fig. 2) which presently are:

– orthoepy (spelling)

– morphological properties (POS, the respective grammatical categories

– description of the meaning (entry definition)

– word formation nest (subnet)

– syntactic properties (most often valencies)

– stylistic, domain and regional features

– semantic relations to other entries (cross-references)

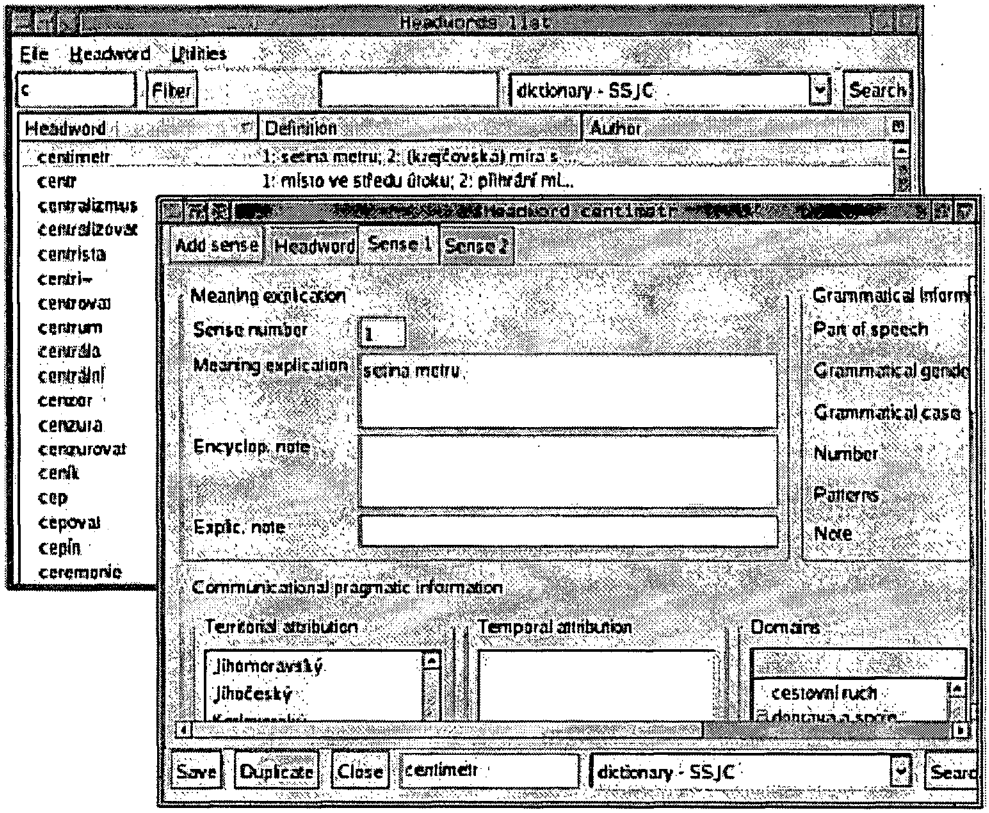
– etymological information

– integration with Czech morphological analyzer

– connection to an external website (Google, Answers.com)

– remarks and additional comments

– integration with the corpus manager Bonito2 and Word Sketch Engine, see Kilgarriff, Rychly, Smrz, Tugwell (2004), which allows a lexicographer to obtain the sorted individual contexts including frequencies and statistical distribution parameters (salience).



c

dictionary - SSJC

Headword	Definition	Author
centimetr	1: setina metru; 2: (krejčovská) míra s ...	
centr	1: místo ve středu útoku; 2: přihrání mí...	
centralizmus		
centralizovat		
centrista		
centri-		
centrovat		
centrum		
centrála		
centrální		
cenzor		
cenzura		
cenzurovat		
ceník		
cep		
cepovat		
cepín		
ceremonie		

Sense number: 1

Meaning explication: setina metru

Encyclop. note:

Explic. note:

Jihomoravský
Jihočeský

cestovní ruch

centimetr

dictionary - SSJC

Figure 2. The PRALED user interface

5 Conclusions

With creation of the DEB2 platform and the individual clients PRALED, DebDict or DEBVisDic lexicographers obtain new tools, which can offer rapid development of necessary lexical resources, i.e. the Czech Lexical Database in particular and a new dictionary of Czech later. There have been no lexicographic station tools available for Czech users so far. Of course, lexicographic stations for foreign languages are available, however, they are not either expensive or in case of open source systems, there are problems with necessary and often complicated modification according to the users' need. Our final aim is to equip the DEB2 platform with enough versatility for all needs, and to make it easily portable to different system installations.

Acknowledgments

This work has been partly supported by the Academy of Sciences of Czech Republic under the projects T100300414 and T100300419, and by the Ministry of Education of CR within the center of basic research LC05113202 and by the project GACR 405/03/0913 of the Grant Agency of CR.

References

A. Dictionaries

Kraus, J., Petrackova, V. et al. (1999), *Akademicky slovnik cizich slov* (Academic Dictionary of Foreign Words), Academia, Praha, electronic version, LEDA, Praha.

Cermak, F. et al. (1983), *Slovnik ceske frazeologie a idiomatiky I-IV* (Dictionary of Czech Phraseology and Idioms), Academia, Praha.

Pala, K., Vsiansky, J. (1994), *Slovnik ceskych synonym (*Dictionary of Czech Synonyms, *SCS)*, Lidove Noviny PublishingPublishers, Praha.

Filipec, J. et al. (1995) *Slovnik spisovne cestiny (*Dictionary of Literary Czech, *SSC)*, Academia, Praha, 1st ed., electronic version, LEDA, Praha.

Petr, et al. (2002), *Slovnik spisovneho jazyka ceskeho (*Dictionary of Written Czech, *SSJC)*, Academia, Praha, 1st edition, electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.

B. Other Literature

Pala, K., Smrz, P. (2004), 'Building Czech WordNet', *Romanian Journal of Information Technology and Science*, vol. 7, No 1-2, Bucharest, pp. 79-88.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D (2004), 'The Sketch Engine', in *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France: Universite de Bretagne-Sud, pp. 105-116 (2004) Manatee, Bonito and Word Sketches for Czech, Trudy mezdunarodnoj konferencii "Korpusnaja lingvistika" – 2004. Izdatel'stvo Sankt-Peterburgskogo universiteta, Sankt-Petersburg, pp. 324-334.

SYN2000. (2000), Ustav Ceskeho narodniho korpusu (the Institute of the Czech National Corpus) FF UK Praha 2000, http://ucnk.ff.cuni.cz

Sedlacek, R., Smrz, P. (2001), 'A New Czech Morphological Analyser Ajka', in *Proceedings of the4th International Conference on Text, Speech and Dialogue*, Zelezna Ruda, Springer Verlag, Berlin, pp. 100-107.

Balkanet Project Website: http://ceid/upatras.gr/Balkanet/ (Patras 2004).