New Clients for Dictionary Writing on the DEB Platform

Aleš Horák, Karel Pala, Adam Rambousek, and Pavel Rychlý
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
{hales,pala,xrambous,pary}@fi.muni.cz

Abstract. In this presentation, we offer an overview of the new clients based on XML database system called DEB. Thanks to the versatile nature of the XML format this platform enables us to develop various applications, namely the management (editing, browsing and other functions) of the electronic readable dictionaries, wordnet-like lexical databases as well as ontologies for Semantic Web applications. First, we characterize the main features of the whole DEB dictionary writing platform and then the implementation strategies of both server and the client part of the DEB platform are briefly described.

Second, we will present the following tools/clients: 1. DEBVisDic which allows to handle different lexical resources and can be used as an appropriate tool for future standardization of wordnet-like databases, 2. PRALED – a client for building Czech Lexical Database, 3. DEBDict – browser for parallel viewing of several electronic dictionaries, 4. DEB CPA browser and editor – a client for development of corpus patterns of verbs, and 5. DEB TEDI tool – a client for building Czech Terminological Dictionary.

For each of the mentioned DEB clients we will give their main features and briefly describe their functionality together with their demonstration.

1 Introduction

There is a need to handle various lexical resources that take the form of wordnets, ontologies, valency lexicons, framenets and others. For this purpose researchers seek for software systems that are able to store dictionary-like data using XML as the core element. Many dictionary publishing houses operate large systems with the complex functionality of so called lexicographic stations that manipulate XML in the last years (DPS Longman [1], TshwaneLex [2], iLEX [3] or ShoeBox [4]). However, these and similar tools are not always able to efficiently merge and manipulate resources obtained from data-driven NLP applications. Therefore, they cannot provide a universal environment for lexical database management as well as semantic networks and ontologies. They also represent rather large systems that are quite complex which is not always an advantage. And, last but not least, some of them are not so cheap. That is why we decided to build
a DEB II platform on which the individual clients can work – in our view this solution is quite modular and flexible since the clients can be adapted for the particular purpose in a short time. One of the reasons for this solution is the fact that some well known lexical resources in the NLP field take the form of semantic networks – the best examples are the Princeton WordNet [5], multilingual EuroWordNet 1, 2 projects (1998-99) [6] and also Balkanet project (2001-4) [7] in which the wordnets for 13 languages have been developed (English, Dutch, Italian, Spanish, French, German, Czech, Estonian, Bulgarian, Greek, Romanian, Serbian and Turkish). In the course of the Balkanet project’s work the specialized software tools for browsing and editing wordnets have been designed and implemented, without whose the job could hardly have been performed – the editor and browser VisDic [8]. The tool has its limitations – it was designed as a local tool only and its flexibility is rather limited.

2 The Main Features of a Common XML Platform

The DEB platform (DEB II, i.e. its second version) follows a strict client-server architecture. The actual development of applications within the DEB platform can be divided into the server part (the server side functionality) and the client part (graphical interfaces with only simple functionality). The server part is built from small parts, called servlets, which allow a modular composition of all services.

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 [9]). The data are transported (using plain HTTP) in RDF, generic XML or plain-text formats or they are marshalled using JSON (JavaScript Object Notation) data structure encapsulation.

The actual data storage backend on the server side is provided by Berkeley DB XML, which is a native XML database providing XPath and XQuery access into a set of document containers. The metadata are stored in widely-used Berkeley DB embedded database which runs on many systems and devices ranging from Linux and Windows operating systems to mobile phones. Berkeley DB XML comes in form of a C++ library with interfaces to many scripting languages.

Since the client applications are mostly oriented to the graphical user interfaces (GUI), we have decided to adopt the concepts of the Mozilla Development Platform [11]. Firefox Web browser is one of the many applications created using this platform.

The Mozilla Cross Platform Engine provides a clear separation between application logic and definition, presentation and language-specific texts. The application design is simple and allows the possibility of concurrent work of different team members which leads to significant time savings.

The main “programming language” used for the GUI design of the DEB clients is called XUL (XML User-interface Language, pronounced “zool”). XUL is a user interface description language based on XML. It allows relatively simple
creation of cross platform applications with possibility of easy customization of design, texts and localization. XUL itself is aimed only on creation of user interface, e.g. windows, buttons or toolbars, but it incorporates wide range of standardized technologies:

- Cascading Style Sheets (CSS) for the visual style of the application,
- JavaScript as a programming language for simple application logic,
- Document Object Model (DOM), XSLT and XPath to work with HTML and XML documents,
- DTD for easy localization,
- RDF as data source.

2.1 Why Client-Server Architecture?

In the client-server environment, the server provides different interfaces using the same data structure and these interfaces can be reused by many client applications. For example, several client applications are using the same interface to query XML dictionaries (with different underlying structure).

One of the main benefits of developing a new dictionary tool on the DEB platform is the homogeneity of the data structure and presentation. If the tool developer commits a change in the data presentation, this change will automatically appear in each client software. And of course, any data flaws discovered can be instantly corrected, there is no need to change the client software or provide new data files to each client.

The data sources can even be implemented with different structures, that the server transforms seamlessly to a homogeneous form, which is then provided to client applications.

2.2 The Clients – Users’ Interfaces

The DEB clients are written in XUL and JavaScript and integrate with Mozilla Firefox Web browser. This allows us to use both Mozilla’s user interface engine and its HTML/XHTML rendering engine as well as built-in components for interaction with filesystem on client computers, XPath interpreter, RDF processor etc.

Due to the feature-rich client architecture the developers may decide whether certain operations should be done on the server or on client parts – e.g. XSLT transformation can be done on both sides.

The particular DEB clients that are currently being implemented within the DEB platform include DEBDict, DEBVisDic, PRALED, DEB CPA and DEB TEDI. We will shortly describe each of them in the next paragraphs.

DEBDict – general dictionary browser. This simple DEB client demonstrates several basic functions of the system:

- multilingual user interface (English, Czech, others can be easily added)
Fig. 1. The DEBDict common interface to several dictionaries with different structures.

- queries to several XML dictionaries (of different underlying structure) with the result passed through an XSLT transformation
- connection to Czech morphological analyzer
- connection to an external website (Google, Answers.com)
- connection to a geographical information system (display of geographical links directly on their positions within a cartographic map) or any similar application

The version of DEBDict that is currently running on our server provides a common interface to 7 dictionaries (see the Figure 1):

- the Dictionary of Literary Czech Language (SSJC, 180,000 entries)
- the Dictionary of foreign words (46,000 entries)
- the Dictionary of Literary Czech (SSC, 49,000 entries)
- the Dictionary of Czech Synonyms (thesaurus, 23,000 entries)
- two dictionaries of Czech Phrasal Words and Idioms (4,000 and 10,000 entries)
- the Diderot encyclopedia (90,000 entries)

As an addition, DEBDict features an interconnection to several web systems and the geographical system with the list of the Czech towns and cities.

DEBVisDic – wordnet editor. DEBVisDic has been conceived as a reimplementation of the previous tool for wordnet semantic networks editor – VisDic. VisDic already exploits the XML data format thus making the wordnet-like databases more standard and exchangeable. Moreover, thanks to its general configuration, VisDic can serve for developing various types of dictionaries,
i.e. monolingual, translational, thesauri and multilingually linked wordnet-like databases. The experience with the VisDic tool during Balkanet project has been extremely positive [8] and it was used as the main tool with which all 6 Balkanet national wordnets were developed.

Within the development of DEBVisDic we pay attention to the relations between wordnets and Semantic Web. DEBVisDic uses a new windowed interface (see the Figure 2) that allows a user to arrange the client layout without any limitations. Of course, DEBVisDic contains all the main features that were present in VisDic, like multiple views of multiple wordnets, hypero-hyponymic tree browsing, inter-dictionary linking or synset editing. With the help of the DEB platform reusability, DEBVisDic is supplemented with a number of new features that were so far accessible only as separate tools or resources such as a connection to a morphological analyzer (for languages, where it is available), language corpora including Word Sketches statistics, access to any electronic dictionaries stored within the DEB server or searching for literals within encyclopedic web sites.

The client-server architecture allows an easy connection of other existing applications to the DEB wordnet server. An example of such application is a direct interface to the VisualBrowser [12] tool that now displays the graphical repre-
sensation of the semantic network from the same database which is displayed in the DEBVisDic tool.

**PRALED – Czech Lexical Database tool.** PRALED is a browser and editor designed for the preparation of the Czech Lexical Database, CLD. It serves as a main tool in preparation of the new comprehensive and exhaustive database of lexicographic information for Czech language. Nowadays, the user’s part of PRALED is under the development in the Institute of Czech Language, Czech Academy of Sciences, Prague. Here DEB is used as a full blown dictionary writing system platform. Thus the main part of the interface consists of the form for lexicographers who can use it for writing the individual entries. The form contains the following fields:

- variants characterized by the appropriate features
- morphological information
- syntactic information in the form of valency frames
- sense definitions
- word derivation information
- semantic relations (hyperonymy/hyponymy, antonymy, cohyponymy, ...)
- etymological information (where it is relevant)
- morphological analyzer/module, which is not a field but a link

The forms can be easily linked to the corpus manager Manatee/Bonito and the Word Sketch Engine [13]. PRALED, in fact, serves as a complete lexicographers station.

**DEB CPA editor and browser.** Corpus Pattern Analysis (CPA, [14]) is a new technique for mapping meaning to words in text. No attempt is made in CPA to identify the meaning of a verb or noun directly, as a word in isolation. Instead, meanings are associated with prototypical sentence contexts. Concordance lines are grouped into semantically motivated syntagmatic patterns. Associating a “meaning” with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns.

CPA editing tool (see the Figure 3) displays the list of verb entries, along with the information who and when updated the entry. Each entry consists of several patterns (the number of patterns is not limited) and it is possible to freely modify their order and content. The main part of the tool, the pattern editing window, allows to enter and modify all the information about one pattern. The form is very versatile, e.g. it allows to add any number of subject/object alternations. The tool is connected to an on-line resource – it is possible to look up subject and object semantic type in Brandeis Semantic Ontology [15] which is hosted on a web server at Brandeis University. Examples documenting the pattern are taken from BNC using a modified version of Bonito2 corpus manager that is integrated to the DEB CPA tool.
DEB TEDI terminological dictionary tool. The DEB TEDI client is the main tool used for preparation of a new big terminological dictionary of Czech. This work is a joint project of the Czech publisher NLN and Masaryk University. The aim of the project is to build a terminological database consisting of about 250,000 dictionary entries. Several printed dictionaries of different size will be generated from that database.

3 Conclusions and Future Directions

We have described the current state of development of clients (dictionary tools) based on the DEB II platform. This platform offers a common implementation base for client/server architecture. Thanks to its high modularity, configurability and flexibility it can be easily adapted for various lexicographic tasks. Using this basis, new individual and powerful dictionary writing tools (clients) such as DEBVisDic are implemented.

In our view the DEB platform is being (and will be) thoroughly tested with its clients, namely the DEBVisDic is currently being prepared for the Dutch Cornetto project and for Hungarian Wordnet project. We also discuss the possibility of DEBVisDic being used as the main wordnet tool in the near future as well, namely in the preparation of the World WordNet Grid (Ch. Fellbaum, personal communication, May 2006).

The PRALED client is used in the Institute of Czech Language, Czech Academy of Sciences (Prague) as a dictionary writing system for building Czech
Lexical Database which is a large project planned for about 6 years from now. The goal is to develop a lexical database of contemporary Czech containing approximately 100,000 entries. An important new feature here is that PRALED will be linked to the Manatee/Bonito corpus manager and the Word Sketch Engine.

Acknowledgements

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 LC0513202.

References