Subject-field-specific Ontologies and Terminologies for the Web Community

Klaus-Dirk Schmitz

University of Applied Sciences Cologne
Institute for Information Management, Department of Modern Languages
Mainzer Str. 5, D-50678 Köln, Germany
klaus.schmitz@fh-koeln.de

Abstract

A terminological thesis written in the Department of Modern Languages at the University of Applied Sciences Cologne contains descriptive terminology of a limited domain. These systematic, concept-oriented terminological data are available in electronic form (MultiTerm database format). The WebTerm Project aims to consolidate and convert the terminological data to a web-based system allowing efficient and free access to these terminologies. Special attention is paid to the dynamic representation of the system of concepts and the ontological relations of these data collections. The paper describes the structure and content of the terminological theses and the terminology contained, as well as the web-based dynamic interface to the ontologies and terminologies developed in the framework of the WebTerm Project.

1. Introduction

In recent years, a growing number of theses and dissertations in the field of terminology has been written at university departments for translation and interpretation. The tremendous effort students investigates in elaborating concept-oriented mostly bilingual terminological collections of a limited domain remained largely ignored because no one outside the respective university knows about these existing collections. And if an interested user outside the respective university is informed about the existence of domain-related terminological collections, it sometimes turns out to be quite difficult to have access because the theses are only available in one paper copy and many university departments are not equipped and allowed to disseminate and „sell" the information. But more and more terminological theses are produced by using terminology management software; therefore the terminological data the user is interested in is also available in electronic form which allows an easier and more efficient way for dissemination and (re-)use.

2. Terminological Theses

At the Department of Modern Languages of the University of Applied Sciences Cologne, more than 150 terminological diploma theses have been elaborated over the last 8 years. The main objective of these theses is to elaborate the terminology of a limited domain in two or sometimes three languages by a descriptive, concept-oriented and systematic approach. Each concept is fully documented by terms (including synonyms, abbreviations, orthographic variants etc.), grammatical and usage information, context examples, definitions, sources and notes. The main focus is the development of a systematically organized concept system of all concepts of the examined domain; the position of each concept within the concept system is represented by a notation ("local classification") indicating the ontological relation to other concepts in the domain.

Figure 1 and 2 show two pages of the printed version of a terminological thesis with a part of the concept system and a typical concept entry.

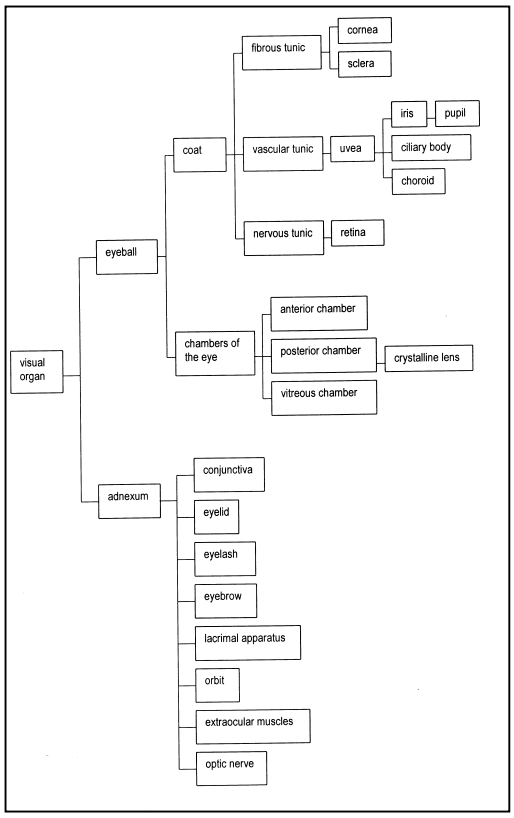

Figure 1: Sample pages with (part of) the concept system from a printed thesis

3. The MultiTerm Database

Since 1995, the terminology management system Trados-MultiTerm is used for elaborating the terminological data for the theses at the University of Applied Sciences Cologne. The definition of the database follows the needs for terminological theses and covers all the necessary data categories including an indexed field for the notation (see Figure 3). Since 1997, all source references of the terminological work are stored in bibliographical entries in the same MultiTerm database, and since 1999 the subject fields and domains of the theses are classified using the Lenoch-Code-Classification of the EU Commis-

sion. MultiTerm export definitions and routines are provided to the students for generating parts of the printed version of the thesis, especially for the terminological entries, the bibliography and the alphabetical bilingual indices with references to the notation.

Stand: 29.02.2000 - 16:12:47
Fachgebiet: Medizin
Lenoch-Code: *#ME*: O
Autor: So-Heong Park
Notation: 1.1.1.1.2.1.1.1

D: Pupille *<f.>*
Quelle: ***Jäckle-Kirchhoff.1995***
Definition: das kreisrunde, dunkel erscheinende Sehloch in der Regenbogenhaut (etwas exzentrisch nach unten-nasal), von den Irismuskeln je nach physiologischem Bedürfnis (Akkommodation) in der Größe verändert (ϕ 1,5 bis 8 mm)
Quelle: ***Jäckle-Kirchhoff.1995***
Kontext: Das zentrale Loch der Iris, die Pupille, ist die Blende des optischen Systems.
Quelle: ***Grehn.1998***

D: Sehloch *<n.> <Synonym>*
Quelle: ***Wahrig.1996***
Kontext: In der Mitte der Iris befindet sich eine rundliche Öffnung, das Sehloch oder die Pupille.
Quelle: ***Kaden.1993***

E: pupil *<sub>*
Quelle: ***Millodot.1997***
Definition: Aperture within the iris, normally circular, through which light penetrates into the eye.
Quelle: ***Millodot.1997***
Kontext: The central opening, the pupil, is derived from the Greek pupa, a small doll-like figure that was reflected from the center of the eye when one looked at someone's eyes.
Quelle: ***Stein.1987***

Figure 2: Sample pages with a terminological entry from a printed thesis

Figure 3: Terminological entry in MultiTerm

4. Converting and Consolidating Data

From November 2000 to February 2002 the WebTerm Project was realized in Cologne with a co-financing of the German Ministry for Research and Technology aiming to consolidate and convert the MultiTerm data and develop a web-based application which allows internet users to have access to this type of terminological data.

The consolidation work comprised the correction of databases errors, the testing and completion of hyper links between terminological entries and hyperlinks to bibliographical entries, the completion of lacking bibliographical sources and lacking Lenoch-Codes for subject field classification, and the consolidation and unification of the data categories and their content. Special attention was paid to include almost empty "ordering concepts" in MultiTerm containing only one term for each language (in parenthesis) and the notation. These "ordering concepts" represent more ontological levels than real concepts and are necessary to group specific subordinated concepts under a heading (see example "1.1.1 (data sources standardized by law)" in Figure 4). Since the graphical representation of the web application is automatically generated from the notation data category of the MultiTerm database, the "ordering concepts" must be included in the source database.

Special handling was necessary for terminological entries with two or more notations representing concepts that appear more than once in the ontology; these entries were duplicated with only one notation for each copy. Another problem occurred with terminological gaps, i.e. the concept and the corresponding term does not exist in one of the languages. Since the term field was lacking in such cases it was added to the MultiTerm entry with the content "term not existing". Both procedures were necessary for the automatic export and conversion of the MultiTerm data.

5. The WebTerm Implementation

After some tests with the Trados Muwa (MultiTerm Web Access) software, designed for the presentation of MultiTerm entries in web applications, a specific software for an own web interface was developed, allowing not only to present the terminological and bibliographical entries but also to display an user-expandable view of the concept system, generated from the notation field of the MultiTerm database. A C++ program converts ordinary text files produced by the export facility of MultiTerm into several lookup-tables and XML files. These files are interpreted by Java scripts to generate the web page with the input template for the search and select functions, the indices for the terms and the bibliographical codes, the graphical representation of the concept system, the terminological and bibliographical entries as well as the hyperlinks between the entries (see Figure 4, 5 and 6).

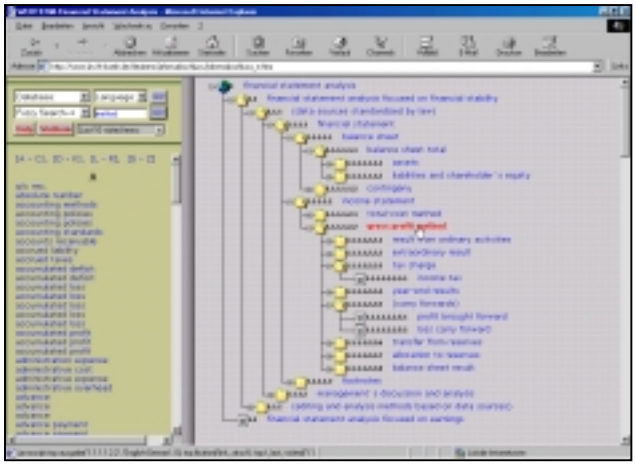

Figure 4: Web interface: concept system

At the moment, the web application is (only) running under Microsoft's Internet Explorer. Other web browsers are not able to dynamically expand the concept system or to follow the hyper links

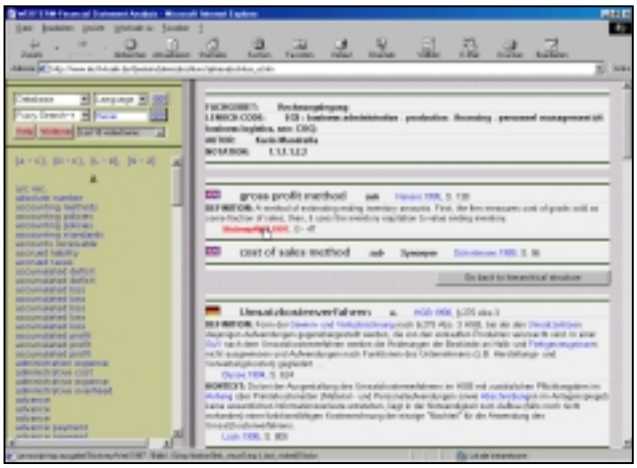

Figure 5: Web interface: terminological entry

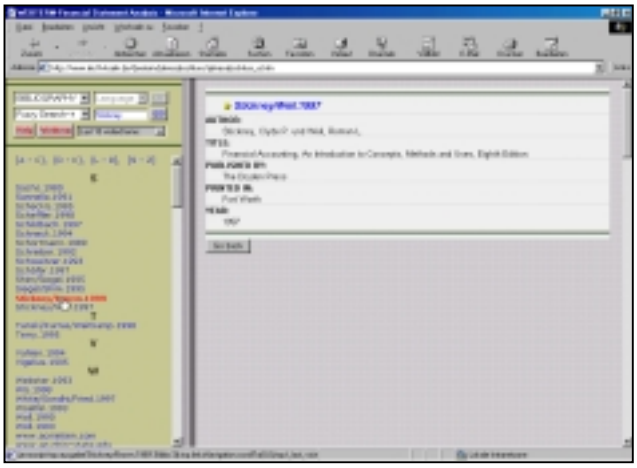

Figure 6: Web interface: bibliographical entry

6. Selecting and Using the Terminology Collections

On top of the different terminology collections, several web pages were designed to allow the internet user to find and access the terminology he or she wants. The bilingual (German and English) start page of the WebTerm site (www.iim.fh-koeln.de/webterm-auswahl_d.htm) allows to select a project description (only in German), a short guideline for the usage of the terminology collections, a disclaimer ("Attention – read carefully before use!") indicating the varying quality of the data, and the link to the databases with the terminology.

When selecting the link to the databases a huge list of terminology collections is offered grouped into subject fields. Each terminology collection is named by the domain that is covered and documented by the languages included and the date (month and year) of finalization. In addition to these data, the number of terminological entries (concepts) and bibliographical entries (references) is indicated (see Figure 7).

If a specific terminology collection has been selected, the whole data of the collection are loaded which may last a certain time. After loading, the web page shows in the left frame a small dialog box for the user interaction and an index with terms or bibliographical reference codes. The right frame is use to display either the ontology of the domain (concept system), a terminological entry or a bibliographical entry for a source (see Figures 4, 5 and 6).

Figure 7: Selecting terminology collections (part of the screen)

The dialog box in the left upper corner of the page (see figure 8) allows the user to select one of the two or three languages as a search and ordering language; the concept system is displayed in this language and the terms in this language are displayed on top of the terminological entry.

Figure 8: Dialog box for user interaction

The "Database:" field offers to switch between the terminological and the bibliographical part of the database, and to unfold or hide the full structure of the ontological system of the domain (see Figure 9).

Besides the ontological access to the terminological entries of the database, a more or less direct access to a certain concept is possible by entering a searchword (term) in the selected search and ordering language. Via a fuzzy search algorithm all matching terms are display in the index frame (see Figure 10); the user can click on the term he or she is interested and the terminological entry is displayed in the right frame of the screen. The "fuzziness" of the search can be chosen by the user in the "Fuzzy Search ->" box.

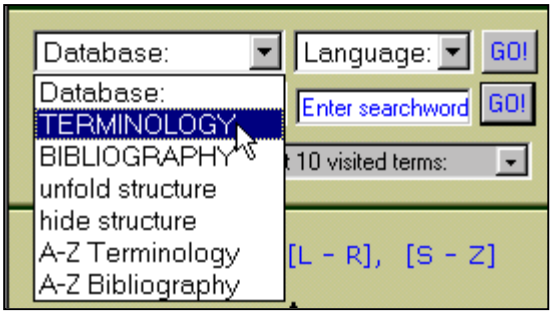

Figure 9: "Database" field options

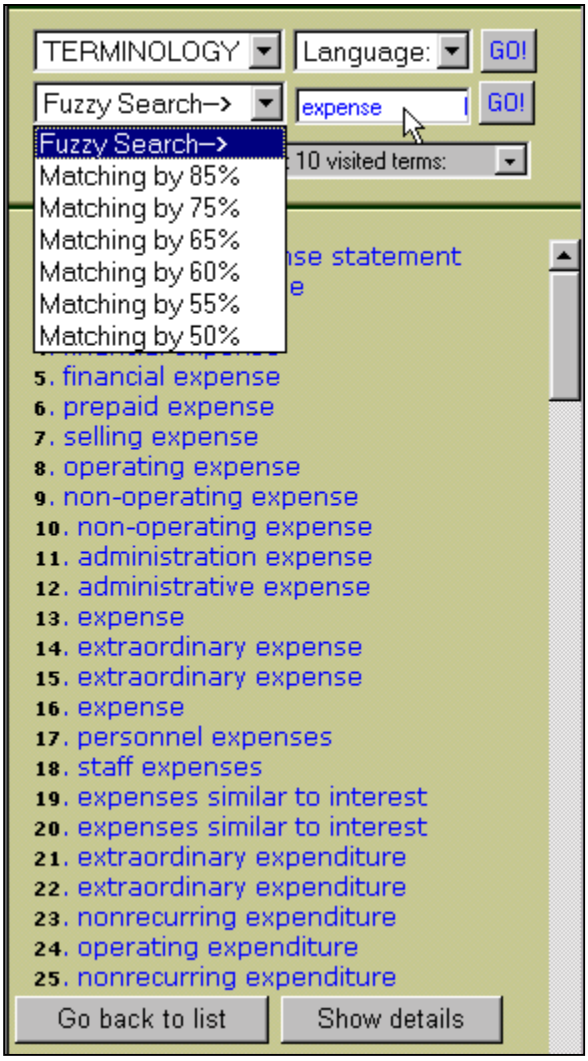

Figure 10: Fuzzy Search

7. Future Tasks

The WebTerm Project was finalized at the end of February 2002 with about 80 terminology collections included. Since all tools are implemented and documented and since a guideline for consolidation of the MultiTerm database was elaborated, further data collections from terminological diploma theses can easily be added. A link to the WebTerm pages are included in the German Terminology Portal (DTP = Deutsches Terminologie-Portal), another project to be realized at the University of Applied Sciences Cologne. The WebTerm inventory will be extended by a list of topics and domains that are under development in theses not yet finished or are proposed by industrial partners. Therefore the WebTerm site will function as a marketplace for new topics and a co-operation platform between industry and university.

It is also planned to enlarge the project idea by creating a network of universities, where terminological theses are being elaborated. The more these universities follow the same approach for terminological theses using the same or a very similar data model for the (MultiTerm) databases the easier the conversion to the WebTerm XML structure will be.

The list of data collections available at the WebTerm site will be provided to terminology servers like ETIS (European Terminology Information Server) using the TeDIF format (Terminology Documentation Interchange Format) developed by the University of Applied Sciences Cologne during the EU co-funded TDCnet Project (Terminology Documentation Centre Network).

8. References

Raupach, I. (2002). WebTerm – Terminologiesammlungen aus Diplomarbeiten im Internet. In F. Mayer, K.-D. Schmitz, and J. Zeumer (eds.), *eTerminology* (pp. 239-245). Köln: DTT.

Raupach, I. & Schmitz, K.-D. (2002). WebTerm – Terminologiesammlungen im Internet. To be published in MDÜ, 2(2002).

Schmitz, K.-D. (1999). Using MultiTerm for Systematic Terminology Work in an Academic Environment. In TermNet (Ed.), *TAMA'98, Proceedings of the 4th TermNet Symposium "Terminology in Advanced Microcomputer Applications - Tools for Multilingual Communications"* (pp. 161-166). Wien: TermNet.