Database of Bavarian Dialects (DBÖ) electronically mapped (*dbo@ema*). A System for Archiving, Maintaining and Field Mapping of Heterogeneous Dialect Data for the Compilation of Dialect Lexicons

Eveline Wandl-Vogt
Österreichische Akademie der Wissenschaften

Christian Kop
Günther Fliedl
Institut für Wirtschaftsinformatik und Anwendungsysteme / Alpen-Adria Universität Klagenfurt

Jost Nickel
Philipps Universität Marburg

Johannes Scholz
Technische Universität Graz

dbo@ema *is a system for the archiving, handling and mapping of heterogenous dialect data for dialect dictionaries. Within this software presentation:*

a) the users should get known to the general project aims of dbo@ema, that are:

- *developement of a webbased, interactive data base*
- *development of a webbased, interactiv tool to map dialect data and background information of a dialect dictionary*
- *developement of a specific, free font for the phonetic transcription of dialect data in digital surroundings (further information see* http://www.wboe.at*)*

b) the users should get known to special tools of the software developed to compiling a dialect dictionary

c) the users should get known to how geoinformation aids the compilation of a dialect dictionary

d) the users should get known to the project Wörterbuch der bairischen Mundarten in Österreich *(WBÖ) (Dictionary of Bavarian dialects in Austria) and*

e) the project Datenbank der bairischen Mundarten in Österreich *(DBÖ) (Data base of bavarian dialects in Austria) that are both mother-projects to the project dbo@ema.*

1. Motivation

The Institute of Lexicography of Austrian Dialects and Names (I DINAMLEX)[1] of the Austrian Academy of Sciences in Vienna stores a tremendous collection of Bavarian words and idioms together with their linguistic context and cultural background. This information is gathered in the last 100 years which in sum are about 5 million documents.

[1] For further information see *http://www.oeaw.ac.at/dinamlex* (28.03.2008).

Figure 1. Voucher to be digitized in the DBÖ

Most of these data is already stored electronically. However, the so called *Database of Austrian Dialects and Names* (DBÖ) does not meet the needs of modern electronic storage of data. Therefore and for the reason to create new possibilities of data-access the project *dbo@ema* was designed by the project leader in 2006[2]: A new system is going to be established and data are transformed into this new system.

The project *dbo@ema* is financed by the FWF[3]. The interdisciplinary project is led by Eveline Wandl-Vogt (I DINAMLEX), cooperation-partners are the Alpen-Adria Universität Klagenfurt, the Technische Universität Graz, the Phillipps-Universität Marburg (Germany) and the Slowenische Akademie der Wissenschaften und Künste Ljuljana (Slowenia).

2. Dimensions of the software

Managing millions of Bavarian idioms (lemmas) is a difficult task. These words are embedded into a context that has to be considered. Thinking about all certain aspects that have to be handled, we came to the conclusion that these idioms can be described along the following dimensions:

- Content.
- Spatial (Geographic).
- Multimedia.
- Scientific Dimensions.
- Administrative.
- Others (Historical, Language).

The *content dimension* deals with the idiom itself. For each word it is necessary to know from which source it was collected. Such a source could be either any Bavarian dialect literature or a historical Bavarian document, or derived from answers of questionnaires. Especially in the case of the I DINAMLEX questionnaires make up a tremendous contingent within the sources for idioms. The reliability of the source is also an important question in that context. Particularly a literature has an author and the answers of questionnaires were collected by special persons (collectors). The importance and value of a certain idiom which is defined in the answer of a questionnaire therefore also strongly depends on some characteristics collector e.g.:

- was it a professional, or an interested layperson,
- what kind of school did the person visit (e.g. is she/he an academic or non-academic person),
- what kind of profession does the person have? (e.g. a teacher or a pastor in a certain region).

From such characteristics the scientific relevance of a certain idiom or the correctness and completeness of its descriptions can be concluded.

[2] For detailled information about the project see Wandl-Vogt (2008); former studies about mapping dialects see Wandl-Vogt (2006 a) and Wandl-Vogt (2006 b).

[3] See *http://www.fwf.ac.at* (28.03.2008).

Idioms itself can be classified into main lemma, reference lemma and relationships between different lemmas (e.g. the root, the siblings of a certain lemma).

Specific problems of the *geographic* dimension were already mentioned in the last section. For the latter the database must have structures (tables and attributes) which allow storing geographic coordinates.

Also for the storage and representation of *multimedia* data, specific additional database structures are needed.

The *scientific* dimension deals with the fact that the whole purpose of the information system is to support the scientist during search and retrieval of words. The result of these enquiries is used to compile articles for the *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ). This work belongs to one of the many tasks the institute has received from the Austrian government in order to document language and cultural background of the country. The dialect data is also a solid basis for many other scientific explorations and papers. It should be clear that the database structure must be optimized according to these needs.

All these aspects discharge into the question how these different data can be *administrated* by different users, i.e. not only the scientist, who wants to search for information but also qualified employees, who have to enter quickly data into the database. Furthermore the scientists currently are also responsible to check the entered data according to their scientific criteria. In situations where one and the same idiom or other information of an idiom can be described differently, thus they help the qualified employees to make the right decision. However, this implies that the data structure (tables, table attributes) within the database is able to support a minimal workflow.

Last but not least the idioms and their regional as well as their linguistic context were gathered over a *long period of time*. This information should be reconstructed at least implicitly from questionnaires or from the birth dates respectively dates of death of the collecting persons. Furthermore nationalities change over time. Names of location which were used in ancient times thus change too. This must be reflected somewhere in the database.

3. Software architecture

The systems consists of the modules database module, GIS-module, Web-Query-module and maintenance module.

Database Module

This Module includes two databases, a MySQL and a PostGIS database. The main database is the MySQL database. It contains all the data belonging and related to a Bavarian word or idiom. In order to provide also a GIS (geographical) view of the data an additional special PostGIS database is used. PostGIs is a GIS extension that is based on Postgres.

GIS-Module:

The GIS-Module allows reading access to the GIS data and provides geographical maps as JPEG images to the calling modules (Web-Query-Module and Maintenance module).

Maintenance Module:

With this module, special users at the *I DINAMLEX* can administer and maintain the database.

Web-Query module:

This module will be open for the public in the end of the project (planned: 01.2009). It will show the several relationships between the words and their background information (e.g. who has collected the word, in which geographical region is it located etc.)

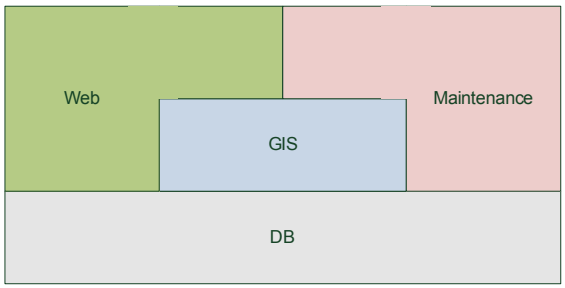

Figure 2. Software Architecture: Modules

In order to achieve this Software architecture with a minimum of costs we used the following platforms: Free MySQL Database (Port 3306); Free PostGIS Database (Port 5432); Free GIS-Web-Mapping Server; Free Apache + PHP Web server; Java Runtime Environment 6.

4. The underlying database

Summarizing all these information has led to the following database schema (have a look at picture 3).

The database schema is represented using a specific diagrammatic database schema specification language. However, this schema only gives an overview of the most important terms.

The content dimension is the largest one starting with the term *Lemma (idiom)* [light green]. These idioms can be related to each other. Next the idioms is related to *vouchers/document (Belegzettel)* via the notion *Beleg (evidence)* [blue]. This makes up the core of the content dimensions. The meaning of these relationships can be summarized in the following way: One or more idioms are documented in the system with a certain semantic.

From each voucher, important related information like sources of idioms and involved *persons* [light grey] are referenced. In particular *Excerptor* (*excerpting person*), *Autor* (*author*), *Gewährsperson (informant),* describe which person has introduced a certain idiom in a certain way. Thus these notions can be seen as aspects of the content dimension.

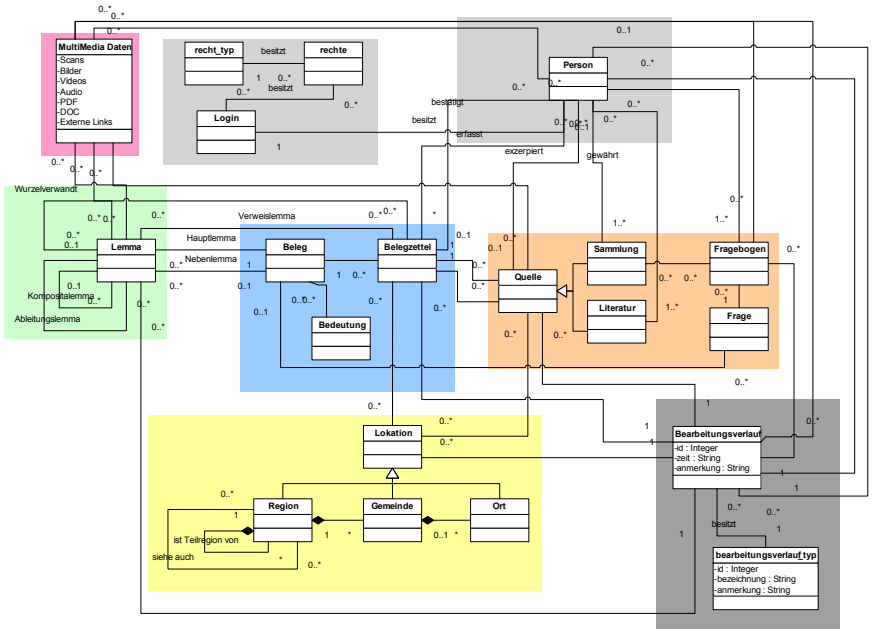

Figure 3. Database schema

The multimedia dimension is represented with the term *Multimediadaten* (multimedia data) [magenta] which is responsible for the storage of the multimedia content. The geographic dimension is represented using the terms *Region* (region), *Gemeinde* (municipality), *Ort* (documented place). It is possible that regions have sub regions. A municipality is always the smallest region and contains cities. These different spatial categories are summarized to the general term *location* [yellow].

Location relates the geographical information to the rest of the dialect data and contains information that is valid for all geographic categories.

The administrative and scientific dimensions contain all the terms and notions. The workflow aspect is treated in the term *Bearbeitung* (work) [dark grey].

The last dimension history and language is implicitly implemented in the detailed structure (attributes) of the database.

5. Further Information

Further information on the project dbo@ema and the tool can be found at: *http://wboe.at/de/*.

References

Wandl-Vogt, E. (2006 a). "Mapping dialects. Die Karte als primäre Zugriffsstruktur für Dialektwörterbücher". In *Wiener Schriften zur Geographie und Kartographie* 17. 89-87.

Wandl-Vogt, E. (2006 b). "Von der Karte zum Wörterbuch. Überlegungen zu einer räumlichen Zugriffsstruktur für Dialektwörterbücher. Dargestellt am Beispiel des Wörterbuchs der bairischen Mundarten in Österreich (WBÖ)". In Corino, E.; Marello, C.; Onesti, C. (eds.). *Atti del XII Congresso internazionale di Lessicografia. Torino, 6-9 Settembre 2006*. Vol. 2. Torino: Alessandria. 721-732.

Wandl-Vogt, E. (2008). "An der Schnittstelle von Dialektwörterbuch und Sprachatlas: Das Projekt *Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)*". *Germanistische Linguistik*. To appear.