Published in: Schmidt, Sara/Siemens, Ray/Kumar, Amit/Unsworth, John (eds.): Digital Humanities 2007.
Conference Abstracts - Urbana-Champaign: University of Illinois, 2007. pp. 164-166.

Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System

Timm Lehmberg, Christian Chiarcos, Erhard Hinrichs, Georg Rehm, and Andreas Witt

1 Introduction and Overall Concept

Most metadata standards used for corpus linguistic purposes (TEI, OLAC, IMDI etc., for a complete overview see Lehmberg and Wörner 2007) require elements that contain legal information about the rights holder to the particular resource and/or its accessibility. Normally these metadata elements are kept very abstract and do neither distinguish between the different types of personal rights nor do they consider the option of multiple holders of copyright.

The legal situation upon which the evaluation of linguistic data to be used for scientific purposes is based is clearly defined, but too complex to be understood completely by non-experts. Furthermore, it varies from one country to the other and is in a constant state of flux.

In the framework of our joint sustainability initiative (see the introduction to this session), a large number of heterogeneous corpora have been acquired from multiple sources and multiple projects, and processed with regard to different individual requirements (Schmidt et al. 2006). This heterogeneity is responsible for the problem that the legal metadata that need to be collected strongly vary with regard to the respective corpus and data situation. Only for a small number of projects associated with our sustainability initiative are detailed sets of legal metadata that inform a potential user of the corpus about, for example, stipulations or copyright holders, readily available. For the majority of projects and corpora, this task has to be performed retroactively.

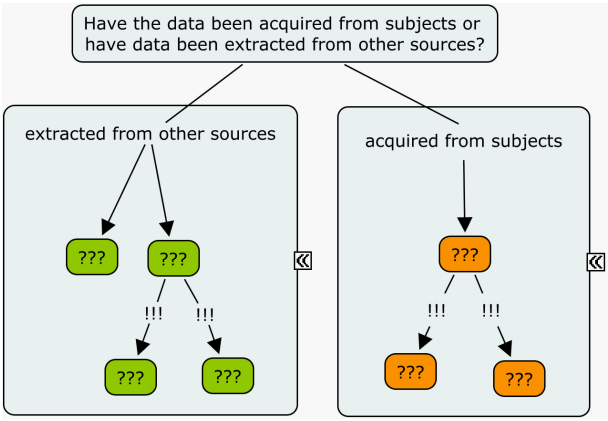

Figure 1: A concept map visualising the query structure

Facing the complexity of the legal context (see Zimmermann and Lehmberg, in this session), it is almost impossible for non-experts to evaluate the situation of their language data and to extract the relevant metadata without professional advice.

To reduce the complexity of this task, concept maps were created with the goal of making the legal situation as well as the legal terminology transparent and understandable to non-professionals. Unlike mindmaps that are primarily used for the (often spontaneous and intuitive) mapping of ideas and processes, the technique of concept mapping is intended more for knowledge modelling: concepts are represented by nodes, links represent the relations between them.

As a utility to create the concept maps modelling the legal situation within our joint sustainability initiative we used *CmapTools*, a program distributed by the Institute for Human and Machine Cognition (IHMC). IHMC CmapTools provide a client/server architecture that allows users at different locations to work collectively on Concept Maps and to discuss their structure and content online.

Based on these schemata and following the principles of decision-trees, we built an additional concept map representing the query structure of a questionnaire. Digressing from the original principles of concept mapping mentioned above, in this map queries are represented as nodes whereas responses are represented as links between them. The primary query given in the centre node (see figure 1) corresponds to two central aspects of law (data protection and copyright, see Zimmermann and Lehmberg, in this session). Each response leads to a large number of additional queries that again, depending on the users' response, have subordinated queries. Further sections of the concept map deal with the accessibility of the data as well as their respective principles and standards of data processing.

In same manner we modelled the query structure that surveys the meta information that ideally has been collected in connection with the compilation process of corpora. Therefore it contains queries asking for established metadata standards (TEI, DC, OLAC, IMDI etc.) that may have been used, and if necessary asks for additional information.

Due to the fact that the IHMC CmapTools provide an export of concept maps into an XML-based format, the content and structure of the concept map can be processed automatically to create the web based questionnaire that is described in the following section.

The complete concept map structures will be demonstrated in conjunction with example scenarios in our presentation.

2 Implementation

As the questionnaire has to be accessible from different research project locations, it has been implemented using a XAMP (any operating system, Apache, MySQL and PHP) architecture to create a user-friendly, web-based interface. The conceptual structure represented by the concept map is transformed into a relational database model. Accordingly, it is possible both to model the tree structure of the queries (Celko, 2004) and to save responses to these questions within the database. Additionally, the database includes user data (as well as user access control data) and links them to the metadata sets of the resource being acquired by the questionnaire.

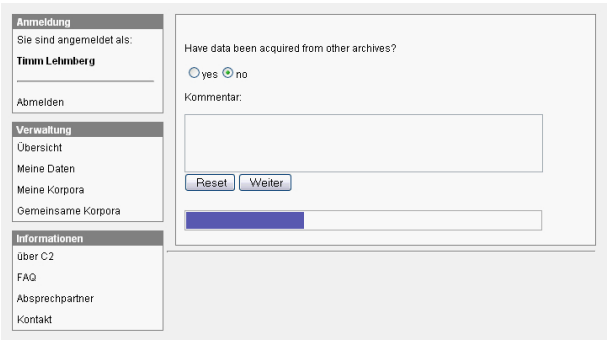

Figure 2: A web-based wizard guides the user through the questionnaire

The user interface is generated by a script that parses the database and guides the user though the questionnaire tree with

the help of a web-based wizard (see figure 2). This architecture has several advantages:

Figure 3: An overview page gives information about the data collection progress

- Subordinate queries that refer to specific details of some legal aspects automatically can be skipped if they become superfluous. For instance, there is no need to query contractual agreements with subjects if there is no personal data contained in the corpus.
- The data model provides users with the option of registering multiple corpora and running the questionnaire wizard individually. Furthermore, users can share the data they entered into the system with other registered users so that it is possible to edit the data across project locations (for example, queries can be skipped, answered later, or left to other users).
- Should the structure or content of the questionnaire tree be changed, the database will be modified accordingly. If the change leads to unanswered queries, this will be indicated to the user in a status page. For this reason, every user account has an overview page that gives information about the state of progress of every registered resource (see figure 3).
- The questionnaire includes queries about metadata content and standards that already have been applied to the registered corpora, so that users do not have to insert redundant information already contained in existing metadata sets.
- Administrator users have unlimited access to all data in the database, so that users can be provided with support, if needed.

We are currently in the process of collecting legally relevant metadata from about 60 different research projects with the aid of the questionnaire system described in this paper. Content and structure of the concept maps is available on our project homepage at <http://www.sfb441.uni-tuebingen.de/c2/>.

Bibliography

Celko, Joe. *Trees and Hierarchies in SQL for Smarties.* San Mateo: Morgan Kaufmann, 2004.

Lehmberg, Timm, and Kai Wörner. "Annotation Standards." *Corpus Linguistics, Handbücher zur Sprach- und Kommunikationswissenschaft (HSK).* Ed. Anke Lüdeling and Merja Kytö. Berlin: de Gruyter, In press.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources." *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art, East Lansing, Michigan, June 2006.* 2006.