Typological Database System

Ontology mediated integration of typological databases for linguistic research

Menzo Windhouwer[*], Adam Saulwick[*], Rob Goedemans[#], Alexis Dimitriadis[+]

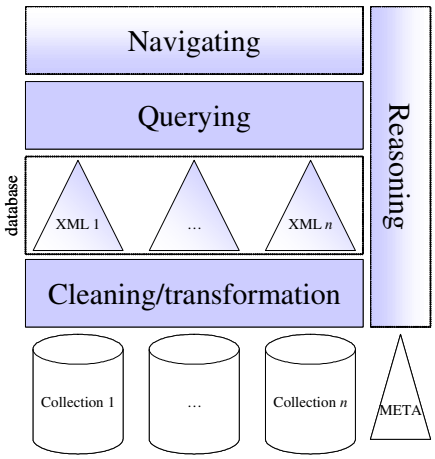
[*]University of Amsterdam, [#]Leiden University, [+]Utrecht University

The field of linguistic typology may be defined as follows [1]:

> The study of the similarities and differences between languages, regardless of any genetic relation, and the resulting categorization of language into 'types'.

This kind of linguistic research involves the collection of information about linguistic phenomena for a representative sample of the languages of the world. Many researchers have stored collections in a digital form, *i.e.* databases. The purpose of the Typological Database System (TDS) project is to make these typological collections of participating institutes available through a unified interface and to allow sophisticated searches across collection boundaries.

At the lower system level this poses a classical data integration or warehousing problem. At the level of semantic integration the aim of the system is closely related to the vision of the semantic web. Concepts in the various databases may represent divergent theoretical perspectives on linguistic phenomena. Hence the TDS system has to address the issue of conflicting conceptualizations of both contributors as well as users. Because knowledge representation technologies, *i.e.* ontologies, allow for the description of the (partial) relationships between concepts, the TDS system uses an ontology to assist the user in conducting typological research. The challenge of the TDS project is to bridge the inevitable gap between the conceptual models of the domain envisaged by contributors and end users, while keeping the meaning of the data intact.



The figure on the right shows the TDS system architecture. The TDS receives stable snapshots of the databases on which to operate. These snapshots are converted to XML documents with a common format and common concepts (where possible), using lossless transformations. The documents are then merged into one XML document based on a common key, *i.e.* a SIL language code [2] or a TDS specific code.

The concepts present in the global TDS XML document are mapped to semantic classes formulated in the linguistic ontology, which is constructed in OWL format using Protégé. This ontology is primarily built bottom-up: classes in the ontology are created on the basis of concepts present in the participating collections. The metadata contains extensive descriptions of the meaning the concepts (in many cases still strongly biased to the theoretical view of a specific database) and the rationale behind the mapping between a concept and an ontology class. The provision of multiple description levels gives a user full specification about a concept in both its source and in the context of the ontology.

The first demonstration system is now online [3]. Future enhancements include the use of the linguistic ontology to allow “smart searches”, *e.g.* extending queries with semantically related concepts and assessing the viability of queries.

References

[1] *OLAC Linguistic Subject Vocabulary*, Open Language Archives Community (OLAC)

http://www.language-archives.org/REC/field.html#typology

[2] *Ethnologue, Languages of the World*, Summer Institute of Linguistics (SIL)

http://www.ethnologue.com/

[3] *Typological Database System*, Netherlands Graduate School of Linguistics (LOT)

http://languagelink.let.uu.nl/tds/