Integrated access to diverse linguistic databases with the Typological Database System

Language typology is a data-intensive discipline, and typological databases are a valuable tool for the typologist. While there are a number of broad-coverage typological databases intended for general use, such as WALS (Haspelmath, Dryer, Gil & Comrie 2005), the great majority of typological databases constitute personal or small-group data collections that are focused on a particular research agenda. Increasingly, such collections are being made available to the linguistic community over the internet, providing the potential for enormous increases in the power of exploratory typological investigation. However, effective utilization of these databases is hindered by the fact that they are a) distributed over a great number of locations, and b) heterogeneous in theoretical background and terminology, organization of the data, and user interface.

The Typological Database System (TDS) is a web-based service allowing unified queries over multiple independently developed typological databases. It currently contains information on circa 1,000 languages from nine integrated databases. The component databases contain data on a range of linguistic features, including agreement, parts-of-speech, word order, predication phenomena, reflexivization, phonological segment inventories, and stress placement. While most data is in the form of "analytical variables" that describe the language as a whole, some databases include primary linguistic data in the form of glossed sentences. We describe the TDS service, the problems involved in integrating diverse typological databases, and the solutions adopted by the TDS. The server, which is operational but is still under development, can be accessed at *http://languagelink.let.uu.nl/tds/.*

The problem of managing and presenting information becomes ever more important as the information available on the internet grows (Abiteboul et al., 2000). As more typological databases become accessible to users other than their creators, colleagues, and others already familiar with them, the following tasks become more challenging:

1. **Resource discovery.** This is simply the step of finding a datasource that contains information on some given topic.
2. **Correct and effective use.** Databases use varying terminology, notation, organization of the data, and search commands. Even if these are documented in detail, they can be quite difficult for a new user to assimilate and employ properly.
3. **Efficiency of resource utilization.** As the amount of online information grows, the time and effort involved in searching databases one by one and collating the results becomes an obstacle to their efficient utilization.

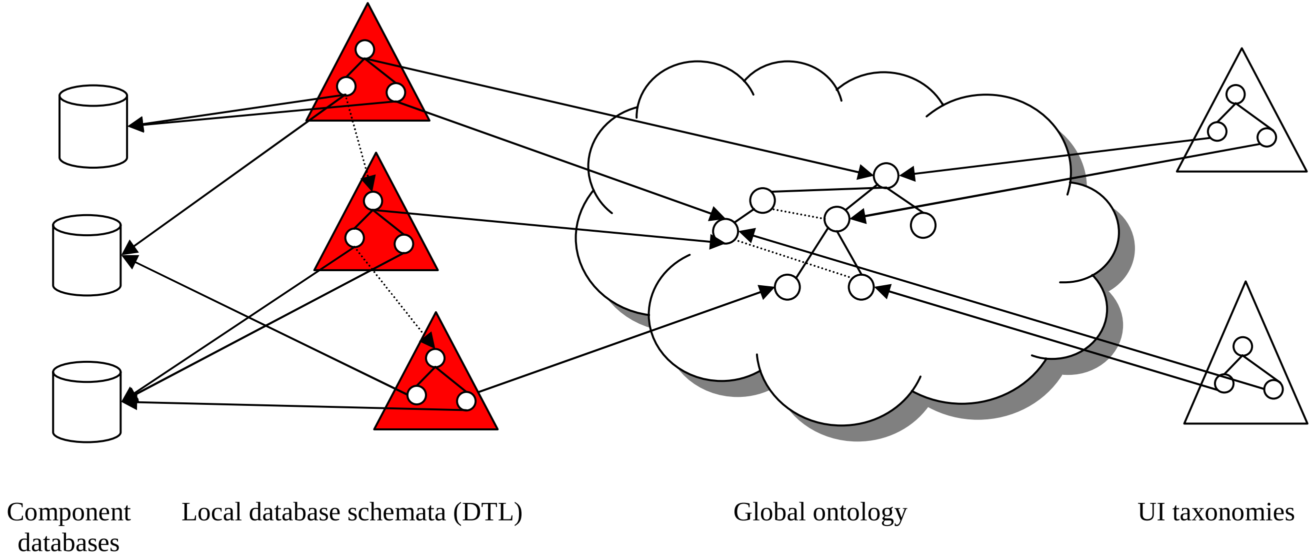
The first of these problems is being addressed through various new metadata standards and initiatives for improved resource description and discovery, including the Dublin Core Metadata Initiative (general), and the Open Language Archives Community (linguistics specific metadata and harvesting protocol). In a related development, standards for the encoding and structuring of linguistic resources, (e.g., the proposed standard of the ILSE Metadata Initiative), allow improved interoperability and integration of conforming resources.

The TDS project directly addresses the second and third challenges, for a moderately-sized collection of independently developed typological databases. The unified interface of the TDS allows speedy combined searches through a single user interface, employing as much as possible a consistent terminology and data organization. The system is oriented toward thorough, highly accurate integration rather than breadth of coverage: Our plans call for the integration of dozens rather than hundreds of databases.

The TDS approach distinguishes between variation in structure or encoding, which is judged to be a design choice of no inherent linguistic significance, and variation in the choice of linguistic terms and (especially) categories and distinctions. While the initiatives mentioned earlier may one day reduce the diversity in structure and encoding among databases, they will have no effect on the divergence of theoretical viewpoints and research traditions that constitutes the most intractable source of heterogeneity. For example, the notions *Subject* and *Object* are understood in several different ways by linguists; and there are alternative ways of expressing structural relationships, e.g., the S/A/R/P distinction as recently applied in Haspelmath (2005). These diverse viewpoints are not only dearly held by their practicioners: They are the subject matter of linguistic analysis, and cannot, and *should not,* be replaced by any uniform, agreed-upon framework. Moreover, it is in the nature of such systems that they are not, as a rule, interconvertible without loss of information.

Accordingly, the TDS approach is to compensate, as much as possible, for purely notational variation in structure or encoding; but to preserve and highlight theoretically salient differences, and alert the user to them by making available detailed documentation for each data field in the database. The data catalog of each component database is mapped, by means of the custom-developed "Data Transformation Language", into a uniform hierarchically-organized space (the "DTL schema"). The schema includes common, database-independent notions as well as database-specific ones, allowing flexibility in the treatment of imported data. The system is supported and cross-linked by means of a global ontology of linguistic concepts, which also allows for multiple alternative perspectives on the same phenomenon. (This is therefore a "two-level" architecture; cf. Stuckenschmidt & van Harmelen, 2005). The ontology also supports the process of finding relevant database fields among the very large number of integrated fields (already in the thousands).

Figure 1: The TDS metadata architecture



References

Abiteboul, Serge, Peter Buneman, & Dan Suciu (2000). *Data on the web: From relations to semistructured data and XML*. San Francisco: Morgan Kaufman.

Haspelmath, Martin (2005). Argument marking in ditransitive alignment types. *Linguistic Discovery* 3.1.

Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (2005). *World Atlas of Linguistic Structures*. Oxford: Oxford University Press.

Stuckenschmidt, H., & van Harmelen, F. (2005). *Information Sharing on the Semantic Web*. Berlin: Springer.

Websites

[DCMI] Dublin Core Metadata Initiative. http://dublincore.org/.

[IMDI] ISLE Metadata Initiative. http://www.mpi.nl/IMDI/.

[OLAC] Open Language Archives Community. http://www.language-archives.org/.

[TDS] Typological Database System. http://languagelink.let.uu.nl/tds/.