abstract 131

Crowdsourcing WALS

Theme session Linked Data in Typology

WALS (Haspelmath et al. 2005) is a hallmark of typology and has revolutionized the way how linguists do typology (Donohue 2006).
Languages are squared with features. At the intersection, we find feature values. E.g. the language English has the feature 'word order' with the feature value 'SVO'. WALS lists 192 features and 2678 languages. However, the resulting data matrix is very sparse, and instead of the possible 514176 data points, there are only about 68000, or 13%.
Already at 13% filling, the matrix is interesting for typologists, and many researchers make use of WALS data, leading to many papers, some of them very influential, e.g. Atkinson (2011). A less sparse matrix is a clear desideratum for typology.

The WALS team do not have the capacity to fill in all the data points, but there are many projects out there, without institutional links to WALS, which collect data according to features defined by WALS and languages as defined by WALS. These projects either add additional feature values for languages which were lacking a value for a certain feature. Or they add a novel feature and add feature values for languages, using WALS codes for the languages. The problem is how these additional data can be made available to other users of WALS.

The WALS database is closed. There is no interface and no process to ingest new data into WALS. The reason for this is that MPI-EVA will not allow write access to its databases from the outside, even less so for random people. Furthermore, WALS has a very high scientific reputation, which could be tarnished by low quality contributions.

The solution for these problems is that data producers store their data points in their own web space according to Semantic Web principles. The data points are then registered and made available to the WALS project for harvesting, including provenance data. The WALS project then makes available the harvested data on the WALS site under a specific label, e.g. "WALS community", in addition to the WALS core already available. Users can then choose whether they want to query only the curated data by the WALS editors, or whether they are also interested in data from other data providers.

This procedure has the following advantages

- Every project manages its own data points. No security risks for central WALS core server
- Clear indication of provenance
- Data producers do not have to care about web server and database administration. They only have to provide their data points in the specified format, which is reasonably trivial
- Shared implementation allows for easy aggregations
- Clear definition of work flow allows automation of processes

This talk will explain the general setup of the crowdsourcing project and show a proof of concept where "WALS core" and "WALS community" are accessed in one query.

References

Atkinson, Quentin. 2011. “Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa”. *Science* 332, 346-349.

Donohue, Mark. 2006. *Review: Typology: Haspelmath, Dryer, Gil & Comrie (2005).* http://linguistlist.org/pubs/reviews/get-review.cfm?SubID=71168.

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