Virtual Language Observatory

Explore the world of language resources and technology fom different perspectives

- Geographical Browsing
- Facetted Browsing
- Catalogue
- LRT Inventory

A portal to the language resources and tools universe

www.clarin.eu/vlo

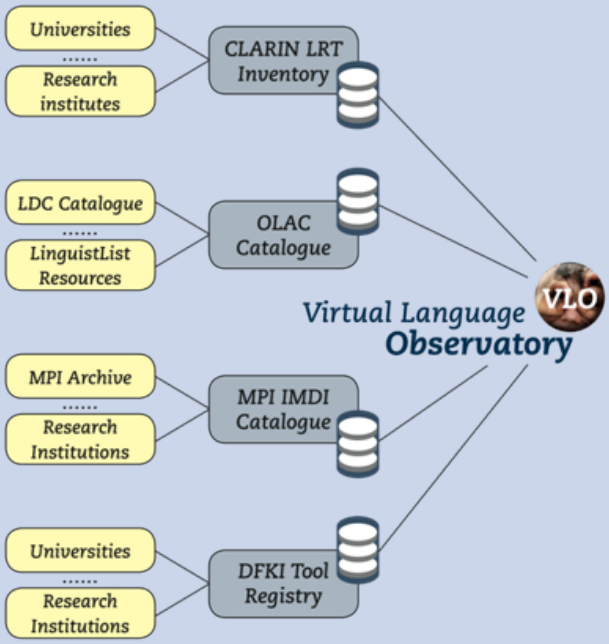
Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, Mariano Gardellini
Max Planck Institute for Psycholinguistics · Email: {firstname.lastname@mpi.nl}

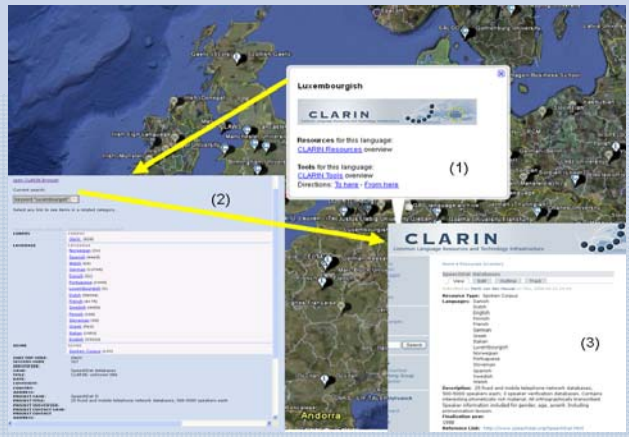
Introduction

The *Virtual Language Observatory* provides multiple views on metadata for linguistic data and software tools. In analogy with the astronomical virtual observatories, it tries to give a consistent online overview of the data and tools that are available at a large variety of resource centres worldwide. **Metadata must be open to allow everyone to find useful resources and tools**. VLO is simply one portal that is harvesting all available metadata and giving the credits to the providers.

Facet browser

All harvested metadata sources are brought together by mapping their respective field descriptors to two metadata sets: for describing resources and tools. From these two metadata sets, a small number of field descriptors are chosen as facets, providing users with well-defined entry dimensions to all resources and tools via a faceted browser, using the Flamenco toolkit. The six facets to which all of the metadata records are mapped are currently *country*, *continent*, *origin*, *language*, *organization*, *genre* and *subject*. Faceted search allows users to find resources in an intuitive manner: each facet selection reduces the number of resources or tools that fall into the selected categories, sharing those common properties. In contrast to traditional hierarchical browsing, faceted browsing offers many different access path to a resource or tool of interest.





By clicking on a language marker (1) the user gets information about a language, together with links to the relevant entries in the faceted search interface (2), where one of the options is to see the resource metadata record in its original context (3), in this example the CLARIN LRT inventory.

Future work

- Curate the metadata: correct typographical errors, add information, use consistent terminology, etc.
- Process CMDI- and ISOcat based metadata
- Use (emerging) standards to refer to persons, projects and resources in a persistent and interoperable way

Data sources

Metadata has been collected from the CLARIN LRT inventory, the MPI IMDI portal which includes a large variety of data providers amongst which is the DOBES collection about endangered languages, linguistic archives distributing their metadata within the Open Language Archives Community (OLAC), the DFKI softwareregistry, a sample of the ELRA catalogue and data from the World-Atlas of Languages (WALS). CLARIN is committed to maintain this VLO portal to serve the interested community.

The GIS perspective

We have also created a Google Earth overlay, combining geographic information with metadata-based information. This work is partially based on the Language-Sites collection and has been extended with links to typological information about the languages from the WALS database and all DELAMAN research centres.

An important aspect is the interaction with the aforementioned faceted search: the user can reach a particular facet-view by clicking on a point associated with a language. From everywhere the user can access the catalogues or even the resources if pointers are provided.

Links

CMDI: http://www.clarin.eu/cmdi
DELAMAN: http://www.delaman.org/
DFKI: http://registry.dfki.de/
DOBES: http://www.mpi.nl/dobes
ELRA: http://catalog.elra.info/
Flamenco: http://flamenco.berkeley.edu
ISOCAT: http://www.isocat.org/
Language-Sites: www.languagesites.org
LRT inventory:
http://www.clarin.eu/view_resources and
http://www.clarin.eu/view_tools
IMDI: http://www.mpi.nl/imdi
MPI catalogue: http://corpus1.mpi.nl/
OLAC: http://www.language-archives.org/
WALS: http://wals.info/