Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

February 2010

Relevance

CLARIN community

for all communities

Virtual Language Observatory

What is it?

The Virtual Language Observatory (VLO) is meant to be the open market place where users can find metadata descriptions of all language resources and tools/services which we can harvest from any useful and trusted source. Currently VLO contains more than 230.000 resources and more than 400 tools already. Different user interfaces are maintained to allow users to find and select resources such as a GoogleEarth overlay for geographic browsing, a facetted browser for easy search and browsing along major criteria and a normal catalogue. The VLO machinery is ready to harvest various types of metadata that is offered via the OAI-PMH protocol. It currently is harvesting data from OLAC, DFKI Tool registry, DOBES, DELAMAN partners, MPI registry, ELRA catalogue and the CLARIN Language Resource and Technology inventory which was meant as a simple registry for resources and tools from CLARIN members. VLO is based on the principle that metadata needs to be open.

What is it for?

The Virtual Language Observatory wants to help researchers to easily find suitable language resources and tools to carry out their research work. They can do this by searching, browsing and navigating geographically. Once they have found a useful resource they can then easily find tools which may work on it. The purpose is that users may directly access the resources or services they have found, given they have the necessary permissions.

Currently the landscape is rather fragmented so that it is very difficult to find useful data and tools. Also the ways in which data resources and tools are described by metadata are very heterogeneous: most resources and tools are not registered at all; many of them are just listed on web sites without using an agreed system; many are described by metadata, but different element sets and vocabularies are used. VLO tries to bring all information together by using the widely agreed OAI-PMH metadata protocol where possible and by mapping the different element sets. Finally all metadata used by VLO will be mapped to the new metadata element set which has been registered in the ISOcat concept registry.

All information in VLO will be available to other metadata harvesting services, i.e. other interested initiatives can create their own portals with special selections, for example, to help users.

Who can use it?

The Virtual Language Observatory can be used by researchers from all disciplines and in fact VLO is interested in harvesting even metadata descriptions about language resources and tools that have been created in other disciplines. Of course we will need to invent suitable dimensions such as “language spoken” or “genre” to structure the huge amount of data in the open VLO market place. This, however, will only work efficiently when all metadata descriptions fulfill some minimal quality requirements which will require much curation effort. Efficient searching will also depend on the familiarity of the user with the semantics of the elements.

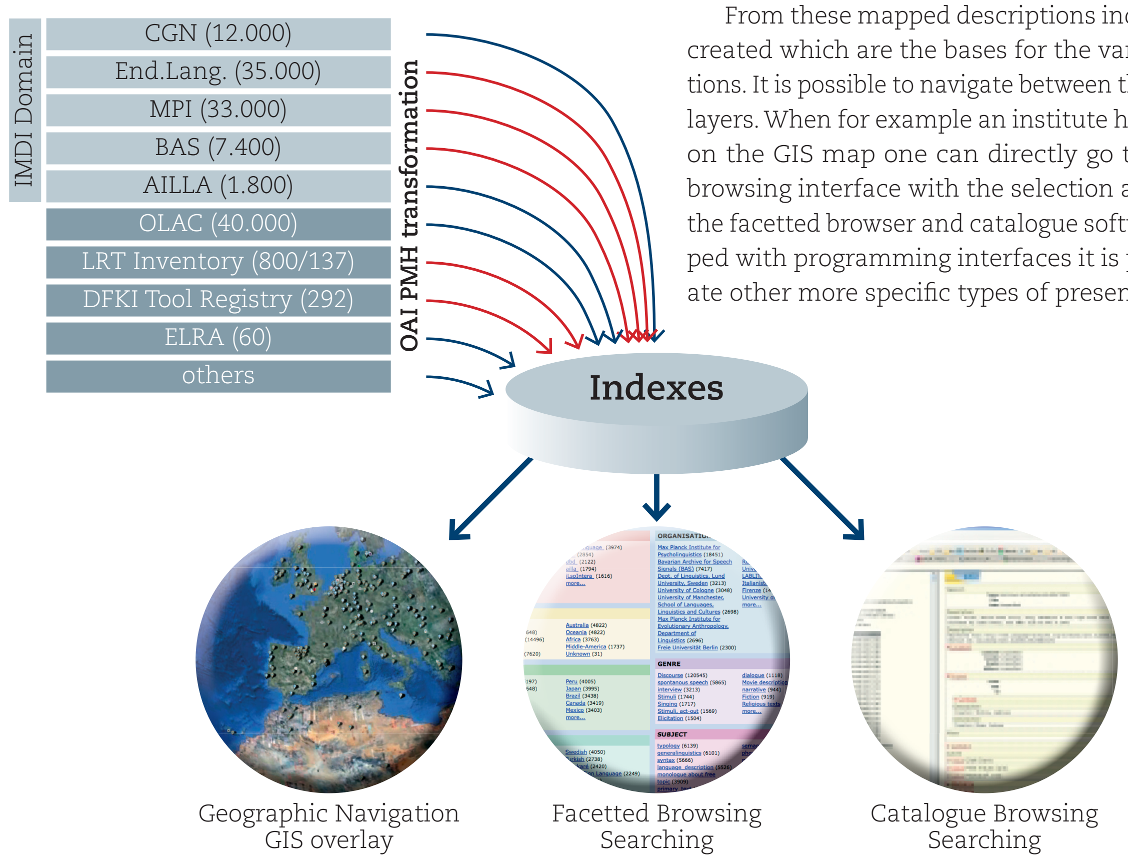
When can it be used?

It can be used now and it will be improved stepwise. We invite everyone to visit VLO, try it out and help in improving it. The basic harvesting, mapping and presentation machinery is in operation. However, the major effort for offering better search and facetted search will not be achieved by improving the technology, but by improving the quality of the metadata descriptions. Yet too little effort is taken by the researchers. To expect that social tagging, i.e. shifting responsibility to the user side, will lead to efficient searching is a trap. Social tagging mainly works if many people do some tagging on the resources which cannot be assumed for language resources in general.

How does it work?

Metadata descriptions from a wide variety of sources are being harvested. According to the descriptions of their element semantics they are currently mapped to IMDI categories. In late 2010 IMDI will be replaced by CMDI which is based on the element set registered in ISOcat. Currently, in many cases manual curation is necessary to semantically align the descriptions to allow a harmonized mapping. In particular the vocabularies for elements such as “institute names” and “genre” are not harmonized. Since in most cases no geographical coordinates are being provided mapping to a geographic system also requires manual operation.

From these mapped descriptions indexes are being created which are the bases for the various presentations. It is possible to navigate between the presentation layers. When for example an institute has been chosen on the GIS map one can directly go to the facetted browsing interface with the selection as a filter. Since the facetted browser and catalogue software are equipped with programming interfaces it is possible to create other more specific types of presentations.



Who is responsible?

VLO is currently maintained by the MPI team, but as indicated all software and metadata is open, i.e. anyone else can make an even better portal to the open market place of language resources and technology. For the quality of the metadata descriptions the data and tool providers are responsible.

Whom to contact?

For the CLARIN infrastructure initiative the official website has the most recent information: http://www.clarin.eu

For all matters you can contact (Alexander König): Alexander.Koenig@mpi.nl

Where to find more information?

The official CLARIN web-site is the source of most information:

CLARIN: http://www.clarin.eu

CLARIN Metadata Requirements Document:
http://www.clarin.eu/deliverables

CLARIN Metadata Short Guide:
http://www.clarin.eu/documents/short-guides

ISOcat doumentation: http://www.isocat.org

Responsible for the content:
Alexander König
MPI
Website: www.clarin.eu
Email: Alexander.Koenig@mpi.nl