Virtual Language Observatory:

the portal to the language resources and technology universe

Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, Mariano Gardellini
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, the Netherlands
Email: {firstname.lastname@mpi.nl}

Abstract

Over the years, the field of Language Resources and Technology (LRT) has developed a tremendous amount of resources and tools. However, there is no ready-to-use map that researchers could use to gain a good overview and steadfast orientation when searching for, say corpora or software tools to support their studies. It is rather the case that information is scattered across project- or organisation-specific sites, which makes it hard if not impossible for less-experienced researchers to gather all relevant material. Clearly, the provision of metadata is central to resource and software exploration. However, in the LRT field, metadata comes in many forms, tastes and qualities, and therefore substantial harmonization and curation efforts are required to provide researchers with metadata-based guidance. To address this issue a broad alliance of LRT providers (CLARIN, the Linguist List, DOBES, DELAMAN, DFKI, ELRA) have initiated the Virtual Language Observatory portal to provide a low-barrier, easy-to-follow entry point to language resources and tools; it can be accessed via http://www.clarin.eu/vlo

1. Introduction

CLARIN (Common Language Resources and Technology Infrastructure) is an [ESFRI] project to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities. CLARIN is offering its services to (1) the different communities of linguists to optimize their models and the tools to the benefit of all using language material, (2) the humanities scholars in the broad sense to facilitate access to language resources and technology and (3) the society to enable lower thresholds to multicultural and multilingual content.

At the time of writing the CLARIN consortium comprises 33 partners and 174 member institutions in 33 countries. About 25 centres within CLARIN are offering numerous language resources and tools to the research community . To foster the sharing of these resources and tools but also to increase visibility to a wider community, we have “opened” the Virtual Language Observatory.

2. Virtual Language Observatory

The *Virtual Language Observatory* provides multiple views on metadata for linguistic data and software. In analogy with the astronomical virtual observatories [Virtual Observatory], it tries to give a consistent online overview of the data that is available at a variety of computing centres. At the time of writing, metadata has been collected from the CLARIN [LRT inventory], the [MPI] IMDI archive (this includes the [DOBES] corpora of endangered languages), linguistic archives distributing their metadata within the Open Language Archives Community [OLAC], the [DFKI] software registry and a sample of the [ELRA] catalogue .

The resulting data sources are then brought together by mapping their respective field descriptors to two metadata sets: one for describing resources, and one for describing 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 facetted browser, using the Flamenco toolkit [Flamenco]. The six facets to which all of the metadata records are mapped are currently *country*, *continent*, *origin*, *language*, *organization*, *genre* and *subject* – as illustrated in figure 1. 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.

The VLO faceted browser displays all metadata (from the various content providers) in a uniform format; in addition, whenever possible, links were created that point to a resource/tool in its original context (*e.g.*, all of the IMDI data in the VLO have backlinks to the IMDI metadata browser, allowing users to continue their search in a tree-structured way). Moreover, we have started to provide links from other metadata viewers (*e.g.,* the geographical browser discussed below) to the faceted browser.

Using faceted browsing on our metadata proved very beneficial for data curation; looking at facet values (e.g., values for the facet 'organization') revealed many data inconsistencies or errors (e.g., many different spellings for our organization, the 'Max Planck Institute for Psycholinguistics'). To support users to signal such inaccuracies to archive management, we plan to augment the faceted browser with an easy-to-use reporting mechanism.

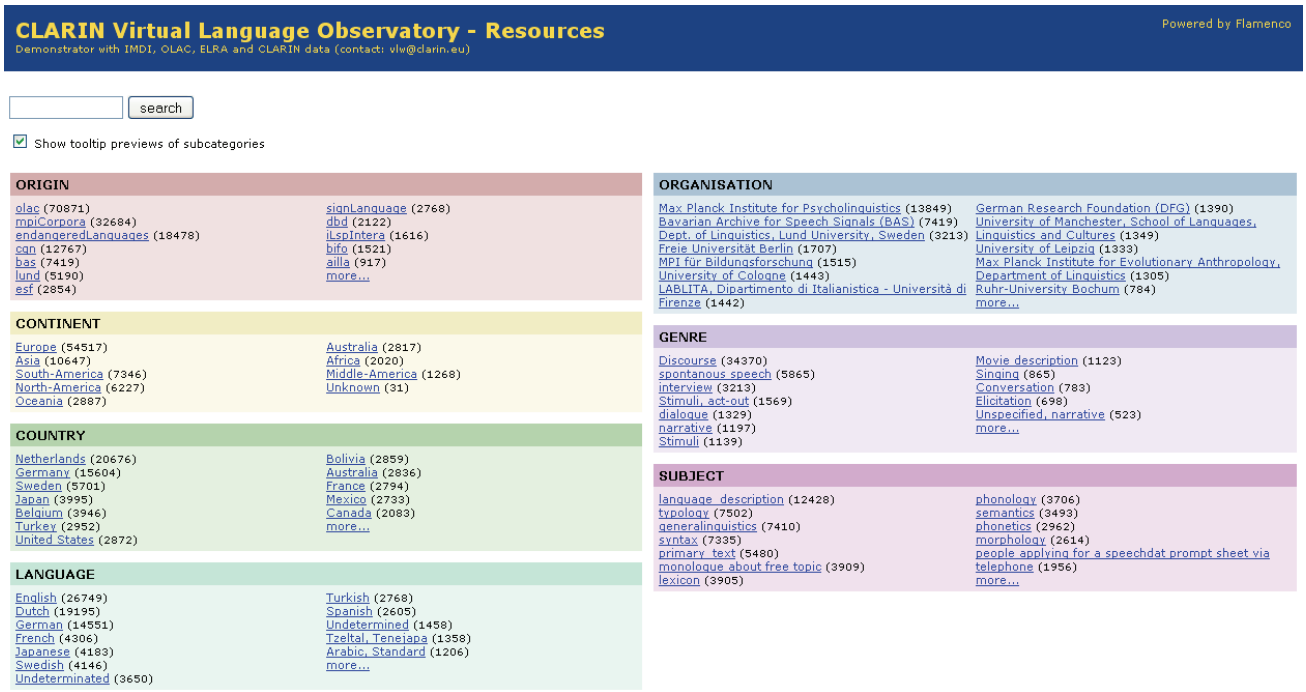

Figure 1: The faceted search interface for language resources at the top level of the Virtual Language Observatory. The number after each facet value indicates the number of metadata records corresponding to that value.

3. 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 [language-sites] and has been extended with links to typological information about the languages from the WALS database [WALS] and DELAMAN research centres [DELAMAN].

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. This aspect is illustrated in figure 2.

4. Services, Consultancy, Information

Next to the VLO, CLARIN also started to offer an overview of relevant web services to the research community, *e.g.,* tokenizers and parsers. A consultancy section, containing FAQs, training material and a list of experts on several subjects complements this – see figure 3 for an impression. Finally all relevant documents produced within the CLARIN research infrastructure [CLARIN], including the so-called short guides explaining a broad range of issues on one A4 paper, are made available. All of this can be found at www.clarin.eu, in the Services and Publications sections.

5. Future work

The Virtual Language Observatory portal is up and running, but there is potential for improving the site and its content. First and foremost, the metadata needs to be further curated. This includes not only the correction of obvious (typographic) errors, the removal of outdated or double entries, the actualisation of metadata, and the filling in of missing information but also a convergence towards controlled vocabularies for descriptor values and the harmonization of language-dependent fields. One of the biggest challenges in this respect is the distributed nature of metadata storage (in particular, with regard to the OLAC providers); here, curation needs to be coordinated and pursued at various sites at once.

The consistent use of persistent identifiers to address all resources and tools is high on the wish list, but also the use of standards for referring to entities such as persons, organizations, publications, etc. The electronic libraries community [DRIVER], and other communities, are facing similar problems and it would be beneficial to seek cooperation with them. The use of [CERIF] (Common European Research Information Format) may solve some of the problems at hand, but more work is needed to judge whether it is the best answer to address all issues. The same goes for the controlled vocabulary service as currently being created by the [CATCH-PLUS] project[1].

Our approach to map the various metadata schemes to each other is *ad hoc.* CLARIN addresses the issue of semantic interoperability by introducing a component-based metadata framework [CMDI], which

[1] We are currently exploring if the integration with CMDI for providing the user with suggestions while entering metadata is an option.

is tied with the ISO Data Category Registry [ISOCAT]. In future work, we would like to take-up the CMDI avenue for bringing together the various distributed metadata for language resources and tools in a more systematic way. Having a well-thought infrastructure in place will realise VLO's vision – to provide the research community with a telescope to the wonderful universe of language resources and tools.

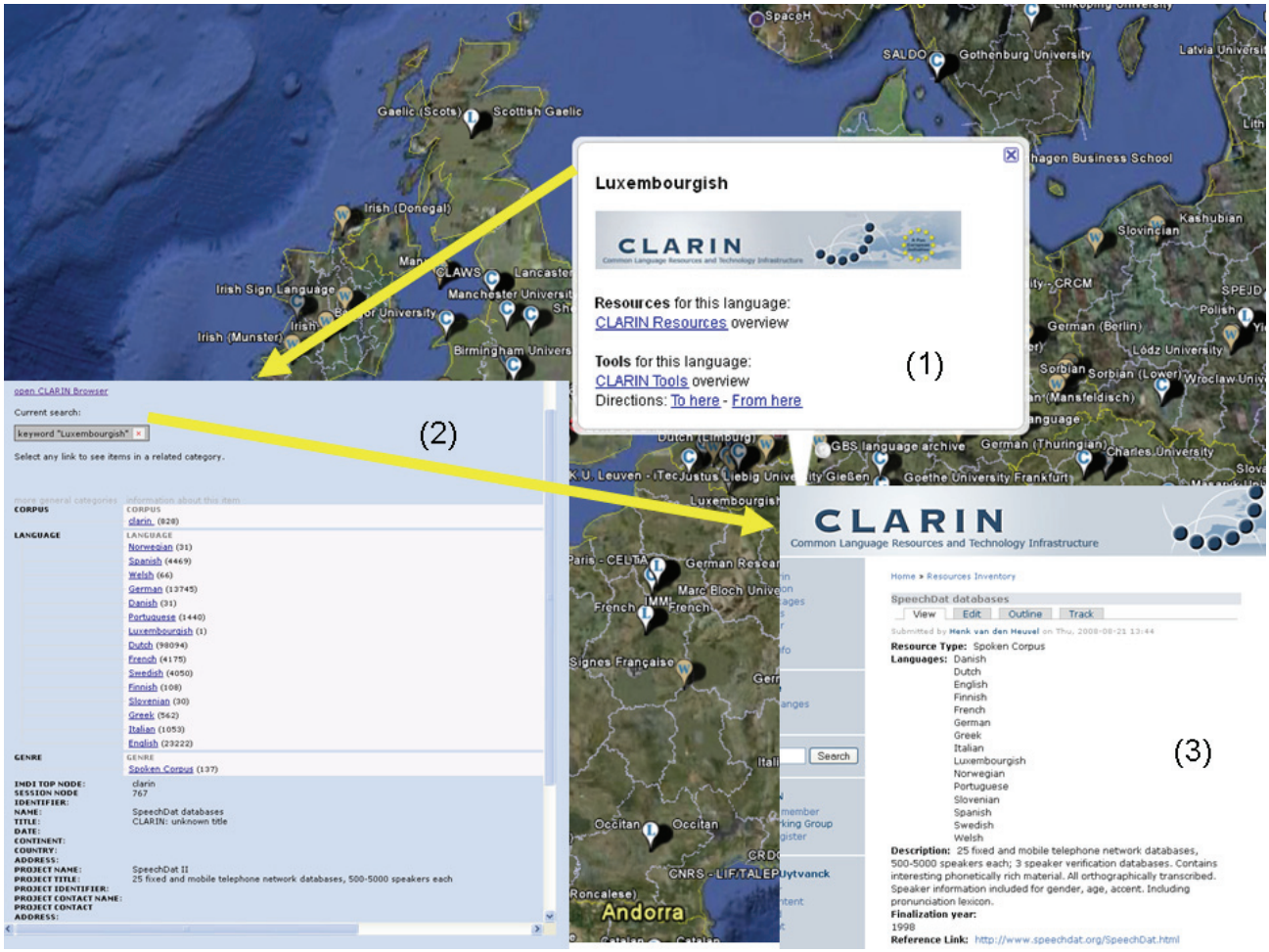

Figure 2: By clicking on a language marker (1) the user gets information about a language, together with links to the relevant entries in the faceted search interface (2), where one of the options is to see the resource metadata record in its original context (3), in this example the CLARIN LRT inventory.

Figure 3: Links to FAQs, training material and human experts.

6. References

[CATCH-PLUS] http://www.catchplus.nl/en

[CERIF] Jörg, B., Jeffery, K. G., Asserson, A., and van Grootel, G. (2009). CERIF 2008-1.0 Full Data Model (FDM): Introduction and Specification. euroCRIS. See http://www.eurocris.org/cerif/cerif-releases/cerif-2008/

[CLARIN] Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., and Koskenniemi, K. CLARIN: Common language resources and technology infrastructure. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[CMDI] Broeder, D., Declerck, T., Hinrichs, E., Piperidis, S., Romary, L., Calzolari, N., and Wittenburg, P. (2008). Foundation of a Component-based Flexible Registry for Language Resources and Technology. *Proceedings of LREC 2008*. See: http://www.clarin.eu/toolkit

[DELAMAN] http://www.delaman.org/

[DFKI] Declerck, T., Jachmann, A. W., and Uszkoreit, H. (2000). The new Edition of the Natural Language Software Registry (an initiative of ACL hosted at DFKI). In *Second International Conference on Language Resources and Evaluation, LREC'00*, pages 1129-1132. See: http://registry.dfki.de/

[DOBES] Wittenburg, P., Mosel, U., and Dwyer, A. (2002). Methods of language documentation in the dobes project. *Proceedings of LREC*, pages 34-42.

[DRIVER] Van der Graaf, M. (2007). DRIVER: Seven Items on a European Agenda for Digital Repositories. *Ariadne*, 52. See: http://www.driver-repository.eu/

[ELRA] http://catalog.elra.info/

[ESFRI] http://ftp.cordis.europa.eu/pub/esfri/docs/esfri-roadmap−report−26092006en.pdf

[Flamenco] Stoica, E. and Hearst, M. A. (2007). Automating creation of hierarchical faceted metadata structures. In *In Procs. of the Human Language Technology Conference,* http://flamenco.berkeley.edu

[ISOCAT] Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. (2009). ISOcat: remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261-276. http://www.isocat.org/

[language-sites] Van Uytvanck, D., Dukers, A., Ringersma, J., & Trilsbeek, P. (2008). Language-sites: Accessing and presenting language resources via geographic information systems. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

[LRT inventory] http://www.clarin.eu/view_resources and http://www.clarin.eu/view_tools

[MPI] Broeder, D. and Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119-132.

[OLAC] Simons, G. and Bird, S. (2003). Building an open language archives community on the OAI foundation. *Library hi tech*, 21(2):210-218.

[Virtual Observatory] Virtual Observatory. (2009, December 3). In *Wikipedia, The Free Encyclopedia*. Retrieved 16:08, March 12, 2010, from http://en.wikipedia.org/w/index.php?title=Virtual_Observatory&oldid=329417615

[WALS] Cysouw, Michael. 2008. Inclusive/Exclusive Distinction in Independent Pronouns. In: Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, Bernard (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. See http://wals.info/