On using the META-SHARE repository software to support a CLARIN centre

Neeme Kahusk
University of Tartu
Tartu, Estonia
neeme.kahusk@ut.ee

Stelios Piperidis
R. C. "Athena", ILSP
Athens, Greece
spip@ilsp.gr

Juli Bakagianni
R. C. "Athena", ILSP
Athens, Greece
julibak@ilsp.gr

Jussi Piitulainen
University of Helsinki
Helsinki, Finland
jpiitula@ling.helsinki.fi

Keywords: META-SHARE, CMDI, PID, repository software, SSO, OAI-PMH

1. Introduction

CLARIN (Váradi et al., 2008) and META-NET[1] are two initiatives, that share pretty much complementary and partially overlapping objectives. CLARIN promotes advanced ICT and language technology empowered research in humanities and social sciences, META-NET and META (the Multilingual Europe Technology Aliance), support a joint international effort for furthering Language Technology as a means towards realising the vision of a Europe united as one single digital market and information space.
Both initiatives stress the importance of language resources infrastructures, their availability and sustainability.
The CLARIN vision is based on the following pillars: language coverage, legal issues, integration of data and services, preservation, ease of access, and crossing borders. META-NET activities fall into three main divisions: compiling a strategic research agenda META-VISION, building a distributed resource infrastructure META-SHARE, and collaboration in multilinguality and machine translation research (META-RESEARCH).
In this paper we are going to share some experience on using META-SHARE and the associated software to support a CLARIN centre language resource repository.

2. Requirements for CLARIN certified centres repositories

There are several types of CLARIN centres, the most common type being B centres, that provide a stable repository with CMDI metadata, provided with PIDs and single-sign-on login.
There are certain requirements to be fulfilled to become a CLARIN centre (Wittenburg et al., 2012). Some of them concern organisational matters (e.g. sustainable financing), but there are certain requirements for repository software as well. Shared resources should be sustainable and accessible.
In short, to fulfill the requirements for a CLARIN centre language resource repository, the following features are essential: SSO, CMDI, PID, OAI-PMH, and FCS.
Single-sign-on (SSO) is required mainly for ease of use: after logging in to use one service, the other services in the network become accessible as well.
The Component MetaData Infrastructure (CMDI) is a flexible system for describing metadata of resources (Broeder et al., 2010).
To ensure sustainability of data and proper references, Persistent Identifiers (PID) are used, that remain the same even if the url of underlying resource should change.
The CLARIN Virtual Language Observatory (VLO) (Uytvanck et al., 2010) uses Open Archives Initiative Protocol[2] for Metadata Harvesting (OAI-PMH) for collecting metadata from all centres repositories, and Federated Content Search[3] (FCS) is used for corpus queries in several repositories.

3. META-SHARE infrastructure and repository software

The META-SHARE infrastructure and the associated repository software (Piperidis, 2012) has been provided and used by META-NET Network of Excellence. The software is based on Django, a Web framework written in Python[4].
A META-SHARE node functions both as metadata registry and resource repository, providing both metadata and resource files. META-SHARE provides a convenient metadata editor, that enables the user to edit, import and export metadata. The META-SHARE infrastructure comprises distributed META-SHARE nodes in such a way, that some nodes export their data, while others, the META-SHARE managing nodes, harvest metadata and provide users with a rich collection of metadata originating from many nodes in different locations.
There are local users at every META-SHARE node. For each node, there is a Administrator user, the responsibility of whom is to create user accounts, accounts for organisations and editor groups. Administrators can assign rights to other users to create and update resources and manage organisations and editor groups.
For every node, there is also at least one special user for metadata harvesting. Harvesting is carried out only for metadata, the resources themselves stay at their original locations and can be downloaded, if appropriate licences are agreed to, and if the authenticated user has the rights to download.

[1] http://www.meta-net.eu

[2] http://www.openarchives.org/OAI/openarchivesprotocol.html

[3] http://www.clarin.eu/content/federated-content-search-clarin-fcs

[4] https://www.djangoproject.com

There is a useful helper for META-SHARE node users, that is the metadata editor. Having the roots in a web application form, it provides mandatory entries as well as recommended and optional ones. The metadata editor saves the result in an XML file and provides validation functionalities as well.

4. Adapting META-SHARE software for CLARIN centre

Being an open-source BSD-licenced software, the META-SHARE repository software is easily configurable and adaptable. From the features needed for a CLARIN centre, SSO and CMDI are in top priority.

In its original form META-SHARE does not have SSO functionality, but there is a module for Django using SAML protocol, djangosaml2[5], that is adjusted for the needs of a CLARIN centre.

To be functional with SSO, the djangosaml2 module was installed to a META-SHARE node and configuration files were changed in order to manage users entering via the corresponding identity federation (Kahusk, 2014). Still, just now, the local user IDs are not merged with remote ones, while this is the responsibility of the site administrator to grant appropriate rights to the federated users. This is a one-time effort for a production server that has local users already; for a new setup it should be done anyway, for local or remote users.

The metadata schema used in META-SHARE is not identical to CMDI, but due to flexible nature of the latter it does not cause much trouble.

The basic nature of CMDI (Gavrilidou et al., 2012) is in profiles built from components, so it is natural to define one or even more profiles that are compatible with the META-SHARE metadata schema. Currently there is a requirement for CLARIN centres that metadata components must have links to ISOcat entries. This is mostly true for metadata in META-SHARE as well, and if not, it is not an technical issue, but an organisational one, as it boils down to finding the appropriate ISOcat category and adding it to the profile description.

As metadata in both META-SHARE and CMDI are represented in XML format, and they are in principle compatible, the problem of converting them is a matter of translating one xml schema to another via an xsl stylesheet.

5. OAI-PMH protocol implemented in META-SHARE

A new ready to be released version of META-SHARE supports the harvesting of metadata with the OAI-PMH protocol. The implementation of the OAI-PMH protocol in the META-SHARE repositories enables not only the harvesting of metadata between META-SHARE nodes, but also the connection with other OAI-PMH compliant repositories like CLARIN VLO.

In terms of OAI-PMH, a META-SHARE node is both a Data Provider, exposing its metadata via OAI-PMH, and a Service Provider, harvesting metadata. The OAI-PMH implementation in META-SHARE supports the exposure and collection of metadata that are formatted in the following metadata schemata: META-SHARE v3.0, CMDI_META-SHARE and OLAC.

As a META-SHARE node is a Data Provider, it responds to all OAI-PMH service requests and exposes its metadata in all the already mentioned metadata formats. The OAI-PMH implementation in META-SHARE supports the grouping of its records into hierarchical sets. As the resources in META-SHARE are language resources, the top-level sets consist in the resource types (corpus, tool/service, language description and lexical conceptual resource). There are also lower-level sets, which are specific to each resource type and give the opportunity to a Service Provider to harvest resources from META-SHARE that are specific to its needs.

When a META-SHARE node acts as a Service Provider, it can harvest a single record, a set of records or all the records of a Data Provider. If a record is newly added to the Data Provider, it will be harvested; if it is updated, the META-SHARE node will update its copy of the record accordingly, and if it is deleted, it will mark this record as deleted.

6. Towards a better repository software

The Estonian CLARIN centre CELR[6] is currently using META-SHARE software for repository[7]. The original software has undergone minimal modifications, most notable of them is integration with djangosaml2 module to provide SSO functionality.

There are further issues that can be addressed if META-SHARE software is to be used as a production resource node for a CLARIN centre. There are fields for PID in the META-SHARE metadata editor, but the responsibility to provide PIDs is on resource manager or site administrator. It would be much more elegant if PIDs would come up in some automatic way. There is an API for handles that can be used to automate the PID assigning process and it is certainly possible to integrate it with the META-SHARE editor.

There are fields for version information in the META-SHARE editor, and this is of great usefullness. But just now, the function of getting repeated information from one version to another is very limited. Actually it can be done only so, that the metadata is exported from one version and then imported to the next version, to be fine-tuned with the editor then. Certainly, this process can be done in a better way, but this is more a matter of ease of use than a real need.

7. References

Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A data category registry- and component-based metadata framework. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 19–21. European Language Resources Association (ELRA).

[5] https://bitbucket.org/lgs/djangosaml2

[6] http://ee.clarin.eu

[7] https://metashare.ut.ee

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE metadata schema for the description of language resources. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 23–25. European Language Resources Association (ELRA).

Kahusk, N. (2014). How to: META-SHARE SSO. How to documentation, University of Tartu, Tartu, Estonia, May 27. https://metashare.ut.ee/site_media/metashare-simplesaml-howto.pdf.

Piperidis, S. (2012). The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 23–25. European Language Resources Association (ELRA).

Uytvanck, D. V., Zinn, C., Broeder, D., Wittenburg, P., and Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 19–21. European Language Resources Association (ELRA).

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., and Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Wittenburg, P., Uytvanck, D. V., Zastrow, T., and Offersgaard, L., (2012). *Clarin Centre Types*. CLARIN. http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-77.