Semantic metadata mapping in practice: the Virtual Language Observatory

Dieter Van Uytvanck, Herman Stehouwer, Lari Lampen
{firstname.lastname}@mpi.nl
Max Planck Institute for Psycholinguistics

Introduction

In the era of the digital data deluge, a researcher needs efficient ways to navigate to the language resources that really matter, whatever the selection criterion is. A plethora of resource inventories and catalogues has been proposed to address this need. However, almost all of them are based on a single metadata scheme, forcing the resource providers to trade off accuracy in favor of compatibility.

The Component MetaData Infrastructure (CMDI, see [1]) tries to address this by composing a tailored metadata schema that relies on pre-canned components with explicit semantic declarations. The challenge that comes with this approach is providing a uniform and easy to use interface to search in the resulting metadata records. The CMDI facet browser that forms the backbone of CLARIN's Virtual Language Observatory does so. In this paper we explain how we gather a large collection of varied metadata records and make them accessible using the CMDI infrastructure as the semantic backbone.

Figure: interface of the VLO facet browser (source: www.clarin.eu/vlo)

Overview of the process

The dissemination of language metadata descriptions is something that happens in a distributed manner. All resource-providing centers create XML descriptions and offer these via the HTTP-based OAI-PMH protocol [3]. Traditionally the format used for metadata is OLAC, an extension of the ubiquitous Dublin Core

schema, targeted towards the linguistic community. Although useful, OLAC lacks a deep semantic definition and relies mostly on best practice guidelines [4]. To overcome these shortcomings various language resource centers are providing CMDI descriptions. This format allows its users to define a set of elements with links to the ISOcat [12] data category registry to ensure semantic interoperability.

Once harvested, the existing relevant OLAC records are converted to CMDI files as well. This approach allows for an elegant and uniform way of importing the metadata into a data store. In previous versions of the VLO [2] this import step relied on an ad-hoc mapping with manually defined conversion steps per metadata format. With the introduction of CMDI, it is only necessary to support one import mechanism, albeit a flexible one. The responsibility for the mapping (i.e. the ISOcat links) now fully lies at the side of the resource provider. This approach is also a guarantee for scalability: adding more CMDI-defined metadata schemes comes at little additional cost.

The ingester then comes into action. This import module reads and analyzes the input files. It examines the ISOcat links and generates the corresponding XPaths in the CMDI-files. Then it loads the facet values into Apache SOLR. This generic facet search engine is the backend of the facet browser within the VLO.

Finally, users can browse through the data as processed by SOLR using a self-developed web front-end. When viewing a single record the CMDI source file is processed and turned into a web page with links to the resources that are described. Whenever a Persistent Identifier is detected (currently handles and URN:NBNs are supported) the viewer turns this into a clickable link with the help of the corresponding resolver.

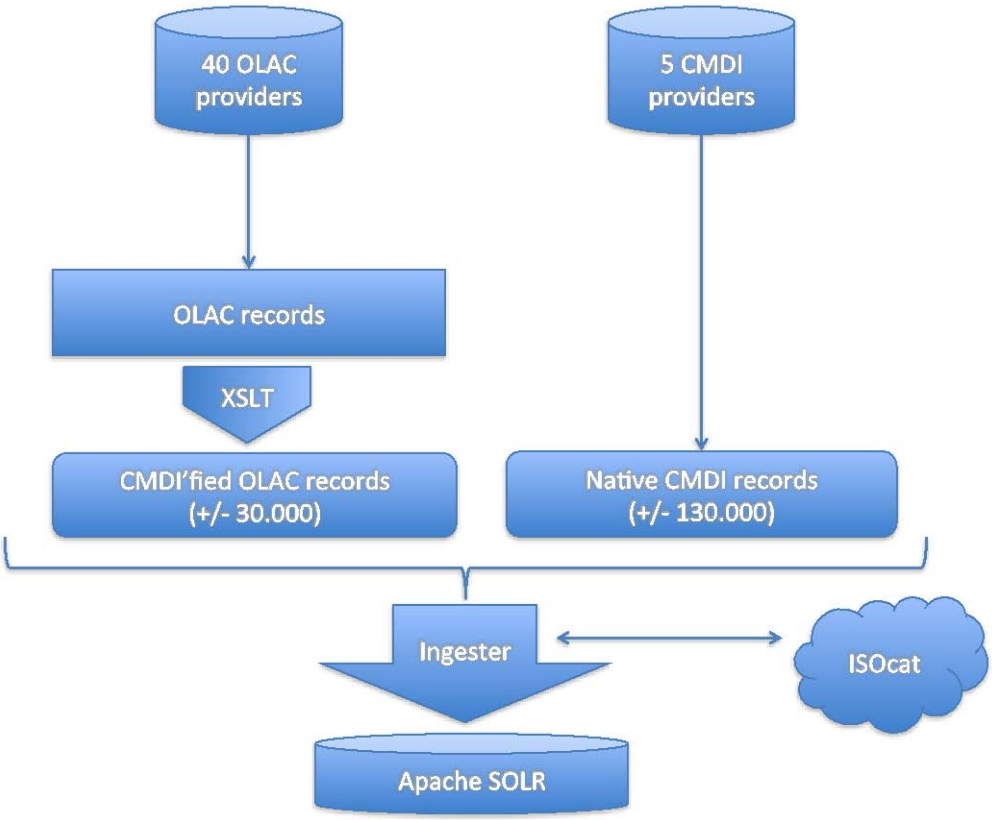

Figure: An overview of the metadata harvesting, conversion and ingestion process

Harvesting the metadata

A harvester, developed internally but in part based on libraries from OCLC [5], is used to collect the metadata from participating providers using the OAI-PMH protocol. The providers comprise two groups: some provide metadata in CMDI format, which is harvested and extracted unchanged, while others provide metadata in OLAC, which is translated into CMDI format by the harvester, as described in more detail below.

Some complications are introduced into the process by transient errors, such as network problems and temporary overload situations. It is effectively impossible to harvest tens of thousands of metadata records without encountering these issues. Consequently, much of the difference between subsequent harvested sets of metadata records consists of newly occurring errors, or errors that have ceased to appear.

To maintain the records in the VLO up to date, it is naturally necessary to repeat the harvesting process periodically. While the OAI-PMH protocol allows incremental harvesting (where only changed and new metadata records are fetched), many providers do not keep track of deleted records. Therefore, to maintain consistency of the harvested records with the source, it is necessary to periodically fetch all records from a provider.

Mapping to CMDI

A migration to a new information architecture paradigm like CMDI never happens at once. Some conversions are necessary. As these could take place at different times and locations we look at some examples.

Post-hoc XSLT conversion: OLAC

Lots of language resource providers still rely on OLAC as a lingua franca. These files are converted immediately after harvesting with the help of an XSLT stylesheet. This maps the elements onto a CMDI profile that contains all of the possible fields in an OLAC record.

A minor amount of curation also takes place during the conversion stage. In particular, a variety of language codes is in common use, including the (now obsolete) SIL and three versions of the ISO standard 639. During the conversion phase, the language codes are unified by translation into ISO 639-3 codes.

Transparent XSLT conversion: IMDI

The metadata repository at The Language Archive [6] uses the IMDI scheme [7] as a backend but delivers CMDI when harvesting over OAI-PMH. To make this possible a CMDI profile with the IMDI elements has been defined. An XSLT processor then converts the IMDI into CMDI instances whenever a request for such a file is received. This conversion thus happens transparently at the provider's side: the VLO is not aware of the IMDI-history of the data it receives.

From a relational database: LRT inventory

The CLARIN LRT inventory [8] is a low-barrier inventory for language resources that was intended to capture information that was not available in an institutional repository. It is powered by the Drupal CMS on top of a relational (MySQL) database. Again a profile was made to hold the information that is stored into the LRT inventory. Then a python script processes the records that are dynamically exported as one CSV-file and generates a CMDI-file for each entry in the database.

Ingesting the CMDI files

After harvesting, the VLO fills a SOLR database. CMDI is a fairly flexible XML format. This means that the information required for the facets can be encoded in different parts of the XML, depending on the repository or even depending on the resource.

In order to alleviate this problem of flexible input formats the VLO relies on ISOcat data categories. These ISOcat data categories are mapped to the matching XPath for the specific CMDI file, based on the XSD of the CMDI profile.

To illustrate this mechanism, consider the mapping to the facet *Language* from the following two different CMDI profiles:

- The CLARIN LRT inventory profile (and the derived XSD)
- The IMDI profile (and the derived XSD)

The *language* facet is semantically connected to the data category Language ID, containing the following definition: *Identifier of the language as defined by ISO 639 that is included in the resource or supported by the tool/service.*

Upon ingestion of a CMDI file that is based on any of the above profiles, the XSD that is generated out of the profile is inspected[1]. Then 2 XPaths are detected that contain a reference to the data category *Language ID*:

- /CMD/Components/LrtInventoryResource/LrtCommon/Languages/ISO639/iso-639-3-code (for the LRT profile)
- /CMD/session/mdgroup/content/content-languages/content-language/Id (for the IMDI profile)

After locating the language element in both profiles the CMDI files are parsed, ingested into the VLO and the facet language is filled for each record using the extracted XPaths. In case no matching data category is found, the VLO allows configuring XPaths directly. This should be considered however as a temporary workaround for profiles that lack ISOcat links.

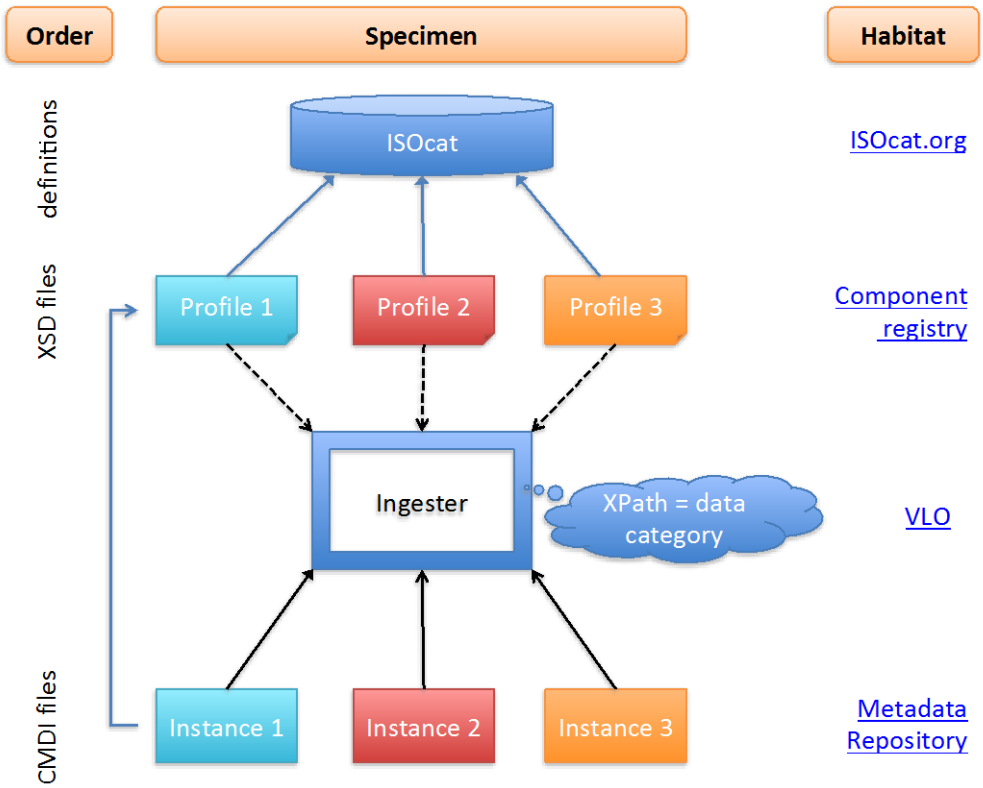

Figure: elements in the CMDI-based ingestion of metadata into the VLO (blue arrows = URLs, black arrows = ... is input for ...)

[1] This is possible because each CMDI metadata file contains a link to its XSD.

Post-processing and import

After the ingest step there is the possibility for a post-processing step. For this step it is possible to register a post-processor for a facet. For each facet the list of all registered post-processors for that facet are executed in turn.

The mechanism of the registered post-processors is used for instance for converting the different language information data formats to the ISO-639-3 language codes. During this step all detected 639-1 and 639-2 language codes are transformed into 639-3 codes.

Afterwards a different post-processor converts the 639-3-code to a human-readable name of the language and makes a link of it to the CLARIN language information website. The latter features links to relevant data sources (e.g. the WALS typological database and the Linguist List site) about that language.

The obsolete ISO-639-2 code DUT is e.g. transformed by the following steps, by which the output of the previous step serves as the input for the next one.

- Iso-639-2 code: DUT
  - Iso-639-3 code: NLD
    - Language label: Dutch
    - Language link: http://www.clarin.eu/external/language.php?code=nld

User experiences

A questionnaire [10] among users of the VLO (the older version as described in [2]) led to the following conclusions:

- Users want to have direct access to the language resources. This was addressed by adding such links.
- Quite some resources can only be accessed after providing a username and password. Some suggestions to make this clear at an early level are given in the next section.
- Many repositories serve records that are hardly relevant in a scenario of electronically enhanced research, e.g. descriptions of books in a library without any ISBN-number or further information. During an iterative quality control procedure they were removed as a data source[2].
- There are some search requirements that cannot be fulfilled with a facet interface. In this case users should be informed about other ways of querying the metadata (e.g. the CLARIN Metadata Browser [11]).

Thorough inspection of the metadata in the VLO led to some additional observations.

- Many metadata records were outdated, pointing to no longer existing online resources. These too were removed as far as possible. A feedback button in the VLO is planned to attend the responsible repository administrator about such issues.

[2] Initially all OLAC providers listed at language-archives.org were harvested.

- Some records contained erroneous information. Due to the distributed nature of the metadata this is a problem that is hard to solve. However we hope that an easy-to-use feedback option could at least enhance the awareness of the providers.

Figure: a metadata record as seen in the VLO's facet browser

Future work

At the time of writing the VLO contains some 190.000 records. Yet there are many more sources of metadata to be added, including the German CLARIN-D [9] centers and the Dutch national library [13].

Another possibility for improvement that was voiced in the user questionnaire is a facet to indicate openness: can any user access a resource, does (s)he have to register beforehand or is it only available to a small circle of people? Although it is hard to come up with a 100% waterproof label a good indication would already improve the usability significantly.

Currently the list of metadata providers is maintained manually. CLARIN plans to initiate a center registry where providers could add (among other things) their OAI harvesting gateway. This list will be sent automatically to the OAI-harvester that powers the VLO.

At the quality and consistency side a controlled vocabulary service has the potential to improve the quality of the metadata descriptions. Such a list could for instance contain organization names and mime types and guide users when generating metadata. After all, creating high-quality metadata is better and cheaper than locating and repairing errors afterwards.

References

[1] Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., and Witt, A. (2011). A pragmatic approach to XML interoperability — the component metadata infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.

[2] Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 900-903). European Language Resources Association (ELRA).

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

[4] http://www.language-archives.org/NOTE/usage.html

[5] http://www.oclc.org/research/activities/past/orprojects/harvester2/harvester2.htm

[6] www.mpi.nl/tla

[7] 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.

[8] http://www.clarin.eu/inventory

[9] http://de.clarin.eu

[10] http://www-sk.let.uu.nl/u/D5R-4.pdf

[11] http://clarin.aac.ac.at/MDService2/

[12] 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/

[13] http://www.kb.nl