Foundation of a Component-based Flexible Registry for Language Resources and Technology

Daan Broeder, Thierry Declerck, Erhard Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari, Peter Wittenburg

MPI for Psycholinguistics, DFKI, University of Tuebingen, ILSP, MPDL, ILSC.
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: daan.broeder@mpi.nl, declerck@dfki.de , eh@sfs.uni-tuebingen.de, spip@ilsp.gr, laurent.romary@mpdl.mpg.de, glottolo@ilc.cnr.it, peter.wittenburg@mpi.nl

Abstract

Within the CLARIN e-science infrastructure project it is foreseen to develop a component-based registry for metadata for Language Resources and Language Technology. With this registry it is hoped to overcome the problems of the current available systems with respect to inflexible fixed schema, unsuitable terminology and interoperability problems. The registry will address interoperability needs by refering to a shared vocabulary registered in data category registries as they are suggested by ISO.

1. Introduction

In the 90-ties it became apparent that the number of digital language resources will increase dramatically due to continuous technological innovation and paradigm shifts. This motivated librarians to start working on the Dublin Core (DC) [1] metadata standard. Amongst others also in the linguistic discipline experts were aware of these trends. In 1998 a first simple electronic form of open metadata description were put into operation to prevent a data cemetery. In January 2000 the ISLE project [2] started with the ISLE Metadata Initiative (IMDI) work and presented the basic ideas at the LREC 2000 conference in Athens [3, 4]. Later in 2000 also the OLAC work was presented at the LDC conference in Philadelphia [5]. In the same time period DFKI started with the NLP software registry [6] and defined a first number of useful descriptors for language technology tools. All three initiatives were started in the domain of linguistics and can be seen as complementary.

IMDI started from scratch, and provided first for a careful analysis of Dublin Core, TEI [7] header tags and initiated a broad discussion in particular amongst European language technology researchers and experts involved in the documentation of endangered languages. The idea was to formulate the needs of the different sub-communities and understand the terminology that is used to come to one integrated metadata set. OLAC on the other hand started with the assumption that the semantics of the DC vocabulary should be re-used and extended where necessary to be suitable for language resources. In the mean time both sets stabilized based on broad discussions and increasing experience in the discipline and they are used by several projects and institutes. While IMDI is seen primarily as an option to gather as much (metadata) documentation as possible and to formulate more detailed research questions, OLAC, like DC itself, can be seen as an umbrella to easily integrate the holdings of different archives. Both metadata sets are formally described and there are XML schemas available to allow verification.

One important difference in focus can be found back in the design of the infrastructure supporting the two sets. OLAC consequently focused on a proper harvesting technique which is based on the OAI PMH protocol [8] and on a search portal. In addition an editor was developed to allow users to create OLAC metadata records. In the case of IMDI where resource management, resource retrieval and resource grouping played equally important roles an editor was developed that allows to create not only metadata descriptions of resource bundles, but also of vocabularies and hierarchies to group resources. A special XML-Browser [9] was developed allowing users to work off-line and to directly process the XML-based descriptions. In addition, a web-application was developed to allow navigating in the linked structure of metadata descriptions with normal www-browsers and structured as unstructured search was supported as well [10]. Also the IMDI portal supports harvesting, but focused on other IMDI type of metadata descriptions. To fit into the DC/OLAC model the IMDI portal also offers a mapping of IMDI to DC and OLAC semantics and supports the OAI PMH protocol allowing various service providers such as OLAC and libraries to harvest all 50.000 IMDI metadata descriptions. In both cases also the infrastructure components stabilized during the first years from 2000 on.

2. Current Problems

Based on the years of experience so far and discussions in the community we can state that there are five major concerns with respect to the current metadata praxis:

- people want to create and use their own schema tailored specifically towards the requirements of the project and simplify the usage of tools in particular for the creation of metadata
- people want to use the terminology that the specific (sub-)community is used to
- people want to mix vocabularies from various initiatives such as to extend IMDI by TEI header elements

- people want to have options to carry out automatic abstractions to create alternative hierarchies and classifications
- it must be easy for institutions to create metadata portals for selections of LRT components

While the first three points address the metadata set, the last two require an extension of the tools infrastructure. Although IMDI is extendible by the key-value pair possibility at various places and although it is possible to group certain extensions into typical profiles, it seems obvious that some feel overloaded by the sheer number of existing fields and that others miss the flexibility to construct their own extensions in an easy way. Also for different linguistic data types such as annotations and lexica the current separation does not seem to be satisfying, in particular since additional units such as for example lexical schema and data category registries need to be described as well. The naming of the linguistically meaningful atomic object is a point of concern. For the domain of lexica the atomic unit is a lexical instance, for annotated media the experience is that describing the bundle of the recording together with the annotations has shown its usefulness as basic unit, for large semi-structured texts received from publishers it is not yet clear what the optimal unit is.

3. Towards a Component Based Metadata Infrastructure

The CLARIN research infrastructure [11] that is going to be set up in Europe and that wants to closely collaborate at the international level will need to build a unified registry of all language resources and technology components as one of their central elements. It will be the central market place to offer all kinds of services, i.e. the metadata registry has not only to give information to find useful components, but also to point to interface descriptions which applications can use. The new registry needs to be based on the experience made with metadata during the last decade and needs to address the new requirements described for example by W3C [12] when specifying web services. Also we need to bring in the knowledge about unique and persistent identifiers in the setup of the new registry.

It is now widely agreed that schemas are not the anchor point for interoperability, there will be an enormous increase of schemas to suit all possible types of interests. The anchor point for interoperability is concept registries that include various terminologies, ontologies that establish useful relations between them and infrastructure components that force users to make use of the existing concepts. A community wide discussion process is required to register all relevant metadata concepts and to include all relevant terms. Here we can make use of the emerging ISO DCR framework initiated by ISO TC37/SC4 [13] and which is dedicated to the establishment of a categories infrastructure for potentially all fields of language resources.

When designing a new metadata infrastructure we need to include sufficient flexibility but also take care for it to remain manageable and sustainable. Therefore we can envisage the following strategy. A core set of metadata elements that is minimal in size and scope and a large set of predefined components (bundles of metadata elements) that can be used as extensions to the core set, these components describe domain typical documentation fragments as for example “location” or “participant” or distinct linguistic data types. The components are registered in a central registry where users can inspect them and use those that they need. Projects can still design their own components but need to register them to allow

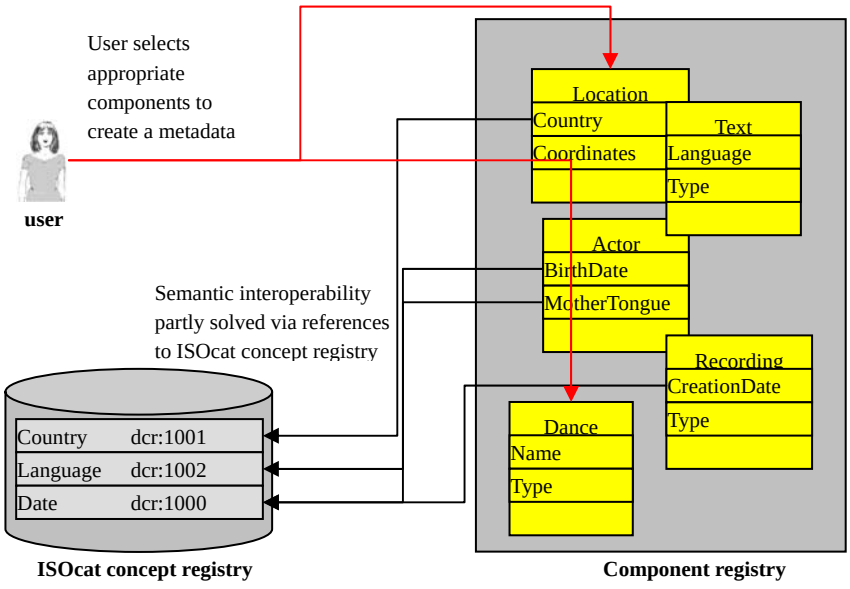

Figure 1 Selecting metadata components from the registry

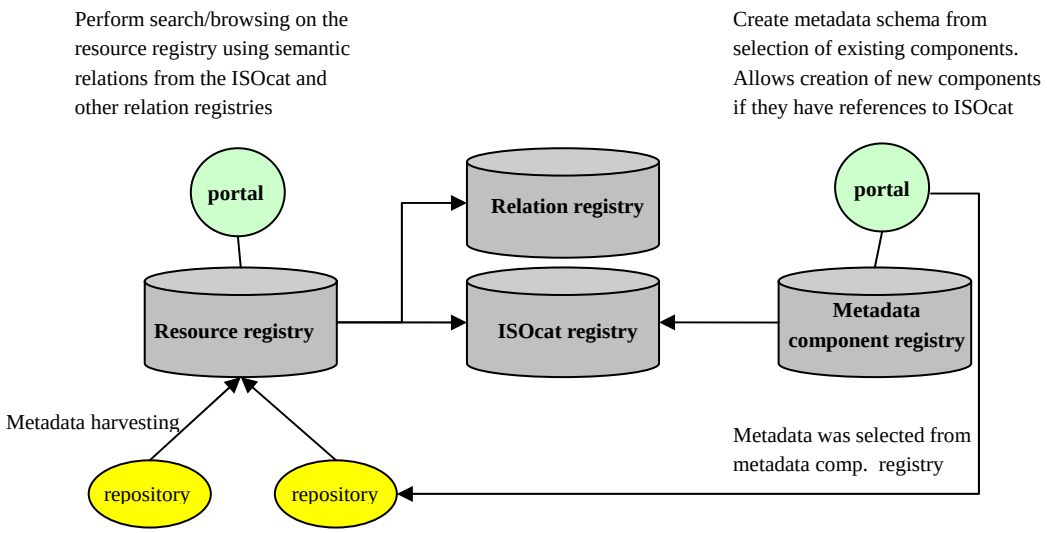

Figure 2. The component registry in perspective.

checks, metadata harvesting and other operations and they should preferably link their terminology to the ISO DCR, non ISO-linked components would first need balloting to be accepted. Existing metadata schemas such as IMDI, OLAC, DC and catalogues such as ELRA's will be included to maintain backward compatibility. The TEI (sub-) schemas will also be available where appropriate. This approach generalizes the one used in designing the Lexical Markup Framework (LMF) [14], also the way LMF uses TEI/ODD [15] as a persistency format has to be considered as a possibility.

A comprehensive taxonomy of language resources and tools will be the basis of a proper requirement analysis of the variety of LRT components. This will lead to the vocabulary of one of the central elements, the one describing the type of the linguistic component. For each identified component the characteristics need to be described as it was already done for lexica in the MILE project, for annotated media in the IMDI initiative and for tools in the DFKI tool registry. These early specifications need to be evaluated on the basis of the experiences that were made. For other resource types that will include, for example, schema registries for various purposes from metadata to lexical components, concept registries, ontologies drawing relations between registered concepts, catalogues of published corpora and general types of ontologies storing arbitrary relations between all sorts of resources expert groups need to be formed to determine the characteristics. Based on the results it will be determined which general core concepts are shared by all descriptions. An analysis will then reveal which concepts need to be included into the CLARIN registry. The CLARIN registry will first be developed separately, but stepwise it will be suggested to be merged into the ISO DCR.

4. Infrastructure

Parallel to the specification work, CLARIN will need to design an optimized infrastructure that is based on the experiences of the last decade. This infrastructure should overcome the limitations of the current systems and anticipate the requirements emerging from using this registry for the domain of web services we want to establish in CLARIN. This is not the place to give a detailed overview; instead we want to give a few dimensions we have to consider:

- The metadata editor needs to be able to incorporate registered components and concepts and apply smart presentation strategies to offer simple user interfaces despite the varying schemas that may be included.
- The infrastructure or its catalogue system needs to have automatic mechanisms to easily create different metadata hierarchies, categorizations and views dependent on the user's or institute's criteria to create selections and browsable domains.
- It must be easy for researchers to create virtual collections crossing institute boundaries.
- It must be easy for researchers to register his/her metadata description or to register a whole set of descriptions.
- It must be made easy for an interested institute to set up a metadata portal and to include those resources that are of interest in its catalogue. We can easily imagine two interests: one is based on national or language considerations and another one will be based on the linguistic resource type. Other typical selections can be thought of. All these portals need to be set up such that authorized persons can make changes in the

metadata information and that these changes are propagated to the portal.

- For any selection the researcher should get a quick overview of how much is still available (free to access and general).
- An API needs to be supported so that interested parties can build their own visualizations.

5. Conclusions

CLARIN's aim is to establish a rich landscape of language resources and technology where the LRT component registry will become the central market place where all researchers can offer their products and where application programs can find the locations of the APIs to make use of services. This requires a re-thinking of our metadata practices in all respects. CLARIN will analyze the experiences made so far, but also look to new requirements that emerged during the last years. A component-based approach leading to a variety of schemas is the only way to meet the wishes. Interoperability needs to be based on a shared vocabulary registered in data category registries as they are suggested by ISO.

CLARIN's broad institutional coverage is a promising basis to meet the advanced needs.

6. References

[1] http://dublincore.org/

[2] http://www.ilc.cnr.it/EAGLES/isle/ISLE_HomePahe.htm

[3] Broeder, D.G., Brugman, H.,Russel, A., and Wittenburg, P., (2000), A browsable Corpus: accessing linguistic resources the easy way. In Proceedings LREC 2000 Workshop Athens.

[4] http://www.mpi.nl/IMDI/

[5] http://www.language-archives.org/documents/overview.html

[6] http://registry.dfki.de/

[7] http://www.tei-c.org/index.xml

[8[http://www.openarchives.org/OAI/openarchivesprotocol.html

[9] http://www.mpi.nl/IMDI/tools/

[10] http://www.lat-mpi.eu/tools/imdi/browser/

[11] http://www.clarin.eu

[12] http://www.w3c.org/

[13] ISO DIS 12620 (2007). Terminology and other language resources - Data categories - Specification of data categories and management of a Data Category Registry for language resources, ISO.

[14] Francopoulo, G., M. George, et al. (2006). Lexical Markup Framework (LMF). Language Resources and Evaluation, Genoa, Italy.

[15] http://www.tei-c.org/release/xml/tei/odd/