A Large Metadata Domain of Language Resources

Daan Broeder, Thierry Declerck*, Laurent Romary^, Markus Uneson, Sven Strömqvist**, Peter Wittenburg**

Max-Planck-Institute for Psycholinguistics, *Saarland University, ^LORIA, **Lund University
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
daan.broeder@mpi.nl

Abstract

The INTERA and ECHO projects were partly intended to create a critical mass of open and linked metadata descriptions of language resources, helping researchers to understand the benefits of an increased visibility of language resources in the Internet and motivating them to participate. The work was based on the new IMDI version 3.0.3 which is a result of experiences with the earlier versions and new requirements coming from the involved partners. While in INTERA major data centers in Europe are participating, the ECHO project focuses on resources that can be seen as part of cultural heritage. Currently, 27 institutions and projects are active with the goal of having a large browsable and searchable domain by the summer of 2004. Experience shows that the creation of high quality metadata is not trivial and asks for a considerable amount of effort and skills, since manual work alone is too time consuming.

Introduction

At LREC 2000 in Athens the first workshop[1] about metadata concepts for making language resources visible in and discoverable via the Internet was organized by some of the authors. At LREC 2002 two groups demonstrated operational frameworks for creating metadata for language resources and for working with them for management and discovery purposes. While OLAC[2] (Open Language Archives Community) started from a Dublin Core point of view with the goal to create a set that allows for the description of all types of language resources, software tools, and advice, the IMDI[3] (ISLE Metdata Initiative) activities started with a slightly different approach. The focus was primarily on multimedia/multimodal corpora and a more detailed set was worked out that can be used not only for resource discovery but also for exploitation and managing large corpora. Most importantly, IMDI allows its metadata descriptions to be organized into linked hierarchies supporting browsing and enabling data managers to carry out a variety of management tasks.

The two years since 2002 have been used to improve the metadata sets based on the experience and feedback of the communities. They have also been used to create an interoperable domain, i.e., a mapping schema was worked out between the IMDI and OLAC sets and the IMDI domain acts as an OLAC data provider. IMDI records can be searched for from the OLAC domain.

IMDI Metadata Set 3.0.3

Based on the experiences and on a broad discussion process including field linguists, corpus linguists and language engineers, the IMDI set 3.0.3 was designed as part of the INTERA project[4]. It is available as an XML-Schema. It was adapted to simplify the content description and the artificial distinction between collectors and other participants probably influenced by Dublin Core was removed. Additionally, three major extensions were applied: First, it is now possible to describe written resources that are not annotations or descriptions. This was necessary, since language collections may contain or even entirely consist of written resources, in the form of field notes, sketch grammars, phoneme descriptions, etc. Second, as a result of long discussions with participants in the MILE lexicon initiative, it is now possible to describe lexica with a specialized set of descriptor elements.

Third, it is now possible to define and add project-specific profiles. Earlier versions of IMDI supported the possibility of extensions at various levels in the form of user-defined key/name–value pairs, i.e., the user was able to define a private category and associate values with it.

This feature was used by individuals and also projects to include special descriptors. However, these descriptors were not fully supported by the IMDI tools. In the new version, projects or sub-domains can define a set of important categories and these are supported while editing or searching. Some of them have done so already, e.g., the Dutch Spoken Corpus project and the Sign Language community.

In conclusion, IMDI consists of a set of core definitions that have to be stable to guarantee users that their work will be exploitable even after many years, and of sub-community specific extensions, which nevertheless are result of discussion processes.

IMDI Framework

Further, the IMDI initiative came up with a whole bunch of professional aids and tools[5] for the latest metadata set version 3.0.3, such as

- a professional and mature Editor that allows users to create fully IMDI compliant metadata descriptions and that supports all IMDI features such as controlled vocabularies and project specific profiles

[1] http://www.mpi.nl/ISLE
[2] OLAC: http://language-archives.org
[3] IMDI: http://www.mpi.nl/IMDI
[4] Integrated European language Resource Area: http://www.elda.fr/index.html

[5] All tools are Open Source and available at the sites: http://www.mpi.nl/tools or http://www.mpi.nl/IMDI

- a handy browser that allows navigating in the distributed domain of linked metadata XML files supporting searching as well as browsing, the setting of bookmarks etc (fig. 1)
- a tree-builder that allows the user to create new user-specific virtual trees by linking arbitrary metadata descriptions and creating arbitrary nodes
- for large archives with a web-server on-the-fly transformed HTML presentation of the metadata files that allow users to browse in the linked metadata domain with normal web-browsers (fig. 2)
- a full-text search tool that gathers all metadata information and treats it like unstructured information in particular for untrained users
- an access management system fully integrated with the metadata domain (fig. 3)
- a wrapper that offers IMDI records according to the OAI MHP (Metadata Harvesting Protocol)

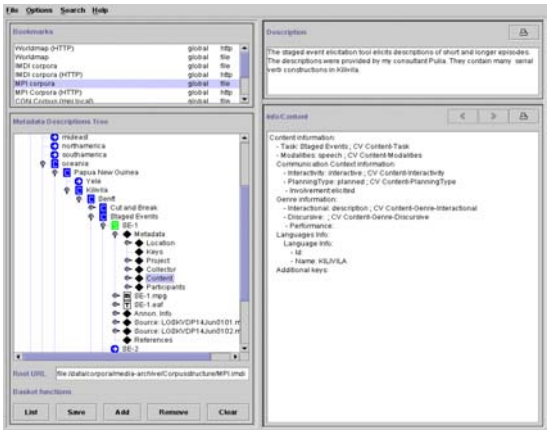

Figure 1 shows the interface of the XML-based special browser that offers advanced functionality.

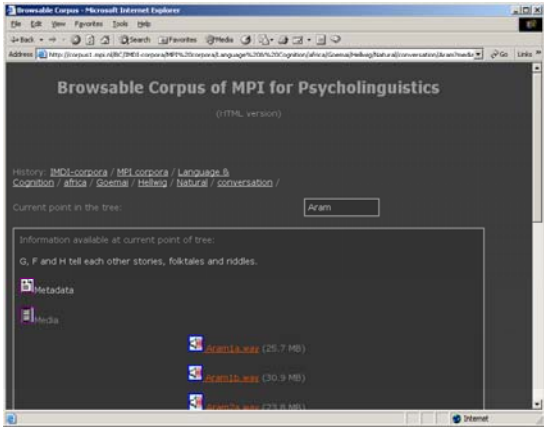

Figure 2 shows the interface for browsing in the metadata domain with the help of normal HTML presentations. The location is indicated by showing the path.

Distributed Metadata

Metadata that is going to be used for discovery purposes has to be distributed, i.e., the individual descriptions can be stored at different servers. The OAI model defines data and service providers both related via the metadata harvesting protocol that defines the interaction pattern and the packaging. The data providers have to provide Dublin Core records to achieve semantic interoperability. However, the OAI protocol also allows to send records specified by another schema, such as IMDI. Based on this information service providers can build services for example for searching that cover a large group of different repositories working internally with different metadata sets.

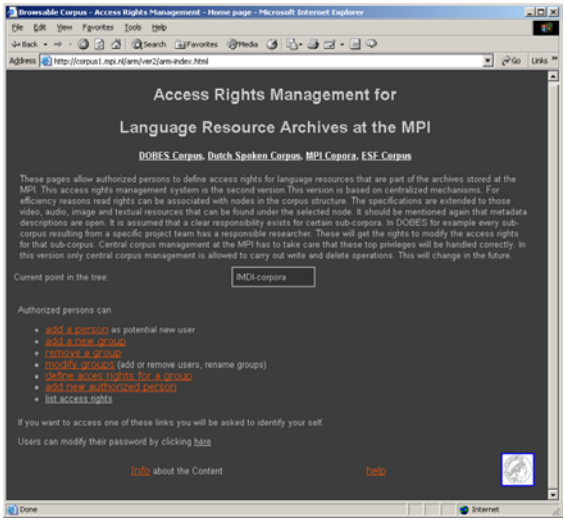

Figure 3 shows the home page that allows managing users, groups and access rights. This management tool is closely related to metadata to provide administrative efficiency.

Although the OAI protocol is comparatively simple to implement, the common praxis is still to harvest XML-files. The ECHO project has shown that most of the institutions are not yet prepared to support OAI. The IMDI metadata infrastructure assumed from the beginning that metadata records can be located at various institutions – even on the notebook of remotely working fieldworkers. Therefore, IMDI metadata records can be linked in a simple way – similar to web sites. The browser only needs a registered URL to integrate the IMDI descriptions into the domain. For searching, the IMDI tools will scan all known metadata links and create indexes that can then be exploited.

However, all IMDI tools expect IMDI type of metadata records, i.e., IMDI is not a concept for establishing interoperability between different metadata sets. Within the ECHO project an integrated metadata domain was built that includes ten different repositories from five different disciplines. It was shown that interoperability at the structural level was mainly achieved by harvesting XML-structured files and at the semantic level by creating special mappings (Wittenburg, 2003/2004). The Dublin Core approach reduces too much of the semantic richness of the provided information. Therefore, it is seen as just another view on the data.

Critical Mass of Metadata

Within the INTERA project the task was to convince various data centers and projects to participate in building a distributed IMDI domain. Typically, these data centers have language resources from the area of language

engineering. In the mean time the metadata is generated by the following institutions: European Language Resource Agency (Paris), Institut National de la Langue Francais (INALF, Nancy), German Center for Artificial Intelligence (DFKI, Saarbrücken), University of Saarland (Saarbrücken), Bavarian Speech Archive (Munich), Meertens Institute (Amsterdam), University of Florence, Institute for Language and Speech Processing (Athens), Instituto Linguistica Computationale (Pisa), University of Ljubljana, University of Sofia and the Max-Planck-Institute for Psycholinguistics (Nijmegen).

In the ECHO project[6] it was one of the tasks to motivate researchers and institutions to create metadata descriptions of resources that can be seen as part of our heritage. Here the following institutions can be mentioned: University of Helsinki, Phonogrammarchiv Vienna, University of Groningen, Kotus (Helsinki), Sweden's national Dialect Archive, European Sign Language Communities (Stockholm, London, Netherlands, Germany), University of Utrecht, University of Uppsala, University of Stavanger, University of Lund, DOBES Programme (Nijmegen).

This new emerging domain including the activities of about 27 partners includes textual corpora, national speech corpora, multimedia/multimodal corpora, parallel corpora, lexicons and various types of written resources.

Yet we don't have a final estimate about the number of individual resources that will be described and available at the end of 2004 when the two mentioned projects will be finished. At the Max-Planck-Institute there are currently about 30.000 sessions described by metadata. Large corpora such as SMARTKOM, the Dutch Spoken Corpus, the LABLITA corpus and the ATILF corpus will be part of the new domain, so we can expect that there will be many more resource units described and therefore searchable.

It is hoped that this emerging domain is large enough to demonstrate the usefulness of metadata for discovery purposes and that it will inspire others to participate. The ENABLER[7] overview has clearly indicated that there is a lack of visibility of language resources in the Internet and that their accessibility is even worse. Therefore, the creation of metadata must be a high priority program to foster re-usage. In a declaration agreed upon at the ENABLER meeting in Paris in 2003 it was stated that the funding agencies should make the generation and integration of proper and openly available metadata descriptions according to one of the two currently existing standards (OLAC or IMDI) obligatory.

Metadata Creation Process

In the first phase of INTERA and ECHO various European data centers and research institutions were asked whether they are interested to participate in creating an integrated metadata domain. The initiative had good response, i.e., most reacted in a positive way. However, the knowledge about the principles and goals of metadata creation and the expectations were very different. Some expected a larger amount of funding support and did not see that metadata is not meant to clean up the state of their repositories.

Most of the data centers that finally participated were aware of the relevance and concept of metadata. Therefore, there was no need for intensive training programs. However, since these centers with large corpora were already using header type of information or some internal database, it was not evident for them that IMDI not only requires metadata records. To create a browsable domain as well it is necessary to create a linked hierarchy of metadata descriptions and meaningful nodes that represent abstract concepts such as "language", "genre" and "age". It would be possible in IMDI to just deliver metadata records, simply create one node representing the institution and link all descriptions to this one node. But that would lead to long and unstructured lists that are not useful for browsing. To help creating meaningful hierarchies programs would be necessary to create abstractions from the metadata descriptions semi-automatically.

The experiences with projects and institutions in the ECHO project were different. Here training courses and introductions were necessary to inform the researchers about all aspects of standardized metadata. In general these groups had to start from scratch, since they had not worked with formal metadata beforehand. Metadata creation then means a considerable amount of work, since interviews may be required and analysis work is needed to fill in the values for the metadata elements.

In special cases such as the Sign Language community a discussion process was initiated that led to additional categories that were absolutely necessary. Only with categories such as "Father.deafness" metadata would be easily exploitable by the members of that specific community. Therefore, the concept of project or community specific profiles was introduced.

Problems

The efforts needed to create metadata descriptions varied considerably, as well as the available skills to write scripts to semi-automatically create basic information that can be enhanced manually. Although the IMDI infrastructure offers an editor with useful options to increase the efficiency such as storing and re-using blocks of information, manual metadata creation is very time-consuming and often not feasible.

The experience showed that it is much easier to use spreadsheet tools such as EXCEL for researchers to create and manipulate a large set of records. The same is true for experienced people that prefer to use scripts to create the metadata records. However, these techniques in general create metadata of bad quality. The following types of problems were encountered:

- There is no guarantee that scripts produce well-formed XML files.
- The character encoding is often not UNICODE.

[6] European Cultural Heritage Online: http://www.mpi.nl/echo

[7] ENABLER: http://www.enabler-network.org

- Most problematic is that the tools used do not provide support for the controlled vocabularies leading to typo errors, spelling variants and many others.

It is the service provider who has to invest time to check the correctness of the produced metadata records and to improve the metadata records in collaboration with the data providers. The OAI[8] model that requires a validation at the moment of registration and simply points to the errors is in general not sufficient. Without additional help many of the data providers would stop.

Improving the content of the metadata descriptions is very important for successful searching. Two phenomena can be observed: (1) Since metadata creation is a hard job, even in evident cases elements are not filled in. (2) As already indicated all kinds of variations can be found, since the creators partly do not make use of controlled vocabularies.

First, in a very large collection it is a problem to identify such errors or missing values. Second, we need to know how to correct them without starting time consuming interactions with the various data providers. To detect errors and variants it makes sense to first run a validation against the controlled vocabularies. Until now, however, the errors have to be corrected manually. Methods that use a formal closeness (one character difference) or other type of heuristics have not yet been tested. Variants that occur due to language differences (for example Afrique, Afrika, Africa) could be corrected if one would have suitable online dictionaries or terminology databases.

Third, filling in empty elements is even more difficult, since there can be many reasons why elements were not used. Until now these cases were identified by accident, e.g., someone inspecting metadata records, finding that for example the country is filled in but not the continent. In such a simple case, a script using geographic thesaurus information could very easily add information. If the "genre" field, however, is not filled in there is no simple way to identify this except by producing long lists. Still it would not be evident how such fields have to be filled in, since only the researchers can do this.

Another aspect that was found during the metadata creation work is that many institutions are looking for institutions that can store their collections. They don't have the human resources to organize them and maintain them in a proper state so that others can use them. So we need ingest tools that easily allows researchers to hand over their data to another institution in an easy way. At the MPI such a system is currently in work. Ingestion will be tightly combined with metadata creation.

Future

Much effort is taken to create and maintain metadata descriptions and it is expected that projects such as INTERA and ECHO will help to increase the awareness that metadata is very important. Therefore, we have to assure that the investments will be maintained over a long period.

All IMDI categories have been registered within the emerging ISO TC37/SC4 data category repository. In doing so semantic definitions are carried out in a widely agreed and machine-readable way. It is expected that also OLAC and TEI categories will be entered in the same way. This would give all definitions a higher degree of stability. It would also allow us to make the semantic mapping between the categories explicit. It would also open the possibilities that researchers create their own mappings between categories and even develop own metadata sets by re-using the existing and well-defined categories.

It is expected that creating metadata will also become more attractive when new applications will become available. The INTERA project has as one other goal to link the domain of language resources with that of tools that operate on such resources. The MIME type concept is not new, however, the requirements go far beyond this. Bundles of resources have to be processed by tools combining several of them in one step. Characteristics of resources such as their annotation schemes are relevant to detect the most useful tool. Within the INTERA project an interaction between the IMDI domain and the ACL tool registry[9] is being developed that is based on the open Language Resource Exchange Protocol (LREP).

Conclusions

In this paper we have presented the metadata creation work in the INTERA and ECHO projects and the experiences that were made. The creation of high quality metadata descriptions in general costs more effort than was originally expected. The fact that many researchers still see metadata creation as an overhead, makes infrastructure projects of this sort a difficult, but nevertheless important enterprise.

A sufficiently large metadata domain is expected to become available this year. In order to convince other institutions and individuals to contribute to this domain more utilities have to be developed to easily create large sets of metadata descriptions, to derive corpus-structured semi-automatically and to enrich the content.

References

Wittenburg, P. (2003). WP2-TR16-2003 Version 3 Note on ECHO's Digital Open Resource Area. http://www.mpi.nl/echo/tech-report-list.html

Wittenburg, P. (2004). WP2-TR17-2004 Version 1 Note on an ECHO Ontology. http://www.mpi.nl/echo/tech-report-list.html

[8] Open Archives Initiative: http://openarchives.org

[9] ACL Software Registry: http://registry.dfki.de