

Interaction of Tools and Metadata-Descriptions¹ for Multimedia Language Resources

Daan Broeder

Max-Planck Institute for
Psycholinguistics
Wundtlaan1 6525 XD
Nijmegen
The Netherlands

daan.broeder@mpi.nl

Peter Wittenburg

Max-Planck Institute for
Psycholinguistics
Wundtlaan1 6525 XD
Nijmegen
The Netherlands

peter.wittenburg@mpi.nl

Abstract

The increasing amount and complexity of multi-media/multi-modal language resources (MMLR) poses a problem in many respects. This paper wants to discuss metadata descriptions that can be used to easy find and locate suitable MMLRs in the Internet and how these descriptions may be used to discover and apply suitable tools on the data.

1 Introduction

We succeeded in reaching a consensus within a representative part of the linguistic community in Europe about a standard for such metadata descriptions. A machine readable implementation of this standard will then allow us to build up a searchable and browsable space. Our presentation is based on the work executed within the framework of the international EAGLES/ISLE [1,2] project that is named IMDI (ISLE Metadata Initiative), on practical work with meta descriptions at the MPI for Psycholinguistics, on a collaborative enterprise to create a browsable corpus demo of material 7 European institutes and on suggestions with

respect to metadata within the DOBES [3] and the CGN [4] projects.

2 Metada for Language Resources

The idea of describing a whole document with the help of a few characteristic metadata elements is not new. Well-known corpora such as Childes [5] have used header information to describe the content, the speakers and the language being spoken etc. The Text Encoding Initiative [6] and the CES group [7] have specified in detail the tag set with which a whole text document can be described. However, all early initiatives were not meant to be a general standard for the description of MM LRs and allow the formation of a searchable and browsable space on the Internet IMDI desires. This is what recent initiatives in other domains such as Dublin Core (DC) [8] and MPEG7 [9] want to achieve: XML-based machine-readable information about certain documents that is openly accessible in the net such that easy retrieval is possible. New initiatives by the W3C such as RDF [10] support these intentions.

¹We distinguish metadata from annotation data, knowing that many don't make this difference. While metadata in this context is meant to describe the whole language resource, the annotation is a time synchronous description of what is happening and is spoken during a recording

Within IMDI we have made an overview about header and metadata elements used so far by the language resource community. This overview and the concrete needs within large European projects will be used to develop and test a first proposal on the way to come to a hopefully widely accepted standard. Compliance with the standard has to guarantee that metadata descriptions created by different people at different locations adhere to the same syntax and to the same semantic definitions of the metadata elements included. The standard has to offer possibilities of adding metadata elements defined by sub communities, projects or even individuals. From other initiatives we know that these goals can only be achieved if the set of metadata elements is not too exhaustive. This does not mean that only limited information can be stored. For instance the metadata description standard certainly includes an element to enter the name of the language spoken, but other very elaborate information about that language can be made available in other data types pointed to by hyperlinks to other data perhaps conforming to more specialised schemas.

IMDI is now entering a phase where the metadata element categories and the metadata elements to be included have been discussed with interested members of the MMLR community for about a year and become stable. Two resource types were selected to start with: (1) multimedia corpora and (2) lexicons, the discussion about the lexicon resources is at the moment less far developed than that concerning the corpora. Within the IMDI initiative we started the search for a suitable set of metadata elements by trying to identify the characteristics of such resources that people such as researchers, developers, students, or even the general public would choose to use to find exactly those resources they are looking for. Very helpful was the study of the creation process and the construction of a structured metadata set as a reflection of an ontology of these resources. We know that resources themselves are not openly available, but at least the metadata description should inform the community about their existence, about intellectual property rights and modes of usage.

Two main categories of metadata can be identified:

- Basic information on the content of the resource: the content language of the resource, and administrative information about the resource.
- Resource descriptions that define the type and structure of the resource.

A full listing of all IMDI elements is given in Appendix A, but for definitions and substructures we refer to [1]. The relevant elements for the resources themselves are:

Resource Link (c)
Media Resource Link (c)
Annotator (string)
Date (c)
Type (ov)
Format (ov)
Content Encoding (string)
Character Encoding (c)

Table 1 elements for media files

Resource Link (c)
Size (string)
Type (ccv)
Format (ov)
Quality (ccv)
Recording Conditions (string)
Position (c)

Table 2 elements for annotation units

C: constrained
OV: open vocabulary
CCV: closed constrained vocabulary

For annotation units multiple units may reside in one file. The relevant elements for characterising the resources in a way that is important to tools (a discussion that we will come to later) are:

- Reference to the resource itself
- Size (of media file, if the tool has a limit)
- Format (for media files somewhat more simple than for annotation units)
- Type (for annotation unit the type of analysis result e.g. morphology, phonetics ...)
- Different encodings

3 Strategies for Metadata Standards

The way IMDI has developed its metadata vocabulary can be described as bottom up. IMDI chose to try to first understand the linguistic community's needs by making an overview of metadata used by different projects and corpora, speak with representatives of many institutions and try to distill a metadata set from it that focuses on retrieval aspects. For IMDI the needs of the creators are the start and end point since the creators are also the major consumer group of language resources. So the question for IMDI posed itself was "how to enable resource discovery of useful language resources that can be used for certain studies etc". This approach leads to a metadata set whose terminology fits the domain and a vocabulary that is considerable richer than the for instance the DC set. Interesting enough another initiative named OLAC [11] that wants to create metadata for language resources has taken the DC set as a starting point. The OLAC approach can be called a top-down one and seems motivated by the wish to join the "very important" Open Archives Initiative (OAI) [12] without having too much work in mapping different metadata sets. OLAC wants to use a slightly more specialised version of the OAI metadata set and because OAI uses Dublin Core as default metadata set the choice of an extended DC set for OLAC is understandable. Of course the question remains if this is sufficient to characterise language resources in a sufficient specific manner.

The discussion showed that both approaches are important especially when the ontology of the domain is not very well understood. IMDI starts with analyzing the domain and leads to a more narrow and specialised categorization scheme. DC on the other hand offers very broad categories the semantics of which are often sloppily defined. Both approaches lead to specific inherent retrieval problems. We have to consider two views; (1) People from inside the domain searching for resources (2) People from outside the domain. People from inside have intimate knowledge of the domain ontology and want more specific categories. People from

outside need broader categories to assure that the resources they search fall in the larger "hit list". The discussion about OLAC DC qualifiers led partly to the same discussions that were carried out in IMDI. This is not surprising, since OLAC somehow has to address the needs of the field and the participants at the meeting were mainly linguists. The OLAC top-down approach which starts with a smaller vocabulary than IMDI will have less problems when addressing the interoperability with metadata sets such as DC. However this advantage disappears when OLAC will add more specific elements and qualifiers to accurately describe the domain.

4 Tools and Metadata

It was the initial idea that the metadata descriptions could also have elements that describe specific tools that can be used to act on the resources themselves. However since resources and tools form orthogonal dimensions it is better to have the metadata description only describe resources and not a set of tools that will change in time anyway. A more elegant solution is to describe the type and structure of the resources in sufficient detail so that "browser" tools used to access the metadata description can decide which ones of the available tools are suitable to handle the data. This can be either based on local user configurable information or on some sort of remote tool registry. At the moment IMDI is experimenting with a scheme of (semi) mime-types to characterise language resources. We foresee that users will want to customise the mapping of tools to resource types to their own taste just as they are able to do with WWW-browsers.

Needed for such a scheme is that tool repositories note the types (mime-types) and encoding, character encoding for which the tools are suitable (see the lists in table 1 and 2). It has to be investigated in detail how far tool registries and resource collections structured by metadata descriptions can be created in a way such that especially naïve users can overcome the frustrating problems of accessing the right resources with suitable tools. This problem is not solved and is one of the greatest obstacles for increasing the reusability of the huge

treasure of resources. IMDI has taken limited tests with a number of tools to study the interaction between mime-type tagged resources and selecting from a tool palette. We have no doubt that this is the way to go.

A special question is the form of the infrastructure. Where will we store the metadata descriptions and/or resources and how are tool registries such as from DFKI [13] made known to the distributed resource universe? During the IMDI project a preliminary solution is found for creating a registry authority for metadata. This registry authority has to build a web-portal, check the quality of the produced meta descriptions, create intuitively understandable browsable hierarchies based on the meta descriptions and link the meta descriptions to other type of information and resources. The registry authority will also provide tools such as a constrained editor that allows the user to create meta descriptions and a suitable browser which can operate on the metadata description files. The IMDI project will also work on requirements for the registry authority and the metadata tools.

At the moment the time has come for IMDI to investigate if and how the metadata description browser can access remote software registries to assist users in the choice of tools to use for resources. This would be a logical extension to the local configurable mapping of tools on resource types that is needed anyway for non-networked situations.

References

- [1] <http://www.mpi.nl/ISLE>
- [2] P. Wittenburg, D. Broeder & B. Sloman: Meta-Descriptions for Language Resources - EAGLES/ISLE - A Proposal for a Meta-Description Standard for Language Resources. http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf.
- [3] DOBES <http://www.mpi.nl/DOBES>
- [4] Corpus Gesproken Nederlands <http://www.nwo.nl/gw/introductie/>
- [5] CHAT <http://childes.psy.cmu.edu/>
- [6] TEI <http://www.uic.edu/orgs/tei/>
- [7] CES <http://www.cs.vassar.edu/CES>
- [8] DC <http://purl.org/dc/documents/>
- [9] MPEG7 <http://mpeg7>
- [10] RDF <http://www.w3.org/RDF/>
- [11] OLAC <http://www.language-archives.org/>
- [12] OAI <http://www.openarchives.org/>
- [13] DFKI-softwareregistry <http://registry.dfki.de/>

Appendix A: IMDI Metadata Vocabulary for Language Resources.

Session	
Name (string)	Resources (group)
Title (string)	
Date (c)	
Continent (ccv)	
Country (ccv)	
Region (string)	
Address (string)	
Description + (sub)	
Keys (sub)	
Project (group)	
Name (string)	
Title (string)	
Id (string)	
Contact (group)	
Description + (group)	
Collector (group)	
Name (string)	
Contact (sub)	
Description + (sub)	
Content (group)	
Communication Context (group)	
Interactivity (ccv)	
Planning Type (ccv)	
Involvement (ccv)	
Genre (group)	
Interactive (ovl)	
Discursive (ovl)	
Performance (ovl)	
Task (ocv)	
Modalities (ocv)	
Languages (group)	
Description + (sub)	
Language + (sub)	
Description + (sub)	
Keys (sub)	
Participants (group)	
Description + (sub)	
Participant (group)	
Type (ov)	
Name + (string)	
Full name (string)	
Role (ov)	
Language + (sub)	
Age (c)	
Sex (ccv)	
Education (string)	
Anonymous (ccv)	
Description + (sub)	
Keys (sub)	
Media File + (group)	
Resource Link (c)	
Size (string)	
Type (ccv)	
Format (ov)	
Quality (ccv)	
Recording Conditions (string)	
Position (c)	
Access (sub)	
Description + (sub)	
Annotation Unit + (group)	
Resource Link (c)	
Media Resource Link (c)	
Annotator (string)	
Date (c)	
Type (ov)	
Format (ov)	
Content Encoding (string)	
Character Encoding (c)	
Access (sub)	
Language (sub)	
Anonymous (ccv)	
Description + (sub)	
Media Carrier +	
Id (string)	
Format (ov)	
Quality (ccv)	
Position (c)	
Access (sub)	
Description + (sub)	
Anonymous (group)	
Resource Link (c)	
Access (sub)	
References (group)	
Description + (sub)	