Using Profiles for IMDI Metadata Creation

Daan Broeder, Peter Wittenburg, Onno Crasborn +
Max-Planck Institute for Psycholinguistics , Nijmegen The Netherlands
PO Box 310, NL-6500 AH Nijmegen, The Netherlands
daan.broeder@mpi.nl, peter.wittenburg@mpi.nl
+ Department of Linguistics, University of Nijmegen, The Netherlands
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
o.crasborn@let.kun.nl

Abstract
In this paper a system to support the creation of extended IMDI metadata records is presented. It is based on bundling definitions of the in the IMDI system user definable key-name/value pairs in a profile. The possibility of using inheritance of profiles in a corpus structure is explored. Profiles Can be created and used by the IMDI Editor, a tool specially designed to create IMDI metadata records.

1. Introduction

Since the introduction of the first IMDI metadata set [1] for the description of language resources, much experience has been acquired about the ways researchers want and need to specify metadata for the language resources they create. The willingness of researchers to provide metadata is very much dependent on the effort involved and the benefits the metadata gives to that researcher, and also, metadata creation should be made as easy as possible by providing the right tools. Although there are also other parties involved that profit from the creation of metadata, like the organisation that subsidizes the collection of the resources of course the linguistic research community as a whole can profit from the reuse of resources that can be easily identified and located.

Limiting the amount of typing a user has to do is the easiest way to bring down the effort of producing metadata. But because the end result should still be as complete metadata records as possible, we can only try to avoid the typing of unnecessary and already specified data. For this the IMDI metadata environment offers an editor to create metadata records that makes use of different types of predefined records.

The use of controlled vocabularies (CVs) within the IMDI tools to simplify the entry of controlled metadata has already been described [2].

Although the IMDI set is flexible enough to allow users to add their own metadata descriptors. It was suggested that such an extension mechanism should be available in a more structured way. Specialised sub-domains and projects require the availability of predefined sets of these descriptors so that they can be shared and not every individual needs to invent them again. We call these predefined sets of metadata descriptors "profiles". They serve as extensions to the IMDI "core-set" of descriptors and can be general enough to serve the needs of a whole sub-domain of linguistic research such as sign language studies[3], or cater for an individual project such as the profile developed for The Spoken Dutch Corpus[4]

At the moment the use of profiles is integrated into the IMDI tools as one of the two possibilities of reuse of metadata

2. Facilitating reuse

2.1. Reuse of partial descriptions

Using the IMDI-Editor, users are able to save often-used parts of IMDI metadata descriptions for reuse. For instance, a researcher who is often working with a specific consultant is able to save all metadata information relating to this person in a "template" that is stored in a user-specific repository. The editor allows the template later to be later inserted in new IMDI metadata descriptions. Different researchers can also share these templates by having the editor import templates into their template repository.

2.2. Use of IMDI profiles

Because the domain of language resources is very broad, flexibility is built into the IMDI set to cater for the requirements of sub-domains, special projects and the specific requirements of individual researchers.

At different levels of the IMDI set a user may define his own set of key-name/value pairs and where the type of the value may be constrained by a user defined controlled vocabulary (CV). For instance, if a user would like to be able to describe the fact that a speaker in one of his recordings is blind he can define a metadata descriptor "Actor.Blind" that has an associated value type IMDIBoolean {True, False, Unknown, Unspecified}[1]

The first versions of the IMDI-Editor left the use of these key/value pairs free, there was no way to reuse and share often used combinations other than by reusing already defined sessions as templates.

To accommodate the reuse of sets of key-name/value pairs and also of predefined values for the existing "core" IMDI descriptors, an IMDI profile can be defined that can be shared with others. At this moment a profile for sign

[1] One may question if this still may be called Boolean. However we find the possibility to be able to also specify Unknown and Unspecified helpful.

language studies [3] is available that was developed within the ECHO project, just as there is a special profile for the Spoken Dutch Corpus [4]. As an example the complete set of key-name value pairs is shown in table 1, to demonstrate that a profile can be an elaborate yet very specific set of descriptors that cannot be accommodated in a more general set. Profiles for multi-modal domain and bilingual studies are being developed. Figure 1 shows the IMDI-Editor using the sign language Profile where the predefined key-name/value pairs for a speaker-signer are visible.

Both IMDI templates and profiles are XML files supported by the IMDI schema and can be created with the IMDI Editor. Often used standardized profiles will be included in the software distribution of the IMDI-Editor while others can be stored on the local file-system. The IMDI Editor can also create modifications of the included profiles as local specialisations.

3. Profile Inheritance

Based on IMDI profiles and the tree structure of IMDI described corpora we are currently developing an inheritance mechanism for IMDI profiles. This mechanism is based on the tree structure of IMDI structured corpora where resources are the leaves of these corpus trees.

A node in such a tree represents the metadata commonalities of all sub-corpora and resources beneath it. For instance a corpus can be divided in male and female speaker subcorpora that again can be subdivided into subcorpora representing different age groups. A more realistic hierarchy is shown in figure 2. , an example from the field linguistics domain.

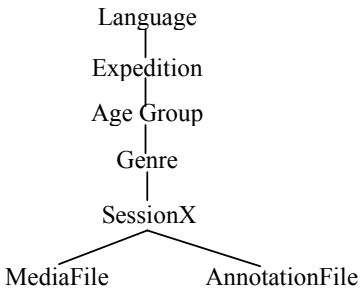

Figure 2 shows a typical hierarchy from field linguistics

If we link profiles to the sub-corpus nodes of a corpus tree we can simplify the creation of metadata considerably by the adopting inheritance rules:

- When a user creates an IMDI metadata record as a child of a corpus node, the profile of such a node automatically is enforced. The profile sets values for metadata descriptors whose value is not yet specified in the metadata record.
- The closest ancestor of the corpus node that also has an associated profile is found and the same procedure is applied again, filling in more metadata fields that are still empty. This rule is repeated until the top node of the corpus is reached or until an ancestor node is reached that has a special profile that forbids further inheritance.
- Multiple inheritance introduces a problem when a session has multiple ancestors that provide competing information for a metadata descriptor. This can be solved by assigning a preferred parent to a node. As a pragmatic solution that would be the parent to which the node was linked first.

Support for this procedure is being implemented in both the IMDI-Editor and the IMDI-TreeBuilder, a tool specifically created for the creation of IMDI corpora trees.

The inheritance mechanism described above can just as easily be applied to the user defined metadata templates from 2.1. It will encourage researchers to create corpus hierarchies that allow reuse of metadata as much as possible.

Figure 3 shows how the combination of profile and template defined metadata flows down to a metadata description for a session

All these predefined records can be distributed in several ways: 1) As part of the editor itself, this procedure is followed for the basic profiles and CVs. 2) Distributed by a web server for use by distributed research groups and projects. 3) Stored on a local file system for use by individual researchers.

In general, we can say that the profile concept enables IMDI to offer facilities to users that have specific wishes and needs and already were using their own metadata sets. A future challenge will be to enable search interoperability between the different profiles outside that offered by the shared IMDI core set of descriptors.

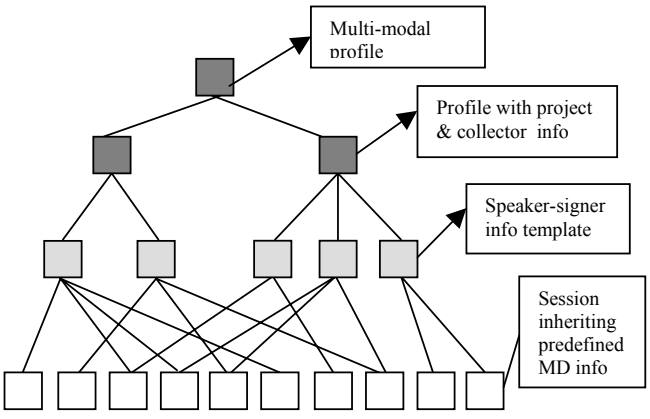

Figure 3. Profiles linked into the corpus hierarchy

4. References

[1] IMDI, http://www.mpi.nl/IMDI

[2] D.Broeder, F. Offenga, D. Willems Metadata Tools Supporting Controlled Vocabulary Services. Proceedings LREC 2002, 1055-1059

[3] O. Crasborn, T. Hanke 2003 Additions to the IMDI metadata set for sign language corpora. Agreements at an ECHO Workshop, metadata for sign language corpora, May 8-9, 2003 University of Nijmegen. http://www.let.kun.nl/sign-lang/echo/docs/SignMetadata_Oct2003.doc

[4] Oostdijk, N. and D. Broeder. 2003. The Spoken Dutch Corpus and its Exploitation Environment. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*. 14 April, 2003. Budapest, Hungary.

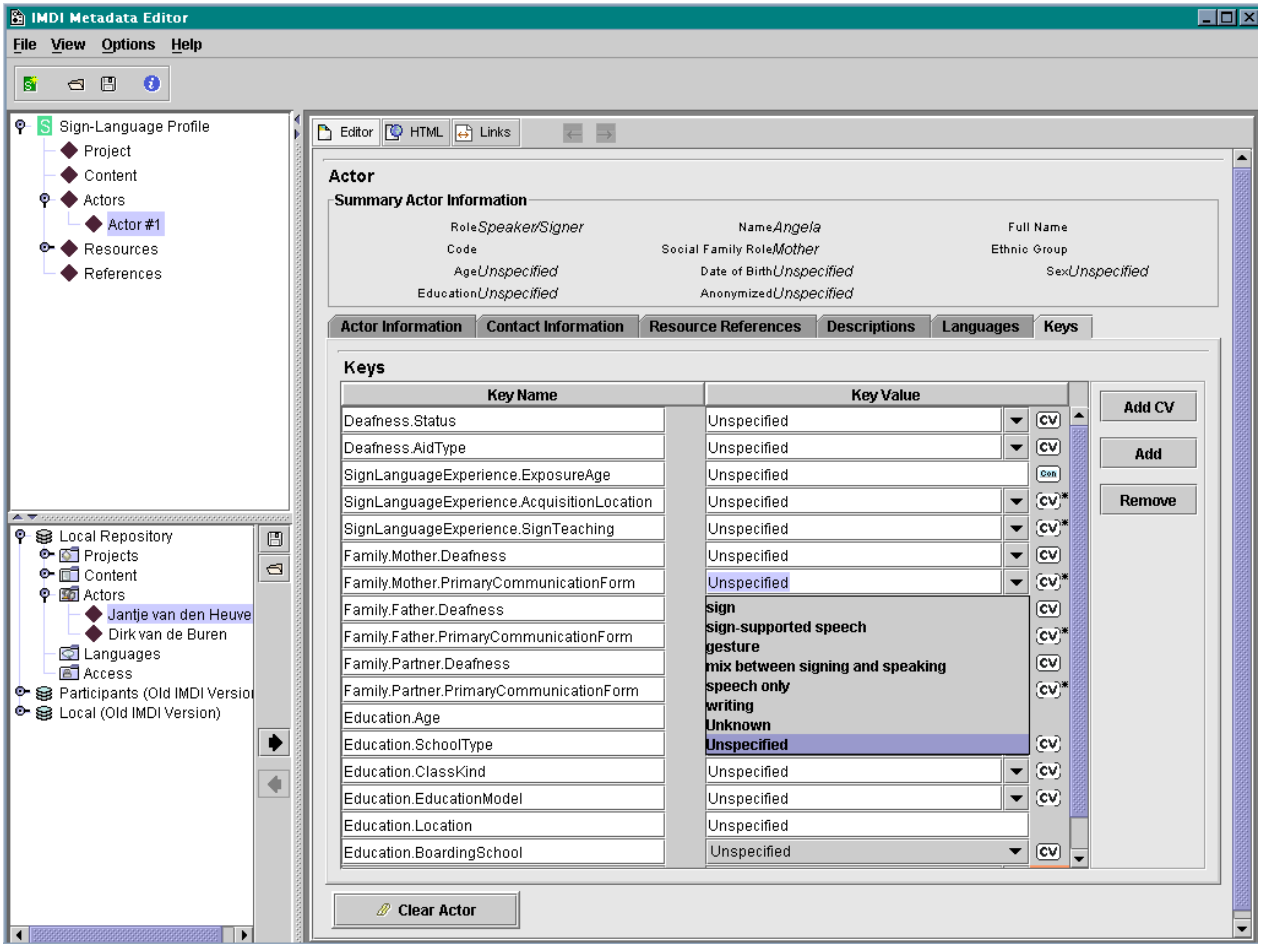

Figure 1. Screen dump IMDI Editor with Sign-Language Profile

IMDI Sign-Language Profile		
Special Key-name/value parts for the content part of		
Elicitation Method	A characterization of specific prompts used for eliciting language production.	no prompt single picture prompt picture story prompt / written language prompt / sign language prompt / video prompt / unknown
Iterpreting . Source	Source modality and language type	sign language, speech / sign supported speech / text / finger spelling / unknown / unspecified
Iterpreting . Target	Target modality and language type	sign language / speech / sign supported speech / text (subtitling) / finger spelling / unknown / unspecified
Interpreting . Visibility	Visibility of the interpreter in the video recordings	not visible / in view during whole session / in view during part of session, unknown, unspecified
Interpreting . Audience	Presence and nature of an audience that the interpreter is signing for.	Audience not present (signing to camera) / audience known to the interpreter / heterogeneous group partly known to the interpreter / anonymous audience (e.g. theatre) / unknown / unspecified
Special Key-name/value pairs for the actors		
Deafness . Status	Actor's ability to hear.	hearing / hard-of-hearing / deaf
Deafness . Aid Type	Type of hearing aid the actor has.	none / conventional / CI
SignLanguageExperience. ExposureAge	Age at which exposure to sign language and sign language use started	c (years;months)
SignLanguageExperience. AcquisitionLocation	Place where sign language was learnt.	home from family/home from tutor/ preschool teachers / teachers / family beyond home / friends
SignLanguageExperience .SignTeaching	Amount of experience with teaching sign language.	none / some / extensive
Family . Mother . Deafness	Describes mother's deafness status	deaf / hard-of-hearing/ hearing / n.a.
Family . Mother . Primary Communication Form	Describes mother's language input towards the actor.	sign / sign-supported speech / gesture / mix between signing and speaking / speech only / writing
Above two keys are repeated for father and partner		
Education . Age	Describes the age during which the school was attended	c(start age, dash, end age)
Education . School Type	Describes the age during which the school was attended	Bilingual home programme / kindergarten / preschool / primary school / vocational training / college / university
Education . Class Kind	Describes the kind of class in the school	deaf / hard-of-hearing / deaf class in hearing school / individually integrated
Education.EducationModel	Describes the education model used at the school	Bilingual / oral / mixed / sign monolingual / oral with interpreter
Education.BoardingSchool	Is the school a boarding school?	yes / no

Table 1. The IMDI Sign-Language Profile