

# Metadata Set and Tools for Multimedia/Multimodal Language Resources

P. Wittenburg, D. Broeder, F. Offenga, D. Willems

Max-Planck-Institute for Psycholinguistics  
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands  
peter.wittenburg@mpi.nl

## Abstract

Within the ISLE Project about International Standards for Language Engineering the IMDI Metadata Initiative developed a complete environment for creating, maintaining and using metadata descriptions for multimedia/multimodal language resources. This environment includes a proposal for a suitable metadata set, tools to create, browse and search in IMDI metadata domains and suggestions about how to organize centers acting as metadata repositories. By using the IMDI approach a formulation in RDF is intended which enable the IMDI set to be integrated in Semantic Web activities.

## 1. Introduction

In 1999 the Max-Planck Institute for Psycholinguistics started using metadata to organise its multi-media corpora [1]. This project was called "Browseable Corpus" (BC) because it not only used metadata for resources in order to make them locatable by automatic procedure, but it also used metadata for creating a hierarchical structure that can be browsed for the purpose of corpus exploitation. This was achieved by recursively structuring corpora in ever-smaller sub-corpora structures with each one described by its own metadata description pointing to the metadata descriptions of its sub-corpora. Creating browsable structures this way which creates space to integrate many other types of information such as project notes, also formed a basis for efficient corpus management.

The basic concepts of BC were used as one of the inputs to the ISLE Metadata Initiative (IMDI) [2] founded in early 2000. IMDI aims to reach consensus within a representative part of the linguistic community on a standard for metadata descriptions for multimedia/multimodal language resources. The IMDI metadata set is currently being applied within projects such as DOBES [3], the CGN corpus [4] and, of course the MPI's own corpora. Its relevance was checked for several other multimedia corpora such as the SmartKom [5] corpus. A preliminary showcase combined corpus data from 6 European institutions into one browsable and searchable domain.

## 2. Using Metadata Descriptions

A key issue in the IMDI approach is that a metadata set should be used for corpus discovery and corpus management as well as corpus exploitation. This implies that the metadata set should be able to describe the resources in sufficient detail to allow the resolution of relevant queries for the domain. It also implies that linked networks of metadata descriptions should be available, generated either automatically or manually and that it should be possible to include human readable texts or files with the metadata descriptions that can assist the user when browsing through a corpus. Corpora organized in this way can be easily integrated into bigger domains and they are an extremely useful facility for corpus managers to group all relevant information and knowledge together to facilitate corpus management. In this domain of linked

metadata descriptions the user would be able to browse and search and as a result find a single resource or a sub-corpus to work on. Consequently the user is likely to want to start a suitable tool for analysis, i.e. the metadata must contain information which indicates which operations can be executed on the resources found. Within IMDI it was anticipated that each user has his own view on corpora, therefore it was concluded that the IMDI environment should provide users the possibility of creating their own hierarchies so that several views can co-exist in parallel.

Of course, metadata will always exist as a source of information distributed via Internet, therefore all resources including the metadata descriptions themselves have to be specified as URLs. In this way metadata descriptions and connected resources can be accessed on the Internet by using standard HTTP. This simplifies the connection of different corpus domains to one super-domain. To support global searches via, for example, Dublin Core [6] based service providers, the IMDI domain is available for metadata harvesting in compliance with the Open Archives Initiative protocol [7].

Although the concept of metadata descriptions is still fairly new, the community is becoming aware that metadata descriptions will facilitate re-usage of valuable resources. Currently, most of the many resources are hidden in the storage containers of the various institutions and companies. Only few of them are visible via web-sites each having its own style of description. Since metadata are available to everyone, a domain of unified descriptions form an ideal way of informing others about available data even if the resources themselves are not directly accessible.

## 3. IMDI Metadata Set

IMDI's guiding principles when defining a metadata set have been that the best way to describe linguistic resources is to be able to describe the events and/or performances that are involved in their creation and usage by the community. The descriptions need to contain as much detail as necessary for a user who needs to easily discover resources, quickly check their usefulness and immediately exploit them. This bottom up approach can be compared with the approach in the media and film community which defined the MPEG7 standard [8]. It can and will lead to a more extensive and structured set than, for instance, the Dublin Core set. In taking such an

approach, the metadata set found can be seen as a first step towards a more complex domain ontology.

Some argue that it is necessary to have a low-overhead metadata set, since users may not want to spend too much time in providing all the information defined by the proposed IMDI element set. For IMDI the solution is that efficient tools are provided and that almost all fields are optional. So the overhead argument in case of more elaborate metadata sets does not hold, if elements are optional as in the IMDI case. Flexibility of the set of elements was one of the recurrent requirements, since we deal with a large number of different projects all recording multimedia material. In IMDI, flexibility was introduced by allowing user definable keyword/value pairs at several levels in the metadata structure.

The IMDI set for sessions<sup>1</sup> contains the necessary elements to describe the project a resource belongs to, the responsible scientists who created it, date and location of the recording, its content, its media files and annotations and if available the its derivative source. In the following a list of all elements is given. It is not the purpose of this paper to explain in detail what all the elements represent. For this we refer to the IMDI web-site: <http://www.mpi.nl/ISLE>. An attribute specifies whether the element is just a string, constrained (c), associated with a closed vocabulary (ccv) as in the case of “Continents” or with an open vocabulary (ov) which is open for extensions, or refers to a sub-block of information (sub).

#### Session

Name	str
Title	str
Date	c
Location	
Continent	ccv
Country	ccv
Region +	str
Address	str
<u>Description</u> <sup>2</sup>	sub
<u>Keys</u> <sup>3</sup>	sub
Project	
Name	str
Title	str
ID	str
<u>Contact</u>	sub
<u>Description</u> +	sub

<sup>1</sup> Sessions are the leaves in a corpus tree and cover units of linguistic analysis or performance including their media and annotation files. The IMDI initiative has defined a few other very similar metadata sets for corpus nodes, published corpora and lexica. They are not discussed in this paper.

<sup>2</sup> Descriptions are a field which the annotator can use to enter prose text intended for quick inspection by the user.

<sup>3</sup> Keys are those fields which guarantee flexibility. Each project or even user can define extensions in form of key-value pairs.

Collector	
Name	str
<u>Contact</u>	sub
<u>Description</u> +	sub
Content	
CommunicationContext	
Interactivity	ccv
PlanningType	ccv
Involvement	ccv
Genre	
Interactive	ovl
Discursive	ovl
Performance	ovl
Task	ovl
Modalities	ovl
Languages	
<u>Description</u>	sub
<u>Language</u> +	sub
<u>Description</u> +	sub
<u>Keys</u>	sub
Participants	
<u>Description</u> +	sub
Participant+	
Type	ov
Name+	str
FullName	str
Code	str
Role	ov
<u>Language</u> +	sub
EthnicGroup	str
Age	c
Sex	ccv
Education	str
Anonymous	ccv
<u>Description</u> +	sub
<u>Keys</u>	sub
Resources	
MediaFile+	
ResourceLink	c
Size	c
Type	ccv
Format	ov
Quality	c
RecordingCondition	str
Position	c
<u>Access</u>	sub
<u>Description</u>	sub
AnnotationUnit+	
ResourceLink	c
MediaID	c
Annotator	str
Date	c
Type	ov
Format	ov
ContentEncoding	str
CharacterEncoding	str
<u>Access</u>	sub
<u>Language</u>	sub
Anonymous	ccv
<u>Description</u>	sub

Source+

ID	str
Format	ov
Quality	ccv
Position	c
<u>Access</u>	sub
<u>Description</u>	sub

#### References

It is important to mention here how multimedia and multimodality can be described in IMDI. The IMDI set allows the user to describe the *Content* of a session which refers to a unit of analysis in the corpus. Each session is associated with the media and annotation resources belonging together. The IMDI set has elements to describe the *Communication Context*, the *Genre*, the *Task*, the *Modalities*, the *Languages involved*, and to add other useful project specific elements.

In most instances the associated vocabularies clarify what the definition of the element is although IMDI has already provided careful definitions. The element *Task* stands for typical experimental tasks occurring in language engineering and field-linguistics such as *info-kiosk situation*, *route description*, *wizard-of-oz experiment*, *frog-story*. The element *Modalities* has, of course, a vocabulary which includes, amongst others, *speech*, *gesture*, *sign*, *facial expression*.

As can be seen, the IMDI set has elements not only to describe content, but also to describe the *Media Files* (type of data, format of file, quality of material, conditions of recording, etc), the available *Annotations* (type of annotation, format of file, etc), and the *Original Media* (cassette, MD, etc) if available. To give the user immediate feedback on accessibility, IMDI contains elements to describe the access rights and whom to contact to obtain the resources.

As already indicated, Controlled Vocabularies (CVs) associated with elements are an important component of the IMDI metadata set and its tools, since they will guarantee that elements are used coherently by researchers and that search operations will provide the correct resources.

To achieve interoperability with Dublin Core (a more general set of 15 partially vaguely defined elements used to describe web resources used by the general public) a mapping document was created. Based on DC, another set (OLAC [9]) was created to achieve interoperability in the language resource domain. IMDI repositories will be open to OAI [7] type of metadata harvesting to implement the interoperability with DC and OLAC.

The IMDI set is defined in all respects through an XML Schema which is available at the IMDI web-site. All tools generate and operate on these XML files.

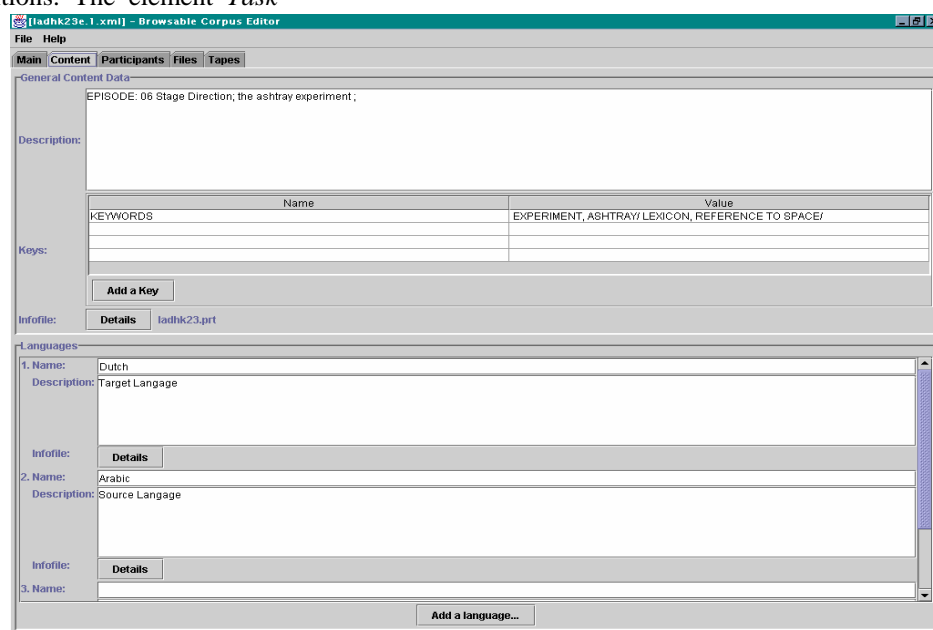
## 4. IMDI Tools

The tools that support the IMDI metadata set and infrastructure are:

- The IMDI BCEditor that is used to create IMDI metadata descriptions.
- The IMDI BCBrowser. A viewer for the IMDI metadata descriptions that allows navigating the universe of connected IMDI metadata descriptions.
- The IMDI Search tool that allows the user to specify a query for specific resources in the IMDI universe.
- A number of scripts allowing to work efficiently

All tools were programmed in Java and Perl for platform independence and are downloadable from the web-site: <http://www.mpi.nl/tools>.

Figure 1 shows a screenshot from the IMDI Editor



The editor presents all the IMDI metadata elements in a structured GUI to the user. It supports the use of Controlled Vocabularies and user definable keyword/value pairs that the IMDI set allows for user or project specific extensions. Also it enforces constraints on the values for some metadata elements where applicable and practical. To aid working efficiency the editor allows the re-usage of a number of element blocks which will recur in many metadata descriptions such as biographical data of the informants and collectors. The editor is programmed to synchronize with repositories providing controlled vocabularies on user command if the computer the editor is running on is connected to the web. This mechanism ensures that the user can download and use the most recent definitions, e.g. of the names of countries. Internationally agreed notation conventions allow differences between different vocabularies. For example, the ISO language lists contain only a few hundred language names and the Ethnologue list [10] contains more than 4000 names. In fact users can add their own

lists but searching would become a problem if there is no mapping definition.

One of the very important functions of the browser is that it offers the user a set of appropriate tools for further analysing resources once they have been located and it allows for operation in a distributed scenario where all resources are indicated by URLs. Each user or group of users can create a configuration file containing information on how to immediately start a tool and pass over the necessary parameters to start the tool with the discovered resource(s). The browser offers a selection from which the user can choose.

The search tool is the most recent IMDI development. It allows the user to specify a query for sessions whose metadata complies with the specified constraints. The UI offers the user an easy way to specify a query compliant with the IMDI element set, the elements value constraints and CVs used.

Results are presented in the form of URLs for the session metadata description files that comply with the query. The user may make these sessions visible in the IMDI-BCBrowser for further inspection or a special corpus label can be created containing all these

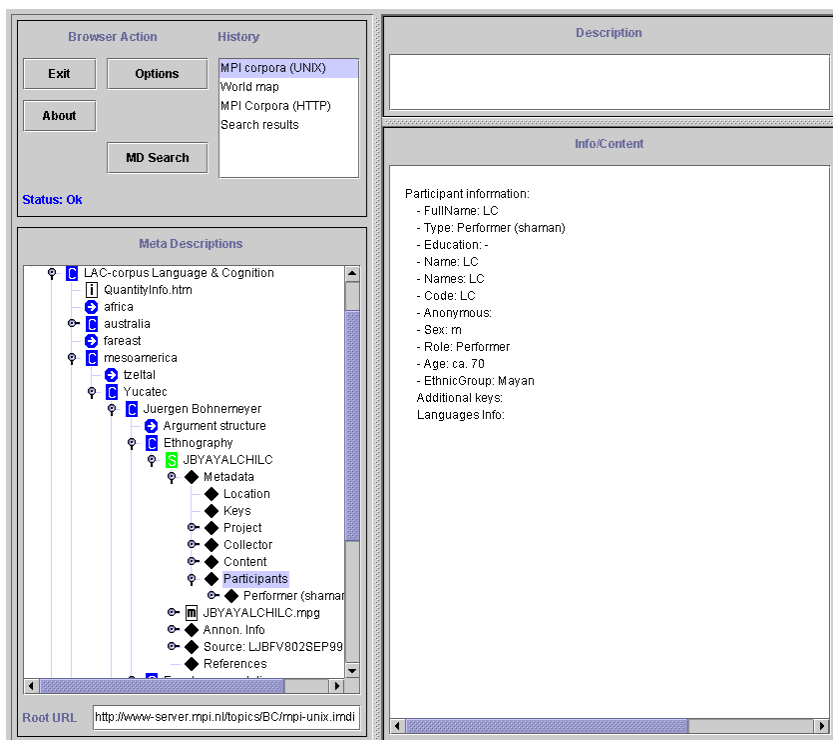


Figure 2 shows a screenshot of the IMDI browser.

The IMDI BCBrowser is the central tool for exploiting the IMDI metadata infrastructure. It allows navigation in the domain of linked IMDI metadata descriptions by clicking on corpus links. The browser keeps track of its position in the browsable corpus structure and displays the metadata and human readable descriptions associated with the sub-corpus in focus. It allows the user to set bookmarks so that easy navigation is facilitated.

The browser is also capable of displaying HTML formatted or PDF files that are often provided as extra documentation for corpora. It is possible to link in such HTML pages or PDF files in the corpus tree. From the HTML pages there may be links back to metadata descriptions making it possible to mix classical HTML browsing with browsing the IMDI corpus universe.

An interesting application of this is a world map that was created as a portal of the MPI corpora. This world map is viewable as an HTML file but has, at the appropriate places, links to metadata descriptions for corpora that correspond to those locations. We are presently engaged in trying to incorporate a professional geographic information system since the HTML world map is not completely satisfactory. The worldmap is just one other alternative view on a corpus since it is organized according to geographical principles.

sessions that can be saved for future reference and processing. The search tool can, of course, be started from the IMDI-BCBrowser. The search tool has to be extended to support the distributed architecture underlying the IMDI concept and it has to be checked as to how it can support harvesting of other metadata repositories by, for instance, using the OAI protocol. Currently, two teams are working on an improved search tool working in fully distributed scenarios.

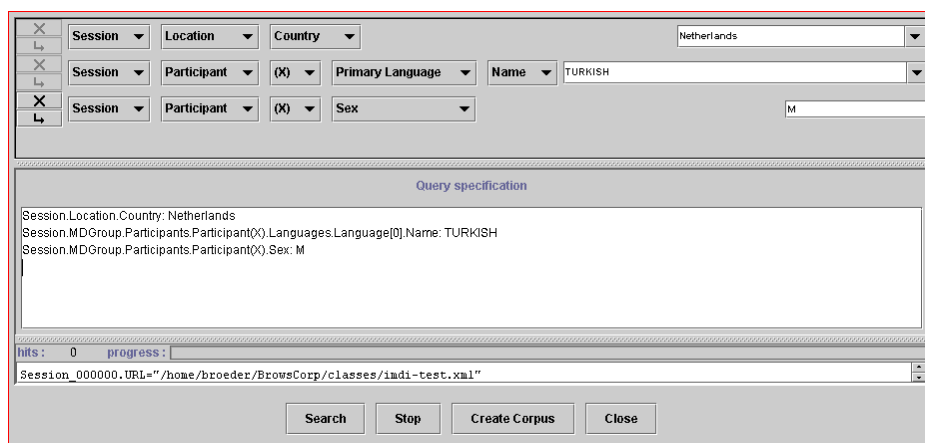


Figure 3 shows a screenshot from the search component.

The IMDI team also created a number of scripts which allow users to efficiently work with IMDI type of metadata descriptions. One such tool is provided to add or change element values in a whole range of MD descriptions by one command. Another allows the user to create metadata descriptions from spreadsheet documents, although this has proved problematic. Spreadsheet entries are not guided by constraints or controlled vocabularies

therefore conformity has to be checked very carefully. There are a few other minor scripts which will hopefully become obsolete when the editor or browser have been extended.

## 5. IMDI Corpora

At present we have available as IMDI tagged corpora:

- the MPI corpora of the “Acquisition” and “Language and Cognition” group which contains more than 2 TB of media data and more than 7000 multimedia sessions;
- a large second learner language acquisition corpus also containing audio recordings;
- the data of the DOBES project about endangered languages where also audio and video recordings form the basis;
- the data of the CGN (Spoken Dutch Corpus) project.

Furthermore we have been experimenting with converting parts of existing corpora to see if the IMDI set is applicable. These tests range from the well-known “ChilDes” corpora [11] to language engineering corpora as “TIMIT” [12] and “SmartKom”. An interesting project was also the construction of a distributed corpus with examples of (parts of) corpora of six different European institutes. This was demonstrated as a first distributed IMDI scenario during the official opening ceremony of the “European Year of the Language” in Lund in 2001.

## 6. Future Developments

As a preliminary solution and part of the IMDI showcase, the MPI serves as a focal point maintaining the IMDI web portal as a starting point for the IMDI universe and maintaining the IMDI metadata Schema and CV definitions. However, the MPI does not have ambitions to perform this task in the long run. Such hosting activities are better performed by organisations such as BAS [13], ELRA and LDC. The maintenance of the IMDI set and the related tools by the MPI has been secured for many years by using them in different long-term projects. Besides these organisational problems, there is also a need for further tool development, such as a tool offering users a graphical interface for creating alternative “personal” corpus trees. Maintenance tools are required that allow users to copy parts of corpus trees to other portable media such as CDROM and DVD. In this way they can work under field conditions or make personal archive copies.

A major revision of the IMDI metadata set is expected to occur in 2002, therefore comments on how to improve it are welcome. According to the most recent discussions, it can be concluded that the MD set in general is very mature and stable with the exception of a very few elements such as “Anonymous”. But the elements and vocabularies which were defined to describe the content of the

resources have to be modified after a year of experience. Here, the elements define the dimensions of descriptions and the vocabularies the values along these dimensions. Although the current definitions are based on linguistic experience, it is obvious that not all contents can be described equally well with them.

Currently, the IMDI definitions are specified with the help of an XML Schema, i.e. the relations between concepts are implicitly defined in the structured IMDI set. To open up the way to the Semantic Web these implicit relations will be explicitly defined with the help of RDF [14]. All RDF Schemas will be put into open RDF repositories so that they can be re-used. It has to be checked whether it will be possible to make use of already existing descriptions within the IMDI set.

## 7. References

- [1] 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.
- [2] ISLE/IMDI: <http://www.mpi.nl/ISLE> & [http://www.mpi.nl/world/ISLE/documents/papers/white\\_paper\\_11.pdf](http://www.mpi.nl/world/ISLE/documents/papers/white_paper_11.pdf) & [http://www.mpi.nl/ISLE/documents/draft/ISLE MetaData 2.5.pdf](http://www.mpi.nl/ISLE/documents/draft/ISLE_MetaData_2.5.pdf)
- [3] DOBES: <http://www.mpi.nl/DOBES>
- [4] CGN: <http://www.now.nl/gw/introductie>
- [5] SmartKom: <http://smartkom.dfki.de>
- [6] DC: <http://www.dublincore.org/>
- [7] OAI: <http://www.openarchives.org/>
- [8] MPEG7: <http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>
- [9] OLAC: <http://www.language-archives.org/OLAC/>
- [10] Ethnologue Language List: <http://www.ethnologue.com>
- [11] ChilDes: <http://chilDes.psy.cmu.edu>
- [12] TIMIT: <http://www ldc.upenn.edu/Catalog/LD93S1.html>
- [13] BAS: <http://www.phonetik.uni-muenchen.de/Bas/BasHomeen.html>
- [14] RDF: <http://www.w3.org/RDF> & <http://www.w3.org/sw>