

# The IMDI Metadata set, its Tools and accessible Linguistic databases

**Daan Broeder**  
**Freddy Offenga**  
**Don Willems**  
**Peter Wittenburg**

Max-Planck Institute for  
Psycholinguistics  
Nijmegen, The Netherlands  
Contact: d.broeder@mpi.nl

## Abstract

The ISLE Metadata Initiative (IMDI) developed a metadata set for describing multi-media/multimodal language resources. It is supported by special tools for creation and exploitation and is used by some large projects in Europe to organise their archives and make the data more accessible.

## 1 Introduction

In 1999 the Max-Planck Institute for Psycholinguistics started using metadata to organise and structure its multi-media corpora [1]. This project was called “Browseable Corpus” (BC) because it did not only use metadata for resources in order to make them locatable by automatic procedure, but also by using 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.

The basic concepts of BC were also present in the ISLE Metadata Initiative (IMDI) [2] that was founded in early 2000. IMDI aims to reach consensus within a representative part of the linguistic community on a standard for metadata descriptions for language resources. Besides the development of a metadata vocabulary it will also deliver a showcase containing tools for creating and exploiting IMDI standard metadata

descriptions. A dataset demonstrating the variety of the domain of applicability of IMDI will also be part of this showcase.

The IMDI metadata set is currently being applied within projects such as DOBES [3], the exploitation software for the CGN corpus [4] and, of course the corpora of the MPI itself.

## 2 Using Metadata Descriptions

As a key issue we think that a metadata set should be used for corpus discovery 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 it should be possible to supply human readable texts or files at different levels of the metadata descriptions that are able to guide and inform a user when browsing through a corpus.

Of course browsing and searching procedures do not exclude one another. Often a certain subcorpus will be located by browsing, after which a search will be performed on that subcorpus to locate specific resources.

Also when a suitable resource has been located, be it through browsing or a direct search procedure, the user is likely to want to start suitable tools for analysis. If this is to be done transparently by means of the browse or search tool, the metadata should describe the resource in relevant detail for the browser or search tool to decide what analysis tools are available for a specific resource.

Another central concept in our metadata framework is that of a *Session*: the assemblage

of all language resources associated with one linguistic performance or event (e.g. annotation

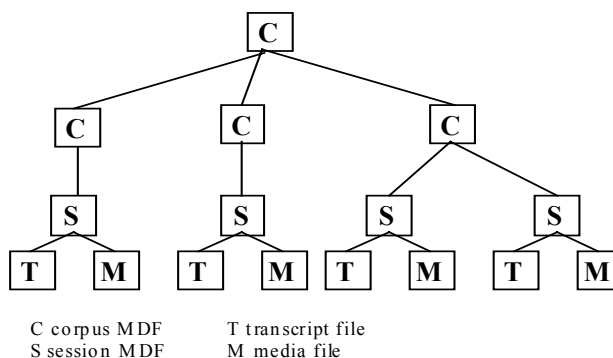


Figure 1

documents, audio and video recordings, photographs and field-notes). The sessions are the leaves in the corpus tree while the subcorpora are its branches (see figure 1). Most metadata pertaining to language resources is shared by all resources in a session but there are a few items such as file format, access restrictions etc. that are specific for an individual resource.

### 3 IMDI infrastructure

In addition to the requirements following from the use of metadata sketched above, a few practicalities had to be taken into account.

First there is the division between metadata and annotation data, which is partly based on existing conventions (e.g. the header of an annotation file and the annotation itself) and partly on the wish to handle existing annotation formats. In MPEG7 en TEI [5,6] for example these two dimensions are integrated in one single file format. The division we use however has the advantage to be able to have separate access restrictions for metadata and annotations. This leaves, for instance, the question of access to resources themselves to the resource creators.

It would have been unthinkable not to take the Internet into account as a medium for accessing the metadata and resources. Therefore all the specifications of resources and metadata descriptions are in the form of URL's. This makes it possible to have a distributed connected web of IMDI metadata descriptions and

connected resources that are made available on the Internet by standard HTTPD servers.

Another requirement taken into account is that of the appropriate technology support level. Our idea is that the IMDI metadata infrastructure should prove as usable for a small project group without many resources or even an individual researcher working on a home-pc as for a large institute such as the MPI. This has led us for example in the case of metadata search to develop tools that work without special database software meanwhile not excluding the use of database software by sites that can support it.

### 4 IMDI Metadata Set

Our guiding principle has 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. This bottom up approach can lead to a very extensive and complicated set (see for instance MPEG7 [5]) that may become unwieldy. However the set should be able to describe the resources in sufficient detail and while this may be a subjective measure only its use will tell us if the set meets the requirements. While the IMDI set is large, its should be remembered that only in rare cases all the metadata fields will be filled in for a resource. Also, IMDI tools for the creation of metadata descriptions support the reuse of (parts of) existing descriptions thus minimising typing work.

Flexibility has been introduced by allowing user definable keyword/value pairs at several levels in the metadata structure. These keyword/value pairs may be linked to "controlled vocabularies" (CV's).

Please see [2] for further details of all the IMDI metadata elements and their vocabularies.

### 5 Controlled Vocabularies

Controlled Vocabularies (CV's) are an important component of the IMDI metadata set and tools. The value of a metadata element can be constrained by a CV in the sense that its value can only be a choice from a limited set. In IMDI we discern the following possibilities: A CV may be closed or open. Here, open means that the user may provide his own value for the element and ignore the set of values. In this case

the vocabulary is just a recommendation. Closed vocabularies are always mandatory. In fact open CV's are not real CV's but could become so after a metadata set has been used sufficiently to identify all values users need. A CV may be a CV list or not. With a list CV a user may choose any number of values from the vocabulary but every value can be chosen only once.

All the IMDI tools have the possibility of connecting to CV servers on the Internet in order to download the latest CV definitions the tools encounter in the metadata descriptions. These definitions are then stored in a cache for speedy future reference and in case no Internet connection is possible as often occurs in field conditions.

## 6 Comparison to other Metadata sets

When trying to create a metadata set to describe a domain several approaches are possible. It is possible to design a set from the bottom up or take an existing metadata set (from another domain) and try to adapt it to your needs. Here IMDI has taken the first approach, although basing its work on existing conventions and "legacy" metadata systems. The other metadata initiative for the linguistic domain OLAC [7] has clearly taken the opposite direction. The OLAC initiative adopted the DC [8] metadata set as a low overhead set and extended this with one extra element and special qualifiers. Thus taking advantage of DC being a well established metadata set and the fact that the important OAI initiative [9] whose concepts and protocols OLAC adopts, chose DC as its metadata set.

The choice of OAI for DC as a metadata set for exchanging information about all possible domains is not disputed. It is however a legitimate question whether DC even with qualifiers and a few extra elements can describe the Linguistic domain in sufficient detail to answer specific questions.

As far as we can tell there will be information loss when trying to map IMDI metadata on the OLAC. Then there is the question of terminology. A metadata set that was specifically created for a domain will use terminology that the user community

understands. When using a set from another domain this will not always be the case.

Another legitimate question is if describing a domain with a specialised metadata set isn't always problematic because of complex mapping problems other metadata sets. Currently we see standards and techniques from the "Semantic Web" such as RDF [10] developing that make such mappings possible. Therefore there is no reason to avoid creating a very specific domain tied set.

In our opinion IMDI is the appropriate set to use if you need to describe or query resources in very specific ways. Possibly OLAC could match this by adding more elements and qualifiers but at the same time losing its low overhead and simplicity.

The advantage of OLAC and DC for answering general questions is clear; therefore we will provide a mapping service for OLAC metadata harvesters, thus serving as an OAI/OLAC data provider. This should occur before the end of the IMDI project.

## 7 The 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.

All tools were programmed in Java and Perl for platform independence. The relevant programmes that are part of the IMDI showcase will be made available to the user community at the end this year when the ISLE project ends. However since they are used in some long-term projects so they will be maintained after the end of the ISLE project.

## 7.1 The IMDI BCEditor

This 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 purposes. Also it enforces constraints on the values for some metadata elements where applicable and practical.

Some subsets of metadata descriptions such

as the biographical data of informants or investigators or project administrative data are often repeated in other sessions. The editor allows the storage of such substructures in separate files for reuse.

When we started on the development of the editor we had the hope that we could develop an editor that could configure itself on the basis of the XML-Schema or DTD for the IMDI set. This proved impractical due to the fact that although XML-Schema do include information on the structure of an IMDI metadata description and the constraints on its element values, it does

The screenshot shows the IMDI BCEditor interface. At the top, there is a menu bar with 'File', 'View', 'Window', and 'Help'. Below the menu bar is a tabbed interface with tabs for 'General', 'Project', 'Collector', 'Content', 'Participants', 'Resources', and 'References'. The 'General' tab is active, showing the 'Session' section with three input fields: 'Session Name' containing 'JBYAYALCHILC', 'Session Title' containing 'u payal chi' don Luz', and 'Recording Date' containing '1997-09-02'. Below this is another tabbed interface with tabs for 'Descriptions', 'Location', and 'Keys'. The 'Descriptions' tab is active, showing a 'Language' field, a large 'Text' area, and a 'Link' field. To the right of the 'Text' area is a vertical list box containing the word 'Unknown'. Below the list box are 'Add' and 'Remove' buttons. At the bottom of the window is a 'Clear Session' button with a trash icon.

Figure 2

not contain information in how to layout these elements in an acceptable GUI. It was therefore decided to hardwire the connection between IMDI metadata elements and the relevant UI elements of the editor in the editor programme. This requires some reprogramming whenever the IMDI set is modified. While this occurred frequently at the start of the IMDI project, the set has become more stable and work on the editor is now more concerned with the refining and perfecting of the UI. A picture of the IMDI BCEditor showing the modular setup of the UI that mirrors the modular structure of the IMDI metadata set can be seen in figure 2.

## 7.2 The IMDI BCBrowser

The IMDI BCBrowser is the central tool for exploiting the IMDI infrastructure. It allows navigation of the universe of linked IMDI metadata descriptions by clicking on corpus links. The browser keeps track of its position in browsable corpus structure and shows the

metadata and human readable descriptions associated with the subcorpus in focus.

The browser is also capable of showing HTML formatted files that are often provided as extra documentation for corpora. It is possible to link in such HTML pages in the corpus tree. From these 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.

A very important function of the browser is that it offers the user a set of appropriate tools for further analysing resources once they have been made visible. As already mentioned, the

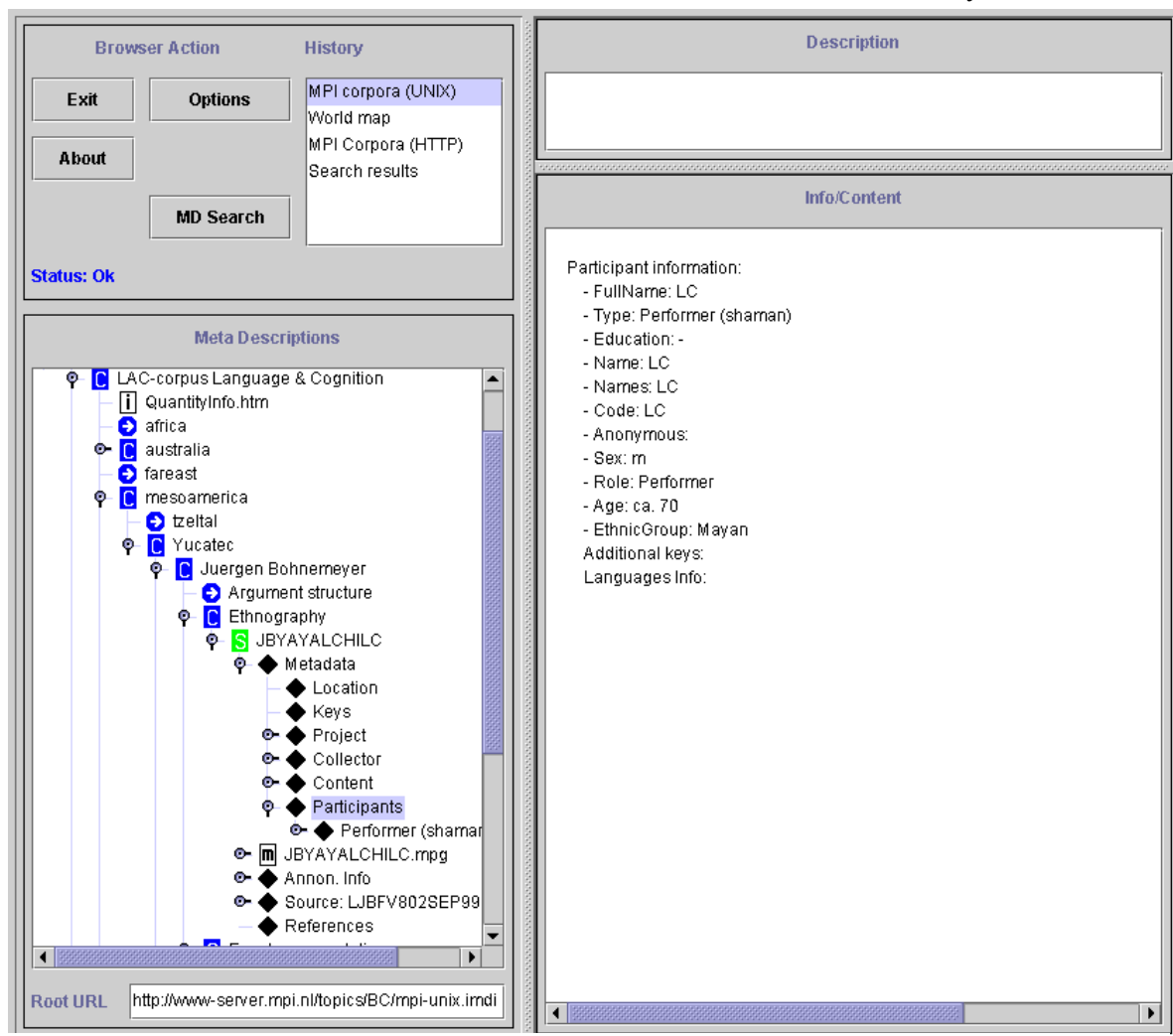


Figure 3

IMDI set has some elements that identify the type and format of a resource. The browser uses this information and also the knowledge of other available resources for a session to choose the appropriate tools. Researchers wishing to refine or modify the metadata descriptions may do so by starting the IMDI-BCEditor from the browser, since this is also one of the tools offered. At present this connection between resource type and format is hardwired in the browser, something we will change for user configurability in the very near future.

More flexible would be a scheme that allows a tool repository to be queried for the availability of appropriate tools after which these could then be downloaded. This requires a special protocol and new standards. There are plans for working this out and implementing it in a future project [10].

A picture of the browser showing part of the MPI corpus for Yucatan is shown in figure 3.

### 7.3 The IMDI BCSearch Tool

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 CV's used.

Results are presented in the form of URL's 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 can be created containing all these sessions that can be saved for future reference and processing. The search tool can of course be started from the IMDI-BCBrowser.

As previously stated, we aimed at developing a system that would function without taking recourse to large database systems. We first considered a "crawler" like approach were a search programme would look through the corpus tree following the links until all the relevant sessions were found. This method, although simple is much too slow due to the repeated resolving of the subcorpus links. At the moment we transform all the session metadata descriptions in a flat structure and put them all into a single file. For consistency purposes, this search-file has to be updated whenever the metadata content of the corpus is modified. A query is now translated in a perl script searching through this file. This mechanism seems adequate for corpora up to 100,000 sessions. We do not exclude the need for database technology for large sites hosting a significant higher number of sessions.

An issue that has not yet been tackled, at

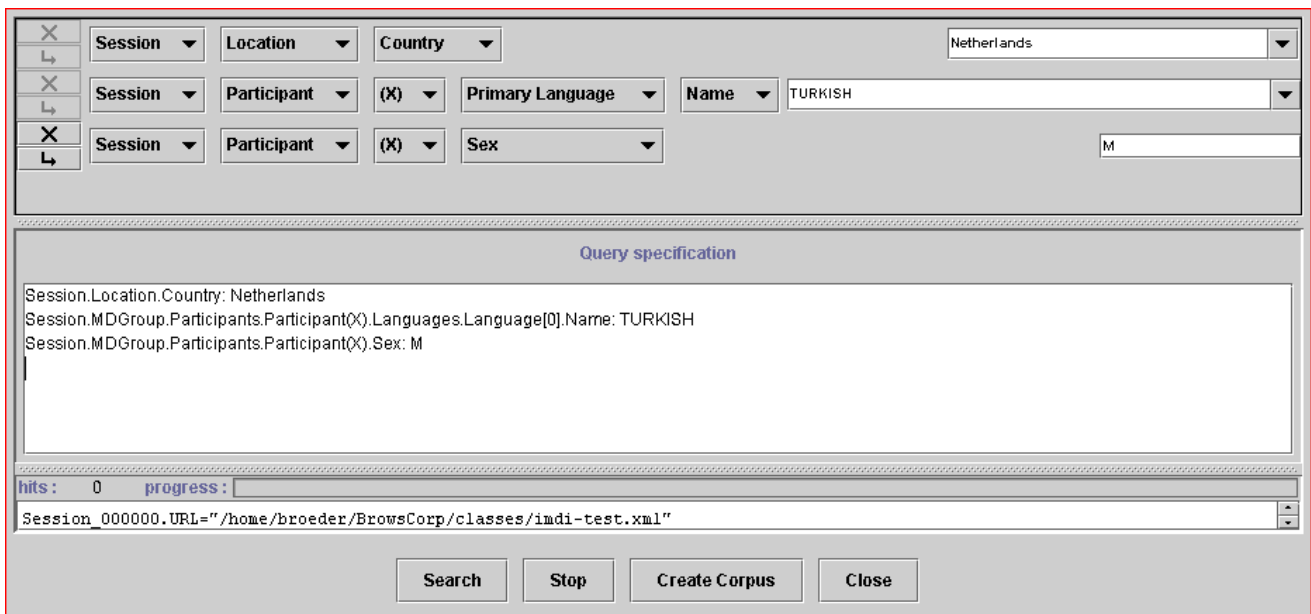


Figure 4

least not through an implementation, is how this search mechanism works in a distributed situation where a user would like to query a number of corpora at different sites. It appears not too difficult to be able to have a site answer a remote IMDI search query, but it might be preferable to concentrate all search-files at one site in the same way OAI/OLAC service providers harvest OAI/OLAC data providers. It should however never be the case that a remote service provider has to be available for a query on local metadata.

A picture of the IMDI Search Tool specifying a simple query is shown in figure 4.

## 8 IMDI Corpora Data and its maintenance

At the moment we have available as IMDI tagged corpora or are in the conversion process of:

- The MPI corpora of the “Acquisition” and “Language and Cognition” group.
- ESF A second language acquisition study corpus.
- The data of the DOBES project.
- The data of the CGN project.

Further more 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 to language engineering corpora as “TIMIT” and “SmartKom”. An interesting project was also the construction of a corpus with examples of (parts of) corpora of six different European institutes. This was demonstrated during the opening of the “Year of the Language” in Lund earlier this year.

On the effort of the creation and maintenance of IMDI metadata we can make the following statements. The biggest step is the conversion of “legacy” non-XML metadata to a version of IMDI. This almost always requires handwork.

Once the step to an XML format is taken, subsequent conversions to higher IMDI versions can be done using XSLT. It is very important to administrate the metadata Schema versions and

the accompanying tools and conversion scripts. Any errors there will lead to severe problems.

## 9 Future developments

A special question is the organisation of the infrastructure. Where will we store the metadata descriptions and/or resources and who will maintain the standards and tools? 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 as ELRA and LDC. But MPI sees itself as a tool builder and the maintenance of these tools has been secured for some time by using them in different long-term projects.

Beside these organisational problems, there is also a need for further tool development. Such as a tool offering the users a graphical interface for creating alternative “personal” corpus trees. Maintenance programmes and scripts are needed 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.

## 10 References.

- [1] Broeder, D.G., Brugman, H., Russel, A., and Wittenburg, P., (2000), A Browsable Corpus: accessing linguistic resources the easy way. LREC 2000 Workshop, Athens.
- [2] <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] MPEG7: <http://www.cselt.it/mpeg/standards.html>

- [6] TEI: <http://www.tei-c.org/>
- [7] OLAC: <http://www.language-archives.org/OLAC/>
- [8] DC: <http://www.dublincore.org/>
- [9] OAI <http://www.openarchives.org/>
- [10] D. Broeder and P. Wittenburg: Interaction of Tools and Metadata-Descriptions for Multimedia Language Resources, in Proceedings of the ACL/EACL Workshop Sharing Tools and Resources, Toulouse, 2001.
- [11] RDF: <http://www.w3.org/RDF> & <http://www.w3.org/sw>