

Metadata Proposals for Corpora and Lexica

P. Wittenburg, W. Peters⁺, D. Broeder

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

Abstract

A number of metadata proposals appear to be relevant to establish a searchable and browsable domain of language resources so that users can easily discover suitable resources on the Web. These proposals differ in their approach, in their descriptive detail, in the set of linguistic data types supported by specific elements and the supporting tools. The IMDI initiative, in particular, has worked out not only a set for (multimedia) corpora, but also for lexica. All initiatives have declared their commitment towards interoperability where Dublin Core will play a role in the near future. For the long term we foresee much effort to make the metadata sets compliant with the trends of the Semantic Web and to allow an increasing re-usage of existing sub-schemas and data categories that will probably be formulated with RDF.

1. Introduction

The enormous increase in the amount and complexity of language resources requires new management and discovery approaches. The normal search engines operating on the normal web-site texts do not offer the precision to allow efficient discovery for the professionals looking for specific language resources. Therefore, in many areas initiatives were started to define metadata sets, i.e. a set of descriptor elements allowing a resource to be described sufficiently well for easy discovery. Differences in approach can be seen in so far that some sets only focus on easy retrieval by supporting search functionality while others also address the issue of resource management and the creation of browsable domains.

For language resources, metadata sets are proposed by Dublin Core (DC), the Open Language Archives Community (OLAC), the ISLE Metadata Initiative (IMDI) and MPEG7 [1,2,3,4]. While Dublin Core provides 15 elements which can be used to describe any type of authored web resource, OLAC and IMDI offer special elements and refinements for language resources. OLAC is based on DC, it wants to cover all types of language resources by applying the 15 elements and essentially adding one element to describe the language covered by the resource. Further OLAC has introduced refinements to narrow down the broad semantic scope of the DC elements.

IMDI has taken a completely different approach as it wanted to describe language resources in greater detail in order to enable domain specialists to find useful resources more easily. To that purpose IMDI derived structured metadata sets based on a number of previous suggestions (for an overview see [4]) and which offer specific elements per data type, i.e. a difference is made between multimedia corpora and lexica.

MPEG7 is another important initiative which has links with the language resource domain. It is driven by the media and film industry and has to operate in the future scenario described by the MPEG4 documents about the user-driven integration of multimedia objects. MPEG7 has to describe multimedia resources in such detail that it is possible to search for media content, but also to filter out certain streams at the decoder. MPEG7 offers a very

detailed list of features which to a large extent represent annotations of low-level and high level characteristics of the streams. Description schemes (DS) can be defined for many types of information groups. MPEG7 already covers more than 100 of such DS. MPEG7 does not offer the elements needed from the specialists working in the language resource domain, although it would be possible to define a special DS within MPEG7.

The harmonization initiatives are of great relevance. Currently, they accept understandably that Dublin Core will form the common set to be used by services offering general services for various resources for the casual web-user who does not rely on a full text search engine. For MPEG7 and IMDI, mapping schemes have been developed which map elements of the respective sets to Dublin Core elements. In close interaction with the Dublin Core specialists, the MPEG7 specialists decided to apply a very restrictive mapping of elements, i.e. only where the element semantics are clear. For example, for the element “creator” or “author” a mapping was proposed. This guarantees that the semantics of the DC elements will not be blurred. The IMDI specialists have described in detail what kind of mappings are possible (see mapping document in [4]). Also they will apply a restrictive mapping first.

2. IMDI Metadata Set for Corpora

For a detailed description of the IMDI metadata set for corpora we refer to documents developed by the IMDI initiative (see [4]). Here we want to consider only its main construction principles.

Session	(Name, Title, Date)
Location	(Continent, Country, Region, Address, Description ¹ , Keys ²)

¹ Descriptions are a field which the annotator can use to enter prose text intended for quick inspection by the user.

² Keys are those fields which guarantee flexibility. Each project or even user can define extensions in form of key-value pairs.

Project

(Name, Title, ID, Contact, Description)

Collector

(Name, Contact, Description)

Content

(CommunicationContext, Genre, Task, Modalities, Languages, Description, Keys)

Participants

(Type, Name, FullName, Code, Role, Language, EthnicGroup, Age, Sex, Education, Anonymous, Description, Keys)

Resources**MediaFile**

(ResourceLink, Size, Type, Format, Quality, RecordingCondition, Position, Access, Description)

AnnotationUnit

(ResourceLink, Annotator, Date, Type, Format, ContentEncoding, CharacterEncoding, Access, Language, Anonymous, Description)

Source

(ID, Format, Quality, Position, Access, Description)

References

The elements in bold face denote the main dimensions of the descriptions. Within the IMDI scheme they represent sub-schemas in analogy to MPEG7. The terms in brackets are the elements within these sub-schemas which can be associated with values. A few of these represent other sub-schemas for separate purposes such as "Access" which refers to a block with access information.

3. IMDI Metadata Set for Lexica

With respect to lexica there was no detailed set of metadata available. With DC all resources can be described, but on an insufficient level for the user from the linguistic domain; also OLAC does not address these special linguistic needs. Within the ISLE Metadata Initiative and a number of projects such as DOBES [7] a need was indicated to create a useful and detailed metadata set that describes all types of lexica such as simple wordlists, multilingual or monolingual dictionaries, concordances and many others. A first proposal has been formulated out within the IMDI project (see [4]). As opposed to corpora there were no earlier examples of such metadata descriptions. Even the Text Encoding Initiative [8] has not made explicit statements about this. Therefore IMDI had to analyze the results of existing lexicon initiatives. The work within initiatives such as EAGLES/ISLE [9], OLIF [10] and the terminology oriented projects such as MARTIF [11] and SALT [12] gave much insight into the ways lexica could be described at metadata level. This analysis resulted in a comprehensive report about lexicon structures [13]. A recent workshop about multilingual lexica within the ISLE

project [14] devoted time to discussing metadata issues. Peters [15] and Gibbon [16] both described principles and elements of how lexica could best be described to satisfy the wishes of the community. While Peters focused on the elements needed to describe linguistic content such as main categories and sub-categories, Gibbon focused on structural aspects. The two proposals can be seen as complementary.

The IMDI proposal unifies the existing IMDI approach for corpora with the suggestions from Peters and Gibbon. The structural analysis has shown that lexicon structures are highly dependent on the languages studied and on linguistic theories used in the modelling process. Therefore lexical metadata should be defined at a sufficiently high and theory-neutral level of linguistic abstraction. For more detailed inspections it should offer the possibility to access more detailed levels of linguistic content description such as a resource schema definition. It was agreed, therefore, to define main domains of linguistic description (e.g. "Orthography") and associate controlled vocabularies with them. A typical content description would contain a list of such main categories, each of them having a set of values. In the case of "Orthography" it could be, for example, "spelling, syllabification, hyphenation".

Other metadata types that have been taken into account concern, among others, the language covered by the resource, the language the resource uses to describe its entries, the modality of the linguistic content (e.g. sound, video, graphics), formal aspects of its storage format, and administrative data such as the creator and the date of creation. This can be done in a similar way to how corpus resources are described in the existing IMDI set.

Lexicon Object

(Name, Title, Date, Version, LexiconType)

Creator

(Name, Contact, Description)

Project(Name, Title³, ID, Contact, Description)**Object Languages**

(Description, MultilingualityType, Language)

Meta Languages

(Description, Language)

Lexical Entry

(Modality, Headword type, Orthography, Morphology, Morphosyntax, Syntax, Phonology, Semantics, Etymology, Usage, Frequency)

LexiconUnit

(Format, AccessTool, Media, Schema, Character Encoding, Size, No Lexical Entries, Access, Keys, Description)

Source**References**

³ In the published version at the IMDI site this element was not mentioned due to an omission.

This first proposal for a lexicon metadata set was designed bottom up in accordance with the user requirements. A distinction had to be made between ObjectLanguages and MetaLanguages. While the first refer to the languages relating to the lexicon, the second are used to describe the languages which are used to describe the meaning and other aspects. The linguistically relevant information is contained in the block named “LexicalEntry”. Here the given terms specify sub-dimensions which can be associated with a number of values.

An example may illustrate the principle. Given that the lexicon contains the following morphosyntactic information *Part of Speech, Inflection, Countability, Gradability and Gender*. Then independent of the structural embedding of the information in the lexicon, the metadata description would contain under “Morphosyntax” exactly that list as values, announcing the availability of the appropriate kind of information. It was decided not to include details of the microstructure in the metadata description since there are too many differences between various lexica. It is up to the user to look at details.

Another point of concern is the obvious fact that there is some overlap with the metadata set for corpora/sessions. When further discussions turn out to support the approach for the definition of the lexical metadata set, efforts have to be taken to formulate sub-schemas so that at least some are the same for the two sets. For some main dimensions such as Creator/Collector, Project, AnnotationUnit/LexiconUnit and References this seems to be possible. Of course, the dimensions intended to describe the lexical content are different from the content descriptors for corpora. Further, there is no “Participants” concept for lexica. These sub-schemas have to be different dependent on the type of metadata set.

4. Comparison of Metadata Environments

For the language resource community it may be useful to compare the existing metadata proposals. A comparison would simplify the task of choosing the most appropriate one a certain project. In the following table a number of points are mentioned. It is not apparent whether these points can be seen as advantageous under all circumstances.

Feature	DC	OLAC	IMDI	MPEG7
language resource specific	no	yes	yes	no
subschema per data type	no	no	yes	no
bottom up design	no	no	yes	yes
size of element set	small	small	middle	large
level of detail	low	low	high	high
extensible by user/project	no	no	yes	yes
overhead ⁴ in creating	low	low	high	high
housekeeping elements	no	no	yes	?
Resource bundling	no	no	yes	?
structure	flat	flat	struct	struct
direct DC compliance	-	yes	no	no
DC mapping available	-	yes	yes	yes
XML Schema Definition	yes	yes	yes	?

⁴ Overhead involved indicates time to spend for creation if all elements are used. In IMDI however only few elements are mandatory.

spec. XML sub-schemes	no	no	not yet	yes
discovery and selection	search	search	s & b ⁵	s, b & f ⁶
management support	no	no	yes	?
editor with cv ready	no	no	yes	no
browser ready	no	no	yes	no
Search ready	yes	yes	yes	no
Search on linguistic details	no	no	yes	no
Efficiency tools ready	?	?	yes	?
Direct tool start	no	no	yes	?

From the outset the IMDI environment was perceived as having to support LR specialists in not only discovering but also managing resources. Therefore, LR of different data type can be integrated into linked domains of metadata descriptions. Resources in these linked metadata universes can be discovered by browsing and/or searching. Once found, a suitable resource can be directly manipulated. Users can themselves configure which tools they want to be able to start immediately from the browser on such resources.

To aid the specialists, the IMDI set offers more detailed description possibilities than the DC or OLAC set. Some see this as a disadvantage since more overhead is involved in the creation process. IMDI solves this by making only few elements mandatory and by offering efficiency tools which allow the modification of whole sub-trees of metadata descriptions with one command. In doing so it offers a kind of spreadsheet functionality.

With respect to design, IMDI and MPEG7 started similarly –in that they both investigated the needs of the communities and followed a bottom-up design. OLAC started from the perspective that its set should be very close to the DC set. Both approaches have their pros and cons. While the mapping from the OLAC set to DC is simple, the other two mappings are not apparent and information is lost. Nevertheless, MPEG7 and IMDI have declared to support the OAI (Open Archives Initiative) type of harvesting protocol [17]. This will allow the casual web user to also run his queries across the MPEG7 and IMDI descriptions - of course with limited detail.

5. Tools for Metadata Operations

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.

⁵ Searching and Browsing

⁶ Searching, Browsing and Filtering

- A number of scripts allowing to work efficiently

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

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.

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 thus facilitating easy navigation.

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.

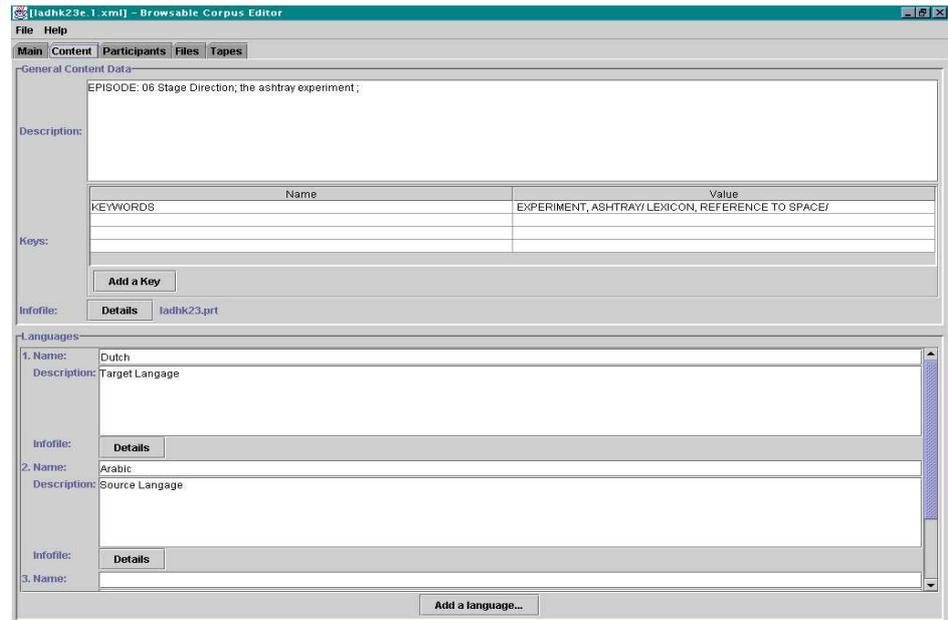


Figure 1 shows a screenshot from the IMDI Editor

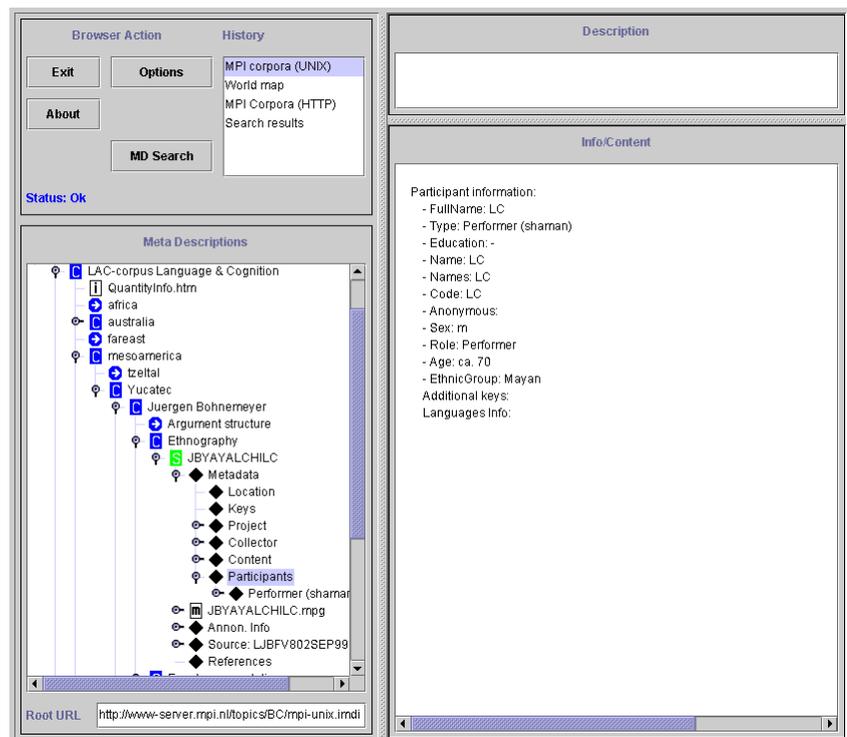


Figure 2 shows a screenshot of the IMDI browser.

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.

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 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.

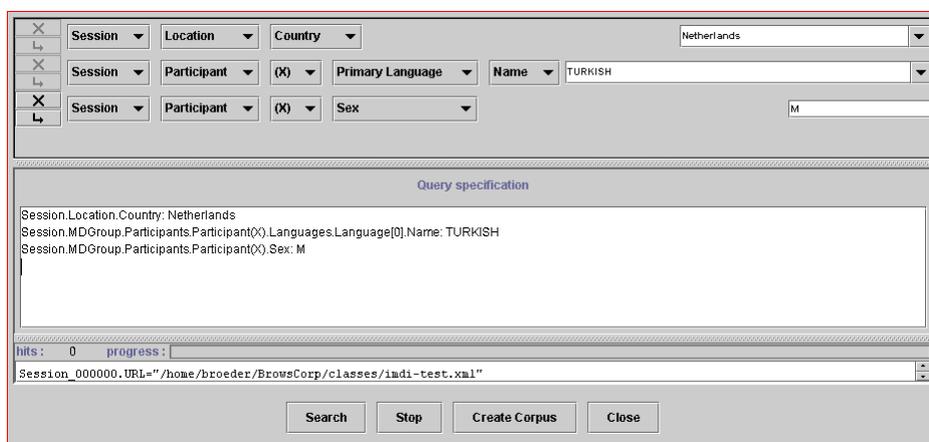


Figure 3 shows a screenshot from the search component.

The IMDI team also created a number of scripts which allow users to work efficiently with IMDI type of metadata descriptions. One such tool is provided to add or change element values in a whole range of MD descriptions using 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.

The next step is to support the lexicon metadata specifications in the editor, which involves the creation of a sub-block for lexicon specific elements. The browser extension will not create large problems since this tool is programmed so that it can easily adapt to different schemas.

6. Future Perspectives

Although unified metadata searchable and browsable in the Web is a fairly new concept, we already have an excellent insight into the requirements for the future.

First we need to convince researchers and developers in universities and industry to participate and to create metadata domains with a critical mass of data so that all the many and valuable resources existing on the disks and tapes will become viewable for the whole domain or at least their metadata. In Europe a first step in this direction has been taken by setting up the INTERA project (Integrated European Resource Area). Its goal is to have various data centers working together to create an IMDI-based metadata domain and to integrate language resource and tool repositories.

The interoperability between different domains has already been addressed by the OAI. We expect even more different metadata initiatives to emerge where both approaches will be perceived: a bottom-up approach driven by the requirements of the discipline and top-down approaches starting from compatibility aspects with DublinCore.

Driven by the needs of the Semantic Web [18] and the requirements of re-usage of

already existing definitions of data categories and structural definitions we notice that there is a great trend towards uniformity at a higher level. The emergence of terminology repositories will create highly reliable namespaces where metadata elements and controlled vocabulary will be specified in a standard and machine readable fashion. This will allow people to use and integrate them into their

definitions. Further, RDF (Resource Description Framework) [19] allows us to define structures and therefore semantic relationships between existing data categories to create new more complex ones. Also these RDF schemas will be available in open repositories enabling their re-use by other initiatives.

Another related development will hopefully enforce international standardization. The ISO organization has set up a sub-committee devoted to the standardization of terminology in language resource management (TC 37/SC4). Metadata will play an important role here.

7. References

- [1] *Dublin Core*: <http://dublincore.org/>

- [2] OLAC Initiative: <http://www.language-archives.org/>
- [3] 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.
- [4] IMDI: <http://www.mpi.nl/world/ISLE/index.html>
- [5] MPEG7: <http://mpeg.csel.it>
- [7] Documentation of Endangered Languages:
<http://www.mpi.nl/DOBES/>
- [8] TEI: <http://www.tei-c.org/>
- [9] EAGLES:
<http://www.ilc.pi.cnr.it/EAGLES96/home.html>;
ISLE:
http://lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
- [10] OLIF: the Open Lexicon Interchange Format
(<http://www.olif.net/>)
- [11] MARTIF; <http://www.ttt.org>
- [12] SALT: as Standards-based Access service to multilingual Lexicons and Terminologies
(<http://www.ttt.org/salt/index.html>)
- [13] Wittenburg, P. (2001) *Lexical Structures*. DOBES internal document. MPI Nijmegen
- [14] ISLE workshop on the multilingual lexical entry, Pisa, April 2001.
- [15] Peters, W. (2001) *Metadata for Lexical Resources*. ISLE CLWG Meeting, Pisa
- [16] Gibbon, D. (2001) *Notes on Lexicon Metadata*. ISLE CLWG Meeting, Pisa
- [17] OAI: <http://www.openarchives.org/>
- [18] Semantic Web: <http://www.semanticweb.org/>
- [19] RDF: <http://www.w3.org/RDF/>