Project ref. no.	IST-1999-10647
Project title	ISLE MetaData Initiative

Deliverable status	Public
Contractual date of delivery	January 2002
Actual date of delivery	30.8.2002
Deliverable number	D10.2
Deliverable title	EAGLES / ISLE Metadata Set for Multimedia/Multimodal Language Resources
Type	Report
Status & version	Final Report - Version 4
Number of pages	174
WP contributing to the deliverable	WP10
WP / Task responsible	Peter Wittenburg
Author(s)	Daan Broeder, Freddy Offenga, Don Willems, Peter Wittenburg
EC Project Officer	Brian Macklin
Keywords	Metadata, language resources, language engineering
Abstract (for dissemination)	The ISLE Metadata Initiative worked out metadata sets to characterize major multimedia/multimodal language resources, which will facilitate their management and retrieval. The metadata sets were designed to make them useful for professional users in language engineering and linguistics. Therefore, their structure, elements and vocabularies are much more exactly defined and more adequate for language resources than for example the Dublin Core set that was made to describe all sorts of web-documents and that is directed towards the occasional web user.

ISLE Metadata Initiative (IMDI)

Final Report[1]

Deliverable 10.2

Metadata Sets

August, 2002

IMDI[2] Technical Report

Max-Planck-Institute for Psycholinguistics
NL, Nijmegen

[1] Since the ISLE Metadata Initiative started its work in September 2000 the final report D10.2 is delivered in August 2002. Due to the delay of the INTERA project which will build on IMDI Standards the results of a workshop about experiences with the IMDI sets and possible new requirements will not be part of this document, but part of the INTERA project funded by the EC as well.

[2] For information about the ISLE Metadata Initiative, please look at the following web-site: www.mpi.nl/ISLE

INDEX

1 INTRODUCTION 4

2 IMDI SESSION SET 8

2.1 ELEMENTS 10
2.2 CONTROLLED VOCABULARIES 11
2.3 CONSTRAINTS 12
2.4 XML SCHEMA 12

3. IMDI CATALOGUE SET 12

4. IMDI LEXICON SET 13

5. IMDI-DC-MAPPING 14

6. IMDI DOCUMENTS 15

7. COMPARISON 15

8. CONCLUSIONS 16

9. REFERENCES 18

10. ACTIVITIES 19

Appendix 1: Metadata Elements for Session Description 25
Appendix 2: Metadata Elements for Catalogue Descriptions 72
Appendix 3: Metadata Elements for Lexicon Descriptions 87
Appendix 4: Mapping IMDI Session Descriptions with OLAC 101
Appendix 5: Vocabulary Taxonomy and Structure 119
Appendix 6: XML Schema 125
Appendix 7: IMDI Flyer 147
Appendix 8: Paper presented at a LREC Preconference Workshop 150
Appendix 9: Paper presented at the ISO TC37/SC4 Foundation Meeting 165

1 Introduction

General

Language Resources (LR) are collections of data representing examples of language being used, either directly, as in corpora, or as derived data, as for example in lexicons and ontologies. Fundamental and applied linguistic research has a long history of generating and using text-based language resources. More recently multi-media/multimodal language resources are being exploited in linguistics and related areas such as sign language, anthropology, computer linguistics, artificial intelligence, phonetics, psychology, speech recognition, multi-modal research and man-machine interface design. They are used for a variety of purposes. Linguists use them to create and test new linguistic hypotheses; speech recognition engineers use them to test speech recognition devices and to set recognition parameters. Increasing amounts of money are being spent on the creation of new language resources and extending language resources to combine a variety of inputs (sound, video, eye tracking, …) and to incorporate multi-modal annotations.

People have always referred to such language resources in terms of basic global characteristics such as "the resource includes speech by a 6 year old male Tamil speaker born in a farming environment" or "this resource includes pointing gestures and speech utterances recorded when people were asked for directions to the railway station". We call this kind of data, which briefly summarizes or characterizes the content of the resource, its metadata-description. Usually language resources include this data in "file headers" that are part of the resource itself or it is put in separate files that have a corpus-specific format. Every project defined its own header structure and content, appropriate to the goals of the project. Special tools or simple ASCII editors were used to display and/or access the metadata.

The development of the Internet and World Wide Web provides the possibility of online language resources, but to be able to exploit this we need a unified way of organizing, structuring and advertising them. We envisage a universe of linked metadata-descriptions offering the interested community information about existing language resources. This universe should be accessible via the Internet with appropriate tools for browsing and searching. Such a system could save researchers and industry a lot of time in discovering the appropriate LRs.

We are sure that the general addition of Metadata-Descriptions to Web-accessible data will eventually revolutionize the way the Internet is used. The reasons for this are two-fold: (1) The increasing amount of available information makes it often impossible to find the information one is looking for incidentally. (2) Structured resources such as language resources cannot be searched for by search engines picking up terms from the text and applying some statistics. Structured metadata descriptions will provide especially the professional user with much better methods to find the resources he is looking for. Thus the language resource community should test these new mechanisms at an early stage, both to confirm that the mechanisms and procedures now becoming available are capable of providing the services required by the language resources community, and to select the specific mechanisms and procedures which best match the needs of the community.

To enable exploitation of corpora by browsing in addition to direct search, IMDI makes it possible to link metadata descriptions into browsable hierarchical structures. This requires the development of a standard for the structure and semantics of these metadata-descriptions. But language resources vary considerably and there is a heterogeneous user community, so we have to ask ourselves whether the requirements for handling such a variety of applications can be captured within a single standard. Therefore, the IMDI (ISLE Metadata Initiative) initiative has investigated which kind of metadata descriptions were used until now. This work was reported in D10.1.

Recent developments (OLAC, MPEG7, METS, ISO TC37/SC4) support us in our initial assumptions about the need of metadata infrastructures for the domain of multimedia language resources. Since this report is not meant to compare the different approaches in detail, we would like to refer especially to two papers. The first was recently presented at an international LREC pre-conference workshop (see Appendix 8). It compares the different approaches that in general appeared later than the IMDI proposal and it discusses how metadata infrastructures are likely to develop due to the requirements of the Semantic Web. The second paper was presented at the foundation meeting of the ISO TC37/SC4 committee (see Appendix 9) and largely accepted as the basis of metadata related work. It also presents ideas how metadata sets will develop in future.

Goals

Based on the overview and evaluation the goal of IMDI was to derive a metadata framework for multimedia/multimodal language resources which has to serve three major functions:

- easy and fast discovery of interesting and relevant resources;
- fast inspection of parameters to check whether a resource is appropriate;
- description of resources with also human readable prose text to assist browsing and exploitation of corpora by non-experts.

The first function can be used in formal queries since they describe exactly the wishes of the user such as the language spoken by the consultants. Another type of elements is not so much oriented towards queries but is useful for quick inspection. If the user has found a suitable resource, but it is not directly available the user wants to know the telephone number of the contact person. No one will ever search on such a descriptor. The third point is aimed at providing information to users that are not acquainted with the corpus and want to familiarize themselves with its structure and content by browsing.

The IMDI metadata framework is built on 2 pillars: (1) a metadata set definition suitable for usage by language domain professionals and (2) a showcase including tools and data. The work to develop a metadata set is described in this report (D10.2). The showcase work will be described in D10.3.

Scope

Multimedia/multimodal language resources are used in many disciplines as indicated above. This ranges from rather complex corpora as they are created, for example, in cognitive anthropology to more simple corpora for a specific industrial purpose where the annotations are often determined by the need to solve one particular problem. Further, there are multimedia language resources in the multi-media industry. However, here the requirements for metadata descriptions are very different, since the media community also wants to code process information (how a certain multi-media production was created, which components were combined etc) and they want to include extensive low-level information about the segments (cuts, camera switching, color index changes, etc). The IMDI project has to define a scope that is in line with the ISLE project. It was decided that IMDI has to limit itself to the needs of the community that is made up by the "linguistics" discipline as the core one and its closely related disciplines. It also focuses on multimedia/multimodal resources and not on complex written material since this is the domain of the catalogue systems of librarians. IMDI should address both researchers and engineers.

Another basic decision was that IMDI should develop primarily a metadata framework for multimedia/multimodal corpora and present a proposal for metadata for lexica (including multimedia/multimodal extensions). Other data types such as for example grammatical and phonetic descriptions should either be integrated into a browsable and searchable metadata domain setup for corpora and lexica or should be described with the help of simple reduced metadata fields. Until now these linguistic data types are not very well described and their structure and content is very heterogeneous.

Formal Basis and Domain Ontology

It was decided very early that all definitions within IMDI should be done in an accepted formal way following the practice set forward by the Dublin Core initiative [1]. The metadata sets and the necessary structure specs were defined in an XML Schema[3]. Also all definitions and constraints were integrated in the schema to support an operational version that serves not only to validate created metadata files but can also be used for tool support.

From the start the developers of IMDI had in mind that creating metadata descriptions is a first step towards a general ontology of the language resource domain. This was one of the main reasons to not adopt from the beginning the metadata vocabulary used to describe general web-resources: the minimal and flat Dublin Core Metadata Element Set (DCMES). This vocabulary was designed basically by librarians to describe all sorts of authored resources. It does not reflect the special aspects of specific other domains and certainly not those of the language resource domain. Therefore it does not have the potential to fulfill the needs of professionals from the language resource domain. Dublin Core cannot be seen as a major contribution to defining an ontology of language resources, since it did not start out with the needs of the language resources domain in mind. More detailed arguments can be found in appendix 8.

Language resources can appear as many different data types such as annotated multimedia resources, dictionaries, wordlists, encyclopedia, ontologies, grammar descriptions, descriptions of phonemes of languages, typology databases, NLP tools and many others. Most of these data types cannot yet be described very well in terms of an underlying and agreed system. This makes it difficult if not impossible at this moment to describe all these types with the help of metadata categories. Therefore, it was decided to follow a heuristic approach and to first focus on corpora with annotated multimedia resources and lexica. For these two data types the domain specialists have started making generalizations (see the Abstract Corpus Model [2], the Annotation Graph Model [3] and various concepts for abstract lexicon descriptions [4]). Therefore, IMDI was able to make statements about categories which can describe such resources in detail. We foresee that in the future we will have a modular framework where groups of metadata elements re-occur in descriptions for many types of language resources (such as IPR statements or addresses). These groups can be flexibly assembled to describe new types of resources. All together this emerging network of metadata element groups will emerge to a contribution to a general Ontology of the Language Resource Domain.

This vision of the IMDI set reusing other sets and being reused by others implicates that metadata descriptions will have to be formalized with the help of frameworks like RDF (Resource Description Framework) [5]. In addition to an XML Schema an RDF Schema would allow the developer to define semantics in an accepted and machine readable fashion that also allows reuse. This also would form a step in the direction of the Semantic Web where agents can process the definitions and relations seamlessly to the user. This, of course, presupposes that all schemas will be available through open repositories. The RDF Schema for the IMDI metadata sets has yet to be developed in the realm of the follow-up project called INTERA (INtegrated European language Resource Area).

Interoperability

Of course, interoperability of metadata sets between different domains and within a domain is necessary, i.e. if a user wants to look for documents which were created for example by a certain author the search engine should allow him to search for resources using the Dublin Core set as well as using the IMDI set. Therefore the IMDI team has created a mapping description (see chapter 5) that maps IMDI elements to OLAC[4] [6] elements. Due to its top-down approach the

[3] see appendix 6

[4] Open Language Archives Community - an initiative from SIL and LDC at Philadelphia. OLAC proposed a set that is based on Dublin Core, but incorporating increasingly more qualifications and sub-elements. Some OLAC specialists expect themselves a merging between IMDI and OLAC in the long run.

mapping from OLAC to DC elements is comparatively simple. This mapping supports that services that use the Open Archives Initiative protocol to harvest metadata records will be able to locate IMDI metadata records although only queries of limited (DC) complexity can be answered. This approach assures that on the one hand the professional users can search in sufficient detail what they are looking for using services that work directly with the IMDI set and that on the other hand the general user can search through the IMDI domain as well using general services. The media community also chose this approach when the MPEG7 [7] standard was defined. Dealing with multimedia resources as IMDI does, MPEG7 clearly stated that Dublin Core is not useful, because it is only designed for simple web objects. They also provided a very restrictive mapping to achieve interoperability with DC.

It should be added here that during meetings at the LREC conference the OLAC and IMDI specialists agreed to open up their repositories such that it would be possible that people using the IMDI search environment can access the OLAC metadata records and vice versa.

Formal Definitions
All IMDI components such as metadata element sets, constraints and controlled vocabularies are described by formal definitions. The most important part of IMDI is the documented definitions (see Appendix 1), since these include the formal agreement of the meaning of the metadata elements. An XML Schema (see Appendix 6) is used for the implementation of the formal definitions.

Schedule
Officially the ISLE project was started in January 2000 in the 5th Framework Program in the Information Science Technology directorate of the EC. One part of the ISLE project (workpackage 10) was dedicated to the creation of a metadata environment for multimedia/multimodal language resources. In September 2000 the subcontracts were finally signed which allowed the ISLE Metadata Initiative (IMDI) to start the actual work. Although the basic work has been carried out, the original date of delivery (January 2002) could not be matched. A preliminary report was presented in February 2002. It was planned by the IMDI team to use the experiences of a year of full usage and the start of a follow up project (INTERA) to do a major revision of the set. Since the start of INTERA has been delayed twice due to administrational reasons the intended meetings cannot be scheduled before October, i.e. their results will not be part of this report.

Chapters
The IMDI initiative created a number of documents for different purposes.

- The document “IMDI Metadata Elements for Session Descriptions” describes multimedia/multi-modal corpora which are based on media recordings. The set is described in chapter 2
- The “IMDI Metadata Elements for Catalogue Descriptions” contains those additional elements which are necessary to describe published corpora. The set is described in chapter 3
- The “IMDI Metadata Elements for Lexicon Descriptions” should be used to describe a selected number of lexicon types such as wordlists, concordances, dictionaries. It is described in chapter 4
- The mapping between the IMDI and OLAC/DC sets to achieve interoperability is described in chapter 5
- Other IMDI documents of relevance are described in chapter 6

Further this document contains a comparison between the different metadata sets (chapter 7), draws some conclusions (chapter 8) and includes a list of activities (chapter 9).

2 IMDI Session Set

In this chapter the essentials of the IMDI set for multimedia corpora will be explained. For details we refer to appendix 1. We distinguish between the structure of and the elements used in the IMDI set which are the dimensions along which the resources are described, the element value vocabularies which are the values the dimensions can take in the case of a predefined selection and the constraints which enforce certain restrictions on the values of elements.

Session Concept

First the term "session" has to be explained. Sessions are the leaves in a corpus tree, i.e. they refer to a single unit of linguistic analysis or linguistic action. Often a single recording session where for example one interview or experiment is made with one subject can be called a "session". In multimedia corpora in general one has an assembly of tightly related tracks such as video tracks (one or more cameras), sound tracks (one or more microphones), signals for example from eye trackers, and several tracks of annotations which can cover very different types of information (transcription, morphosyntax, gesture encoding, ...). All these tracks have a close relationship since they are coupled not only to the same event, but also to exactly the same time line. Therefore, the session concept in IMDI models this relationship by implicit structure.

The IMDI set for Sessions can be used to describe the whole bundle of information that is tightly linked in a session. It can also be used to describe higher order nodes in a corpus tree where these higher order nodes represent a certain layer of abstraction from the individual sessions e.g. all sessions with a female speaker. Therefore, a whole corpus can be described following the proposed metadata element set. The IMDI concept allows the interested community to define abstract nodes on levels which even abstract from the details of a single corpus, i.e. it can be used to form a world-wide browsable and searchable metadata domain of language resources.

Structure

In contrast to the Dublin Core set which is a flat list of elements[5], the IMDI set contains structure. Since the Dublin Core is now used to describe a great many types of resources it was a wise decision to simply provide a flat list, since structure would always imply a bias towards certain types of resources. The IMDI Session set, however, was developed for describing linguistic corpora; therefore it can and should be biased to facilitate the work of the creators of the language resources. Including structure can also improve the expressive power of the set as shown in the following two examples: It is often the case that sessions have two audio tracks - one is used to record the interviewer and the other uses a high quality microphone for the subject. Accordingly one may have two sets of annotation tracks - one for each person or different types of annotation dependent on the quality. Of course, it should be noted somewhere in the metadata which of the audio tracks is related to a certain annotation tier. This directly allows the speech engineer to select the right track for his operations. This relation between sound track and annotation track is implicitly stored in the IMDI set[6]. To formulate these complex relationships with the help of the flat Dublin Core set would require to explicitly formulating them that hardly will be done.

In what follows we provide another example that demonstrates the greater expressional power of the IMDI set. In the IMDI set it is no problem to have a number of participants and associate with each of them a number of attributes such as age or languages spoken. In a flat list this is not possible except by misusing the refinement concept that will change the semantics of the element and therefore will not be officially supported.

---

[5] Since recently, the architecture working group of the Dublin Core initiative came up with first suggestions about implementing DCMES within XML and RDF.

[6] Actually this extension was a requirement from professionals working with speech and multimodal resources.

Another example of inherent structure is that it is specified which elements can occur several times and which cannot. A participant can only have one age, while there can be several participants. There can be several sources, but each source can only have one unique identifier. Therefore, in IMDI it is constrained which elements can be used several times.

Flexibility

From the many discussions the IMDI team had with researchers from different sub domains, it was clear that projects and individuals might need flexibility to express aspects of their resources as categories that are not explicitly included in the proposed set. Partly, these needs require possibilities to integrate prose text and partly other user or project specific categories are required. The IMDI set accounts for both cases. At many layers in the IMDI set it is possible to enter descriptions that are just (formatted) text or URL links to files containing such text. Secondly, at various places in the IMDI set it is possible to add name-value pairs, i.e. the user defines himself a category, gives it a name and adds a value to it. Clever designed search engines can extend their search dimensions to such categories. It is the intention of the IMDI initiative to check regularly whether certain typical descriptors are used as new "names". This could indicate that the actual descriptor set has deficiencies.

Low-Overhead vs. High-Overhead

Metadata sets such as IMDI or MPEG7 have more elements than the Dublin Core set. Filling in all metadata elements especially for legacy resources that need to be converted to IMDI can become so time-consuming that people don't want to invest the necessary time. An argument sometimes heard is that the Dublin Core set is called a low-overhead set, while IMDI is called a high overhead set. This view is not correct. It is important to know that in IMDI actually just one element is obligatory to be filled in: the metadata description has to have a name and a unique URL. This means that it is for the user to decide how much he is willing or able to describe. Essentially it is a matter of tool support whether users can be efficient when creating metadata descriptions. This point will not be discussed here, but in deliverable D10.3.

Process Description

The current IMDI set is a result of many discussions. The basic ideas go back to an initiative at the Max-Planck-Institute to organize its immensely growing multimedia/multimodal corpus and other related initiatives such as DC and TEI. These were subject of the overview in deliverable D10.1. At the beginning of the ISLE project a workshop with international participation was organized at the LREC conference 2000 to present the basic goals, concepts and to get feedback. Once the basic elements turned out to become clear (March 01) a special IMDI workshop with international participation was organized to present and discuss the proposal. Further, many presentations were given at conferences and workshops to get feedback from various communities. Moreover, the IMDI team participated in ISLE, DC and SALT [8] workshops that focused on related topics. Further meetings were organized with specialists in building up archives with multimedia/multimodal content in language engineering.

Meetings with specialists who work with complex textual (not multimedia) resources showed that the IMDI set can also be used to describe that type of data. However, it cannot be overseen that some element labels need to be adapted and that a few more elements would be necessary to satisfy the needs of that domain.

Currently, the IMDI set is used actively or has been used as a test to describe metadata for a number of large corpora:

- Multimedia Corpus of the Max-Planck-Institute in Nijmegen (already now more than 7000 sessions), Nijmegen, NL [9]
- Second Learner Corpus of the European Science Foundation, Nijmegen, NL [10]
- Dutch Spoken Corpus project (where additional fields are used as extra name-value pairs), NL [11]

- The Childes Corpus (maintained by the CMU in Pittsburgh) in a test setup at MPI which is further tested in cooperation with CMU [12]
- DOBES project to document endangered languages, Nijmegen, NL [13]
- Dutch Speech Style Corpus, NL
- AILLA project, Austin, USA [14]

Further checks were done and test setups were realized for various corpora known from language engineering such as the TIMIT speech corpus [15] and the large multimodal SmartKom corpus gathered by DFKI (Saarbrücken) and BAS (Munich) [16].

Further, the IMDI set was applied by 6 European institutions on a limited scale to create the IMDI preliminary Showcase for the Opening Event of the European Year of the Languages in Lund in 2001:

- Max-Planck-Institute for Psycholinguistics, Nijmegen, NL
- Max-Planck-Institute for Evolutionary Anthropology, Leipzig, D
- University of Lund, Se
- University of Helsinki, Fi
- CNRS-Paris, F
- ELRA, F

The IMDI set is intended to become the basis for two great European projects that cover almost all European countries and will contain language resources from different disciplines. These projects together with input from the AILLA project will be the primary source for all modifications in the IMDI sets.

At every step the intended revisions were communicated with the Steering and Advisory Boards of the IMDI initiative. Feedback was taken seriously and integrated to achieve consensus.

A major revision is planned for 2002after an extensive usage of the set and after new projects have stated their requirements. It will improve the possibilities for describing textual resources and another attempt with international specialists will be made to improve the content description section. A first suggestion was already made at an LREC pre-conference workshop [17]. The process will remain the same: all intended revisions will be communicated with the committed specialists.

Current Version
The current version of the IMDI set of Metadata Elements for Session Descriptions is V2.6 from January 2002. It is explained in detail in appendix 1 which is also available on the IMDI web site: www.mpi.nl/ISLE. In this document also exact statements about the revision history are made.

2.1 Elements

The element set can be split into seven major dimensions:

Session

direct attributes	the session as a whole such as Date
Project Description	the project such as Project.Title
Collector Description	the collectors involved such as Collector.Name
Content Description	the content such as Content.Genre
Resource Description	the resources such as Resources.MediaFile
Participants Description	the participants involved such as Participant.Role
Reference Description	references to the resource such as publications

These dimensions can be seen as fairly orthogonal although collectors often are also participants. It was long debated in the community whether “Creator” would be a better alternative instead of

“Collector”. The term “Creator” has a legal connotation, i.e. all persons who participated in the creation process are creators. For language resources such a definition is not useful. One wants to distinguish between the persons who carry out for example the research and who are interviewed. Therefore the neutral term “Collector” was preferred in contrast to the “Participants”. Their role can be separated easily. Until now these major dimensions turned out to be very stable. Each of the dimensions includes a number of sub-dimensions or elements. The content is described in 5 sub-dimensions:

Content

CommunicationContext	what is the communicative context of the recording
Genre	type of genre
Task	which typical task such as Wizard of Oz experiment
Modalities	which modalities are involved in the recording
Languages	which languages are spoken

The content part turned out to be the most difficult part within the IMDI Session Set. The reason for this is that it is impossible to define orthogonal dimensions. While the elements Task, Modalities and Languages seem to be stable, the first two sub-dimensions can be expected to undergo a revision after some time of usage.

Also the Resources are subdivided into three major sub-dimensions:

Resources

MediaFiles	all information concerning the media description
AnnotationUnit	all information relevant to describe the annotations
Source	elements referring back to the originals

These dimensions turn out to be very stable. Also the elements with which the resources are described in detail turn out to be sufficient for easy discovery or quick inspection. The “source” element can be used for all types of sources, i.e. it can be an original tape that the element refers to or a book in case of a textual corpus.

For further details we refer to the definitions in appendix 1.

2.2 Controlled Vocabularies

Controlled vocabularies define the values an element can have. To achieve uniformity at the level of values of elements in non-metric spaces controlled vocabularies are of great relevance. They will prevent typo errors and limit the set of possible values. Therefore, controlled vocabularies form the basis for successful querying.

The IMDI set contains two different types of controlled vocabularies:

- fixed controlled vocabularies
- open controlled vocabularies

Fixed controlled vocabularies can be used in those cases where the designers know the possible or sufficient set of values. The continents of the world represent such a closed set of values. In some of these cases the XML Schema defines the fixed controlled vocabulary (CV). In other cases such as the country or language names we refer to external definitions, although the form of these external definitions are defined by an XML schema. Every CV definition is just an instantiation of this schema. It is important to note that if the CV would be defined in the IMDI Schema it cannot be changed without changing the schema. When the CV is defined by a reference to an external definition this definition can be changed without any such need. Some of these external definitions should be open repositories that can be relied upon such as the ISO country name lists or the SIL Ethnologue language codes. The tools have to support these controlled vocabularies so the user is forced to select one of the values. At the moment all IMDI CV definitions are available from http://www.mpi.nl/IMDI/Schema.

In many cases, however, we don't yet know what the sufficient set will be. The concept of a general metadata set for this domain is too new to make reliable statements. In these cases we integrate values that are used in the discipline, but the user is allowed to add and use other values as well. It is known that such a scheme will only work when the set of values offered is updated regularly to improve the user guidance.

In some cases (content description) it was argued that only a hierarchy would satisfy the user. This poses a problem that is a known problem from mathematics: a too narrow specified hierarchical system of categories tends to be over-specified, i.e. during coding and retrieval errors the amount of deviations will increase. The IMDI discussions led to a limited depth of structure (2 levels) and maximal guidance by offering at the very beginning a reasonable set of values.

In some cases such as country codes, it is known that even the Ethnologue list [17] does not cover all the approximately 7000 languages of the world. Therefore, it is necessary to have different repositories that can exist in parallel.

Closed controlled vocabularies are used for continents, countries, some aspects of content description, languages, sex of participants, type and quality of media file, quality of the source (medium). Open vocabularies are used for some aspects of content description, type and role of participants, format of media files, type and format of annotation units, and format of source (media).

The controlled vocabularies are described in detail in appendix 1. It is intended to investigate the usage of the elements in summer 2002 as well to adapt the current lists that are used in the case of the open controlled vocabularies. Moreover, the OLAC and IMDI specialists have agreed to define one XML schema to communicate controlled vocabularies between tools and repositories and to re-use existing ones as extensively as possible.

2.3 Constraints

Defining constraints is another option to increase uniformity and to improve searching procedures. Constraints are useful to guarantee that for example dates are specified in one uniform style and that no typo errors are made. Therefore, a number of fields are associated with that sort of formal constraints. They are specified in the XML Schema for the IMDI set and can therefore be used by tools.

2.4 XML Schema

An XML Schema describes the whole IMDI Session Metadata Set. It specifies the elements, their structural relations, the structure of controlled vocabularies and the constraints. The Schema is available on the IMDI web site so that everyone can use it and develop tools to process IMDI data records. For the details we refer to appendix 6.

3. IMDI Catalogue Set

During the preparation of the IMDI Session set it was admitted that there are a number of categories that may optimize the description of a "Published Corpus" as a whole. An example is the usage of a corpus such as the well-known TIMIT corpus. That sort of descriptors are mainly used in catalogues of language resource agencies that sell resources. It was argued that such categories should not overload the IMDI Session set, since categories like "Pricing" should not be associated with an individual recording or a set of recordings covering a certain participant.

The catalogue set is a comparatively small flat set that is largely comparable with the Dublin Core set. It includes the typical categories resource agencies need. It is also described by an XML Schema (in fact its definition is currently part of the IMDI Metadata Schema in appendix 6). The current version of the IMDI Catalogue Metadata Set is 2.1. It is described in detail in appendix 2.

4. IMDI Lexicon Set

As was decided in the Steering Committee meeting in May 2000 in Athens, the ISLE Metadata Initiative (IMDI) should make a proposal for a metadata set which is capable to describe lexical resources in sufficient detail to allow retrieval and quick inspection.

Lexica are very important language resources and in a certain sense complementary to the corpora. Corpora represent language as it is spoken, i.e. language is presented in its time dimension. Lexica focus around the words and affixes, i.e. the units in language that bare meaning. There are a great number of different types of resources centered on words or concepts such as wordlists, glossaries, concordances, dictionaries, terminology databases, thesauri, ontologies, encyclopedia and others. All these have different characteristics and may have to be described differently by metadata. Therefore, the IMDI initiative decided to limit the scope of its proposal to those lexicon types which have as main entry a lexical headword and probably rich descriptions of its linguistic characteristics mostly included in complex structures. So, concept oriented databases such as thesauri and ontologies were primarily not focused on. It will be checked in a follow-up project where lexical resources of different sort will be included.

For obvious reasons lexica have been the subject of standardization efforts for a long time. They are at the center of various automatic and manual operations such as syntactic parsing, information extraction and translation. Therefore, the IMDI initiative looked closely to such activities as EAGLES, ISLE/MILE, OLIF, MARTIF, SALT [18,19,20,21]. Especially at a MILE workshop in Pisa in 2001 the topic metadata was addressed and a number of experts represented their views. Based on the presentations of two experts (W. Peters, D. Gibbon) and the knowledge about lexica the first version of the metadata set was developed.

The basis for the design of the lexicon metadata set was the anticipation that the current metadata sets in the area of language resources are just the beginning of establishing a domain ontology. Corpora and lexica form the two most relevant data types, many other data types such as grammar descriptions, phoneme descriptions, typology descriptions and many others are formally not well defined yet. But it is clear that a domain ontology covering all the data types will re-use building blocks of metadata description definitions for characterizing the resources. Many elements such as Creator, Project, Languages and Access can be shared for many language resource types. An obvious choice for a language to be used for defining building blocks and their underlying relation is RDF (Resource Description Framework).

In this respect a metadata set was designed which re-uses many elements and substructures from the IMDI Session set, but which also describes the specific characteristics of lexica. Since the microstructure of lexica reveals a lot of detail that is dependent on the language included and which is often dependent on specific linguistic theories, it was clear to the designers that too much detail should not be included. It would map these differences into the metadata set and therefore decrease its usefulness for example when searching. An approach was chosen which covers the main linguistic categories: Headword Type, Orthography, Morphology, Morphosyntax, Syntax, Phonology, Semantics, Etymology, Usage and Frequency. These main categories are the dimensions along which people can describe the content of the lexicon. Each dimension therefore is associated with a controlled vocabulary that has to be extendable by the user.

The category “Syntax” currently has the following possible values: Complementation, Alternation, Modification, Shallow Parsing, Deep Parsing, Functional Parsing, Collocations, Typology. The semantic of such a description is as follows: If the value “Collocations” is set for the category “Syntax” it will tell the user that the lexicon contains information of that type. The metadata set cannot go into deep detail about how the structural inclusion of this information is done in the lexicon. It makes sense that a lexicon provider adds his schema definition, gives examples and/or provides a link to a document such that the user can carry out a quick inspection on the basis of metadata.

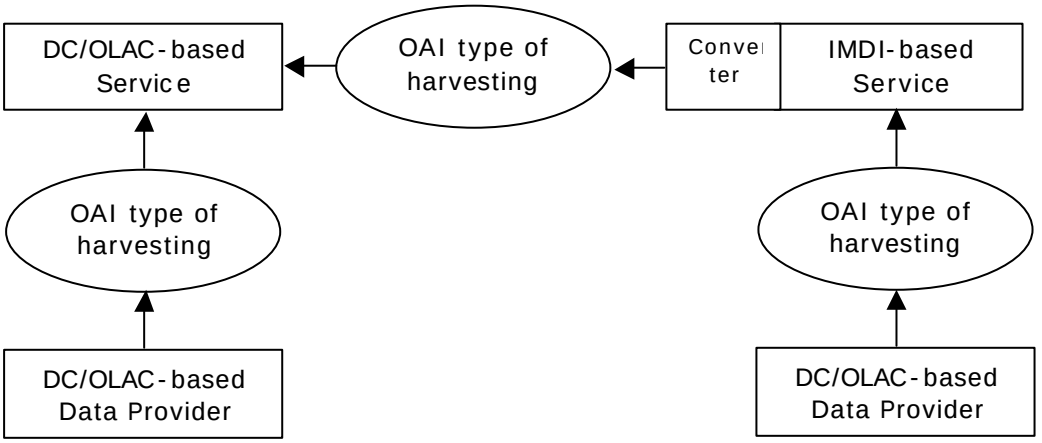
The proposal for a lexical metadata set was launched in December 2001. It is open for discussion. It was presented at the LREC conference [22]. We expect suggestions for modifications. For details we refer to appendix 3.

5. IMDI-DC-Mapping

The emergence of metadata was motivated by the enormous increase of web-based resources of all types. One of the first metadata initiatives was the one by librarians who developed the Dublin Core metadata set. DC should allow any user to describe each type of resource to improve professional retrieval. DC is widely accepted as the general set for web-based documents and it can be expected that services for the general public will rely on this set. Consequently, to facilitate the creation of metadata repositories that store inventories of different archives the Open Archives Initiative (OAI) [22] developed a light protocol that allows metadata harvesting where the default set is the DC one[7].

The OAI designers distinguish data providers, service providers and users. The data providers have repositories with metadata that are used by service providers to offer services. Users then use these services. The service provider must have access to the metadata records offered by the data providers. This is done via the OAI harvesting protocol. Of course, the service provider has to know what kind of metadata record is provided to be able to interpret the data transmitted. In the default case the DC set is assumed and should always be available next to other sets. A descriptor tells the service provider which other metadata sets are available. The service providers may use these alternative sets as they see fit but data providers are not obliged to provide them.

In December 2000 the OLAC (Open Language Archives Community) initiative, driven by SIL (Summer Institute of Linguistics) and LDC (Linguistic Data Consortium), launched its DC-based metadata set, the OLAC Metadata Set, for describing all kinds of linguistic resources. In fact it defined a number of refinements of the rather vaguely defined DC elements and adds two sub-elements: subject.language and type.linguistics. The idea is that some services based on the DC type of packaging can easily adapt to the OLAC set.



To be able to interoperate with OAI service providers and OLAC, the IMDI initiative has written a document “Mapping IMDI Session Descriptions with OLAC” which is the basis for its developments for a converter that converts IMDI records into DC/OLAC records. In doing so DC/OLAC based metadata harvesters could include IMDI descriptions into their search space. Of

[7] The OAI founders understood that many different metadata sets will emerge which will be more suitable for special communities. Therefore, the protocol allows transporting other types of metadata records as well. But any query to a DC-based service is limited to the power of DCMES of course.

course, since the IMDI set is much richer than the comparably poor DC set much information will be lost during the conversion step. For the details of the mapping we refer to appendix 4.

In the case of a mapping with the OLAC set and providing data for OLAC services, it is up to the OLAC service providers to offer all metadata records to full DC-based service providers. The IMDI initiative will provide both unqualified DC records (as OAI requires) and the initial OLAC extensions thereof.
Also the reverse case was studied, i.e. how can DC/OLAC repositories be included when searching from an IMDI service. Details can be viewed in appendix 4

The document has been discussed with the OLAC designers and dependent on the changes in the IMDI and OLAC sets it will have to be adapted.

6. IMDI Documents

The IMDI team created a number of other documents, for example, a flyer that was spread at various occasions. The flyer is included as appendix 7.

7. Comparison

There are a number of metadata sets that can in principle be used by members of the language resource community. This chapter is devoted to compare them and to give an overview about their main characteristics[8]. The following sets will be compared:

- Dublin Core — metadata set to describe all type of documents in the Web independent of their content and form
- IMDI — metadata sets to describe the characteristics of linguistic data types
- OLAC — metadata set to describe all type of language resources with a Dublin Core like set
- MPEG7[9] — metadata set developed by the media community

The Dublin Core Metadata Set was developed mainly by librarians in the expectation that all authored resources such as Web-documents can be described by this set. To achieve this generality the element set had to be small and general enough. A consequence was that the elements are in general vaguely defined and give some space for interpretation. On the other hand the DC documents require that the semantics of its elements may not be extended by its usage in communities. The DC set itself does not prescribe any structure, but a description may contain repetitions of the same element. A document can have several creators, therefore the element “DC:Creator” can occur several times. An architecture group was formed and is busy to make suggestions for implementing DCMES with XML and RDF.

The OLAC set was developed with the intention to stay closely to the Dublin Core set that makes mapping and therefore standard harvesting very simple. This is an advantage for the general user looking in the web for language resources. Since Dublin Core has only an element that allows the user to describe the language a document is written in, OLAC added a sub-element “Subject.Language” which covers the language a document is about. To be more suitable for the language resource community a number of other refinements were defined. For details we refer to the OLAC site (www.language-archives.org).

The IMDI sets (as the MPEG7 set) were designed by first investigating what the community has done before (see deliverable D10.1) and what the community requires in the Internet age. The

[8] As was already mentioned: more can be found in appendix 8.
[9] We still did not get access to the full MPEG7 specifications. So we have to rely on the openly available documents and the discussions with the MPEG7 experts.

outcome of a long discussion process guided by a steering board and an advisory board are the IMDI sets as described in this document. Ignoring the Dublin Core Metadata set at the beginning satisfies the needs of the community, but requires more efforts to achieve interoperability.

Topic	DC	OLAC	IMDI	MPEG7
approach	general	derivedfrom DC set	derived from requirements	derived from requirements
size	15 elements	15 elements	more elaborate	exhaustive
qualifiers/refinements	necessary to narrow down the broad element semantics	a number of refinements of DC for language resources	specific set for LR	
structure	flat - not worked out yet	flat	structured re-usable blocks	structured
bundling of coupled resources	relation concept	relation concept	implicitly possible	
mandatory	Few elements	few elements	few elements	?
re-usable blocks	no	no	yes	yes
differences for different data types	no	no	yes	yes
primary services	searching	searching	browsing and searching	browsing, searching, and filtering
DC mapping	-	simple	difficult - much information is lost	difficult - much information is lost
XML-Schema	available	available	available including constraints and controlled vocabularies	available
controlled vocabularies	no	some	several	?
retrieval	precision problematic	improved due to refinements	improved to comparatively exact definitions from the field and detail of categories	?
editor supporting constraints and controlled vocabularies	no	yes 1. version ready	yes	?
browser supporting constraints and controlled vocabularies	no	no	yes	?
search tool supporting constraints and controlled vocabularies	no	yes	yes	?

8. Conclusions

The management and discovery of language resources with metadata infrastructures that are openly accessible in the Internet is a fairly new concept. The ISLE metadata project was the first initiative to develop a full metadata infrastructure for multimedia/multimodal language resources when the White Paper was officially presented at an International Workshop at the LREC conference 2000 in Athens. Based on an overview about previous work with similar intentions (see Delivery 10.1) and many discussions with experts from the field of Language Engineering and Linguistics a metadata set for sessions which are the basic units of multimedia recordings was developed including tools that support it. Further, a metadata set for corpora was proposed that offers additional elements only useful for registered corpora. Finally, a metadata set for lexica

was proposed that needs further elaboration and discussion. It was decided that it is too early to define metadata sets for other linguistic data types. However, the infrastructure should include possibilities to integrate them.

The IMDI metadata sets are based on the specific nature of the language resources and on the wishes of the experts in the field such that easy management and discovery is made possible. This approach differs from the one taken by the Dublin Core initiative, since here it was the intention to develop a metadata set that can be used to describe all types of web resources and that is directed to the general, occasional web-user. From the overview and the discussions it was concluded that the Dublin Core set would not match with the needs of the community. At the LREC conference the IMDI team gave a broad overview about all that has been achieved in the two years of its activities within the ISLE project [23,…,30].

In the meantime other initiatives with closely related focus presented their metadata proposals:

- OLAC as a proposal to describe all types of language resources including software tools and "best practice recommendations" with a set derived from the flat Dublin Core set.
- MPEG7 as a very detailed proposal allowing the film and movie industry to describe their productions in great detail.

Also in the domain of education (IMS/LOM) and content management (IEC 82045-1) metadata sets were developed. While the OLAC initiative has made a number of agreements with data providers and has gathered a reasonable number of metadata records, the MPEG7 standard was only recently accepted and lacks support. Nevertheless, it can be expected that it will become an important set within the near future.

Basically, all initiatives accept the Dublin Core set as a minimal element set to be supported for the general web user searching for arbitrary documents. Therefore, also IMDI worked out a mapping to DC and offers converted IMDI records with much more restricted information to be harvested by the lightweight harvesting protocol proposed by the Open Archives Initiative.

Until now the IMDI set was used to describe more than 13.000 sessions which is the basic unit of linguistic analysis. These sessions can now be easily discovered. The IMDI metadata framework will be further used and extended and it will also be the basis for new projects to be started soon. The experience of more than one year of work with a fairly stable set are in general very good, the experts can express what they see as important especially since the IMDI set offers flexibility with respect to elements and controlled vocabularies. In the coming months after the INTERA and ECHO projects have started (both projects are funded by the EC) a workshop will discuss the experiences and the potential additional needs. This will lead to a major revision of the current IMDI sets.

Recently, in the framework of the new founded ISO TC37/SC4 future perspectives of metadata especially in the Semantic Web area were discussed. The IMDI work will be able to give valuable input as can be shown by the document that was accepted as the basic paper for the metadata activities [see appendix 9].

9. References

[1] Dublin Core Metadata Element Set. http://dublincore.org

[2] H. Brugman, P. Wittenburg. *The Application of annotation models for the construction of databases and tools*. Proceedings of the Workshop on Linguistic Databases. Philadelphia 2001.

[3] LDC Anno Graph

[4] P. Wittenburg, W. Peters, S. Drude. *Analysis of Lexical Structures from Field Linguistics and Language Engineering.* Proceedings of the LREC 2002 Conference. Las Palmas 2002.

[5] Resource Description Framework. http:/www.w3.org/RDF

[6] Open Language Archives Community. http:/www.language-archives.org

[7] MPEG7. http:/ipsi.fhg.de/delite/Projects/MPEG7

[8] Schema

[9] MPI Corpus: http:/www.mpi.nl/corpus

[10] ESF Second Learner Corpus: http://www.mpi.nl/world/data/esf_archive/html/

[11] Dutch Spoken Corpus: http://lands.let.kun.nl/cgn

[12] Childes: http://childes.psy.cmu.edu

[13] DOBES: http://www.mpi.nl/DOBES

[14] AILLA: http://www.ailla.org/pc/language.html

[15] TIMIT: http://www.ldc.upenn.edu/Catalog

[16] SmartKom: http://smartkom.dfki.de

[17] H. Johnson, A. Dwyer. *Customizing the IMDI Metadata Schema for Endangered Languages.* Proceedings of the International Workshop on Resources and Tools in Fieldlinguistics. Las Palmas 2002.

[18] EAGLES: http://www.ilc.pi.cnr.it/EAGLES96

[19] ISLE/MILE: http://www.ilc.pi.cnr.it/EAGLES96/isle

[20] OLIF: http://www.olif.net/olif2

[21] MARTIF: http://www.ttt.org/clsframe

[22] Open Archives Initiative: http://www.openarchives.org

[23] P. Wittenburg, W. Peters, D. Broeder. *Metadata Proposals for Corpora and Lexica*. Proceedings of the LREC 2002 Conference. Las Palmas 2002.

[24] D. Broeder, F. Offenga, D. Willems. *Metadata Tools Supporting Controlled Vocabulary Services.* Proceedings of the LREC 2002 Conference. Las Palmas 2002.

[25] P. Wittenburg, D. Broeder. *Management of Language Resources with Metadata.* Proceedings of the Workshop on International Standards of Terminology and Language Resources Management. Las Palmas 2002.

[26] P. Wittenburg, D. Broeder, F. Offenga, D. Willems: *Metadata Set and Tools for Multimedia/Multimodal Language Resources*. Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation. Las Palmas 2002.

[27] P. Wittenburg, U. Mosel, A. Dwyer. *Methods of Language Documentation in the DOBES Program.* Proceedings of the International Workshop on Resources and Tools in Fieldlinguistics. Las Palmas 2002.

[28] R. Skiba, H. Brugman, D. Broeder, P. Wittenburg. Corpus Organization and Access in Field Linguistics at the MPI. Proceedings of the International Workshop on Resources and Tools in Fieldlinguistics. Las Palmas 2002.

[29] R. Guirardello-Damian, R. Skiba. Trumai Corpus: an Example of Presenting Multi-Media Data in the IMDI-Browser. Proceedings of the International Workshop on Resources and Tools in Fieldlinguistics. Las Palmas 2002.

[30] D. Broeder, P. Wittenburg, T. Declerck. *LREP: A Language Repository Exchange Protocol.* Proceedings of the LREC 2002 Conference. Las Palmas 2002.

10. Activities

10. 1 List of activities from March until September 2000

1. Preparation and Organization of the 1st EAGLES/ISLE Workshops at the LREC Conference in Athens
   Three workshops about metadata, annotation systems, and tools were organized under the leading role of the MPI in cooperation with Univ. Sheffield, Univ. Odense, and Vassar College. The workshops attracted much interest and were very successful. The ISLE concepts and tasks were explained in detail.
2. Preparation of contributions to these Workshops and the conference
   MPI gave three talks in these workshops, two of them were dedicated to introduce the basic ISLE ideas.
3. Preparing summary statements of these workshops
   Summery statements were written to describe the essential outcome of the workshops.
4. Preparation and participation in the meeting of the Technical Board of the EAGLES/ISLE joint project
   During LREC conference a first meeting of the Technical Committee of the joint EAGLES/ISLE project was held. MPI participated and prepared part of the thematic areas.
5. Preparation of the White Paper for the EAGLES/ISLE MetaData Initiative (IMDI)
   A White Paper was produced by MPI which describes the goals of IMDI and the problems which have to be solved. This White Paper was widely accepted also at the LREC conference as basis.
6. Setup of the Web-Site for the EAGLES/ISLE MetaData Initiative
   A first web-site was designed and setup to prepare the LREC workshops and announce the programs and speakers. After the LREC workshop a new web-site was designed and setup which meets the requirements of the work in IMDI. This is operational since weeks and found much positive feedback. It will be extended continuously to give a broad perspective of the work in the area of metadata. The web-site contains three parts: an open site for the public, a site where only board members have access to, and a site which serves to administer the project at the MPI
7. Setup of an Organizational Framework for the EAGLES/ISLE MetaData Initiative
   A Steering Board was built which has to lead and control the work within the IMDI. It contains many important researchers from different countries in EUROPE and the US. It also covers people form different sub-communities. The Technical Board will be assembled dependent on the concrete questions. Currently it contains two persons with decent knowledge in XML technologies.
8. Establishing stable contacts with many important and relevant organizations, projects and initiatives in the area of metadata
   The Advisory Board of the project gives a good impression about the links the IMDI has established. Firm links have been established with other initiatives such as DublinCore and MPEG7 that deal about metadata. Also links have been established to the W3C as one of the driving forces of future activities in the Internet. Further, a number of links have been established to well-known language resource projects to achieve a broad foundation for the IMDI project. Discussions via the web have been started with several of those initiatives and institutions which all share our great interest in a metadata standard.
9. Organizing a first meeting of the Steering Board of the EAGLES/ISLE MetaData Initiative (IMDI)
   During the LREC conference a first meeting of the IMDI Steering Board was held to synchronize about basic issues of the project. The content of the White Paper was accepted and the organizational structure of the project was agreed upon. The minutes are on the project internal web-site.
10. Starting with an overview about earlier meta data (and header data) work in the IMDI

Work was started to present an overview about earlier metadata oriented work in the area of language resources and multimedia documents. Not much work has been done yet in this respect, although the need for metadata as structuring utility is large. The overview already contains the most relevant approaches and will be extended dependent on suggestions of the interested community. The main work has already been carried out. After some period an evaluation will be done which will serve as basis for the forthcoming work on a proposal for a standard. The overview is available on the web-site.

11. Starting with first concepts for an IMDI Metadata standard
    First brainstorming has been carried out to come up with such a standard a standard. Relevant papers especially from people in the DublinCore initiative have been studied. A list of relevant papers will be added to the web-site. This brainstorming work will have an impact on a first preliminary metadata "standard" to be used in a large project to document endangered languages with the help of multimedia language resources.

Most of the work is described at the web-site "www.mpi.nl/ISLE".

10. 2 List of activities until January 2001

1. Presentation of the IMDI Metadata concept at the Talkbank Technical Workshop. Pittsburgh, October.
   Brugman, H., Wittenburg, P. 'MPI Tools and Software Architectures'.
2. Presentation of the IMDI Metadata concept at the Talkbank Workshop on Web-based Language Documentation &Description. Philadelphia, December.
   Broeder, D., Suihkonen, P., Wittenburg, P. 'Developing a Standard for Meta-Descriptions of Multimedia Language Resources'.
3. Wittenburg and Brugman (MPI) organized the second DOBES Workshop about "Documenting Endangered Languages". Hannover, January. Here the IMDI Metadescription work was presented and discussed.
4. The Overview about Metadata Descriptions and Header Formats was completed and is now available as deliverable.
5. A revised IMDI set was worked out based on the various discussions and was sent to the interested community.
6. A new updated version of the Metadata Editor was produced and given to a number of European institutions (U Lund, U Helsinki, MPI Leipzig, Lacito Paris) to create metadata descriptions.
7. An updated version of the metadata browser was produced and given to the same institutes to view the MD descriptions they have created.
8. Currently an IMDI Workshop is prepared which will take place at March 1/2nd in Nijmegen. It is intended that this workshop will end up with a first Open IMDI proposal to the interested community. It will be added by a request for comments.
9. Currently, also a distributed demonstration of the IMDI concept is in preparation that includes corpora from the above mentioned institutions plus the ELRA catalogue data. It is intended to demonstrate this distributed metadata hierarchy at the opening of the European year of the language in Lund and at the IMDI workshop in March.

We refer to the IMDI website for further details: www.mpi.nl/ISLE.

10. 3 List of activities until July 2001

A. Presentations

1. Wittenburg and Brugman (MPI) organized the second DOBES Workshop about "Documenting Endangered Languages". Hannover, January. Here the IMDI Metadescription work was presented and discussed.

2. Wittenburg (MPI): Demonstration of the IMDI Metadata Initiative at the official opening ceremony of the European Year of the Language in Lund in February 2001
3. Presentation of the IMDI work to scientists working in large projects about Minority Languages, Lund, February
4. Presentation of the IMDI work at the Lund Linguistic Workshop, Lund, February. P. Wittenburg: Multimedia/Multimodal Corpora, Lund, February 2001
5. Organization of the 2-day EAGLES/ISLE Metadata Workshop in Nijmegen, March 2001
6. Several presentations of the IMDI work at the EAGLES/ISLE Metadata Workshop in Nijmegen by Broeder, Wittenburg, Brugman, Willems and Offenga
7. P. Wittenburg: Presentation of the IMDI work to BAS (Bayrisches Archiv für Sprache) and Smartcom project managers, Munich, March 2001
8. P. Wittenburg: Presentation of the IMDI work to corpus project leaders at ILSP, Athens, April 2001
9. P. Wittenburg: Presentation of the IMDI work at the ISLE-NIMM workshop, Stuttgart, April 2001
10. P. Wittenburg: Presentation of the IMDI work at the International E-Meld meeting, St. Barbara, June 2001

B. Work done for Workpackage 10
(most of the material is publicly available at the IMDI web-site - www.mpi.nl/ISLE)

1. Preparation of all presentations and demonstrations
2. Defining the first "frozen" IMDI Metadata set
3. Generating the definition paper: Metadata Elements for Session Descriptions Version 2.5 (which includes formal definitions of the elements and controlled vocabularies)
4. Generating a Reference Cart as flyer describing the IMDI set 2.5 in short
5. Generating a PR flyer for IMDI
6. Defining a metadata set for Catalogues
7. Creation of XML schemas for Session and Catalogue MD descriptions
8. Establishment of an IMDI Showcase for the Opening Ceremony of the European Year of the Language in Lund, this showcase covers metadata from 6 European institutions (ELRA, Lund U, Helsinki U, Max-Planck-Institute Leipzig, LACITO Paris, Max-Planck-Institute Nijmegen), has a world-map to select resources and allows to directly execute tools on retrieved resources.
9. First version of an IMDI-to-DC mapping document (to finally allow IMDI records being harvested by Service Providers following the Open Archives Initiative standards
10. Adapting the IMDI editor to the new IMDI set 2.5
11. Optimizing the IMDI editor to allow reusable blocks
12. Adapting the IMDI browser to the new IMDI set 2.5
13. Writing scripts to create IMDI MD descriptions from Excel sheet type of data
14. Designing the IMDI search functionality (together with IMS and ISSCO)
15. First discussions about Metadata descriptions for Lexica together with W. Peters and D. Gibbon

10. 4 List of activities until January 2002

A. Presentations

1. D. Broeder, P. Wittenburg, "Accessing Multimedia Language Resources the Easy Way", ACL/EACL2001 Workshop on Human Language Technology and Knowledge Management, Proceedings of the ACL/EAC Conference, Toulouse, July
2. D. Broeder, F. Offenga, D. Willems, P. Wittenburg, "The IMDI metadata set, its tools and accessible linguistic databases", Proceedings of the IRCS Workshop on Linguistic Databases, LDC, Philadelphia, December

3. D. Broeder, H. Brugman, P. Wittenburg, “Aspects of Modern Multi-Modal/Multi-Media Corpora Exploitation Environments”, Proceedings of the EUROSPEECH 2001 Conference, Aarhus, September
4. D. Broeder, “Exploring the Dutch Spoken Corpus”, CGN Workshop, Tilburg, OctoberDafydd Gibbon, W. Peters, P. Wittenburg, “Metadata for Lexica”, DOBES Workshop, Bielefeld, November
6. P. Wittenburg, “Corpus construction using Metadata Descriptions”, University Utrecht, January

B. Work done for Workpackage 10
(most of the material is publicly available at the IMDI web-site - www.mpi.nl/ISLE)

1. Definition of closed vocabularies for the IMDI set
2. Preparation of open repositories for the closed vocabularies
3. Definition of constraints for the IMDI set
4. Improve the XML Schema to include
    a. controlled vocabularies
    b. constraints
    c. catalogue elements
5. Create the final version of the IMDI to OLAC/DC mapping document
6. Develop a first version of an IMDI Metadata set for Lexica
7. Further development of the Editor[10]:
    d. integration of closed vocabularies
    e. integration of constraints
    f. code debugging
    g. user interface optimization
    h. make it downloadable via the webstart mechanism
8. Further development of the browser[11]:
    i. integration of closed vocabularies
    j. integration of constraints
    k. allow to add local tools via configuration files
    l. add a bookmarking feature
    m. implement the basket idea to collect various resources (for later operations)
    n. code debugging
    o. make it downloadable via the webstart mechanism
9. Development of a first version of a search tool operating on IMDI metadata descriptions
10. Integrate the Search tool in the browser
11. Search Workshop to exchange about search solutions implemented so far between ISSCO and MPI
12. Generating large amounts of metadata descriptions[12] for
    p. the famous CHILDES corpus
    q. the Dutch Spoken Corpus as the big reference corpus for Dutch and Flemish
    r. the MPI corpus covering more than 7000 Metadata descriptions referring to more than 2 TB of multimedia data including for example gesture recordings
13. writing a Perl script for batch mode changes in serious of metadata descriptions
14. forming an international group (including experts from the US and Australia) to work out a second version of the content description elements
15. Joining the ISO TC37/SC4 sub-committee and taking over the lead of the metadata group (Wittenburg is appointed as Working Group Leader)

---

[10] all tools are downloadable via the webstart mechanism from the MPI tools site
[11] same
[12] The whole metadata tree of the corpus will be made available to the public when final ethical and juridical issues have been solved. The development of the very heterogeneous corpus available at the MPI can be seen as a test case. It will be integrated to the final Showcase corpus.

10. 5 List of activities until August 2002

A. Presentations

1. P. Wittenburg, U. Mosel, A. Dwyer: *Methods of Language Documentation in the DOBES Program*. Proceedings of the LREC 2002 Conference. Las Palmas, May 2002
2. R. Skiba, H. Brugman, D. Broeder, P. Wittenburg: *Corpus Organization and Access in Field Linguistics at the MPI.* Proceedings of the LREC 2002 Conference. Las Palmas, May 2002
3. D. Broeder, F. Offenga, D. Willems: *Metadata Tools Supporting Controlled Vocabulary Services.* Proceedings of the LREC 2002 Conference. Las Palmas, May 2002
4. D. Broeder, P. Wittenburg, T. Declerck: *LREP: A Language Repository Exchange Protocol.* Proceedings of the LREC 2002 Conference. Las Palmas, May 2002
5. P. Wittenburg, W. Peters, D. Broeder: *Metadata Proposals for Corpora and Lexica.* Proceedings of the LREC 2002 Conference. Las Palmas, May 2002
6. P. Wittenburg, D. Broeder: *Metadata Overview and the Semantic Web.* Proceedings of the International Workshop on Resources and Tools in Field Linguistics. Las Palmas, May 2002.
7. H. Brugman, S. Levinson, R. Skiba, P. Wittenburg: *The DOBES Archive: Its Purpose and Implementation.* Proceedings of the International Workshop on Resources and Tools in Field Linguistics. Las Palmas, May 2002.
8. R. Guirardello-Damian, R. Skiba*: Trumai Corpus: an Example of Presenting Multi-Media Data in the IMDI-Browser.* Proceedings of the International Workshop on Resources and Tools in Field Linguistics. Las Palmas, May 2002.
9. P. Wittenburg: *Metadata - Future Perspectives and ISO Tasks.* ISO TC37/SC4 Foundation Meeting. Las Palmas, May 2002
10. P. Wittenburg, D. Broeder: *Management of Language Resources with Metadata.* Workshop on International Standards of Terminology and Language Resources Management. Las Palmas, May 2002
11. P. Wittenburg, D. Broeder, F. Offenga, D. Willems: *Metadata Set and Tools for Multimedia/Multimodal Language Resources.* Workshop on Multimodal Resources and Multimodal Systems Evaluation. Las Palmas, May 2002
12. D. Broeder, B. Hellwig: *Metadata Principles and Tools.* DOBES Workshop. Nijmegen, May 2002
13. R. Skiba, N. Brom: *Corpus Integration.* DOBES Workshop. Nijmegen, May 2002

B. Work done for Workpackage 10

(most of the material is publicly available at the IMDI web-site - www.mpi.nl/ISLE, the tools are available at the web-site: www.mpi.nl/tools)

16. Preparations of the follow-up EC project called INTERA to build up an Integrated European Resource Area
17. Preparations of the follow-up EC project called ECHO to build up a metadata infrastructure for Cultural Heritage
18. Preparations for the E-MELD meeting in Ann Arbor Michigan and discussions with OLAC specialists about the following agreements: (1) Unified schema for controlled vocabulary repositories, (2) Interaction about controlled vocabularies and (3) Integration of the OLAC and IMDI search domains.
19. Creation of a script which maps IMDI to OLAC descriptions
20. Implementing the OAI harvesting protocol
21. Preparing the ISO meeting and building an ISO Task Force to write a requirements document
22. Preparing the basic document for ISO TC37/SC4 WG1/WI2
23. Establish firm contacts with the following initiatives: MPEG7, IMS, ISO 82045
24. Discussions about the Content Description Elements and Vocabularies with International Experts
25. Further development of the Editor:
    a. code debugging and user interface optimization
    b. flexible handling of prefixes for efficient file linking
    c. developing a node construction component

    d. writing and improving the user manual
26. Further development of the Browser:
    a. code debugging and user interface optimization
    b. integration of the EUDICO/ELAN tool for immediate execution
    c. integration of metadata and content search for the Dutch Spoken Corpus[13]
    d. designing a general solution for an integrated search mechanism
    e. writing and improving the user manual
27. Further development of the Search Tool:
    a. Extend searching to distributed scenarios, i.e. now the search component offers a combination of local and distributed search
    b. Discussion with ISSCO and IMS about their database-based searching approach
28. Discussion of Open Source regulations to distribute the IMDI software under Open Source regulations
29. Web Demonstrator
    a. improving the existing metadata descriptions
    b. creating clickable worldmaps, linking them to nodes in the existing IMDI metadata tree and integrating them into the browser
    c. transferring the metadata descriptions created for the official opening ceremony of the European year of the languages to the current IMDI set versions and integrating the data into the demonstrator

C. Work done for Workpackages 8 and 9

1. checking the correctness of the reports
2. creating a demo file with multimodal annotations

[13] This was seen as a test for a more comprehensive approach to combine metadata and content search

ISLE Metadata Initiative (IMDI)

PART 1 A

Metadata Elements for Session Descriptions

NOTE:
some identified open issues are marked with green

Draft Proposal Version 2.7

June, 2002

INDEX

1 INTRODUCTION AND MOTIVATION 27

2 SESSION ELEMENTS OVERVIEW 29

2.1 SESSION SCHEMA 29
2.2 SUB-SCHEMAS 30

3 METADATA ELEMENT DEFINITIONS 31

3.1 SESSION 31
3.2 PROJECT 34
3.3 COLLECTOR 35
3.4 CONTENT 35
3.5 PARTICIPANTS 38
3.6 RESOURCES 41
3.7 REFERENCES 47
3.8 SUB-SCHEMAS 48

4 VOCABULARIES 52

4.1 LOCATION . CONTINENT 52
4.2 LOCATION . COUNTRY 52
4.3 CONTENT . COMMUNICATION CONTEXT 52
4.4 CONTENT . GENRE 54
4.5 CONTENT . TASK 57
4.6 CONTENT . MODALITIES 58
4.7 PARTICIPANT . TYPE 58
4.8 PARTICIPANT . ROLE 59
4.9 ANNOTATION UNIT . TYPE 59
4.10 ANNOTATION UNIT . FORMAT 60
4.11 SOURCE . FORMAT 60
4.12 LANGUAGE . NAME 62
5.1 LANGUAGE IDENTIFIER ENCODING 63
5.2 MEDIA POSITION ENCODING 64

6 REFERENCES 65

APPENDIX A : META TRANSCRIPT 66

A.1 META TRANSCRIPT 66

APPENDIX B : REVISION HISTORY 68

1 Introduction and motivation

This document for a schema of metadata elements is specifically directed towards describing multi-modal multimedia language corpora. There will be a separate schema for catalogue metadata which is used to describe a published corpus. We hope to extend the proposal in the near future with a special scheme for lexicons.

The community needs a more extensive set of metadata elements

We were guided by the desire to enable not only the resource discovery of major resources such as whole corpora but also be able to find individual resources from within corpora. For instance community members not only want to answer the question "find me all corpora with yaminjung speakers" but also "find me all sessions (recordings) with female yaminjung speakers younger then 60". To be able to answer questions like this we cannot use an existing general metadata scheme used for instance for library resource discovery such as Dublin-Core as it is currently defined. The community needs a more extensive set of metadata elements that captures the many needs of the different linguistic domains to easily find suitable resources.

Another guiding principle was the need to be able to browse the descriptions of language resources next to using them for automatic resource discovery. Although the two are similar, browsing capability requires "human readable" descriptions of (sub-) corpora and resources. Therefore you will find that the proposed set offers the possibility to specify these descriptions or link in (URL) references to other such "human-readable" descriptions at many levels.

Access to the metadata descriptions is always free

You will notice that the metadata transcriptions only contain references to real language resources such as audio/video files and transcriptions and annotations. All these references are accompanied by a structure specifying access restrictions for these resources. In our concept the access to metadata in the metadata transcriptions is always free although the metadata referring to individual persons may be rendered anonymous. The access to the resources themselves though may be restricted.

Flexibility for sub-communities to add their own descriptive elements

The possibility to have sub-communities add their own specific descriptions is approached in two ways. At different levels of the session description it is possible to add a list of keys in the form of name/value pairs. This possibility can be exploited by having sub-communities defining their own sets of required keys. Secondly the meta-description is characterised by a metadata description format identification. This identification will tell tools working with metadata descriptions what they can expect with respect to the structure of the metadata descriptions and the set of metadata elements used. The format identification could also be used to inform specifically tailored tools to look for specific extensions to the basic scheme and act accordingly. This functionality is closely connected to the way the metadata elements will be implemented and will pose extra requirements regarding this implementation. For the moment it seems

wise to avoid the matter of structure and implementation and concentrate on discussing the appropriateness and sufficiency of the proposed metadata element set for our purposes.

Editors reduce typing effort and allow re-use

The sheer number of proposed elements may let people believe that it is a heavy burden to have to supply all this information. It should be taken into account that in most projects the metadata descriptions for different sessions vary only in a few fields. The IMDI editors allow users to use existing metadata transcriptions to generate new ones. This will considerably reduce the amount of typing involved.

Only a few elements are mandatory

We need to say something on the set of metadata elements that should be minimally specified. Evidently not all the information that can be specified with the proposed set of metadata elements is always available. This is specifically the case for legacy resources or very specialistic resources. Therefore only those elements should be mandatory that are needed for the correct functioning of tools working with the metadata descriptions. For the session metadata only the session name is needed to distinguish between other sessions in the same corpus or sub-corpus.

Human readable descriptions can be added

At several places in the IMDI set there are keys (attribute name - value pairs) to extend the set with domain specific information. With the appropriate tools it will be possible to search for specific values of a named attribute. This will not be possible (or at least much more difficult) when the same information is entered in a description element, since the description elements are not structured. The description elements are more useful for human readable descriptions.

Although the content part of the IMDI set is currently more focussed on describing speech, discussion is going on about content descriptions of written language so that these descriptions can be integrated in the content group.

It should be noted that in this document only the metadata for sessions is described. Sessions can be grouped to form a corpus or sub-corpus. A corpus can contain sessions and sub-corpora.

2 Session Elements Overview

2.1 Session schema

- Session
  - Name (string) *
  - Title (string)
  - Date (c)
  - Location (group)
    - Continent (ccv)
    - Country (ccv)
    - Region + (string)
    - Address (string)
  - Description + (sub)
  - Keys (sub)
  - Project (group)
    - Name (string)
    - Title (string)
    - Id (string)
    - Contact (group)
    - Description + (group)
    - Keys (sub)
  - Collector (group)
    - Name (string)
    - Contact (sub)
    - Description + (sub)
  - Content (group)
    - Communication Context (group)
      - Interactivity (ccv)
      - Planning Type (ccv)
      - Involvement (ccv)
    - Genre (group)
      - Interactional (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)
      - Code (string)
      - Role (ov)
      - Language + (sub)
      - Ethnic Group (string)
      - Age (c)
      - Sex (ccv)
  - Resources (group)
    - Media File + (group)
      - Resource Link (c)
      - Size (string)
      - Type (ccv)
      - Format (ov)
      - Quality (ccv)
      - Recording Conditions (string)
      - Position (c)
      - Access (sub)
      - Description + (sub)
      - Keys (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 Id (ccv)
      - Anonymized (ccv)
      - Description + (sub)
      - Keys (sub)
    - Source +
      - Id (string)
      - Format (ov)
      - Quality (ccv)
      - Position (c)
      - Access (sub)
      - Description + (sub)
    - Anonymized (group)
      - Resource Link (c)
      - Access (sub)
  - References (group)
    - Description + (sub)

		Education (string)		
		Anonymized (ccv)		
		Description + (sub)		
		Keys (sub)		

2.2 Sub-schemas

Language			Access	
	Id (ccv)			Availability (string)
	Name + (string)			Description + (sub)
	Description + (sub)			Date (c)
				Owner (string)
Keys				Publisher (string)
	Key + (sub)			Contact (sub)
Key			Contact	
	Name = Value (string)			Name (string)
	Vocabulary Link (c)			Address (string)
				E-mail (c)
Description				Organisation (string)
	Text (string)			
	Language Id (ccv)			
	Info Link (c)			

Legend
***** the element is required.
+ indicates a list of one or more elements
string sequence of alphanumeric symbols including spaces and punctuation.
sub sub-schema
group grouping of elements
c the element is constrained by a certain encoding scheme
ccv closed controlled vocabulary - the content of the element must be selected from a closed set of values.
ov open vocabulary - the content of the element can be selected from a predefined set of suggested values or can be user defined. An ov can later be changed into a ccv provided by some repository
ovl open vocabulary list - a list of values for the content of the element can be selected from a predefined set of suggested values or can be user defined. An ov can later be changed into a ccv provided by some repository

3 Metadata Element Definitions

The elements for session descriptions are defined using the following attributes:

- *Element/Group Name*
  A name of the element or grouping.

- *Identifier*
  A unique identifier assigned to the element.

- *Definition*
  A statement that clearly represents the concept and essential nature of the data element.

- *Encoding*
  A statement that describes how the content of the element is encoded.

- *Comment*
  Remarks concerning the application of the data element.
  Dublin Core equivalent: some elements can be mapped with the Dublin Core Metadata Element Set [DCMES]. If this is possible, the Dublin Core equivalent of the IMDI element will be named here.[14]
  Example: sometimes an example helps to clarify the use of the element. If this is the case, the example will be mentioned here.

3.1 Session

Group:	Session
Identifier:	Session
Definition:	The session concept bundles all information about the circumstances and conditions of the linguistic event, groups the resources belonging to this linguistic event, records the administrative information of the event and describes the content of the event.
Encoding:	Session . Name Session . Title Session . Date Session . Location Session . Description + Session . Keys Project Collector Content Resources Participants References
Comments:	If an interviewer questions a consultant the resulting session description does not only contain the recording of that interview

[14] The mapping of IMDI elements to DC elements is done here in a simplified way. While IMDI elements are embedded in a structure, DC only describes a flat list of elements. The consequences of structure are ignored here to keep the mapping simple. More careful statements about IMDI - DC mapping will be made in a follow-up document.

but also the transcription and annotations and also for instance any photo images that were taken of this interview. It may well be that a researcher decides that one interview contains in fact more then one session if for instance the informant is asked to perform different tasks during that interview. This is all at the discretion of the researcher. The session is just a concept that can be used to create order when dealing with many linguistic resources. From a corpus and sub-corpus perspective the session description is any leaf in an arbitrary corpus tree hierarchy.

3.1.1 Session . Name

Element: Session . Name
Identifier: Session . Name
Definition: A short name to identify the session.
Encoding: string
Comments: The session name is typically a short name or abbreviation of one or two words. This identifier distinguishes the session from others in the same (sub-) corpus and is used for quick browsing. The name of the session can be considered a shorthand of the session title.
Example: Fatima 1

3.1.2 Session . Title

Element: Session . Title
Identifier: Session . Title
Definition: A full title for the session.
Encoding: string
Comments: The session title is the complete title of the session without any abbreviations.
Dublin Core equivalent: DC:Title
Example: Interview with Fatima, first session

3.1.3 Session . Date

Element: Session . Date
Identifier: Session . Date
Definition: The date when the primary data of the session was created.
Encoding: The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format. See 5.1.
Comments: In general the primary data of the session is audio or video data.
Dublin Core equivalent: DC:Date
Example: 2000-12-30

3.1.4 Session . Location

Group: Session . Location
Identifier: Session . Location
Definition: Groups the information about the location of where the session was_recorded or originated.
Encoding: Session . Continent
Session . Country
Session . Region +
Session . Address
Comments:

Session . Continent

Element: Session . Continent
Identifier: Session . Continent
Definition: The continent of where the session was recorded or originated.
Encoding: Closed controlled vocabulary 'Session . Continent' (4.1).
Comments:

Session . Country

Element: Session . Country
Identifier: Session . Country
Definition: The country where the session was recorded or originated.
Encoding: Closed controlled vocabulary 'Session . Country' (4.2).
Comments:

Session . Region

Element: Session . Region
Identifier: Session . Region
Definition: The region or sub-region of where the session was recorded or originated.
Encoding: string
Comments: This element can also be used to describe sub-regions.
Examples: europe, the netherlands, gelderland, achterhoek.

Session . Address

Element: Session . Address
Identifier: Session . Address
Definition: The address where the session was recorded or originated.
Encoding: string
Comments: For instance if recording sessions took place at an institution, the address of the institute is meant. There is no constraint on this element, since this element is only used for human inspection.

3.1.8 Session . Description

Element: Session . Description
Identifier: Session . Description
Definition: An elaborate description of the circumstances and conditions of the linguistic event.
Encoding: Description (sub-schema)
Comments: A description of the content is better specified at the level of the "Content . Description" element. Here a relevant description referring to the session as a whole can be given.
Example: A conversation of mother, father and child at the breakfast table.

3.1.9 Session . Keys

Element: Session . Keys
Identifier: Session . Keys
Definition: Name-value pairs to describe domain specific information about the session
Encoding: Keys (sub-schema)
Comments: Should be used to add name-value pairs which are important for searching domain specific attributes of session conditions which are not covered by the session level elements. While the

description elements are free text elements, keys are more formal notations which can also be exploited by search engines.
Example: length = 182

3.2 Project

Group: Project
Identifier: Project
Definition: Groups the information about the project for which the sessions were originally created.
Encoding: Project . Name
Project . Title
Project . Id
Project . Contact
Project . Description +
Project . Keys
Comments: If the session was made within the context of a project, the project element contains information regarding this project. This information is typically reused for many sessions and corpus leafs when they all belong to the same project.

3.2.1 Project . Name

Element: Project . Name
Identifier: Project . Name
Definition: A short name or abbreviation of the project.
Encoding: string
Comments: Example: MUMIS

3.2.2 Project . Title

Element: Project . Title
Identifier: Project . Title
Definition: The full title of the project.
Encoding: string
Comments: Dublin Core equivalent: DC:Title
Example: Multimedia Indexing and Searching

3.2.3 Project . Id

Element: Project . Id
Identifier: Project . Id
Definition: A unique identifier for the project.
Encoding: string
Comments: Dublin Core equivalent: DC:Identifier
Example: IST-1999-10651

3.2.4 Project . Contact

Element: Project . Contact
Identifier: Project . Contact
Definition: Contact information about the person or institution responsible for the project.
Encoding: Contact (sub-schema)
Comments:

3.2.5 Project . Description

Element: Project . Description

Identifier: Project . Description
Definition: An elaborate description of the scope and goals of the project.
Encoding: Description (sub-schema)
Comments: Dublin Core equivalent: DC:Description

3.2.6 Project . Keys

Element: Project . Keys
Identifier: Project . Keys
Definition: A list of name-value pairs used to describe project specific information.
Encoding: Keys (sub-schema)
Comments:

3.3 Collector

Group: Collector
Identifier: Collector
Definition: Groups information about the collector of the session.
Encoding: Collector . Name
Collector . Contact
Collector . Description +
Comments: The collector is the person which actually carried out the data collection. This has to be differentiated from the person or institution responsible for a whole project and from the person playing an active role in the recording. In some cases these persons are one and the same, but in general they are not the same.

3.3.1 Collector . Name

Element: Collector . Name
Identifier: Collector . Name
Definition: The name of the person responsible for the collection of the session data.
Encoding: string
Comments: Dublin Core equivalent: DC:Creator

3.3.2 Collector . Contact

Element: Collector . Contact
Identifier: Collector . Contact
Definition: The contact information about the person responsible for the collection of the session data.
Encoding: Contact (sub-schema)
Comments:

3.3.3 Collector . Description

Element: Collector . Description
Identifier: Collector . Description
Definition: Additional information about the person responsible for the collection of the session data.
Encoding: Description (sub-schema)
Comments:

3.4 Content

Group: Content

Identifier: Content
Definition: Groups information about the content of the session.
Encoding: Content . Communication Context
Content . Genre
Content . Task
Content . Modalities
Content . Languages
Content . Description +
Content . Keys
Comments: The content group is used to describe the content of the session. This is done using four dimensions (communication context, genre, task and modalities). The vocabularies and user entries in the different dimensions are not free of redundancy.
This group will be most heavily debated and IMDI is grateful for every suitable comment.

3.4.1 Content . Communication Context

Group: Content . Communication Context
Identifier: Content . CommunicationContext
Definition: Groups the linguistic features of the session concerning the context of the communication.
Encoding: Communication Context . Interactivity
Communication Context . Planning Type
Communication Context . Involvement
Comments: This group of elements is used to describe the communication context in which the recording took place.

Content . Communication Context . Interactivity

Element: Communication Context . Interactivity
Identifier: CommunicationContext . Interactivity
Definition: Characterizes the degree of interactivity between all the participants in the session.
Encoding: Closed controlled vocabulary 'Content . Communication Context' (4.3.1).
Comments:

Content . Communication Context . Planning Type

Element: Communication Context . Planning Type
Identifier: CommunicationContext . PlanningType
Definition: Indicates in how far the consultant planned the linguistic event.
Encoding: Closed controlled vocabulary 'Content . Communication Context' (4.3.2).
Comments:

Content . Communication Context . Involvement

Element: Communication Context . Involvement
Identifier: CommunicationContext . Involvement
Definition: Indicates in how far the researcher was involved in the linguistic event.
Encoding: Closed controlled vocabulary 'Content . Communication Context' (4.3.3).
Comments:

3.4.2 Content . Genre

Group: Content . Genre
Identifier: Content . Genre
Definition: Lists the conventionalized discourse types of the content of the session.
Encoding: Genre . Interactional
Genre . Discursive
Genre . Performance
Comments: Dublin Core equivalent: DC:Type.

Content . Genre . Interactional

Element: Genre . Interactional
Identifier: Genre . Interactional
Definition: Lists the interactional genre of the session content.
Encoding: Open vocabulary list 'Content . Genre' (4.4.1).
Comments:

Content . Genre . Discursive

Element: Genre . Discursive
Identifier: Genre . Discursive
Definition: Lists the discursive genre of the session content.
Encoding: Open vocabulary list 'Content . Genre' (4.4.2).
Comments:

Content . Genre . Performance

Element: Genre . Performance
Identifier: Genre . Performance
Definition: Lists the performance genre of the session content.
Encoding: Open vocabulary list 'Content . Genre' (4.4.3).
Comments:

3.4.3 Content . Task

Element: Content . Task
Identifier: Content . Task
Definition: The major task carried out in the session.
Encoding: Open vocabulary 'Content . Task' (4.5).
Comments: In areas such as language engineering often typical tasks are carried out or typical situations are dealt with such as "info kiosk task", "wizard of oz" experiment or "frog story". It has to be possible to specify such typical recurring tasks.

3.4.4 Content . Modalities

Element: Content . Modalities
Identifier: Content . Modalities
Definition: Gives a list of modalities used in the session.
Encoding: Open vocabulary 'Content . Modalities' (4.6).
Comments: The element is not used to give an exhaustive list of all the modalities, but should be used to list the modalities which are typical for the task or of interest for the researcher.
Example: in route direction one would typically look at speech and gestures and not at eye-gaze.

3.4.5 Content . Languages

Group: Content . Languages

Identifier: Content . Languages
Definition: Groups information about all the languages used in the session.
Encoding: Content . Languages . Language +
Content . Languages . Description +
Comments:

Content . Languages . Language +
Element: Content . Languages . Language +
Identifier: Content . Languages . Language +
Definition: A list of all the languages used in the session. The main language is the first language in the list.
Encoding: Language (sub-schema)
Comments: Each language used is described by a small sub-schema.

Content . Languages . Description
Element: Content . Languages . Description
Identifier: Content . Languages . Description
Definition: A description of the languages used in the session.
Encoding: Description (sub-schema)
Comments: Note that this description concerns the set of languages as a whole. Language specific descriptions are contained in the language sub-schema.

3.4.6 Content . Description
Element: Content . Description
Identifier: Content . Description
Definition: An elaborate description of the content of the session.
Encoding: Description (sub-schema)
Comments: In opposition to the elements prose text can be used here to describe the content.
Dublin Core equivalent: DC:Description

3.4.7 Content . Keys
Element: Content . Keys
Identifier: Content . Keys
Definition: A list of name-value pairs used to describe the domain specific characteristics of the content.
Encoding: Keys (sub-schema)
Comments: Name-value pairs can additionally be used to describe the content.

3.5 Participants
Group: Participants
Identifier: Participants
Definition: Groups information about all the participants in the session.
Encoding: Participants . Description +
Participant +
Comments:

3.5.1 Participants . Description
Element: Participants . Description
Identifier: Participants . Description

Definition: A description of the interactions and interrelations between the participants in the session.
Encoding: Description (sub-schema)
Comments: Note that this description concerns all participants and should be used to describe interactions and interrelations between participants. Information about specific participants should be described by the description sub-schema in the participant group.

3.5.2 Participant

Group: Participant
Identifier: Participant
Definition: Groups information about one participant in the session.
Encoding: Participant . Type
Participant . Name +
Participant . Full name
Participant . Code
Participant . Role
Participant . Language +
Participant . Ethnic group
Participant . Age
Participant . Sex
Participant . Education
Participant . Anonymized
Participant . Description +
Participant . Keys
Comments:

Participant . Type

Element: Participant . Type
Identifier: Participant . Type
Definition: The functional role of the participant.
Encoding: Open vocabulary 'Participant . Type' (4.7).
Comments: The type is meant as a rough categorization of participants such as; interviewer, consultant, contributor, computer etc.
This is in contrast to the role of a participant which is used for example to describe relations amongst the contributors.

Participant . Role

Element: Participant . Role
Identifier: Participant . Role
Definition: The role of the participant in the session.
Encoding: Open vocabulary 'Participant . Role' (4.8).
Comments: For instance when interviewing part of a family group, “Role” should specify the mutual relations within the group.

Participant . Name

Element: Participant . Name
Identifier: Participant . Name
Definition: The name of the participant as it is used by others in the transcription.
Encoding: string
Comments: This is the name of the participant which is used by others to identify him or her. Note that this is often not the same as the full name of the participant. This name can be blended out to

general users of the metadata to protect the identity. Blending out depends on the logical "anonymized" element.

Participant . Full name

Element: Participant . Full name
Identifier: Participant . Fullname
Definition: The full name of the participant
Encoding: string
Comments: This is the official name of the participant.

Participant . Code

Element: Participant . Code
Identifier: Participant . Code
Definition: Short unique code to identify the participant.
Encoding: string
Comments: Mostly the code is used in the transcription and annotations to identify parts belonging to this specific participant.

Participant . Language +

Element: Participant . Language +
Identifier: Participant . Language +
Definition: Lists the languages the participant is familiar with. The first language in the list is the participants first language.
Encoding: Language (sub-schema)
Comments:

Participant . Ethnic Group

Element: Participant . Ethnic Group
Identifier: Participant . EthnicGroup
Definition: The ethnic group of the participant
Encoding: string
Comments:

Participant . Age

Element: Participant . Age
Identifier: Participant . Age
Definition: The age of the participant
Encoding: The age is encoded as years;months.days from Codes for the Human Analysis of Transcripts [AGECHAT]. See 5.2.
Comments: Especially when children are acting as participants it is important to have detailed information.

Participant . Sex

Element: Participant . Sex
Identifier: Participant . Sex
Definition: The sex of the participant.
Encoding: Closed controlled vocabulary { Unknown, Male, Female, Unspecified }.
Comments: When the data about the sex of the participant is lost or simply not recorded, the sex 'Unknown' should be selected. In case of an artificial participant (a computer) 'Undefined' should be selected.

Participant . Education

Element: Participant . Education

Identifier: Participant . Education
Definition: The education of the participant.
Encoding: string
Comments: Can also be used to describe the literacy of the participant. Due to many expected differences this element is not constraint. Nevertheless, short keyword like indications are recommended.

Participant . Anonymized

Element: Participant . Anonymized
Identifier: Participant . Anonymized
Definition: Indicates whether or not the participant name and full name are replaced by pseudo names to make him/her anonymous.
Encoding: Closed controlled vocabulary { true, false }
Comments: If anonymized is set to 'True', the name and full name of the participant can only be obtained from the 'Anonymized' resource when access is granted.

Participant . Description

Element: Participant . Description
Identifier: Participant . Description
Definition: A description of specific information about the participant.
Encoding: Description (sub-schema)
Comments:

Participant . Keys

Element: Participant . Keys
Identifier: Participant . Keys
Definition: A list of name-value pairs to describe domain specific characteristics of the participant.
Encoding: Keys (sub-schema)
Comments: Sometimes elements are needed to describe specific characteristics of the participant depending on a certain research domain. The keys can be used for this purpose.

3.6 Resources

Group: Resources
Identifier: Resources
Definition: Groups information about all the resources associated with the session.
Encoding: Media File +
Annotation Unit +
Source +
Anonymized
Comments: In general there are three types of resources: original recordings, digitized media files and annotation files. In the future, direct digitization will occur more often.

3.6.1 Media File

Group: Media File
Identifier: MediaFile
Definition: Groups information about the media file.
Encoding: Media File . Resource Link

Media File . Size
Media File . Type
Media File . Format
Media File . Quality
Media File . Recording Conditions
Media File . Position
Media File . Access
Media File . Description +
Media File . Keys

Comments:

Media File . Resource Link

Element: Media File . Resource Link
Identifier: MediaFile . ResourceLink
Definition: A link to the media file.
Encoding: The link is encoded as an Uniform Resource Locator as described by [RFC1738]
Comments: Dublin Core equivalent: DC:Identifier.

Media File . Size

Element: Media File . Size
Identifier: MediaFile . Size
Definition: Human understandable specification of the size of the media file.
Encoding: string
Comments: The size of the media file is not meant to be machine processed. Normally the specification will be given in bytes.

Media File . Type

Element: Media File . Type
Identifier: MediaFile . Type
Definition: The type of the media file.
Encoding: Closed controlled vocabulary 'MediaFile . Type' (4.9). The media file type is encoded as a top-level media type from Multipurpose Internet Mail Extensions (MIME) as described in [RFC2046].
Comments:

Media File . Format

Element: Media File . Format
Identifier: MediaFile . Format
Definition: The format of the media file.
Encoding: Open vocabulary 'MediaFile . Format (4.10). The media file format is encoded as a media subtype from Multipurpose Internet Mail Extensions (MIME) as described in [RFC2046].
Comments:

Media File . Quality

Element: Media File . Quality
Identifier: MediaFile . Quality
Definition: An numeric indication of the quality of the media file.
Encoding: Closed controlled vocabulary { Unknown , 1, 2, 3, 4, 5, Unspecified }
Comments: It is suggested to describe the quality of the recordings with help of a number between 1 and 5 where 1 stands for low and 5 for high quality. It is known that this quality judgement is fairly

subjective and that there are large differences between various disciplines.

Media File . Recording Conditions

Element: Media File . Recording Conditions
Identifier: MediaFile . RecordingConditions
Definition: Describes the technical conditions under which the media file was recorded.
Encoding: string
Comments: Used to describe the equipment used for the recording (e.g. microphone type, amplifier type etc.). This element is not constrained and covers prose text. Nevertheless, short typical descriptions are recommended.

Media File . Position

Element: Media File . Position
Identifier: MediaFile . Position
Definition: The start- and end position of the session in the specified media file.
Encoding: See 'Media Position' (5.5)
Comments: It may occur that a session is just a fragment within the media file.

Media File . Access

Element: Media File . Access
Identifier: MediaFile . Access
Definition: Specifies the access rights of the media file.
Encoding: Access (sub-schema)
Comments:

Media File . Description

Element: Media File . Description
Identifier: MediaFile . Description
Definition: Gives a description of the media file.
Encoding: Description (sub-schema)
Comments:

3.4.7 Media File . Keys

Element: Media File . Keys
Identifier: Media File . Keys
Definition: A list of name-value pairs used to describe properties of the media file.
Encoding: Keys (sub-schema)
Comments:

3.6.2 Annotation Unit

Group: Annotation Unit
Identifier: AnnotationUnit
Definition: Groups information about the annotation unit.
Encoding: Annotation Unit . Resource Link
Annotation Unit . Media Resource Link
Annotation Unit . Annotator
Annotation Unit . Date
Annotation Unit . Type

Annotation Unit . Format
Annotation Unit . Content Encoding
Annotation Unit . Character Encoding
Annotation Unit . Access
Annotation Unit . Language
Annotation Unit . Anonymized
Annotation Unit . Description +
Annotation Unit . Keys

Comments: This group of elements describes all the characteristics of a specific annotation unit. Each unit refers to one layer of annotation or transcription and is independent of whether they are contained in one or more files.

Annotation Unit . Resource Link

Element: Annotation Unit . Resource Link
Identifier: AnnotationUnit . ResourceLink
Definition: A link to a file containing the corresponding annotations.
Encoding: The link is encoded as an Uniform Resource Locator as described by [RFC1738]
Comments: Dublin Core equivalent: DC:Identifier

Annotation Unit . Media Resource Link

Element: Annotation Unit . Media Resource Link
Identifier: AnnotationUnit . MediaResourceLink
Definition: A link to the media file from which the transcription originates.
Encoding: The link is encoded as a Uniform Resource Locator as described by [RFC1738]
Comments: Used to indicate which annotation unit belongs to which media file. For example, when there are two recordings with different microphones, there can be separate annotations for separate media files.

Annotation Unit . Annotator

Element: Annotation Unit . Annotator
Identifier: AnnotationUnit . Annotator
Definition: Name of the person who did the annotation or transcription.
Encoding: string
Comments: When there are more than one annotator, all the names of the annotators can be described by this element.

Annotation Unit . Date

Element: Annotation Unit . Date
Identifier: AnnotationUnit . Date
Definition: The date when the annotation unit was created.
Encoding: The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format. See 5.1.
Comments: Dublin Core equivalent: DC:Date

Annotation Unit . Type

Element: Annotation Unit . Type
Identifier: AnnotationUnit . Type
Definition: The type of the annotation unit.
Encoding: Open vocabulary 'Annotation Unit . Type' (4.11).

Comments: This element allows to specify the type of annotation such as phonetic, morphosyntax etc.

Annotation Unit . Format

Element: Annotation Unit . Format
Identifier: AnnotationUnit . Format
Definition: The file format which is used for the annotation.
Encoding: Open vocabulary 'Annotation Unit . Format' (4.12). Comments: E.g. such formats as CHAT and Shoebox are encoded as experimental (or private) MIME subtypes like; text/x-chat, text/x-shoebox.
Dublin Core equivalent: DC:Format

Annotation Unit . Content Encoding

Element: Annotation Unit . Content Encoding
Identifier: AnnotationUnit . ContentEncoding
Definition: Name of the encoding scheme used for the annotation purpose.
Encoding: string
Comments: Often is may be interesting to know whether for example morphosyntax was encoded following the "Eurotype" guidelines. In that case the element would have the value "Eurotype".

Annotation Unit . Character Encoding

Element: Annotation Unit . Character Encoding
Identifier: AnnotationUnit . CharacterEncoding
Definition: Name of the character encoding used in the annotation unit.
Encoding: The character encoding of the annotation unit is encoded as the charset parameter of the content-type from Multipurpose Internet Mail Extensions (MIME) as described in [RFC2046].
Comments: Example: UTF-8

Annotation Unit . Access

Element: Annotation Unit . Access
Identifier: AnnotationUnit . Access
Definition: Access rights of the annotation unit.
Encoding: Access (sub-schema)
Comments:

Annotation Unit . Language Id

Element: Annotation Unit . Language Id
Identifier: AnnotationUnit . LanguageId
Definition: The language used for the annotation unit.
Encoding: See 'Language Id' (5.4).
Comments: Here the language is meant which is used for the encoding. For and English transcription the value of this element should be "English".

Annotation Unit . Anonymized

Element: Annotation Unit . Anonymized
Identifier: AnnotationUnit . Anonymized
Definition: Specifies whether or not the names in the transcripts are replaced by pseudo names to make them anonymous.
Encoding: Closed controlled vocabulary { true / false }.
Comments:

Annotation Unit . Description

Element: Annotation Unit . Description
Identifier: AnnotationUnit . Description
Definition: Description of the annotation unit.
Encoding: Description (sub-schema)
Comments:

3.4.7 Annotation Unit . Keys

Element: Annotation Unit . Keys
Identifier: Annotation Unit . Keys
Definition: A list of name-value pairs used to describe properties of the annotation unit.
Encoding: Keys (sub-schema)
Comments:

3.6.3 Source

Group: Source
Identifier: Source
Definition: Groups information about the source.
Encoding: Source . Id
Source . Format
Source . Quality
Source . Position
Source . Access
Source . Description +
Comments: These elements are used to describe the original recordings. Often people want to have the reference to the original audio or video tape.

Source . Id

Element: Source . Id
Identifier: Source . Id
Definition: Short code to identify the source.
Encoding: string
Comments: Can be used to look up the source in an audio / video archive. Dublin Core equivalent: DC:Identifier

Source . Format

Element: Source . Format
Identifier: Source . Format
Definition: Physical storage format of the media.
Encoding: Open vocabulary 'Source . Format' (4.11).
Comments: Dublin Core equivalent: DC:Format

Source . Quality

Element: Source . Quality
Identifier: Source . Quality
Definition: Quality of the recorded data of the source.
Encoding: Closed controlled vocabulary { Unknown, 1, 2, 3, 4, 5, Unspecified }.
Comments: It is suggested to describe the quality of the recordings with help of a number between 1 and 5 where 1 stands for low and 5 for high quality. It is known that this quality judgement is fairly

subjective and that there are large differences between various disciplines.

Source . Position

Element: Source . Position
Identifier: Source . Position
Definition: The start- and end position of the source corresponding to the session.
Encoding: See 'Media Position' (5.5).
Comments: It may occur that a session is just a fragment within the media file.

Source . Access

Element: Source . Access
Identifier: Source . Access
Definition: Access rights of the source.
Encoding: Access (sub-schema)
Comments:

Source . Description

Element: Source . Description
Identifier: Source . Description
Definition: Description of the source.
Encoding: Description (sub-schema)
Comments:

3.6.4 Anonymized

Group: Anonymized
Identifier: Anonymized
Definition: Groups information about the name conversion file for persons who are anonymized in the transcript.
Encoding: Anonymized . Resource Link
Anonymized . Access
Comments:

Anonymized . Resource Link

Element: Anonymized . Resource Link
Identifier: Anonymized . ResourceLink
Definition: Link to the file used to convert the pseudo names into real names.
Encoding: The link is encoded as a Uniform Resource Locator as described by [RFC1738]
Comments: Dublin Core equivalent: DC:Identifier

Anonymized . Access

Element: Anonymized . Access
Identifier: Anonymized . Access
Definition: Access rights of the pseudo-name to real-name conversion.
Encoding: Access (sub-schema)
Comments:

3.7 References

Group: References
Identifier: References

Definition: Groups documentation associated with the session.
Encoding: References . Description +
Comments: Here any list of descriptions and references to other notes and publications can be given.

3.7.1 References . Description

Element: References . Description
Identifier: References . Description
Definition: Documentation associated with the content.
Encoding: Description (sub-schema)
Comments:

3.8 Sub-schemas

3.8.1 Keys

Group: Keys
Identifier: Keys
Definition: A list of attribute name-value pairs for domain specific information.
Encoding: Key +
Comments: An example of a name-value pair is; Color = Red, where the name of the attribute is 'Color' and the value of the named attribute is 'Red'.
Keys are especially useful for larger projects to define common keys.

Key

Element: Key
Identifier: Key
Definition: Associate 'Value' with 'Name'.
Encoding: <Name> = <Value>
Comments: An example of a name-value pair is; Color = Red, where the name of the attribute is 'Color' and the value of the named attribute is 'Red'.

Key . Vocabulary Link

Element: Key . Vocabulary Link
Identifier: Key . Vocabulary Link
Definition: Link to a vocabulary of selectable values for a named key.
Encoding: The link is encoded as a Uniform Resource Locator as described by [RFC1738]
Comments:

3.8.2 Language

Group: Language
Identifier: Language
Definition: Groups information about a language.
Encoding: Language . Id
Language . Name
Language . Description +
Comments:

Language . Id

Element: Language . Id

Identifier: Language . Id
Definition: Specifies a unique code to identify the language.
Encoding: See 'Language Id' (5.4).
Comments: Dublin Core equivalent: DC:Language

Language . Name
Element: Language . Name +
Identifier: Language . Name +
Definition: A list of human understandable names of the language.
Encoding: string
Comments: In general the names from the [ETHNOLOGUE] list from SIL International are recommended.

Language . Description
Element: Language . Description
Identifier: Language . Description
Definition: Elaborate description of the language.
Encoding: Description (sub-schema)
Comments:

3.8.3 Access

Group: Access
Identifier: Access
Definition: Groups information about access rights.
Encoding: Access . Availability
Access . Date
Access . Owner
Access . Publisher
Access . Contact
Access . Description +
Comments:

Access . Availability
Element: Access . Availability
Identifier: Access . Availability
Definition: Availability of the resource.
Encoding: string.
Comments: At first the specifications should be made within a prose string. At a later phase more formal descriptions will be suggested.

Access . Date
Element: Access . Date
Identifier: Access . Date
Definition: Date of access rights evaluation.
Encoding: The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format. See 5.1.
Comments:

Access . Owner
Element: Access . Owner
Identifier: Access . Owner
Definition: Name of the owner of the resource.
Encoding: string

Comments:

Access . Publisher

Element: Access . Publisher
Identifier: Access . Publisher
Definition: The name of the publisher responsible for the distribution of the resource.
Encoding: string
Comments: Dublin Core equivalent: DC:Publisher

Access . Contact

Element: Access . Contact
Identifier: Access . Contact
Definition: The contact information of the organisation to obtain access to the resource.
Encoding: Contact (sub-schema)
Comments:

Access . Description

Element: Access . Description
Identifier: Access . Description
Definition: A description of the applied access restrictions.
Encoding: Description (sub-schema)
Comments:

3.8.4 Contact

Group: Contact
Identifier: Contact
Definition: Groups information about a contact person.
Encoding: Contact . Name
Contact . Address
Contact . E-mail
Contact . Organisation
Comments:

Contact . Name

Element: Contact . Name
Identifier: Contact . Name
Definition: The name of the contact person.
Encoding: string
Comments:

Contact . Address

Element: Contact . Address
Identifier: Contact . Address
Definition: The address of the contact person.
Encoding: string
Comments:

Contact . Email

Element: Contact . E-mail
Identifier: Contact . Email
Definition: Specifies an E-mail address of the contact person.
Encoding: The E-mail address is encoded according to [RFC822]. See 5.3.

Comments:

Contact . Organization

Element: Contact . Organization
Identifier: Contact . Organization
Definition: The organization of the contact person.
Encoding: string
Comments:

3.8.5 Description

Group: Description
Identifier: Description
Definition: Groups the elements to supply a human readable description.
Encoding: Description . Text
Description . Language Id
Description . Info Link
Comments:

Description . Text

Element: Description . Text
Identifier: Description . Text
Definition: A human understandable prose text.
Encoding: string
Comments:

Description . Language Id

Element: Description . Language Id
Identifier: Description . LanguageId
Definition: An identifier of the language in which the description was written.
Encoding: See 'Language Id' (5.4).
Comments: Dublin Core equivalent: DC:Language

Description . Info Link

Element: Description . Info Link
Identifier: Description . InfoLink
Definition: A link to a description file.
Encoding: The link is encoded as an Uniform Resource Locator as described by [RFC1738]
Comments:

4 Vocabularies

Several elements are constrained by a limited set of values. These sets of values are defined as 'vocabularies' which are used for the encoding of IMDI elements. There are two types of vocabularies: open and closed controlled. A closed controlled vocabulary consists of a pre-defined set of values as they are provided and maintained by IMDI. An open vocabulary contains a set of suggested values but is not limited to this set. Domain specific values can still be entered by the user. Both types of vocabularies can be restricted to a single value or allow for multiple values in which case they are called a vocabulary list. E.g. a closed vocabulary list is a vocabulary with predefined values from which multiple values can be selected.

4.1 Location . Continent

The following closed vocabulary is used to identify the continent:

- Unknown
- Unspecified
- Africa
- Asia
- Australia
- Europe,
- Oceania
- North-America
- Middle-America
- South-America

http://www.mpi.nl/IMDI/Schema/Continents.xml

4.2 Location . Country

Closed vocabulary.
The country is encoded with a two-letter code as described by [ISO3166-1].
There is a difference between the document and the implementation here. The implementation consists of country names instead of ISO codes!
http://www.mpi.nl/IMDI/Schema/Countries.xml

4.3 Content . Communication Context

To enable searching for particular linguistic features the group of elements 'Communication Context' as proposed in [DOBES6B1] can be used to define properties of participant interaction, the degree of planning of the consultant and the researcher involvement.
The definitions and examples are directly taken from [DOBES6B1]. Some comments are extracted from the definition for consistency.

4.3.1 Interactivity

The following closed controlled vocabulary is used:

- Unknown
- Unspecified
- Interactive
- Non-interactive

- Semi-interactive

http://www.mpi.nl/IMDI/Schema/Content-Interactivity.xml

Value: Interactive
Definition: Speech events consists of verbal interaction between at least two participants.
Comments: The event may or may not include an investigator.
Examples: Many types of narrative; conversation.

Value: Non-interactive
Definition: Speech/song produced without expecting extended verbal responses from hearer(s).
Comments: Corresponds often to monologue.
Examples: many types of oratory and song; some narrativizing. Procedural texts.

Value: Semi-interactive
Definition: Primarily monologic speech punctuated by repeated interjections from the hearer(s).
Comments: -
Examples: An elderly woman tells a myth, and is prompted repeatedly by her grand-daughters. Or: While a speaker is telling a story, a child comes in and is told to be quiet.

4.3.2 Planning Type

The following closed vocabulary is used:

- Unknown
- Unspecified
- Spontaneous
- Semi-spontaneous
- (Consultant/performer-)planned

http://www.mpi.nl/IMDI/Schema/Content-PlanningType.xml

Value: Spontaneous
Definition: Unprompted speech/song.
Comments: Topic not determined from context or observers.
Examples: Conversation, chatting, joke-telling, singing while harvesting.

Value: Semi-spontaneous
Definition: Prompted speech/song.
Comments: Topic directed in some way by an investigator or community member, but participants speak/sing freely within this context.
Examples: Interview; Queries (Investigator asks, "Tell me about the history of your village", or: "Show me how to make Baked Alaska"); Retellings (investigator asks speaker to read or look at something and then re-tell a story, or describe a task in his/her own words); Promptings (children in a local school answer a teacher's question, or read aloud for him/her).

Value: Planned (Consultant/Performer-planned)
Definition: The speaker prepares in detail the structure and content of his/her "performance" in advance

Comments: This differs from 'Elicitation' (involvement), where the performer/consultant is given a framework but does not necessary plan his/her answer.
Examples: Political and ritual speech, poem recitation. Courtroom interactions would be an example of 'Planned' *and* 'Elicited' speech.

4.3.3 Involvement

The following closed vocabulary is used:

- Unknown
- Unspecified
- Elicited
- Non-elicited
- No-observer

http://www.mpi.nl/IMDI/Schema/Content-Involvement.xml

Value: Elicited
Definition: Investigator asks speaker(s) to produce isolated phonemes/ words/ utterances / grammatical structures.
Comments: -
Examples: Speakers asked to pronounce phonemes in different (phonological) environments; responses to morphological or lexical questionnaires. It may be also be possible to *elicit* Semi-spontaneous speech (planning type) if the consultant is asked to respond "as fast as possible without thinking".

Value: Non-elicited
Definition: The researcher does not interfere verbally with the speech event (other than the researcher's mere presence).
Comments: -
Examples: -

Value: No-observer (Observer absent)
Definition: No outside observer is present.
Comments: -
Examples: A tape recorder runs continuously in room while people talk (having been for example set there a half hour earlier by the investigator, with permission of course).

4.4 Content . Genre

The group 'Content . Genre' as proposed by [DOBES6B1] allows open vocabularies for different types of genre. The definitions and examples are directly taken from [DOBES6B1]. Some comments are extracted from the definition for consistency. The suggested values are aiming especially at spoken communication.

4.4.1 Interactional

The following vocabulary list is suggested:
- Unknown

- Unspecified
- Conversation
- Verbal-contest (including debate)
- Interview
- Meeting (gathering)
- Riddles (riddling)
- Consultation
- Greetings
- Leavetakings
- Humor

http://www.mpi.nl/IMDI/Schema/Content-Genre-Interactional.xml

Value: Conversation
Definition: -
Comments: -
Examples: -

Value: Verbal contest (including debate)
Definition: -
Comments: -
Examples: -

Value: Interview
Definition: -
Comments: -
Examples: -

Value: Meeting (gathering)
Definition: -
Comments: -
Examples: -

Value: Riddles (riddling)
Definition: A witty question to a respondent who is obligated to reply.
Comments: -
Examples: -

Value: Consultation
Definition: -
Comments: -
Examples: A visit to shaman; doctor visit.

Value: Greetings
Definition: -
Comments: -
Examples: -

Value: Leavetakings
Definition: -
Comments: -
Examples: -

Value: Humor
Definition: -
Comments: -
Examples: -

4.4.2 Discursive

The following vocabulary list is suggested:

- Unknown
- Unspecified
- Procedure
- Explanation

http://www.mpi.nl/IMDI/Schema/Content-Genre-Discursive.xml

Value: Procedure
Definition: A directive description of the procedures involved in the preparation or production of something.
Comments:
Examples: How to make tortillas, how to make a whip.

Value: Explanation
Definition: Practical, theoretical, or historical reality statements
Comments: -
Examples: How the monkey got its tail.

4.4.3 Performance

The following vocabulary list is suggested:

- Unknown
- Unspecified
- Oratory
- Oral-history
- Historical-narrative
- Narrative
- Oral-poetry
- Song
- Proverb
- Lament
- Insult

http://www.mpi.nl/IMDI/Schema/Content-Genre-Performance.xml

Value: Oratory
Definition: Using speech effectively in a conventionalized format to address an audience within political, legal, ceremonial, or religious settings.
Comments: -
Examples: -

Value: Oral-history
Definition: An account of firsthand experience, recalled retrospectively and communicated to an interviewer for historical purposes.

Comments: -
Examples: -

Value: Historical-narrative
Definition: An secondhand account of the experience of historical figures and events which may be partly or wholly fictional, communicated to both locals and outsiders for both historical purposes and entertainment, cf. *erzählte Welt* (Weinrich, 1964).
Comments: -
Examples: -

Value: Narrative
Definition: A recounting of one or more fictional events by one or more narrators to an audience of at least one.
Comments: -
Examples: -

Value: Oral-poetry
Definition: Spoken or sung or mixed, relatively structured form (in prosody and syntax), often with distinctive language.
Comments: -
Examples: Oral epics, narrative poetry, ballads (shorter, lyrical narratives), and panegyric odes.

Value: Song
Definition: A tune with recognizably-structured lyrics.
Comments: -
Examples: Popular and love songs, lullabies.

Value: Proverb
Definition: A summary of the wisdom of collective experience, often one line long; formulaic.
Comments: -
Examples: -

Value: Lament
Definition: -
Comments: -
Examples: -

Value: Insults
Definition: An insolent verbal act creating animosity.
Comments: -
Examples: -

4.5 Content . Task

The following open vocabulary is used:

- Unknown
- Unspecified
- Info-kiosk
- Wizard-of-oz

- Travel-planning
- Room-reservation
- Frog-story

Implementation is not consistent in the use of "-" and <space>, See:
http://www.mpi.nl/IMDI/Schema/Content-Task.xml

4.6 Content . Modalities

The following open vocabulary of modalities is used:
- Unknown
- Unspecified
- Speech
- Writing
- Gestures
- Pointing-gestures
- Signs
- Eye-gaze
- Facial-expressions
- Emotional-states
- Haptics

http://www.mpi.nl/IMDI/Schema/Content-Modalities.xml

Any chance of getting these definitions from the researchers?

Value: Speech
Definition: -
Comments: -
Examples: -

4.7 Participant . Type

The following open vocabulary of participant types is used:
- Unknown
- Unspecified
- Consultant
- Contributor
- Interviewer

http://www.mpi.nl/IMDI/Schema/Participant-Type.xml

We need definitions, comments and examples here (in the following format?)

Value: Speech
Definition: -
Comments: -
Examples: -

4.8 Participant . Role

The following open vocabulary of participant roles is used:

-
-
-
- Unknown
- Unspecified
- Mother
- Father
- Child
- Boss
- Partner
- Student
- Teacher

http://www.mpi.nl/IMDI/Schema/Participant-Role.xml

4.9 MediaFile . Type

The following closed vocabulary of media file types is used:

- Unknown
- Unspecified
- Audio
- Video
- Image
- Drawing

http://www.mpi.nl/IMDI/Schema/MediaFile-Type.xml

4.10 MediaFile . Format

The following open vocabulary of media file formats is used:

- Unknown
- Unspecified
- MPEG1
- MPEG2
- MPEG4
- WAV

What about these?

- AIFF
- JPEG

http://www.mpi.nl/IMDI/Schema/MediaFile-Format.xml

4.11 Annotation Unit . Type

The following open vocabulary of annotation unit types is used:

- Unknown
- Únspecified

- Morphology
- Orthographic
- Phonemic
- Phonetic
- Morphosyntax
- Syntax
-
- Free-translation
- Literal-translation
- Semantic

http://www.mpi.nl/IMDI/Schema/AnnotationUnit-Type.xml

4.12 Annotation Unit . Format

The following open vocabulary of annotation unit formats is used:

- Unknown
- Unspecified
- CHAT
- Shoebox
- RDBMS
- TRS
- EAF
- AIF
- BAS

The media file format is encoded as a media subtype from Multipurpose Internet Mail Extensions (MIME) as described in [RFC2046]. The media type of this MIME subtype is 'text'.

http://www.mpi.nl/IMDI/Schema/AnnotationUnit-Format.xml

4.13 Source . Format

The following open vocabulary is used:

- CC
- CD
- CD-ROM
- DAT
- DVD
- DVD-ROM
- MD
- Reel
- Hi8
- VHS
- DV
- U-matic

http://www.mpi.nl/IMDI/Schema/Source-Format.xml

Value: CC

Definition: Compact cassette
Comments: -
Examples: -

Value: CD
Definition: Compact Disc
Comments: -
Examples: -

Value: CD-ROM
Definition: Compact Disc - Read-Only Memory
Comments: -
Examples: -

Value: DAT
Definition: Digital Audio Tape
Comments: -
Examples: -

Value: DVD
Definition: Digital Video Disc
Comments: -
Examples: -

Value: DVD-ROM
Definition: Digital Video Disc - Read-Only Memory
Comments: -
Examples: -

Value: MD
Definition: Mini Disc
Comments: -
Examples: -

Value: Reel
Definition: -
Comments: -
Examples: -

Value: Hi8
Definition: High 8 Video tape
Comments: -
Examples: -

Value: VHS
Definition: VHS Video tape
Comments: -
Examples: -

Value: DV
Definition: Digital Video tape
Comments: -
Examples: -

Value: U-matic
Definition: U-matic Video tape
Comments: -
Examples: -

4.14 Language . Name

A closed controlled vocabulary of language names from [ETHNOLOGUE] is used to define the language name.

http://www.mpi.nl/IMDI/Schema/SIL-Languages.xml

5 Encoding formats

5.1 Date

The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format.

Regular expression:
((19[0-9][0-9])|([0-9][0-9][0-9][0-9]))-(1[0-2]|0[1-9])-(3[0-1]|[1-2][0-9]|0[1-9])

5.2 Age

The age is encoded as years;months.days from Codes for the Human Analysis of Transcripts [AGECHAT].

Regular expression:
Unknown|Unspecified|[0-9]+(;[0-1]?[0-9](\.[0-3]?[0-9])?)?

5.3 E-mail

The E-mail address is encoded according to [RFC822].

Regular expression:
(.+@.+\...+)?

5.4 Language Id

The language identifier is encoded as follows:
<namespace identifier>:<language identifier>

Regular expression:
((ISO639:([a-z][a-z][a-z]?)?)|(ISO639(-1:[a-z][a-z]|-2:[a-z][a-z][a-z]))|(RFC1766:(x-sil-[a-z][a-z][a-z]|[a-z][a-z][a-z]?-[A-Z][A-Z])))?

The following namespace identifiers are allowed:

ISO639-1
Specifies the code set for language identification in the form of a two-letter code. See [ISO639-1].

ISO639-2
Specifies the code set for language identification in the form of a three-letter code. See [ISO639-2].

ISO639
Allows both [ISO639-1] and [ISO639-2] code sets for language identification.

RFC1766
Allows both two-letter [ISO639-1] codes and [ISO639-1] combined with [ISO3166-1] country codes. See [RFC1766].
The three-letter codes from the [ETHNOLOGUE] list from SIL International are allowed by using the prefix 'x-sil-' for the three-letter code (See [LANGID] for more information). For example, one could enter the language identifier 'x-sil-dut' to indicate the Dutch language.

Examples:

ISO639-2:ger	*German as specified by ISO639-2*
RFC1766:en-US	*English as spoken in the US specified by RFC1766*
RFC1766:x-sil-dut	*Dutch as specified in the [ETHNOLOGUE] list.*

5.6 Media Position

The encoding of the start- and end positions on media files and media carriers depend on the type of media. The following encoding is used:

5.6.1 Time Position

CD, DAT, MD, Audio files (e.g. on CD-ROM)

Encoding: hh:mm:ss-HH:MM:SS

Description: hh:mm:ss represents the start position in hours (hh), minutes (mm) and seconds (ss) and HH:MM:SS represents the end position in hours (HH), minutes (MM) and seconds (SS).

DVD, Video files (e.g. on DVD-ROM)

Encoding: hh:mm:ss:ff-HH:MM:SS:FF

Description: hh:mm:ss:ff represents the start position in hours (hh), minutes (mm), seconds (ss) and video frames (ff) and HH:MM:SS:FF represents the end position in hours (HH), minutes (MM), seconds (SS) and video frames (FF).

Regular expression:
Unknown|[0-9][0-9]:[0-9][0-9]:[0-9][0-9](:[0-9]+)?|Unspecified

5.6.2 Counter Position

CC, Reel

Encoding: x-y

Description: x is any number of digits to represent the start position and y any number of digits to represent the end position.

Regular expression:
Unknown|[0-9]+|Unspecified

6 References

- [AGECHAT] The age of the participant encoded as years;months.days from Codes for the Human Analysis of Transcripts (CHAT). MacWhinney, Brian. 1991. The Childes Project: Tools for Analyzing Talk

- [DCMES] Dublin Core Metadata Element Set
  http://dublincore.org/documents/dces/

- [DOBES6B1] Dokumentation der Bedrohten Sprachen (DOBES), Metadata Description Recommendations: Content, Draft 03.03.01, Arienne Dwyer and Ulrike Mosel.

- [ETHNOLOGUE] Ethnologue language name index
  http://www.sil.org/ethnologue/names/

- [ISO639-1] Code for the representation of names of languages, International Organization for Standardization (ISO), 1988.

- [ISO639-2]
  Codes for the representation of names of languages - part 2: alpha-3 code, International Organization for Standardization (ISO), 1998.
  http://lcweb.loc.gov/standards/iso639-2/langhome.html

- [ISO3166-1]
  Codes for the representation of names of countries, International Organization for Standardization (ISO), 1997.
  http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/index.html

- [ISO8601] Data elements and interchange formats - Information interchange - Representation of dates and times, International Organization for Standardization (ISO), 2000.

- [LANGID] Language Identification and IT: Addressing problems of linguistic diversity on a global scale, Peter Constable and Gary Simons, SIL International, 2000.
  http://www.sil.org/silewp/2000/001/

- [RFC822] Standard for the format of ARPA internet text messages
  http://www.ietf.org/rfc/rfc0822.txt

- [RFC1738] Uniform Resource Locators
  http://www.w3.org/Addressing/rfc1738.txt

- [RFC1766] Tags for the identification of language
  http://www.ietf.org/rfc/rfc1766.txt
  specifies a two letter code taken from [ISO639-1], followed optionally by a two letter country code taken from [ISO3166-1]

- [RFC2046]
  Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types
  http://www.ietf.org/rfc/rfc2046.txt

- [MIMETYPES] Media Types
  ftp://ftp.isi.edu/in-notes/iana/assignments/media-types/media-types

- [W3CDTF] Date and Time Formats, W3C Note
  http://www.w3.org/TR/NOTE-datetime

Appendix A : Meta Transcript

The meta transcript is a container for different kinds of metadata descriptions, such as: session descriptions, sub-corpus descriptions, corpus descriptions, lexicon descriptions etc.

A.1 Meta Transcript

Group: Meta Transcript
Identifier: Metatranscript
Definition: Groups information about the metadata description itself.
Encoding: Meta Transcript . Date
Meta Transcript . Version
Meta Transcript . Format Id
Meta Transcript . Originator
Meta Transcript . Type
Meta Transcript . History
Comment: These elements serve administrative purposes and are used by tools that work with metadata descriptions.

A.1.1 Meta Transcript . Date

Element: Meta Transcript . Date
Identifier: Metatranscript . Date
Definition: The date of when the metadata description file is created.
Encoding: The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format
Comment: When a metadata editor is used to create a new metadata description file, it should save the date of creation in this element.
Dublin Core equivalent: DC:Date

A.1.2 Meta Transcript . Version

Element: Meta Transcript . Version
Identifier: Metatranscript . Version
Definition: The version of the content of the metadata description file.
Encoding: string
Comments: When metadata in the metadata description file is changed, this version number should be incremented.

A.1.3 Meta Transcript . Format Id

Element: Meta Transcript . Format Id
Identifier: Metatranscript . FormatId
Definition: The format identifier of the metadata description file.
Encoding: string
Comments: The format identifier is used to indicate which metadata schema and revision is used to describe the metadata elements.

A.1.4 Meta Transcript . Originator

Element: Meta Transcript . Originator
Identifier: Metatranscript . Originator
Definition: Indicates how the metadata description file is produced.
Encoding: Closed controlled vocabulary { Automatic, Hand, Hand checked }
Comments: A metadata description file can be generated by a certain tool, by hand or checked by hand after its generated

A.1.5 Meta Transcript . Type

Element: Meta Transcript . Type
Identifier: Metatranscript . Type
Definition: The type of the metadata description.
Encoding: Closed controlled vocabulary { Session | Sub-corpus | Corpus | Annotation | Lexicon }
Comments:

A.1.6 Meta Transcript . History

Element: Meta Transcript . History
Identifier: Metatranscript . History
Definition: Link to the change history of the metadata in the metadata description.
Encoding: The link is encoded as an Uniform Resource Locator as described by [RFC1738]
Comments: When there are modifications in the metadata itself causing a change in information content or loss of information, this can be recorded in an external resource. This link points to that resource.

Appendix B : Revision history

Version: 2.7
Date: july 2002; MPI ISLE Team

Added 'Keys' to 'Project'
Added 'Keys' to 'Media File'
Added 'Keys' to 'Annotation Unit'
Replaced 'Anonymous' with 'Anonymized'

Version: 2.6
Date: january 2002; MPI ISLE Team

Synchronised vocabularies with implementations (chapter 4).
Updated encoding formats (chapter 5).

Version: 2.5
Date: 8 june 2001; MPI ISLE Team

First frozen element set.

Version: 2.4
Date: 7 june 2001; MPI ISLE Team

Major revision to improve formalization. Added the following labels to describe the elements: Element / Group, Identifier, Definition, Encoding, Comments
Separated definition from comments at several places
Cleaned up the element definitions
Added more standard encoding formats
Replaced the element overview table with a one-page version without definitions. This table has links to the element groups for easy look-up
Changed 'Media Id' of 'Annotation Unit' into 'Media Resource Link'
Added info from DOBES technical Report 6B1
Added open/closed controlled vocabularies
Added IMDI encoding formats
Moved meta transcript definitions to appendix

Version: 2.3
Date: 2 april 2001; MPI ISLE Team

Added 'Keys'to Session
Removed 'Type' from 'Content'
Removed 'Register/Style' from 'Content'
Removed 'Channel' from 'Content'
Removed 'Event' from 'Content'
Added group-element 'Communication Context' to 'Content'
Added 'Interactivity' to 'Content - Communication Context'
Added 'Planning Type' to 'Content - Communication Context'
Added 'Involvement' to 'Content - Communication Context'
From 'Content' Replaced element 'Genre' by group-element 'Genre'
Added 'Interactional' to 'Content - Genre'
Added 'Discursive' to 'Content - Genre'
Added 'Performance' to 'Content - Genre'
Added 'Description' to 'Media File'
Added 'Recording Specs' to 'Media File'
Added 'Description' to 'Annotation Unit'
Added 'Media Id' to 'Annotation Unit'
Changed 'Font / encoding table' to 'Encoding' in 'Annotation Unit'
Added 'Description' to 'Media Carrier'
Removed 'Researcher +' with all sub-elements from 'Participants'
Removed 'Consultant +' with all sub-elements from 'Participants'
Removed 'Contibutary +' with all sub-elements from 'Participants'
Added 'Participant +' to 'Participants'
Added the following elements 'Participant+' : Description, Type, Name, Code, Role, First Language, Other

Language +, Ethnic Group, Age, Sex, Education, Link, Keys, Anonymous
Removed 'Address' from 'Collector'
Removed 'Link' from 'Collector'
Added 'Contact' to 'Collector'
Added 'Description' to 'Collector'
Removed '+' from 'Annotation Unit - Type'

Added encoding format section including: W3CDTF, RFC1738, ISO639-2, RFC1766, Ethnologue Language Name Index, ISO3166-1, RFC2046, Media Types

Version: 2.2
Date: 23 january 2001; MPI ISLE Team

Added 'Type' to 'Metatranscript'
Removed 'Institute/affiliation' from project (already in 'Contact')
Added 'Type' to 'Content'
Added 'Register/Style' to 'Content'
Added 'Channel' to 'Content'
Added 'Event' to 'Content'
Removed 'Born' from 'Age/Born' in 'Informant'
Changed 'Interviewer' to 'Researcher'
Changed 'Informant' to 'Consultant'
Added 'Language' to 'Transcription / Annotation File'
Replaced 'Publications' with 'Description +' in 'References'
Added 'Annotator' to 'Transcription / Annotation File'
Changed 'Creator' in 'Collector'
Added 'Age' to 'Interviewer'
Changed 'Transcription / Annotation File' into 'Annotation Unit'
Added 'BOOK' to 'Media Carrier – Storage Format'
Added 'Description' as structured sub-element
Changed comment in 'Language ID'

Version: 2.1
Date: 18 December 2000; MPI ISLE Team (isle@mpi.nl)

'Description' added to 'Participants', 'Description' added to 'Informant'

Version: 2.0
Date: 2 November 2000; MPI ISLE Team (isle@mpi.nl)

First external version

ISLE Metadata Initiative

Session Metadata Reference Card 2.7

Session

Name	str
Title	str
Date	c
Location	
Continent	cv
Country	cv
Region	str
Address	str
Description +	
Keys	

Project

Name	str
Title	str
Id	str
Contact	
Description +	
Keys	

Collector

Name	str
Contact	
Description +	

Content

Communication Context	
Interactivity	cv
PlanningType	cv
Involvement	cv
Genre	
Interactional	cv
Discursive	cv
Performance	cv
Task	cv
Modalities	cv
Languages	
Description	
Language +	
Description +	
Keys	

Language

Language Id	cv
Name	str

Session

Resources

MediaFile +	
Resource Link	c
Size	c
Type	cv
Format	cv
Quality	c
Recording Conditions	str
Position	c
Access	
Description	
Keys	
Annotation Unit +	
Resource Link	c
Media Resource Link	c
Annotator	str
Date	c
Type	cv
Format	cv
Content Encoding	str
Character Encoding	str
Access	
Language Id	
Anonymized	c
Description +	
Keys	
Source +	
Id	str
Format	cv
Quality	c
Position	c
Access	
Description +	
Anonymous	
Resource Link	c
Access	

Legend

+	= One or more
GREEN	= Sub-schema
BLUE	= Controlled Vocabulary (cv)
PINK	= Constraint (c)

Session

Participants

Description +	
Participant +	
Type	cv
Name +	str
Full name	str
Code	str
Role	cv
Language +	
Ethnic group	str
Age	c
Sex	cv
Education	str
Anonymized	c
Description +	
Keys	

References

Description +	

Access

Availability	c
Description	str
Date	c
Owner	str
Publisher	str
Contact	

Description

Text	str
Language Id	c

Contact

Name	**str**
Address	str
Email	c
Organization	str

Appendix 2

ISLE Metadata Initiative (IMDI)

PART 1 B

Metadata Elements for Catalogue Descriptions

Draft Proposal Version 2.1

June, 2001

INDEX

1 INTRODUCTION 74

2 CATALOGUE ELEMENTS OVERVIEW 75

3 METADATA ELEMENT DEFINITIONS 77

3.1 CATALOGUE 77
3.1.1 Catalogue . Name 78
3.1.2 Catalogue . Title 78
3.1.3 Catalogue . Id 78
3.1.4 Catalogue . Description 78
3.1.5 Catalogue . Subject Language 78
3.1.6 Catalogue . Document Language 78
3.1.7 Catalogue . Location 78
3.1.8 Catalogue . Content Type 79
3.1.9 Catalogue . Format 79
3.1.10 Catalogue . Quality 80
3.1.11 Catalogue . Smallest Annotation Unit 80
3.1.12 Catalogue . Application 80
3.1.13 Catalogue . Date 80
3.1.14 Catalogue . Project 80
3.1.15 Catalogue . Publisher 81
3.1.16 Catalogue . Authors 81
3.1.17 Catalogue . Size 81
3.1.18 Catalogue . Distribution Form 81
3.1.19 Catalogue . Access 81
3.1.20 Catalogue . Pricing 81
3.2 SUB-SCHEMAS 81
3.2.1 Access 81

4 VOCABULARIES 82

4.1 CATALOGUE . CONTENT TYPE 82
4.2 CATALOGUE . SMALLEST ANNOTATION UNIT 82
4.3 CATALOGUE . APPLICATION 82

5 ENCODING FORMATS 83

5.1 LANGUAGE IDENTIFIER ENCODING 83

6 REFERENCES 84

APPENDIX A : CATALOGUE METADATA INVENTORY 85

APPENDIX B : REVISION HISTORY 86

1 Introduction

In discussing the current set of IMDI metadata elements with various people we came to the conclusion that some metadata elements that describe “published” corpora at the top level (for instance the number of CD’s) are not present in the current IMDI set or are inappropriate to describe at the Session level. We therefore tried to make an inventory of such elements using existing description formalisms used by institutions that deal with “published corpora” such as [ELRA] and [LDC]. We call the set of metadata elements that describe “published corpora” at the top-level “catalogue” metadata elements for language resources.

This proposal has been put forward after studying the current catalogue structure of ELRA, LDC and the UHLCS metadata requirements (compiled by Pirkko Suihkonen). It takes as a starting point all the fields specified by ELRA / LDC and looks for any matching UHLCS fields (see Appendix A).

What we have done is to categorise the fields and try to determine which of them are already covered in the IMDI proposal for SESSIONS. Also we try to determine which elements of the IMDI proposal for SESSIONS would better fit in this catalogue metadata set or would make sense to be duplicated in the catalogue set.

It would be feasible to implement this catalogue metadata set for language resources in Dublin-Core [DCMES]. This is to be expected since it is metadata about published material without much information on the constituting parts. Therefore we have made a suggestion for DC equivalents with all the proposed catalogue elements. Sometimes it appears that a choice is possible, this has been indicated by a question mark.

Some concluding remarks:

- We are not yet sure that a separate IMDI catalogue metadata set is useful in all cases. It would seem so if you were looking for a corpus you can obtain as a unit and which has probably been produced for your needs (e.g. language technology applications).

- The flat description of a whole corpus by this set seems to make it ideal for easy identification of corpora by less-specific (in comparison to the IMDI Session vocabulary) search such as OAI.

All proposed DC qualifiers are our own except for Subject.language which is [OLAC]’s.

2 Catalogue Elements Overview

Element Name	DC Equivalent	Definition	Encoding	Connection to Session Elements
Name	Title.short	The name of the corpus	string	duplicate
Title	Title	The title of the corpus	string	duplicate
Id +	ID	A Unique identifier for the corpus. For example an ISBN.	string	None
Description + Subject Language	Description Subject. language ([OLAC] qualifier)	A description of the corpus The language subject of analysis	string language identifier	duplicate Summation of
Document Language	Language	The language the document is in	language identifier	Summation of
Location	Coverage	Groups the information about the location of where the corpus content was made	continent, country, region	Generalisation
Continent		The continent of where the corpus content was made	closed controlled vocabulary	
Country		The country of where the corpus content was made	ISO3166-1	
Region		The region or sub-region of where the corpus content was made	string	
Content type	Type? Subject.type?	The type of the corpus	open vocabulary Type = *{ written, speech, terminology}* Subtype = { *Corpus, Monoling. Lex., Multiling. Lex., Telephone speech, Desktop/microphone, Mutimodal/Multimedia, Other speech related}* **See ELRA**	Generalisation
Format	Format	Groups information about the formats used in the corpus	text, audio, video	Implicit in resource description
Text		The format of the text used in the corpus	Text (encoding, ...)	
Audio		The format of the audio used in the corpus	Speech (sample freq, stereo/mono sample width,...)	
Video		The format of the video used in the corpus	Video(MPEG1, Quicktime, ...)	
Quality	Format.quality	Groups information about the quality of the corpus content	audio, video	

	Audio		The quality of the audio data in the corpus	closed controlled vocabulary {1..5}	
	Video		The quality of the video data in the corpus	closed controlled vocabulary {1..5}	
Smallest Annotation Unit		Format.unit	The smallest unit of annotation used in the corpus	open vocabulary {paragraph, utterance, word, phoneme, ...}	In future version part of annotation unit
Application		Type. application ? Subject. application?	The application domain of the corpus	open vocabulary list (LDC set seems pretty exhaustive)	Summation of
Date		Date.Issued	Publishing date	YYYY-MM-DD	none
Project		Creator. project		string	duplicate
Publisher		Publisher	An entity responsible for making the resource available	string	none
Authors		Creator	An entity primarily responsible for making the content of the resource	string	Summation of
Size		Format.Extent	Total size of the corpus	string	total sum
Distribution Form		Type. distribution? Format. distribution?	How are the corpora distributed	open vocabulary list {CD, ftp, ... }	none
Access		Rights	The access conditions of the corpus	sub-schema	none
Pricing		Rights.price	The price of the corpus	string	none

3 Metadata Element Definitions

The elements for session descriptions are defined using the following attributes:

- *Element/Group Name*
  A name of the element or grouping.

- *Identifier*
  A unique identifier assigned to the element.

- *Definition*
  A statement that clearly represents the concept and essential nature of the data element.

- *Encoding*
  A statement that describes how the content of the element is encoded.

- *Comment*
  Remarks concerning the application of the data element.
  Dublin Core equivalent: some elements can be mapped with the Dublin Core Metadata Element Set [DCMES]. If this is possible, the Dublin Core equivalent of the IMDI element will be named here.
  Example: sometimes an example helps to clarify the use of the element. If this is the case, the example will be mentioned here.

3.1 Catalogue

Group:	Catalogue
Identifier:	Catalogue
Definition:	Groups information about a published corpus.
Encoding:	Catalogue . Name Catalogue . Title Catalogue . Id + Catalogue . Description + Catalogue . Subject Language Catalogue . Document Language Catalogue . Location Catalogue . Content Type Catalogue . Format Catalogue . Quality Catalogue . Smallest Annotation Unit Catalogue . Application Catalogue . Date Catalogue . Project Catalogue . Publisher Catalogue . Authors Catalogue . Size Catalogue . Distribution Form Catalogue . Access Catalogue . Pricing
Comments:	

3.1.1 Catalogue . Name

Element: Catalogue . Name
Identifier: Catalogue . Name
Definition: Name of the corpus.
Encoding: string
Comments:

3.1.2 Catalogue . Title

Element: Catalogue . Title
Identifier: Catalogue . Title
Definition: Title of the corpus.
Encoding: string
Comments:

3.1.3 Catalogue . Id

Element: Catalogue . Id
Identifier: Catalogue . Id
Definition: Unique identifier for the corpus.
Encoding: string
Comments: This can be an ISBN.

3.1.4 Catalogue . Description

Element: Catalogue . Name
Identifier: Catalogue . Name
Definition: Description of the corpus.
Encoding: string
Comments:

3.1.5 Catalogue . Subject Language

Element: Catalogue . Subject Language
Identifier: Catalogue . SubjectLanguage
Definition: Language subject of analysis.
Encoding: See 'Language Identifier Encoding' (5.1).
Comments:

3.1.6 Catalogue . Document Language

Element: Catalogue . Document Language
Identifier: Catalogue . DocumentLanguage
Definition: Language the document is in.
Encoding: See 'Language Identifier Encoding' (5.1).
Comments:

3.1.7 Catalogue . Location

Group: Catalogue . Location
Identifier: Catalogue . Location
Definition: Groups information about the location of where the corpus content was recorded or originated.
Encoding: Location . Continent
Location . Country
Location . Region
Comments:

Catalogue . Continent

Element: Catalogue . Continent

Identifier: Catalogue . Continent
Definition: The continent of where the corpus content was recorded or originated.
Encoding: Closed controlled vocabulary { Africa, Antarctica, Asia, Australia, Europe, North America, Oceania, South America }.
Comments:

Catalogue . Country
Element: Catalogue . Country
Identifier: Catalogue . Country
Definition: The country where the corpus content was recorded or originated.
Encoding: Closed controlled vocabulary. The country is encoded with a two-letter code as described by [ISO3166-1].
Comments:

Catalogue . Region
Element: Catalogue . Region
Identifier: Catalogue . Region
Definition: The region or sub-region of where the corpus content was recorded or originated.
Encoding: string
Comments: This element can also be used to describe sub-regions. Examples: europe, the netherlands, gelderland, achterhoek.

3.1.8 Catalogue . Content Type
Element: Catalogue . Content Type
Identifier: Catalogue . ContentType
Definition: The type of the corpus.
Encoding: Open vocabulary 'Catalogue . Content . Type' (4.1).
Comments:

3.1.9 Catalogue . Format
Group: Catalogue . Format
Identifier: Catalogue . Format
Definition: Groups information about the formats used in the corpus.
Encoding: Format . Text
Format . Audio
Format . Video
Comments:

Catalogue . Format . Text
Element: Catalogue . Format . Text
Identifier: Catalogue . Format . Text
Definition: The format of the text used in the corpus.
Encoding: string
Comments:

Catalogue . Format . Audio
Element: Catalogue . Format . Audio
Identifier: Catalogue . Format . Audio
Definition: The format of the audio data used in the corpus.
Encoding: string
Comments:

Catalogue . Format . Video

Element: Catalogue . Format . Video
Identifier: Catalogue . Format . Video
Definition: The format of the video data used in the corpus.
Encoding: string
Comments:

3.1.10 Catalogue . Quality

Group: Catalogue . Quality
Identifier: Catalogue . Quality
Definition: Groups information about the quality of the corpus content.
Encoding: string
Comments:

Catalogue . Quality . Audio

Element: Catalogue . Quality . Audio
Identifier: Catalogue . Quality . Audio
Definition: The quality of the audio data in the corpus.
Encoding: Closed controlled vocabulary { 1 .. 5 }.
Comments:

Catalogue . Quality . Video

Element: Catalogue . Quality . Video
Identifier: Catalogue . Quality . Video
Definition: The quality of the video data in the corpus.
Encoding: Closed controlled vocabulary { 1 .. 5 }.
Comments:

3.1.11 Catalogue . Smallest Annotation Unit

Element: Catalogue . Smallest Annotation Unit
Identifier: Catalogue . SmallestAnnotationUnit
Definition: The smallest annotation unit used in the corpus.
Encoding: Open vocabulary 'Catalogue . Smallest Annotation Unit' (4.2).
Comments:

3.1.12 Catalogue . Application

Element: Catalogue . Application
Identifier: Catalogue . Application
Definition: The application domain of the corpus.
Encoding: Open vocabulary list 'Catalogue . Application' (4.3).
Comments:

3.1.13 Catalogue . Date

Element: Catalogue . Date
Identifier: Catalogue . Date
Definition: The publishing date of the corpus.
Encoding: The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format.
Comments:

3.1.14 Catalogue . Project

Element: Catalogue . Project
Identifier: Catalogue . Project

Definition: Name of the project for which the corpus was originally created.
Encoding: string
Comments:

3.1.15 Catalogue . Publisher

Element: Catalogue . Publisher
Identifier: Catalogue . Publisher
Definition: An entity responsible for making the resource available.
Encoding: string
Comments:

3.1.16 Catalogue . Authors

Element: Catalogue . Authors
Identifier: Catalogue . Authors
Definition: An entity primarily responsible for making the content of the resource.
Encoding: string
Comments:

3.1.17 Catalogue . Size

Element: Catalogue . Size
Identifier: Catalogue . Size
Definition: Total size of the corpus.
Encoding: string
Comments:

3.1.18 Catalogue . Distribution Form

Element: Catalogue . Distribution Form
Identifier: Catalogue . DistributionForm
Definition: How are the corpora distributed.
Encoding: string
Comments:

3.1.19 Catalogue . Access

Group: Catalogue . Access
Identifier: Catalogue . Access
Definition: Groups information about access rights.
Encoding: Access (sub-schema)
Comments:

3.1.20 Catalogue . Pricing

Element: Catalogue . Pricing
Identifier: Catalogue . Pricing
Definition: The price of the corpus.
Encoding: string
Comments:

3.2 Sub-schemas

3.2.1 Access

See the document 'Metadata Elements for Session Descriptions' for the details.

4 Vocabularies

4.1 Catalogue . Content Type

Open vocabulary:
- Written
- Speech
- Terminology

Subtype:
- Corpus
- Monolinguistic Lexicon
- Multilinguistic Lexicon
- Telephone speech
- Desktop/microphone
- Mutimodal/Multimedia
- Other speech related

For example see [ELRA].

4.2 Catalogue . Smallest Annotation Unit

Open vocabulary:
- Paragraph
- Utterance
- Word
- Phoneme
- ...

4.3 Catalogue . Application

Open vocabulary list. For example see the [LDC] set.

5 Encoding Formats

5.1 Language Identifier Encoding

The language identifier is encoded as follows:
<namespace identifier>:<language identifier>

The following namespace identifiers are allowed:

ISO639-1
Specifies the code set for language identification in the form of a two-letter code. See [ISO639-1].

ISO639-2
Specifies the code set for language identification in the form of a three-letter code. See [ISO639-2].

ISO639
Allows both [ISO639-1] and [ISO639-2] code sets for language identification.

RFC1766
Allows both two-letter [ISO639-1] codes and [ISO639-1] combined with [ISO3166-1] country codes. See [RFC1766].
The three-letter codes from the [ETHNOLOGUE] list from SIL International are allowed by using the prefix 'x-sil-' for the three-letter code (See [LANGID] for more information). For example, one could enter the language identifier 'x-sil-dut' to indicate the Dutch language.

Examples:

ISO639-2:ger	*German as specified by ISO639-2*
RFC1766:en-US	*English as spoken in the US specified by RFC1766*
RFC1766:x-sil-dut	*Dutch as specified in the [ETHNOLOGUE] list.*

6 References

- [DCMES] Dublin Core Metadata Element Set
  http://dublincore.org/documents/dces/

- [ETHNOLOGUE] Ethnologue language name index
  http://www.sil.org/ethnologue/names/

- [ELRA] European Language Resources Association,
  http://www.icp.grenet.fr/ELRA/

- [ISO639-1] Code for the representation of names of languages, International Organization for Standardization (ISO), 1988.

- [ISO639-2]
  Codes for the representation of names of languages - part 2: alpha-3 code, International Organization for Standardization (ISO), 1998.
  http://lcweb.loc.gov/standards/iso639-2/langhome.html

- [ISO3166-1]
  Codes for the representation of names of countries, International Organization for Standardization (ISO), 1997.
  http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/index.html

- [ISO8601] Data elements and interchange formats - Information interchange - Representation of dates and times, International Organization for Standardization (ISO), 2000.

- [LANGID] Language Identification and IT: Addressing problems of linguistic diversity on a global scale, Peter Constable and Gary Simons, SIL International, 2000.
  http://www.sil.org/silewp/2000/001/

- [LDC] Linguistic Data Consortium, http://morph.ldc.upenn.edu/

- [OAI] Open Archives Initiative, http://www.openarchives.org/

- [OLAC] Open Language Archives Community, http://www.language-archives.org/

- [RFC1766] Tags for the identification of language
  http://www.ietf.org/rfc/rfc1766.txt
  specifies a two letter code taken from [ISO639-1], followed optionally by a two letter country code taken from [ISO3166-1]

- [W3CDTF] Date and Time Formats, W3C Note
  http://www.w3.org/TR/NOTE-datetime

Appendix A : Catalogue Metadata Inventory

	UHLCS	ELRA	LDC
Name	Corpus Title	Item Name	Item Name
Authors			Authors
Project	Projects Sponsors		Project(s) +
Date		Creation date	
Catalogue Id's		ELRA Id	LDC Catalogue Id. NIST Catalogue Id. ISBN Catalogue Id.
Type / Quality		Type = { *written, speech, terminology, Tools & software}* Subtype = { *Corpus, Monoling. Lex., Multiling. Lex., Telephone speech, Desktop/microphone, Mutimodal/Multimedia, Other speech related*}	Data type = { *Lexicon, Speech, Text*} Data source = { *Broadcast, conversation, microphone, mobile-radio, newswire, parallel, pronounciation, telephone, varied}*
Format / Quality			Sample Frequency Sample Format
Subject Language	Main-language Languages	Language	Language(s) +
Documentation	Description of the corpus		Online documentation Readme file
Application			Application={*discourse analysis, information retrieval, language identification, language modelling, machine translation, message understanding, natural language processing, parsing, pronunciation modeling, speaker identification, speech recognition, speech synthesis, spoken dialogue systems, prosody, tagging, topic detection & tracking*}
Media	Physical Storage format Server OS type Exploitation tools		Number of CD's (or) ftp
Distributor	Distributor name, address	Implicit	Implicit
Pricing		Member price Non-member price	Member price Non-member price Membership year (when free for members)
Access / Licensing	Location Access how-to		Member license Non-member license

Appendix B : Revision History

Version: 2.1
Date: 8 June 2001; MPI ISLE Team

First frozen element set.

Version: 2.0
Date: 4 June 2001; MPI ISLE Team

Smallest Annotation Unit was added
Size of Corpus was added

Version: 1.0
Date: 23 February 2001; MPI ISLE Team

First version

Appendix 3

ISLE Metadata Initiative (IMDI)

PART 1 C

Metadata Elements for Lexicon Descriptions

Draft Proposal Version 1.0

December, 2001

IMDI[15] Technical Report

Peter Wittenburg
Dafydd Gibbon
Wim Peters

[15] For information about the ISLE Metadata Initiative, please, look at the following web-site: www.mpi.nl/ISLE

INDEX

1 INTRODUCTION 89

1.1 LEXICAL METADATA 89
1.2 SCOPE OF THE PROPOSAL 89
1.3 LEXICON INITIATIVES 90
1.4 METADATA, XML AND RDF 91
1.5 APPROACH 92

2 METADATA SUGGESTIONS 93

2.1 PETERS' PROPOSAL 93
2.2 GIBBON'S PROPOSAL 94
2.3 IDE&ROMARY PAPER 95
2.4 OVERVIEW ABOUT LEXICON STRUCTURES 95

3 DESIGN FOR LEXICAL METADATA 96

3.1 INTRODUCTORY REMARKS 96
3.2 LEXICON OBJECT ELEMENTS 96
3.3 LEXICON ENTRY ELEMENTS 98

4 REFERENCES 100

1 Introduction

Within ISLE a work package was defined which has to deliver, among other things metadata standards for multimedia/multimodal language resources. The ISLE MetaData Initiative (IMDI) [1] was formed and many outstanding scientists from the discipline have joined the boards. The IMDI Steering Board decided at its Athens meeting [see 1] that the initiative should restrict itself to developing metadata sets for multimedia/multimodal corpora and lexica. With respect to corpora IMDI established a metadata set and developed tools [see 1]. With respect to lexica it was stated that a first draft proposal should be ready at the end of 2001. With this paper the authors wish to present their ideas about a lexicon metadata set to the field experts.

1.1 Lexical Metadata

Before discussing the task in more detail it is necessary to explain the term "metadata" in the context of lexica. There are often statements such as "all data is metadata" or "lexical data is metadata". Dependent on the perspective both visions are correct, of course. However, when we speak about lexical metadata in this paper we mean data which describes lexical resources which can be easily retrieved suitable ones with the help of simple queries. Even this view does not fully clarify what lexical metadata is, since experts probably have much more detailed questions for finding resources compared to those who are simply looking for an English translation of a German word.

With respect to metadata there are two major differences between corpora and lexica resources. (1) Lexical data are abstractions from corpora data or language use[16]. This means that finding a suitable resource is far less problematic. While there will be hundreds to thousands of annotated recordings for a certain language there will be comparatively few lexical resources. (2) One could argue that given the existence of a schema description for a lexical resource this schema itself describes its linguistic content at a high level of detail. So the tags used in the schema describe the content of the lexicon. If well-agreed tag labels were used this might already be sufficient. However, there are many lexica which don't have a schema description associated with them and many elements seen as being necessary in metadata descriptions (see chapter 3) are not included.

Since we are faced mostly with lexica for which there is no schema definition, since the schema definitions in many cases will offer too much detail, since the lexicon structures are very different from each other as a recent investigation showed [2] and since the tag labels used by different creators are still very different despite all standardization efforts, the authors stick to the idea to create a separate standard for lexical metadata descriptions. Of course, this standard has to correspond closely to what is currently discussed in the various initiatives trying to define generic standards for lexica.

1.2 Scope of the proposal

Firstly we have to describe the scope of the metadata set discussed in this paper, since there are many types of lexical resources such as wordlists, dictionaries, glossaries, concordances, terminology databases, thesauri, and ontologies. At first instance we want to restrict ourselves to those databases which have as main entry a lexical headword and describe its characteristics. Therefore we will exclude concept

[16] Due to this abstraction process some people speak of lexical data as metadata in a general sense.

oriented databases such as thesauri and ontologies which relate the concept entry to other concepts in a language. It may be necessary to add other descriptive elements.

We do not want to distinguish between monolingual and multilingual lexica because some of the multilingual lexica can be broken down into monolingual lexica and a special list containing the SynSets as in Wordnet. Other lexica may have multilingual entries as part of their structure. In this case the content description will describe this appropriately.

1.3 Lexicon Initiatives

Much work has already been carried out on standardizing lexicons especially to facilitate language engineering applications. While TEI [3] does not make detailed proposals for lexical tag sets various projects often executed under the EAGLES/ISLE [4] umbrella worked out concrete lexical structures. GENELEX [5] can be seen as an early attempt to describe a generic lexicon structure with an exhaustive DTD including many attributed. SIMPLE [6] was an attempt to encode multilingual lexica in a uniform way with the result of a 12 smaller example lexica. MULTILEX [7] was another project focusing on the implementation of 15 concrete lexica applying a structure derived from the GENELEX model. The MILE (Multilingual Computational Lexicon) project [8] recently started within ISLE has the task of standardizing multilingual lexica.

Other relevant work was undertaken by the OLIF2 consortium resulting in the OLIF2 proposal (Open Lexicon Interchange Format) [9]. OLIF2 defines a large number of lexical features, but does not make statements about their structural embedding. Each OLIF2 entry is a monolingual entry containing various feature/value pairs, cross-references between entries in the same language lexicon, and transfers defining bilingual transfer relations. The OLIF2 proposal for features describes four main categories: administrative, morphological, syntactic, semantic. The features are similar to those found in other more generic lexicon proposals.

Much work has been done in the area of terminology databases. The MARTIF (Machine Reachable Terminology Interchange Format) [10] work describes a format to facilitate the interchange of terminological data among terminology management systems. This work resulted in the ISO 12200 specifications. Complementary to that ISO 12620 [11] specifies how “Data Categories” which are the basic elements to describe lexical content have to be defined. Lexical terms can be defined by designating a concept entry and being associated with a language. Term related information specifies the type of the terms added. This is done by assigning attributes to the entries such as POS, etymology etc which are labels also found in dictionaries serving purposes other than terminology. Description information sets the terms into relation with domains and points to positions in concept hierarchies. Administrative information can also be added to each term. Among other things it includes typical metadata such as creator name, creation date etc.

Another important initiative is CLS (Concept Oriented with Links and Shared references) [12]. While ISO 12620 only defines the concept an entry belongs to, CLS also defines the structures between data items within entries. In doing so CLS is a framework for defining structure and content of terminology databases. The SALT project (Standards-based Access to Lexicon and Terminologies [13]) was recently initiated mainly driven by the needs from language engineering. SALT suggests the XLT (XML representations of Lexicons and Terminologies) family of formats for representing, manipulating, and sharing terminological data. A core DTD/Schema defining the generic structure can be combined with a particular Data Constraint

Specification which results in a member of the XLT family of formats. The data constraint specification determines which data categories will be used in the particular application and which values these can have. The core structure of SALT is based on the MARTIF proposal.

As shown, a number of these projects introduce categories which can also be seen as typical metadata. OLIF2 presents the most detailed list. However, none of these projects addressed the problem of finding lexical resources on the Web by querying on structured metadata descriptions in particular. Since it will be important to automatically extract metadata categories from the term descriptions, for example in terminology databases, it will be necessary to closely synchronize with the projects mentioned.

1.4 Metadata, XML and RDF

Most of the metadata sets published so far are simply lists of descriptor elements which can be used once or several times to characterize a resource. An example is Dublin Core which defines 15 elements, some qualifiers to refine the elements and some constraints. An example of a structure supporting metadata standard is IMDI. The reasons for supporting structure were to allow the definition of dependencies and in doing so to make queries possible which are of interest to the users.

The following simple example shows that in the structured case it is possible to answer queries such as: “Give me all resources where the age of participants is ‘8’ and where they speak ‘German’”. This is not possible in the unstructured case.

Structured description such as with IMDI	Set-like description such as with DC
participant	contributor=peter
participant.name=peter	contributor=jan
participant.age=6	age=6
participant.language=english	age=8
participant	language=english
participant.name=jan	language=german
participant.age=8	
participant.language=german	

Of course the creators of the DC set knew this problem. RDF (Resource Description Framework) [14] was designed to combine metadata elements from different sets which are stored in some repository such as DC in http://purl.org/DC/ and to relate these elements to each other. RDF syntax allows us also to introduce structure to cope with problems as indicated in the example above.

The need to be able to combine elements from different sets goes back to the Warwick framework [15]. It was recognized that there will be many metadata sets dedicated to domains and even groups within domains. The idea was and remains that there should be mechanisms to combine semantic definitions (elements from the vocabularies) from several metadata descriptions to create a new more complex metadata standard. RDF was amongst other reasons developed to allow this combination. In RDF one first has to specify the name spaces which will be used by referring to URIs where the corresponding repositories can be found. One can then reuse elements from the different repositories for one’s own purposes.

RDF makes use of XML as underlying syntax and was designed before the XML Schema idea arose. There is so much overlap on the structural level that RDF can be seen especially as a framework to specify semantic relationships between elements.

1.5 Approach

In this document we will follow the approach used within the IMDI project. (1) We will develop a metadata set as required by the community and which can be implemented soon. The definition will be available via a public URI. (2) We will describe mappings to other metadata sets such as DC to enable OAI type of metadata harvesting (see www.mpi.nl/ISLE for an IMDI to DC mapping). (3) The vocabulary will be entered into the SALT terminology database.

The described approach is necessary as we first have to understand what the user community requires. Starting from DC as some other initiatives have done does not lead to the necessary understanding of the domain. This approach was also chosen by the MPEG7 initiative.

In the case of exact mappings of the terms used for example in IMDI and in DC it will be possible later to create RDF-based schema definitions of the IMDI sets. This would open another way to achieve interoperability between applications which may evolve during the coming decade. Currently, OAI type of harvesting appears sufficient to satisfy the needs of a very broad community.

2 Metadata Suggestions

During the past year a number of presentations were given which make suggestions about the metadata elements which should be used to describe lexica.

2.1 Peters' Proposal

Wim Peters presented two papers one in September 2000 [16] and another in February 2001 [17]. In the first paper he looks at lexicons having lemma based entries which are associated with linguistic descriptions from 5 main areas: orthography, syntax, morphology, semantics, phonology. Another important area is usage information. Of course, these main areas can be divided into sub-categories which will not be listed here. The way sub-categorial information is represented is different for the various lexica. He also discusses that lexica can be in different formats such as XML-structures, typed-feature structures, relational structures, or project specific idiosyncratic structures. Finally, he refers to the work started within the GENELEX and ISLE projects which try to homogenize the linguistic information in lexical resources. A checklist was created for evaluation purposes which can also serve as a source for defining relevant metadata characteristics. The following slots where defined in a tree-like structure:

1 Headword; lemma; entry
2 phonetic transcription
3 alternative spelling
4 inflection, conjugation
5 cross-reference
6 morphosyntax
6.1 POS
6.2 inflectional class
6.3 derivation
6.4 gender
6.5 number
6.6 mass vs count
6.7 gradation
7 subdivision counter
8 entry subdivision
9 sense indicator
10 linguistic label
11.1 subcategorization frame
11.2 obligatory of elements
11.3 auxiliary
11.4 light or support construction
11.5 periphrastic constructions
11.6 phrasal verbs
11.7 collocator
11.8 alternations
12.1 semantic type
12.2 argument structure
12.3 semantic relations
12.4 regular polysemy
12.5 domain
12.6 decomposition
13 translation
14 gloss; definition
15 near-equivalent
16 example phrase (straightforward)
17 example phrase (problematic)
18 multiword unit
19 subheadword
20 usage note
21 frequency

In his second paper a step towards formulating lexical metadata was made. He distinguishes external (creator, associated tools, ...) and internal information (linguistic content). With respect to external information Peters suggests including the following elements: ID, Date of creation, Title, Creator, Publisher, Contributor, Project, Contact, Format[17], URL, Type[18], Access, Source, Relation, References. With respect to internal information, Peters mentions the following descriptors: Content-Description, Content-Language, Content-Size, Region, Range, Modality[19], Data Format[20].

Further, Peters describes the differences between top-down and bottom-up developed metadata sets for lexica. The advantages of the top-down design are a high-level, underspecified and uniform set of metadata and a high degree of theory independence. The bottom-up approach starts with concrete examples and tries to establish commonalities. If the sample is large enough both approaches come very close. A combination of both approaches is advocated and the results of the MILE project are seen as relevant.

2.2 Gibbon's Proposal

Dafydd Gibbon also makes suggestions about lexical metadata in his MILE presentation [18]. He describes a layered approach to the problem of defining a metadata set for lexica. He distinguishes two levels of lexical objects: (1) The Lexicon Object covering general information about the lexicon as a whole. (2) Information about the Lexical Entries which describe the content. With respect to the Lexicon Object three categories are distinguished: (1) bibliographical data such as creator, publisher, title, date etc; (2) Medium and format aspects; (3) macrostructure type of information such as languages involved, lexicon type (taxonomy), etc.

The description of a Lexical Entry comprises a type description and a microstructure description. The latter gives information about the underlying structure of the entries and the enclosed data category groups. These are the classical type of high level linguistic information categories each having their special list of data categories defined by some user groups or standardization bodies. Gibbon further elaborates by giving an impression of how the structure of the data categories could be described.

- Lexicon Object
  - bibliographical data (creator, publisher, title, date, ...)
  - Medium and format description
  - macrostructure type (languages involved, lexicon type/taxonomy, ...)
- Lexical Entry
  - type (encyclopaedic, multiword unit, ...)
  - microstructure data model
    - structure (flat, tree,...)
    - data category groups
      - data categories - surface
      - data categories - structure
      - data categories - meaning
      - data categories - examples

[17] oracle/MS access/shoebox/...
[18] collection, dataset, software/ ...
[19] linguistic mode such as orthography, sound, gesture, ...
[20] relational, feature structure, relational, list, ...

2.3 Ide&Romary Paper

A paper presented by Nancy Ide and Laurent Romary [19] does not explicitly mention the metadata issue, but due to its focus on a flexible framework for representing computational lexica it is also relevant to the metadata topic. Basically they argue for the definition of common data categories (lexical objects), their representation with the help of RDF, and their embedding in an abstract XML representation. The description of a data category is metadata which is correct in the broader sense of metadata. Giving such definitions it would be possible to automatically extract those descriptors which are called metadata elements in this paper.

2.4 Overview about Lexicon Structures

Recently, Wittenburg presented an analysis of the structures of real lexica [2]. These were presented by the teams in the DOBES project [20], an analysis of structures of written lexica by Bell and Bird [21], and an analysis of the structures of well-known lexica such as GENELEX [5], CELEX [22], and KirrKirr [23]. It was shown that the structures used were very different and that the linguistic sub-domain is not the most relevant criterion. In field linguistics the heterogeneity in structural complexity is naturally the largest. The more complex structures in field linguistics are comparable to what is known, for example, from CELEX. The structure of written lexica is fairly simple, while the GENELEX model had to be exhaustive.

In field linguistics it can be seen that the language under investigation influences the structural choices and the complexity. In highly inflected languages such as Wichita most utterances are lexical entries bearing a rich internal structure. Documenting these languages requires a rich lexicon.

Another observation was that the grouping of data categories (the primitive elements found in lexica) is also very heterogeneous. This leads to the assumption that differences in linguistic theory and languages influence the structural choices.

3 Design for Lexical Metadata

3.1 Introductory Remarks

Already at this moment it seems to be wise to embed lexical metadata descriptions in an RDF schema. This allows us to make use of the various definitions of lexical data categories. As we have seen there are a number of projects which have started to define them. Lexical metadata descriptions have to make use of them as far as possible.

From the overview about lexical structures we can conclude that lexicon metadata can not go into too deep detail when describing the lexical entries. Lexical metadata cannot intend to represent the lexical structure, since there are so many differences and linguists in general don't agree about the details. In this proposal we have chosen an approach to mention the main categories of lexical data and to allow people to add values to these categories which represent the common names for lexical sub-categories. However, the values associated with these main categories have to be selected from an controlled vocabulary and added as a flat list. Only such an approach will promise to not create over-specifications which will prevent successful searching. Due to the differences in linguistic theory which partly represents the different natures of the languages under study sub-categories can be listed under more than one main category.

Example with two main categories:

Orthography:	Spelling, Syllabification, Hyphenation
Morphosyntax:	POS, Inflection, Gender

The semantics of this typical description can be read in the following way: The lexicon includes information about the two main categories orthography and morphosyntax. Orthography incorporates information about spelling, syllabification and hyphenation. Morphosyntax incorporates information about POS, Inflections and Gender. The entries do not make statement about the details of the encoding and whether substructures are used etc. The purpose is that the person searching for inflectional data will find a hit and then has to look in detail into the resource itself to understand how it is done.

As in the case of corpora we have to include flexibility in the metadata set such that researchers can add important fields which are not yet in the proposed element set.

3.2 Lexicon Object Elements

We would like to follow Gibbon's classification and first discuss metadata elements on the level of the lexicon as a whole. Here we let us guide by the proposals from Peters, Gibbon, the current IMDI set for describing multimedia corpora, and the suggestions made in the various lexicon and terminology projects discussed above[21].

Lexicon Object

Name	A short name which identifies the Lexicon
Title	A more elaborated title of the Lexicon
Date+	Date of the creation and major modifications
Version	Version indication
LexiconType	Type following some taxonomy[22] (controlled vocabulary)

[21] It has to be noted here that it is not mandatory to fill in all fields.

[22] This metadata set at first instance was only made for a selection of lexical types such as dictionary | wordlist | glossary | concordance | terminologyDB

Creator +	The responsible persons who created the resource
Name	Name of creators
Contact	Contact address which can be the creators address or a substitution address
Description	A suitable description associated with the set of creators
Project	A block to describe the project
Name	Short name of the project
ID	Unique project identifier
Contact	contact address sub schema
Description +	some space for descriptions to be associated with the project
Object Languages	A block to describe the languages included in the lexicon
Description	some space for a prose description
MultilingualityType	languages can occur in different flavors in lexica, they can occur as multilingual entries in ML lexica, but they also can occur as translations of for example sense descriptions; this difference can be indicated with the help of a controlled vocabulary
Language +	a list of languages included, each language be described in a substructure
Meta Languages	A block to describe the languages which are used to define terms, to describe meaning and similar
Description	some space for a prose description
Language +	a list of languages included, each language be described in a substructure
Lexical Entry	a block which describes the linguistic content, i.e. the attributes the lexicon contains
EntryElements	a possibility to add feature/value pairs describing the linguistic categories used in the lexicon; the feature is taken from a controlled vocabulary; per feature a flat list of descriptors can be specified also taken from a controlled vocabulary; for details see below
Format	a rough indication of the format the lexicon is in such as relational table, structured plain text, some XML format, html format, ...
AccessTool	many lexica are only interpretable via concrete access tools such as Shoebox, ORACLE, FoxPro, Access, Web-Browser, ...
Media	this entry tells whether the lexicon includes audio or video samples or graphics
Schema	name of or ref to the documented structure which could be a DTD, Schema or similar
Character Encoding +	this list should give an impression of the type of fonts needed to render all data included such as UTF-8, ISO-latin
Size	the size of the lexicon in bytes
No Lexical Entries	the number of lexical entries the lexicon includes
Access	sub-schema where access info is given
Keys	a possibility to add feature/value pairs to define new keywords
Source	this entry can't be anything else than a "special description" which describes which sources were used to build the lexicon; this info is relevant, but I don't see any way to make it a formal description
Description +	these are general descriptions which can be associated with the lexicon
References	block to cover references to publications etc
Description +	

Sub-Schemas

Access	
ResourceLink	URL pointing to the resource if it is directly accessible
Availability	codification of terms of access (has to be worked out
Description	prose description associated with access
Date	date of statements about access
Owner	defines the owner of the lexicon

Publisher	defines the publisher of the lexicon
Contact	specifies a sub-schema describing whom to contact

Contact

Name	name of the contact person
Address	address info
Email	email address
Organization	name of an institution

Description

Text	this is the prose text of the description
Language ID	this describes the language the text is in
Info Link	each description can be associated with further information such as web-pages etc

Language

Language ID	formal language specifier from ISO or SIL lists
Name	general name of the language
Description	a description of the language can be associated

3.3 Lexicon Entry Elements

The lexicon entry elements describe the linguistic content covered by each lexicon.

The following 11 main entry categories have been distinguished as a proposal for implementation. Each main category contains zero or more names of object classes that represent subclasses of the linguistic descriptive level captured by the category. Each occurrence indicates whether the corresponding linguistic information is present in the lexicon. Some of these object classes can have subclasses of their own which are not shown and discussed here. The idea is that at a later moment not only the names of these object classes will be available, but that the user can receive more detailed information. The list of object classes is not meant to be exhaustive and can be extended if necessary.

In order to accommodate linguistic annotation in a maximally polytheoretic and flexible fashion, it is possible to duplicate existing subclasses as descendants of other main categories if the need arises. The categories do not make any statement about the details of the encoding and whether subdivisions are used. The purpose is that the person searching for e.g. morphological segmentation data will find a hit in the meta-description of some lexicon, and subsequently has to take a more detailed look into the resource itself to find out about the format and granularity of the available segmentation data.

Modality indicates which mode of communication is captured in the lexicon. Possible values are:

- Spoken
- Written
- Sign

Headword type indication of the linguistic nature of the entry in the lexicon. Possible values are:

- Sentence
- Phrase
- Wordform
- Lemma entry conforming to the unmarked wordform (e.g. infinitive for verbs).
- Abstract Lemma entry not conforming to any wordform of the group subsumed by the lemma.
- Stem
- Affix

Orthography possible values are:

- Hyphenated Spelling
- Syllabified Spelling

	Spelling Variants	orthographic variations with or without preferred spelling information
	Citations	

Morphology	possible values are:	
	Stem	deep or surface stem
	Stem Allomorphy	variations at stem level
	Segmentation	analysis into morphological constituents such as affixes
	Production rules	governing the production of surfaec forms on the basis of stems
	Typology	any classification of entries or morphological entities

Morphosyntax	possible values are:	
	Part of Speech	syntactic class of the entry.
	Inflection	any inflectional or conjugational information
	Countability	pluralization properties
	Gradability	e.g. adjectival comparative/superlative constructions
	Gender	e.g. neuter
	Typology	any classification of entries

Syntax	possible values are:	
	Complementation	Syntactic subcategorization
	Alternation	alternative complementation patterns
	Modification	e.g. adjectival modification patterns
	Shallow Parsing	segmentation into chunks
	Deep Parsing	finer grained analysis below chunk level
	Functional Parsing	syntactic functions such as subject
	Collocations	significant juxtaposed entries/wordforms
	Typology	any classification, e.g. prepositional/phrasal verb

Phonology	possible values are:	
	Transcription	any type of phonetic/phonological transcription
	IPA Transcription	transcription in International Phonetic Alphabet
	CV pattern	transcription in terms of consonant-vocal combinations
	Constituent Structure	segmentation in to phonetic constituents
	Intonation	stress marking, constituent length etc.

Semantics	possible values are:	
	Sense distinction	polysemy and/or homonymy
	Ontological classification	related concepts and conceptual relations
	Gloss	informal description of the sense in natural language
	Definition	formal description of the sense e.g. as a 1st order logic formula
	Connotation	non-denotational information such as pejorative
	Idiom	idiosyncratic use
	Componential Features	formula or list containing a finite set of meaning attributes
	Cross-references	links to other entries/wordforms
	Semantic relations	relations between entries or associated concepts
	Preference	characterization of the arguments in the semantic predicate

Etymology	information about the historical context (morphological, phonological, syntactic, semantic) of a lexical entry or wordform.

Usage	Pragmatic/sociolinguistic information; possible values are:	
	Region	e.g. dialect
	Style	e.g. slang

Frequency	corpus-derived frequency of occurrence

4 References

[1] EAGLES/ISLE Metadata Initiative; http://www.mpi.nl/ISLE

[2] Wittenburg, P. (2001) *Lexical Structures*. DOBES internal document. MPI Nijmegen

[3] TEI; http://www.tei-c.org

[4] EAGLES/ISLE; http://www.ilc.pi.cnr.it/EAGLES96

[5] GENELEX; http://www.ilc.pi.cnr.it/EAGLES96/lexarch

[6] SIMPLE; http://www.ilc.pi.cnr.it/EAGLES96

[7] MULTILEX; http://www.ilc.pi.cnr.it/EAGLES96/lexarch

[8] MILE; http://www.ilc.pi.cnr.it/EAGLES96

[9] OLIF; http://www.olif.net

[10] MARTIF; http://www.ttt.org

[11] ISO 12620; http://www.iso.ch/cate/d2517.html

[12] CLS; http://www.ttt.org

[13] SALT; http://www.ttt.org/salt

[14] RDF; http://www.w3.org/RDF

[15] Warwick Framework; http://www.dlib.org/dlib/july96/07weibel.html

[16] Peters, W. (2000) *Metadata for Lexicons*. MPI Lexicon Workshop, Nijmegen

[17] Peters, W. (2001) *Metadata for Lexical Resources*. ISLE CLWG Meeting, Pisa

[18] Gibbon, D. (2001) *Notes on Lexicon Metadata*. ISLE CLWG Meeting, Pisa

[19] Ide, N., Romary, L. (2001) A Flexible Framework for Representing Computational Lexicons. ISLE CLWG Meeting, Pisa

[20] http://www.mpi.nl/DOBES

[21] Bell, J., Bird, St. *A Preliminary Study if the Structure of Lexicon Entries*. http://www.ldc.upenn.edu/exploration/expl2000/program.html

[22] CELEX; http://www.mpi.nl/world/celex

[23] Manning, C. *KirrKirr.* Lexical Tool; http://www.sultry.arts.usyd.edu.au/kirrkirr

Appendix 4

ISLE Metadata Initiative (IMDI)

PART 2 A

Mapping IMDI Session Descriptions with OLAC

OLAC Version 0.3 , 2001-06-16
IMDI Version 2.5, 2001-06-08

Draft Proposal Version 1.0

August, 2001

IMDI[23] Technical Report

Max-Planck-Institute for Psycholinguistics
NL, Nijmegen

[23] For information about the ISLE Metadata Initiative, please, look at the following web-site: www.mpi.nl/ISLE

INDEX

1 INTRODUCTION.....103

2 MAJOR PROBLEMS 103

2.1 SESSION VS. RELATION..... 103
2.2 STRUCTURED VS. FLAT METADATA SET..... 105

3 OLAC PERSPECTIVE.....106

4 IMDI PERSPECTIVE.....112

1 Introduction

In March 2000 the White Paper of the IMDI initiative was released in which the intentions and the approach of IMDI were explained. The goal of IMDI is to primarily define a metadata set for multimedia/multimodal language resources as part of the ISLE project. In December 2000 LDC and SIL presented the OLAC initiative which has as one of its goals to define a metadata set for all sorts of language resources being used in linguistics and related disciplines.

Since the community IMDI is addressing is a subset of the community OLAC is addressing and since both should offer their metadata records to the even more general DublinCore based community, this document was written. It describes how the mapping between IMDI and OLAC elements can be done. Since OLAC claims to support the DC metadata set and therefore is compliant with the OAI proposals, we will not make special statements how IMDI elements can be mapped to DC elements at first instance.

The IMDI-OLAC mapping will be viewed from two perspectives: (1) the OLAC perspective and (2) the IMDI perspective . For the OLAC perspective we will take each OLAC element and look for a suitable candidate element in the IMDI set. For the IMDI perspective we will take each IMDI element and look for a suitable mapping with OLAC elements. The result has to be a proposal of how to serve IMDI information to OAI type of metadata harvesters and vice versa.

Mapping metadata elements between two worlds is not a completely new task. We would like to refer to the “Harmonization” project which has as goal to harmonize the MPEG7 and the DC worlds. It should be mentioned here that this project has decided to follow a very strict mapping rule, i.e. only very few elements of the rich MPEG7 set will be mapped onto DC elements. The reason for this is mainly to not further extend the semantic definitions of the DC elements which are partly already vague enough.

The purpose of this document is two fold:
1) It will serve as a basis to make IMDI records available for OAI harvesting.
2) It will be used to start discussions with the DC Usage committee.

2 Major Problems

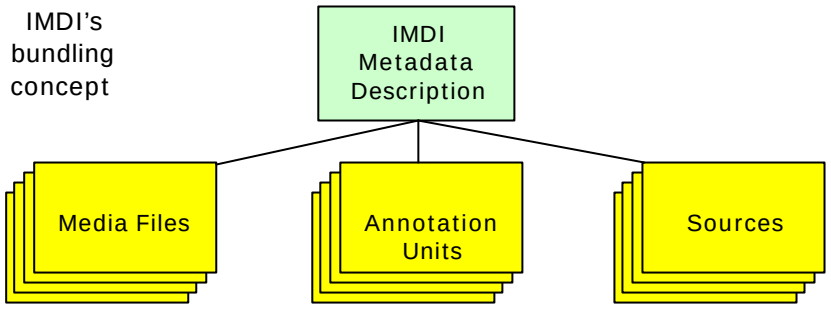
Of course, such mappings cannot be done without severe problems. In this chapter we want to mention two major problems we faced when mapping IMDI to OLAC.

2.1 Session vs. Relation

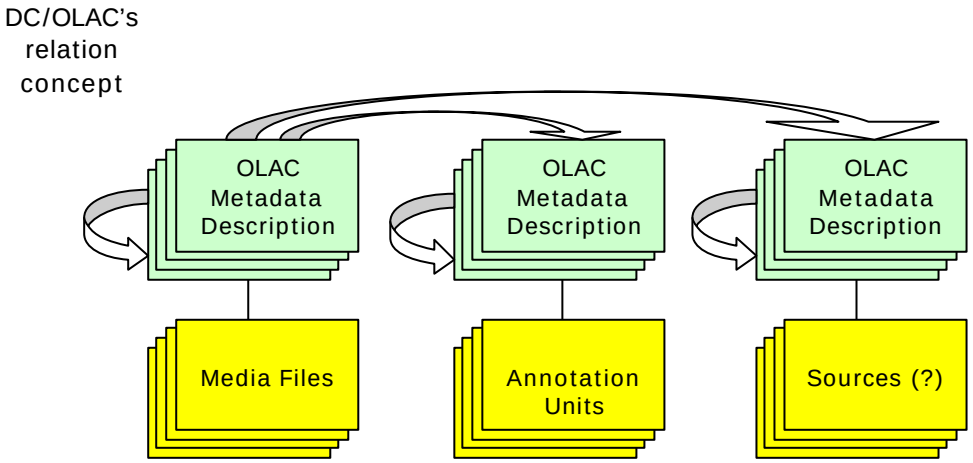
With the IMDI metadata description for sessions several resources can be associated with one metadata description (MD). A single MD describes the recording event in relevant linguistics terms together with the data derived from the recording such as digitized media files, different kinds of annotations of those media files (morphology, phonetics, syntax etc.) and also the actual sources of the linguistic session. Types of sources are for example audio recordings, video recordings, written text, books and so on. Of course, for each of these resources a separate MD could be created.

In many areas of linguistics (Language Engineering, Field Linguistics, ...) it is common to have several cameras and/or microphones or even other equipment such as eye-trackers per recording event. All these sensors result in source files being

later digitized to media files or stored in media files directly. All these media files can result in separate annotation files covering a number of tiers. The result is a 2-dimensional array of tightly related resources all describing the same linguistic event as indicated in the figure below. The IMDI MD set allows the user to group these individual resources together and to express their relationship implicitly.



OLAC uses a flat MD with no implicit way to bundle resources. To do this in OLAC one has to make separate MD files for each of the resources and heavily use the DC:Relation element to bundle them. Since the MD files are separated all related files should refer to each other to make sure that the related files can be found. The second figure suggests just one specific way to do the bundling in the OLAC domain. It is assumed that there is a master file (one of the media files) which has a number of dependent files in both dimensions. Organizing the metadata descriptions this way imposes a certain access structure which will in most cases not be useful. Think of a situation where people have created annotations but where there is no sound file yet.



A more neutral scheme could be implemented if every MD representing one of the resources points to every other MD. This, however, would create much overhead.

For the concrete IMDI-to-DC mapping we have chosen to use the DC:Source and the DC:Relation elements. The semantics are as follows:
- An Annotation Unit has relations to other Annotation Units when they contain different annotation layers all describing the same linguistic event.
- An Annotation Unit has as Source(s) a set of Media Files.

- The Media Files may have as sources IMDI:Sources.

2.2 Structured vs. flat metadata set

Since OLAC as DC only makes statements about metadata elements and therefore implicitly describes a flat MD set, the possibilities to have related or structured elements within a metadata description is limited. The DC:Relation element describes relation between documents. For example, a participant name can not be associated with a description while in IMDI they belong to one and the same participant structure. There are several other examples where the community interviewed by IMDI requested that sort of structural detail such as the relation of dates with different events in the lifetime of a document (various annotation tiers may be created at very distinctive moments in time. When such IMDI structures are flattened, the relevant information is lost.

Example:
IMDI set

```
AnnotationUnit
AnnotationUnit.Type = transcription
AnnotationUnit.Date = 2002-08-08
...
AnnotationUnit
AnnotationUnit.Type = morphosyntax
AnnotationUnit.Date = 2004-08-08
...
```

Flattened OLAC set

```
Type.Data = transcription
Type.Data = morphosyntax
Date = 2002-08-08
Date = 2004-08-08
```

In OLAC structure is sometimes allowed by using the 'refine' attribute. For example, the refinement of "OLAC:Contributor" allows the definition of a role. In IMDI this would have been named "Contributor . Role". The problem will be for all OAI harvesters that these details will not be used while querying.

The “metadata community” was aware of these problems. To tackle these type of problems the Resource Description Framework (RDF) was invented. RDF would allow the designer to embed elements into structures and therefore define semantic relations. In IMDI they are already implicitly coded. The IMDI designers expect that in future richly structured modularized metadata documents will dominate. IMDI already has specified a number of such modules (“data category groups”) which can be treated separately.

3 OLAC Perspective

The OLAC perspective of the IMDI - OLAC mapping must be understood as follows: Every OLAC element is looked at and suitable IMDI elements are identified. Of course, many IMDI elements will not be touched by using this perspective. In the comment field we give a view about the mapping and make a concrete suggestion of how to actually do the mapping to enable OAI type of harvesting of IMDI records.

OLAC Element	OLAC Definition	IMDI Element	IMDI Definition	Comment
Contributor	An entity responsible for making contributions to the content of the resource.	Session . Participants . Participant . FullName	The full name of the participant.	IMDI separates several types of participants from an open vocabulary (researcher, consultant, contributary, computer etc.). The OLAC refine attribute can be used to indicate the role of the contributor. **Action:** For each participant; copy IMDI:Participant.FullName to OLAC:Contributor and copy IMDI:Participant.Type to the OLAC:Contributor refine attribute.
Coverage	The extent or scope of the content of the resource (typically include spatial location, temporal period or jurisdiction; best to be taken from controlled vocabularies)	Session . Country	The country where the session was recorded or originated.	IMDI includes several elements to specify a spatial location (Continent, Country, Region). Since OLAC doesn't allow the 'refine' attribute here, it is not possible to use refinements like 'continent' or 'region'. There is no equivalence for the temporal period in IMDI. **Action:** Copy the IMDI:Session.Country element to OLAC:Coverage.
Creator	An entity primarily responsible for making the content of the resource. (person, organization, service, ...)	Session . Collector . Name	The name of the person responsible for the collection of the session data.	Although the semantics are slightly different a simple match with OLAC is possible. The OLAC refine attribute can be used to allow 'collector' as the role of the creator.

				Note: in OLAC contributors can also appear as creators. **Action:** Copy the IMDI:Session.Collector . Name to OLAC:Creator and set refine to 'collector'.
Date	A date associated with an event in the life cycle of the resource.	Session . Date; Session . AnnotationUnit . Date	The date when the primary data of the session was created; The dates when the probably different annotation units were created.	Both types of IMDI dates are creation dates and can therefore not be distinguished by using the OLAC refine attribute (refine can be: created, issued or modified). Since the IMDI dates are in the same standardized encoding as the OLAC code attribute, the IMDI date can be copied there. **Action:** Copy IMDI: Session.Date to OLAC:Date and set code to Session.Date. For all annotation units copy IMDI:AnnotationUnit.Date to OLAC:Date and set code to AnnotationUnit.Date.
Description	An account of the content of the resource.	Session . Description; ; Session . Content . Description; Session . Resources . MediaFile . Description; Session . Resources . AnnotationUnit . Description;	A human understandable prose text.	IMDI has several descriptions on various levels to allow people to express comments associated to the elements on that level. Links to external descriptions in IMDI are provided by an InfoLink sub-element of Description while in OLAC they apparently can be stored in the description element itself. The language of the description can be stored in the Language sub-element in IMDI and in the lang attribute in OLAC. **Action:** Since in OLAC there exist different MD descriptions for the media files and the annotation files we can copy the corresponding description fields to the various OLAC entries.

				For each IMDI Description copy the corresponding IMDI:Description.Text to OLAC:Description, set OLAC lang attribute to Description.LanguageId. Store IMDI Description . InfoLink at the end of OLAC:Description.
Format	The physical or digital manifestation of the resource. (the code attribute is used to support MIME types)	Session . Resources . MediaFile . Type Session . Resources . MediaFile . Format Session . Resources . AnnotationUnit . Format	The type of the media file; The format of the media file; The file format which is used for the annotation.	The OLAC code attribute can be mapped with the IMDI:MediaFile.Type and IMDI:MediaFile.Format elements, since IMDI:MediaFile.Type is encoded as the major part of MIME and IMDI:MediaFile.Format as the minor part. The IMDI:AnnotationUnit.Format is already encoded as MIME and can be mapped with the code attribute of OLAC:Format. **Action:** For each MediaFile copy IMDI:MediaFile.Type and IMDI:MediaFile.Format in the code attribute of OLAC:Format (in MIME syntax). For each AnnotationUnit copy IMDI:AnnotationUnit.Format to the code attribute of OLAC:Format.
Format . cpu	The CPU required to use a software resource.	-	-	No mapping with IMDI.
Format . encoding	An encoded character set used by a digital resource.	Session . Resources . AnnotationUnit . CharacterEncoding	Name of the character encoding used in the annotation unit.	One to one mapping to IMDI. **Action:** Copy IMDI:AnnotationUnit.CharacterEnc ding to OLAC:Format.encoding.
Format . markup	A markup scheme used by a digital resource.	Session . Resources . AnnotationUnit . ContentEncoding	Name of the encoding scheme used for the annotation purpose.	One to one mapping to IMDI. **Action:** Copy IMDI:AnnotationUnit.ContentEncod ng to OLAC:Format.markup.
Format . os	An operating system required	-	-	No mapping with IMDI.

	to use a software resource.			
Format . sourcecode	A programming language of software distributed in source form.	-	-	No mapping with IMDI.
Identifier	An unambiguous reference to the resource within a given context.	Session . Resources . MediaFile . ResourceLink; Session . Resources . AnnotationUnit . ResourceLink; Session . Resources . Source . Id	A link to the media file; A link to the Annotation Unit; Short code to identify the source	OLAC allows an URI here which maps with IMDI ResourceLink of MediaFile and AnnotationUnit. Non-electronic resources may be described by a local identifier according to OLAC. The IMDI:Source.Id is an example of such a non-electronic resource. Note that OLAC has different MD descriptions for the different resources. **Action:** For each MediaFile copy IMDI: MediaFile.ResourceLink to OLAC:Identifier. For each AnnotationUnit copy IMDI:AnnotationUnit.ResourceLink to OLAC:Identifier. For each Source copy IMDI:Source.Id to OLAC:Identifier.
Language	A language of the intellectual content of the resource.	Session . Resources . AnnotationUnit . LanguageId	The language used for the annotation unit.	Meant is the language the resource is written in. In the IMDI case it is the language used for the annotations for example. The OLAC code attribute maps with the Language. Id from IMDI. **Action:** For each AnnotationUnit copy IMDI:LanguageId to the code attribute of OLAC:Language.
Publisher	An entity responsible for making the resource available.	Session . Resources . MediaFile . Access . Owner; Session . Resources . MediaFile . Access . Publisher; Session . Resources . AnnotationUnit . Access	Name of the owner of the resource; The name of the publisher responsible for the distribution of the resource.	IMDI separates the access rights owner from the publisher. Both are in accordance with the OLAC definition. **Action:** For each MediaFile copy IMDI:Access.Owner to OLAC:Publisher, copy

		. Owner; Session . Resources . AnnotationUnit . Access . Publisher;		IMDI:Access.Publisher to OLAC:Publisher. For each AnnotationUnit copy IMDI:Access.Owner to OLAC:Publisher, copy IMDI:Access.Publisher to OLAC:Publisher.
Relation	A reference to a related resource.	-	-	The problem with such a conversion from structured to a flat metadata set is explained in the text. **Action:** We suggest to use the relation element to indicate whether there are several files of the same type such as for example several media files.
Rights	Information about rights held in and over the resource.	Session . Resources . MediaFile . Access; Session . Resources . AnnotationUnit . Access	Groups information about access rights.	The different resources described by IMDI have their own access policy statements. All information about rights is in the IMDI access structure including: availability, date, owner, publisher, contact and description. **Action:** For each MediaFile; copy text from all Access sub-elements into OLAC:Rights. For each AnnotationUnit; copy text from all Access sub-elements into OLAC:Rights.
Source	A reference to a resource from which the present resource is derived.	Session . Resources . Source . Id	Short code to identify the source.	**Action:** As explained in the text we will indicate that (1) a IMDI:source is the DC:Source of the IMDI:media file and that (2) each IMDI:MediaFile is the source of an IMDI:AnnotationUnit.
Subject	The topic of the content of the resource.	Session . Content . CommunicationContext; Session . Content .	Groups the linguistic features of the session concerning the context of the communication;	IMDI uses several classifications for the content description. **Action:**

		Genre; Session . Content . Task; Session . Content . Modalities.	Lists the conventionalized discourse types of the content of the session; The major task carried out in the session; Gives a list of modalities used in the session.	For each sub-element of IMDI:Content copy the element text to OLAC:Subject.
Subject . language	A language which the content of the resource describes or discusses.	Session . Content . Languages . Language . Name; Session . Content . Languages . Language . Id	A human understandable name of the language; Specifies a unique code to identify the language.	The OLAC code attribute maps with the Language. Id from IMDI. **Action:** For each Language copy IMDI:Langauge.Name to OLAC:Subject.language, copy IMDI:Language.Id to the code attribute of OLAC:Subject.language.
Title	Short name given to the resource.	Session . Title	A full title for the session.	One to one mapping to IMDI. **Action:** Copy IMDI:Session.Title to OLAC:Title.
Type	The nature or genre of the content of the resource.	-	-	No mapping to IMDI.
Type . data	The nature or genre of the content of the resource from a linguistic standpoint.	Session . Resources . AnnotationUnit . Type	The type of the annotation unit.	Each IMDI annotation unit has a linguistic data type which maps with this OLAC element. **Action:** For each AnnotationUnit copy IMDI:AnnotationUnit.Type to OLAC:Type.data
Type . functionality	Software functionality.	-	-	No mapping to IMDI.

4 IMDI Perspective

The IMDI perspective of the IMDI - OLAC mapping must be understood as follows: for each IMDI element it will be checked how best OLAC/DC elements can be mapped. It has to be kept in mind here that in the OLAC/DC philosophy each of the resources bundled as a session in IMDI are separate resources and therefore have different metadata descriptions. This amounts to a consistency check problem when merging different MD descriptions to one. Perhaps it will be wise to leave the descriptions separate. It has to be seen in OLAC descriptions how the Source and Relation elements will be used.

IMDI Session Element	**IMDI Definition**	**OLAC Element**	**OLAC Definition**	**Comment**
Name	Is a short name of a session	Title	Short name given to the resource.	One of these two elements has to be used for mapping with OLAC:Title. IMDI:Session.Name is a required element.
Title	Title of the session			
Date	Date of the recording the session is based on	Date	A date associated with an event in the life cycle of the resource.	OLAC dates are associated with each separate document (media, annotations) and when the creation attribute is used the dates can be mapped with the corresponding date in IMDI. IMDI:Session.Date would be chosen as creation date of the original media file. Others may become the creations dates of the media files and the annotation units.
Continent	Continent where the recording is made	-	-	OLAC:Coverage doesn't specify continent and therefore doesn't match.
Country	Country where the recording is made	Coverage	The extent or scope of the content of the resource	It might be possible to identify country codes in OLAC:Coverage, so these should be mapped to IMDI:Session.Country.
Region	Region where the recording is made	-	-	no mapping
Address	Sometimes it may be required to specify the address	-	-	no mapping
Description	Description of a session	-	-	It may be assumed that an OLAC:Description concerns the content of the resource, which maps

				with the IMDI:Content.Description. So this description can't be mapped.
Keys	Name-value pairs to describe domain specific information about the session	-	-	OLAC doesn't allow additional name-value pairs.
Project . Name	Short name or abbreviation of the project	-	-	There is no real equivalent in the OLAC set to describe the project. One could think of qualifying elements such as OLAC:Title, OLAC:Description, OLAC:Identifier; but it would mean to extend their semantics
Project . Title	Title of the project	-	-	,,
Project . Id	A unique identifier for the project	-	-	,,
Project . Contact	A relevant address associated with the project	-	-	,,
Project . Description +	A description associated with the project	-	-	,,
Collector . Name	The name of the person who was responsible for creating the resources in terms of recording and collecting. It is not meant to be person being interviewed etc.	Creator (refine = 'collector')	An entity primarily responsible for making the content of the resource.	Depends on the OLAC refine attribute which indicates the role of the creator. E.g. when refine is 'collector', this would map to IMDI:Session.Collector.Name.
Collector . Contact	The contact address of the collector	-	-	
Collector . Description +	A description associated with a collector	-	-	
Content . CommunicationContext . Interactivity	elements to allow to classify the content of a session (for details look in IMDI Session Metadata Set)	-	-	
Content . CommunicationContext . Planning Type		-	-	
Content . CommunicationContext . Involvement		-	-	

Content . Genre . Interactional		-	-	The best match for OLAC:Subject seems to be IMDI:Content.Genre.
Content . Genre . Discursive		-	-	
Content . Genre . Performance		-	-	
Content . Task		-	-	
Content . Modalities		-	-	
Content . Languages . Description	Description associated with the whole set of languages since their usage may share certain characteristics	-	-	
Content . Languages . Language . Id	a unique code for a given language being the subject of the recording	Subject.language (code attribute)	A language which the content of the resource describes or discusses.	The code attribute of OLAC:Subject.language maps with IMDI:Content.Languages.Language.Id
Content . Languages . Language . Name	a name in general used to indicate a language	Subject.language	A language which the content of the resource describes or discusses.	The OLAC:Subject.language maps with IMDI:Content.Languages.Language. Name.
Content . Languages . Language . Description	A free-text description associated with the languages used in the recording	-	-	
Content . Description	Free-text description of the content of the recording	Description	An account of the content of the resource.	The OLAC:Description element fits best with IMDI:Content.Description. Text; the corresponding IMDI:Content .Description.LanguageId maps with OLAC's lang attribute.
Content . Keys	List of attribute-value pairs to further describe the content	-	-	
Participants . Description	A free-text description associated with the participants occurring in the recording	-	-	
Participants . Participant . Type	a major classifier of the participants participating in a recording taken from a controlled vocabulary	Contributor (refine = <participant.type>)	An entity responsible for making contributions to the content of the resource.	An IMDI participant is a contributor to the content in OLAC's definition. The participant type can then be mapped with OLAC's refine attribute.

Participants . Participant . Name	Name of the person used in the annotations	-	-	
Participants . Participant . FullName	Full-name of the participant	Contributor (refine = <participant.type>)	An entity responsible for making contributions to the content of the resource.	In case of the right type in OLACs description a mapping is possible
Participants . Participant . Code	All elements further specify the participating researcher. IMDI people say that it is important to have these specifications often for quick inspection only.	-	-	
Participants . Participant . Role	an element to sub classify the role of the different interviewees participating in a recording	-	-	
Participants . Participant . Language . Id	Specifies a unique code to identify the language.	-	-	
Participants . Participant . Language . Name	A human understandable name of the language.	-	-	
Participants . Participant . Language . Description	Elaborate description of the language.	-	-	
Participants . Participant . EthnicGroup	The ethnic group of the participant	-	-	
Participants . Participant . Age	The age of the participant	-	-	
Participants . Participant . Sex	The sex of the participant.	-	-	
Participants . Participant . Education	The education of the participant.	-	-	
Participants . Participant . Anonymous	Indicates whether or not the participant name and full name are replaced by pseudo names to make him/her anonymous.	-	-	

Participants . Participant . Description	A description of specific information about the participant.	-	-	
Participants . Participant . Keys	Mechanism to extend the description of the participant by attribute-value pairs	-	-	
Resources . MediaFile . ResourceLink	Unique identifier which normally is an URL	Identifier	An unambiguous reference to the resource within a given context.	the OLAC:Identifier element points to the media file so it can be mapped with the corresponding IMDI element
Resources . MediaFile . Size	the size of the session's media file	-	-	
Resources . MediaFile . Type	the type of the session's media file such as Photo, Audio, Video	Format	The physical or digital manifestation of the resource.	The OLAC element can be mapped to the two IMDI elements after detailed inspection
Resources . MediaFile . Format	the format of the session's media file such as mpg, jpg, wav, …			
Resources . MediaFile . Quality	the quality of the session's media file in global terms taken from a controlled vocabulary	-	-	
Resources . MediaFile . RecordingConditions	element to describe the recording conditions (amplifier, microphone, ...)	-	-	
Resources . MediaFile . Position	the start and stop time references of the session's media file with respect to its original material	-	-	
Resources . MediaFile . Access	Structured element to describe the access rights of the media file	Publisher Rights	An entity responsible for making the resource available; Information about rights held in and over the resource.	the elements of the IMDI sub-structure can be mapped with two elements found in the OLAC description, the mapping has to decide whether the name mentioned will be mapped with IMDI-Publisher or IMDI-Owner (the first is preferred)
Resources . MediaFile . Description	description associated with the media file	Description	An account of the content of the resource.	the OLAC:Description field of the Media File MD description can be mapped with this IMDI element
Resources . AnnotationUnit .	Unique identifier which normally is an URL	-	-	

ResourceLink				
Resources . AnnotationUnit . MediaResourceLink	Unique identifier to identify the media file which was used for the annotation	-	-	
Resources . AnnotationUnit . Annotator	The person who did the annotations	-	-	
Resources . AnnotationUnit . Date	Date when a certain annotation tier was created	Date	A date associated with an event in the life cycle of the resource.	The date (with creation qualifier) to be found in the OLAC MD description of each annotation file can be mapped to this IMDI element
Resources . AnnotationUnit . Type	the type of the session's annotation file such as orthographic, phonetic, morphologic, syntactic, translation, …	Type.data	The nature or genre of the content of the resource from a linguistic standpoint.	one to one match with OLAC
Resources . AnnotationUnit . Format	the format of the session's annotation file such as CHAT, Shoebox, AIF, …	Format	The physical or digital manifestation of the resource.	both elements are supposed to contain a format specifier which allows a one to one match
Resources . AnnotationUnit . ContentEncoding	document which inguistic type of encoding is used	Format.markup	A markup scheme used by a digital resource.	One to one match with OLAC.
Resources . AnnotationUnit . CharacterEncoding	document which character encoding is used	Format.encoding	An encoded character set used by a digital resource.	One to one match with OLAC.
Resources . AnnotationUnit . Access	Structured element to describe the access rights of the annotations	Publisher Rights	An entity responsible for making the resource available; Information about rights held in and over the resource.	the elements of the IMDI sub-structure can be mapped with two elements found in the OLAC description, the mapping has to decide whether the name mentioned will be mapped with IMDI-Publisher or IMDI-Owner (the first is preferred)
Resources . AnnotationUnit . LanguageId	The languages the annotations are written in; Can be several	Language	A language of the intellectual content of the resource.	one to one mapping with OLAC
Resources . AnnotationUnit . Anonymous	Just a switch to indicate whether real names have to be replaced by pseudo names	-	-	
Resources . AnnotationUnit .	a description associated with the annotation unit	Description	An account of the content of the resource.	the description of the OLAC MD file associated with the annotations could

Description				be mapped with this IMDI element
Resources . Source . Id	Reference to a specific tape with a unique label	Identifier	An unambiguous reference to the resource within a given context.	if there is a OLAC MD description for the original recording which does not have a URI then the pointer could be used to map with this IMDI element
Resources . Source . Format	Element characterizing the media format such as DAT, DV, VHS, Hi-8, ...	-	-	
Resources . Source . Quality	Element characterizing the quality of the signals on the original material taken from a controlled vocabulary	-	-	
Resources . Source . Position		-	-	
Resources . Source . Access	Structured element to describe the access rights of the original media carriers	Publisher Rights	An entity responsible for making the resource available; Information about rights held in and over the resource.	the elements of the IMDI sub-structure can be mapped with two elements found in the OLAC description, the mapping has to decide whether the name mentioned will be mapped with IMDI-Publisher or IMDI-Owner (the first is preferred)
Resources . Source . Description	a description associated with the original recording media set	Description	An account of the content of the resource.	the description of the OLAC MD file associated with the original resource could be mapped with this IMDI element
References . Description	Descriptive element to point to all sort of related documents	Relation	A reference to a related resource.	OLAC has the OLAC:Relation element to enter such references, i.e. these can be mapped to the IMDI:References.Description element

Appendix 5

ISLE Metadata Initiative (IMDI)

PART 3 A

Vocabulary Taxonomy and Structure

Draft Proposal Version 1.1

December, 2001

IMDI[24] Technical Report

Max-Planck-Institute for Psycholinguistics
NL, Nijmegen

[24] For information about the ISLE Metadata Initiative, please, look at the following web-site: www.mpi.nl/ISLE

INDEX

1 INTRODUCTION 121

1.1 TAXONOMY OF VOCABULARIES 121

2 REQUIRED FUNCTIONALITY 122

2.1 INFRASTRUCTURE 122
2.2 VOCABULARY STRUCTURE 123
2.3 IMPLEMENTATION OF CV'S IN IMDI TOOLS 123

APPENDIX A 124

1 Introduction

The term vocabulary as we use it in the IMDI documentation should not be confused with the term "metadata vocabulary". The last term refers to the total set of metadata elements defined for a specific domain or application. Whenever we use that concept we will write "metadata vocabulary" in full. With "vocabulary" we mean the set of string values that can be used to assign a value to a specific metadata element or attribute of an element.

1.1 Taxonomy of vocabularies

We distinguish four classes of controlled vocabularies:

- *Closed Controlled vocabulary*: The value of the metadata element is one and only one element from a finite set of values.
- *Closed Controlled vocabulary list:* The value of the metadata element is a list with a number of elements from a finite set of values.
- *Open Controlled vocabulary*: The value of the metadata element is one and only one element from a finite set of elements **or** is a user specified string.
- *Open Controlled vocabulary list*: The value of the metadata element is a list with a number of elements from a finite set of values **or / and** a number of user specified strings

The "Open CV" and "Open CV list" are there to advise users and not constrain them. This is a requirement for some metadata elements where some users were quite specific about. It is a matter of taste if we allow a metadata element to have a CV list as a value or allow the metadata element to be repeated. We chose the first approach for compactness sake.

If a vocabulary is a list, open or closed is dependent on the context, application or metadata schema itself, it is not something that is determined by the vocabulary itself. It is very well conceivable that a specific vocabulary is a closed CV for one metadata element and an open CV list for another. So we will specify in the IMDI XML-Schema what type of vocabulary a metadata element is connected to when applicable for the element.

2 Required functionality

For the IMDI metadata set "we have and still are refining" a number of vocabularies that need to be flexible and dynamic. That means that the definition of the vocabulary should be defined in a separate file and can be stored on a central vocabulary server. We have already developed a tentative XML schema for such a vocabulary definition and as a transmission protocol HTTP seems sufficient. Efficiency is important because sometimes "large" vocabularies have to be transferred (think of the set of "language identifiers"). Tools using these vocabularies would also need a caching mechanism for speed and offline work.

Vocabularies are downloadable. After they have been downloaded an application should be able to check if the vocabulary is still up to date. Therefore a vocabulary definition needs elements defining the creation date and an URL link to its origin.

2.1 Infrastructure

The important vocabularies that are defined by the IMDI standard should all be available from a central repository server and these definitions should be well maintained by a central authority. However IMDI tools should also be configurable in such a way that a user can link the free definable key/value pairs that are available at several levels within the IMDI session descriptions to specific project bound CV definitions. These CV's can be available on local servers or on the local file system.

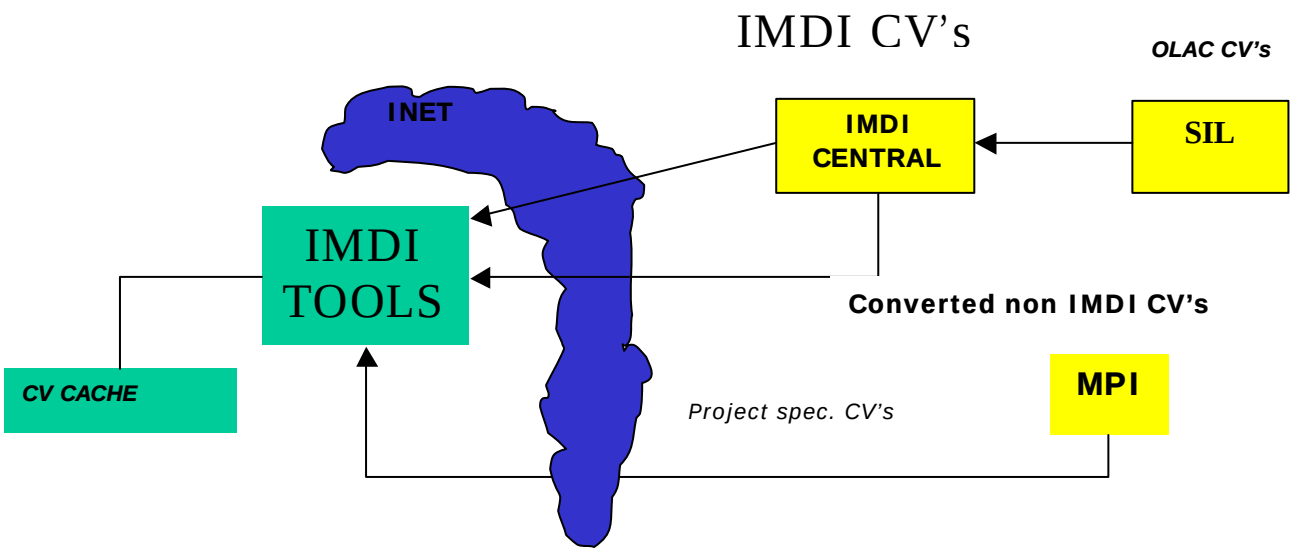


Figure 1 IMDI tools using CV's from different sources

For accessing vocabulary servers that offer vocabularies in non-IMDI formats a bridge could be created in the form of an XSL converter on the central IMDI site (see Figure 1).

2.2 Vocabulary Structure

We have chosen for a definition of vocabularies in the form of an XML file that is an instantiation of the XML-Schema shown here:

```
<xsd:complexType name="VocabularyDefType">
   <xsd:annotation>
       <xsd:documentation>
    The definition of a vocabulary. Attributes: Date of creattion, Link to
    origin. Contails a Description be element to describe the domain of the
    vocabularyand a (unspecified) number of value enries
       </xsd:documentation>
      </xsd:annotation>
      <xsd:sequence>
          <xsd:element ref="imdi:Description” Type maxOccurs=”unbounded"/>
          <xsd:element name="Entry" maxOccurs="unbounded">
              <xsd:complexType>
                  <xsd:simpleContent>
                      <xsd:extension base="xsd:string">
                          <xsd:attribute name="Value" type="xsd:string"/>
                      </xsd:extension>
                  </xsd:simpleContent>
              </xsd:complexType>
          </xsd:element>
      </xsd:sequence>
      <xsd:attribute name="Name" type="xsd:string" use="required"/>
      <xsd:attribute name="Date" type="xsd:date" use="required"/>
      <xsd:attribute name="Link" type="xsd:urlRef" use="required"/>
  </xsd:complexType>
```

Another possibility is to define every vocabulary with its own XML-Schema but then we would obliged to define all mappings between metadata elements and the corresponding vocabularies in the IMDI schema itself, loosing flexibility. As a disadvantage we loose the possibility of having the XML parser check the validity of fixed mappings between metadata elements and corresponding vocabulary. However as stated above vocabularies are often not fixed so that the connection between metadata element and vocabulary can not be defined in the IMDI schema but only in an instantiation of that schema, this is not considered a big disadvantage.

2.3 Implementation of CV’s in IMDI tools

At the moment all IMDI Tools, the IMDI-BCBrowser, IMDI-BCEditor and the IMDI-BCSearch tools implement the use of CV’s although not in a completely user configurable fashion

Appendix A

Example of a continent vocabulary

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- edited with XML Spy v3.5 NT (http://www.xmlspy.com) by Daan Broeder
(Max-Planck Institute for Psycholinguistics) -->
<imdi:VocabularyDef xmlns:imdi="http://www.mpi.nl/IMDI/Schema/IMDI.xsd"
xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI.xsd
D:\users\BROEDER\Documents\DOC\LAPP\ISLE\IMDI.xsd" Name="Continents"
Date="2001-05-06" Origin="httpts://www.mpi.nl/IMDI/Schema/Continents.xml">
    <imdi:Description>
        <Text Language="SIL:xxx">List of linguistic continents </Text>
        <Text Language="SIL:xxx"
link="http://www.mpi.nl/IMDI/Documents/Continents.html" />
    </imdi:Description>
    <Entry Value="Africa"/>
    <Entry Value="Asia"/>
    <Entry Value="America-North"> Not a real continent </Entry>
    <Entry Value="America-South">Not a real continent </Entry>
    <Entry Value="Europe"/>
    <Entry Value="Australia"/>
</imdi:VocabularyDef>
```

Appendix 6

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v4.2 U (http://www.xmlspy.com) by Daan
Broeder (Max-Planck Institute for Psycholinguistics) -->

<xsd:schema targetNamespace="http://www.mpi.nl/IMDI/Schema/IMDI"
xmlns:imdi="http://www.mpi.nl/IMDI/Schema/IMDI"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" attributeFormDefault="unqualified"
version="1">
      <xsd:element name="METATRANSCRIPT">
            <xsd:annotation>
                  <xsd:documentation>The outer element with
administrative data of all metadata description files. Version 1.0 is
based on Session description version 2.5 and Catalogue description
2.1</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType mixed="false">
                  <xsd:choice>
                        <xsd:element name="Session"
type="imdi:SessionType" maxOccurs="unbounded"/>
                        <xsd:element name="Corpus"
maxOccurs="unbounded">
                              <xsd:annotation>
                                    <xsd:documentation>Content
specification for CORPUS has yet to be finished</xsd:documentation>
                              </xsd:annotation>
                              <xsd:complexType>
                                    <xsd:complexContent>
                                          <xsd:extension
base="imdi:CorpusType"/>
                                    </xsd:complexContent>
                              </xsd:complexType>
                        </xsd:element>
                        <xsd:element name="Catalogue"
type="imdi:CatalogueType"/>
                  </xsd:choice>
                  <xsd:attribute name="Date" type="xsd:date"
use="required"/>
                  <xsd:attribute name="Originator" type="xsd:string"
use="optional"/>
                  <xsd:attribute name="Version" type="xsd:string"
use="required"/>
                  <xsd:attribute name="FormatId" type="xsd:string"
use="required"/>
                  <xsd:attribute name="History" type="xsd:anyURI"
use="optional"/>
                  <xsd:attribute name="Type"
type="imdi:MetatranscriptType" use="required"/>
            </xsd:complexType>
      </xsd:element>
      <xsd:complexType name="NamedLinkType">
            <xsd:annotation>
                  <xsd:documentation>link to other resource.
Attribute name is for the benefit of browsing</xsd:documentation>
            </xsd:annotation>
            <xsd:simpleContent>
                  <xsd:extension base="xsd:anyURI">
```

```
                        <xsd:attribute name="Name" type="xsd:string"
use="required"/>
                  </xsd:extension>
            </xsd:simpleContent>
      </xsd:complexType>
      <xsd:complexType name="DescriptionType" block="extension"
mixed="false">
            <xsd:annotation>
                  <xsd:documentation>Human readable description in
the form of a text with language id specification and/or a link to a
file with a description and  language id specification. The name
attribute is to name the link (if present)</xsd:documentation>
            </xsd:annotation>
            <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                        <xsd:attribute name="LanguageId"
type="imdi:LanguageIdType" use="optional"/>
                        <xsd:attribute name="Name" type="xsd:string"
use="optional"/>
                        <xsd:attribute name="Link" type="xsd:anyURI"
use="optional"/>
                  </xsd:extension>
            </xsd:simpleContent>
      </xsd:complexType>
      <xsd:complexType name="ContactType">
            <xsd:annotation>
                  <xsd:documentation>Contact information for this
data</xsd:documentation>
            </xsd:annotation>
            <xsd:sequence>
                  <xsd:element name="Name" type="xsd:string"
minOccurs="0"/>
                  <xsd:element name="Address" type="xsd:string"
minOccurs="0"/>
                  <xsd:element name="Email" type="xsd:string"
minOccurs="0"/>
                  <xsd:element name="Organisation" type="xsd:string"
minOccurs="0"/>
            </xsd:sequence>
      </xsd:complexType>
      <xsd:complexType name="LocationType">
            <xsd:annotation>
                  <xsd:documentation>Information on creation location
for this data</xsd:documentation>
            </xsd:annotation>
            <xsd:sequence>
                  <xsd:element name="Continent"
type="imdi:Vocabulary">
                        <xsd:annotation>
                              <xsd:documentation>The continent where
the session/corpus was recorded or originated</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Country" type="imdi:Vocabulary">
                        <xsd:annotation>
                              <xsd:documentation>The country where
the session/corpus was recorded or originated</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Region" type="xsd:string"
minOccurs="0" maxOccurs="unbounded">
```

```
                        <xsd:annotation>
                              <xsd:documentation>The region or sub-
region where the session/corpus was recorded or
originated</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Address" type="xsd:string"
minOccurs="0">
                        <xsd:annotation>
                              <xsd:documentation>The address where
the session/corpus was recorded or originated</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
            </xsd:sequence>
      </xsd:complexType>
      <xsd:element name="Keys">
            <xsd:annotation>
                  <xsd:documentation>List of a number of key name
value pairs. Should be used to add information that is not covered by
other metadata elements at this level</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType mixed="false">
                  <xsd:sequence>
                        <xsd:element name="Key" minOccurs="0"
maxOccurs="unbounded">
                              <xsd:complexType>
                                    <xsd:simpleContent>
                                          <xsd:extension
base="imdi:Vocabulary">
                                                <xsd:attribute
name="Name" type="xsd:string" use="required"/>
                                          </xsd:extension>
                                    </xsd:simpleContent>
                              </xsd:complexType>
                        </xsd:element>
                  </xsd:sequence>
            </xsd:complexType>
      </xsd:element>
      <xsd:element name="Language">
            <xsd:annotation>
                  <xsd:documentation>A list of languages used in the
session</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType mixed="false">
                  <xsd:sequence>
                        <xsd:element name="Id"
type="imdi:LanguageIdType" maxOccurs="unbounded">
                              <xsd:annotation>
                                    <xsd:documentation>Unique code to
identify a language</xsd:documentation>
                              </xsd:annotation>
                        </xsd:element>
                        <xsd:element name="Name"
maxOccurs="unbounded" type="imdi:Vocabulary">
                              <xsd:annotation>
                                    <xsd:documentation>A list of
human understandable names to identify a language</xsd:documentation>
                              </xsd:annotation>
                        </xsd:element>
                        <xsd:element name="Description"
type="imdi:DescriptionType" minOccurs="0" maxOccurs="unbounded">
```

```
                                <xsd:annotation>
                                      <xsd:documentation>Description
for this particular language.</xsd:documentation>
                                </xsd:annotation>
                          </xsd:element>
                    </xsd:sequence>
              </xsd:complexType>
        </xsd:element>
        <xsd:complexType name="AccessType">
              <xsd:annotation>
                    <xsd:documentation>Groups information about access
rights for this data</xsd:documentation>
              </xsd:annotation>
              <xsd:sequence>
                    <xsd:element name="Availability" type="xsd:string">
                          <xsd:annotation>
                                <xsd:documentation>Availability of the
data. OpenCV</xsd:documentation>
                          </xsd:annotation>
                    </xsd:element>
                    <xsd:element name="Date" type="imdi:DateType">
                          <xsd:annotation>
                                <xsd:documentation>Date when access
rights were evaluated</xsd:documentation>
                          </xsd:annotation>
                    </xsd:element>
                    <xsd:element name="Owner" type="xsd:string">
                          <xsd:annotation>
                                <xsd:documentation>Name of owner
resource</xsd:documentation>
                          </xsd:annotation>
                    </xsd:element>
                    <xsd:element name="Publisher" type="xsd:string">
                          <xsd:annotation>
                                <xsd:documentation>Publisher
responsible for distribution of this data</xsd:documentation>
                          </xsd:annotation>
                    </xsd:element>
                    <xsd:element name="Contact"
type="imdi:ContactType"/>
                    <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded"/>
              </xsd:sequence>
        </xsd:complexType>
        <xsd:complexType name="ProjectType">
              <xsd:annotation>
                    <xsd:documentation>Project
Information</xsd:documentation>
              </xsd:annotation>
              <xsd:sequence>
                    <xsd:element name="Name" type="xsd:string">
                          <xsd:annotation>
                                <xsd:documentation>A short name or
abbreviation for the project</xsd:documentation>
                          </xsd:annotation>
                    </xsd:element>
                    <xsd:element name="Title" type="xsd:string">
                          <xsd:annotation>
                                <xsd:documentation>The full title of
the project</xsd:documentation>
                          </xsd:annotation>
```

```
                  </xsd:element>
                  <xsd:element name="Id" type="imdi:DescriptionType"
maxOccurs="unbounded">
                        <xsd:annotation>
                              <xsd:documentation>A unique identifier
for the project</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Contact"
type="imdi:ContactType">
                        <xsd:annotation>
                              <xsd:documentation>Contact information
for this project</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded">
                        <xsd:annotation>
                              <xsd:documentation>Description for this
project</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
            </xsd:sequence>
      </xsd:complexType>
      <xsd:element name="CounterPosition">
            <xsd:annotation>
                  <xsd:documentation>Position (start (+end) ) on a
oldfashioned tape without time indication</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType mixed="false">
                  <xsd:attribute name="Start"
type="imdi:CounterPositionType" use="required"/>
                  <xsd:attribute name="End"
type="imdi:CounterPositionType" use="optional"/>
            </xsd:complexType>
      </xsd:element>
      <xsd:element name="TimePosition">
            <xsd:annotation>
                  <xsd:documentation>Position in a media file or
modern tape</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType mixed="false">
                  <xsd:attribute name="Start"
type="imdi:TimePositionType" use="required"/>
                  <xsd:attribute name="End"
type="imdi:TimePositionType" use="optional"/>
            </xsd:complexType>
      </xsd:element>
      <xsd:element name="Languages">
            <xsd:annotation>
                  <xsd:documentation>Groups information about the
languages used in the session</xsd:documentation>
            </xsd:annotation>
            <xsd:complexType mixed="false">
                  <xsd:sequence>
                        <xsd:element name="Description"
type="imdi:DescriptionType" minOccurs="0" maxOccurs="unbounded">
                              <xsd:annotation>
                                    <xsd:documentation>Description
for the list of languages spoken by this
participant</xsd:documentation>
```

```
                                </xsd:annotation>
                        </xsd:element>
                        <xsd:element ref="imdi:Language"
minOccurs="0" maxOccurs="unbounded"/>
                  </xsd:sequence>
            </xsd:complexType>
      </xsd:element>
      <xsd:simpleType name="LanguageIdType">
            <xsd:annotation>
                  <xsd:documentation>Language identification either
ISO or SIL</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:pattern value="(ISO639(-1|-2)?:.*)?"/>
                  <xsd:pattern value="(RFC1766:.*)?"/>
                  <xsd:pattern value="(SIL:.*)?"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="emptyString">
            <xsd:annotation>
                  <xsd:documentation>empty string
definition</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:maxLength value="0"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="EmptyType">
            <xsd:annotation>
                  <xsd:documentation>empty type
definition</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:enumeration value="Unknown"/>
                  <xsd:enumeration value="Unspecified"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="TimePositionType">
            <xsd:annotation>
                  <xsd:documentation>Time position in the hh:mm:ss:ff
format</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:pattern value="Unknown|[0-9][0-9]:[0-9][0-
9]:[0-9][0-9](:[0-9]+)?|Unspecified"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="CounterPositionType">
            <xsd:annotation>
                  <xsd:documentation>Counter position in the  nnn
format</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:pattern value="[0-9]*|Unknown|Unspecified"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="VocabularyType">
            <xsd:annotation>
                  <xsd:documentation>specifies the four vocabulary
types</xsd:documentation>
            </xsd:annotation>
```

```
            <xsd:restriction base="xsd:string">
                  <xsd:enumeration value="ClosedVocabulary"/>
                  <xsd:enumeration value="ClosedVocabularyList"/>
                  <xsd:enumeration value="OpenVocabulary"/>
                  <xsd:enumeration value="OpenVocabularyList"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="VocabularyRefType">
            <xsd:annotation>
                  <xsd:documentation>Pointer to a vocabulary
definition</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:anyURI"/>
      </xsd:simpleType>
      <xsd:complexType name="Vocabulary">
            <xsd:annotation>
                  <xsd:documentation>value for an element/attribute
that is a vocabulary</xsd:documentation>
            </xsd:annotation>
            <xsd:simpleContent>
                  <xsd:extension base="imdi:CVSstring">
                        <xsd:attribute name="Type"
type="imdi:VocabularyType" default="OpenVocabularyList"/>
                        <xsd:attribute name="DefaultLink"
type="imdi:VocabularyRefType" use="optional"/>
                        <xsd:attribute name="Link"
type="imdi:VocabularyRefType" use="optional"/>
                  </xsd:extension>
            </xsd:simpleContent>
      </xsd:complexType>
      <xsd:complexType name="AgeType">
            <xsd:annotation>
                  <xsd:documentation>Specifies age of a person with
differerent counting methods</xsd:documentation>
            </xsd:annotation>
            <xsd:simpleContent>
                  <xsd:extension base="imdi:AgeValueType">
                        <xsd:attribute name="AgeCountingMethod"
type="imdi:AgeCountingMethodType" default="SinceBirth"/>
                  </xsd:extension>
            </xsd:simpleContent>
      </xsd:complexType>
      <xsd:simpleType name="AgeCountingMethodType">
            <xsd:annotation>
                  <xsd:documentation>The counting
method</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:enumeration value="SinceConception"/>
                  <xsd:enumeration value="SinceBirth"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="AgeValueType">
            <xsd:annotation>
                  <xsd:documentation>The age of a
person</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:pattern value="Unknown|Unspecified|[0-9]+(;[0-
1]?[0-9](\.[0-3]?[0-9])?)?"/>
            </xsd:restriction>
```

```
      </xsd:simpleType>
      <xsd:complexType name="VocabularyDefType" mixed="false">
            <xsd:annotation>
                  <xsd:documentation>The definition of a vocabulary.
Attributes: Date of creattion, Link to origin. Contails a Description
be element to describe the domain of the vocabularyand a
(unspecified) number of value enries</xsd:documentation>
            </xsd:annotation>
            <xsd:sequence>
                  <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded"/>
                  <xsd:element name="Entry" maxOccurs="unbounded">
                        <xsd:complexType>
                              <xsd:simpleContent>
                                    <xsd:extension base="xsd:string">
                                          <xsd:attribute name="Tag"
type="xsd:string" use="optional"/>
                                          <xsd:attribute name="Value"
type="xsd:string" use="required"/>
                                    </xsd:extension>
                              </xsd:simpleContent>
                        </xsd:complexType>
                  </xsd:element>
            </xsd:sequence>
            <xsd:attribute name="Name" type="xsd:string"
use="required"/>
            <xsd:attribute name="Date" type="xsd:date"
use="required"/>
            <xsd:attribute name="Tag" type="xsd:date"
use="optional"/>
            <xsd:attribute name="Link" type="imdi:VocabularyRefType"
use="required"/>
      </xsd:complexType>
      <xsd:element name="VocabularyDef"
type="imdi:VocabularyDefType">
            <xsd:annotation>
                  <xsd:documentation>Instantiation of a
VocabularyDefType</xsd:documentation>
            </xsd:annotation>
      </xsd:element>
      <xsd:simpleType name="CVSstring">
            <xsd:annotation>
                  <xsd:documentation>Comma seperated
string</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:pattern value="[^,]*(,[^,]+)*"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:complexType name="MDGroupType">
            <xsd:annotation>
                  <xsd:documentation>Type for group of metadata
pertaining to a session</xsd:documentation>
            </xsd:annotation>
            <xsd:sequence>
                  <xsd:element name="Location"
type="imdi:LocationType">
                        <xsd:annotation>
                              <xsd:documentation>Groups information
about the location where the session was created</xsd:documentation>
                        </xsd:annotation>
```

```
                  </xsd:element>
                  <xsd:element ref="imdi:Keys"/>
                  <xsd:element name="Project"
type="imdi:ProjectType">
                        <xsd:annotation>
                              <xsd:documentation>Groups information
about the project for which the session was (originally)
created</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Collector">
                        <xsd:annotation>
                              <xsd:documentation>Groups information
about the collector of the session. This is the person responsible
for the collection of the sesion data.</xsd:documentation>
                        </xsd:annotation>
                        <xsd:complexType>
                              <xsd:sequence>
                                    <xsd:element name="Name"
type="xsd:string">
                                          <xsd:annotation>

      <xsd:documentation>The name of the person  responsible for
creating the session</xsd:documentation>
                                          </xsd:annotation>
                                    </xsd:element>
                                    <xsd:element name="Contact"
type="imdi:ContactType">
                                          <xsd:annotation>

      <xsd:documentation>Contact information for this
collector</xsd:documentation>
                                          </xsd:annotation>
                                    </xsd:element>
                                    <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded">
                                          <xsd:annotation>

      <xsd:documentation>Description for this
collector</xsd:documentation>
                                          </xsd:annotation>
                                    </xsd:element>
                              </xsd:sequence>
                        </xsd:complexType>
                  </xsd:element>
                  <xsd:element name="Content">
                        <xsd:annotation>
                              <xsd:documentation>Groups information
about the content of the session. The content description takes place
in several (overlapping) dimensions</xsd:documentation>
                        </xsd:annotation>
                        <xsd:complexType>
                              <xsd:sequence>
                                    <xsd:element name="Task"
type="imdi:Vocabulary">
                                          <xsd:annotation>

      <xsd:documentation>List of he major tasks carried out in the
session. OVList</xsd:documentation>
                                          </xsd:annotation>
                                    </xsd:element>
```

```
                                    <xsd:element name="Modalities"
type="imdi:Vocabulary">
                                          <xsd:annotation>

      <xsd:documentation>List of modalities used in the session.
OVList</xsd:documentation>
                                          </xsd:annotation>
                                    </xsd:element>
                                    <xsd:element
name="CommunicationContext">
                                          <xsd:annotation>

      <xsd:documentation>This groups information concerning the
context of communication</xsd:documentation>
                                          </xsd:annotation>
                                          <xsd:complexType>
                                                <xsd:sequence>
                                                      <xsd:element
name="Interactivity" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>degree of interactivity.
CCV</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="PlanningType" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Degree of planning of the event.
CCV</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Involvement" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Degree of involvement of the researcher.
CCV</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                <xsd:sequence>
                                          <xsd:complexType>
                                    <xsd:element>
                                    <xsd:element name="Genre">
                                          <xsd:annotation>

      <xsd:documentation>Groups information of the conventionalised
discourse types of the session content</xsd:documentation>
                                          </xsd:annotation>
                                          <xsd:complexType>
                                                <xsd:sequence>
                                                      <xsd:element
name="Interactional" type="imdi:Vocabulary">
```

```
      <xsd:annotation>

      <xsd:documentation>Lists the interactional genres of the
session content. OVList</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="Discursive" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Lists the discursive  genres of the session
content. OVList</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="Performance" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Lists the performance genres of the session
content. OVList</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                  </xsd:sequence>
                                            </xsd:complexType>
                                      </xsd:element>
                                      <xsd:element
ref="imdi:Languages"/>
                                      <xsd:element ref="imdi:Keys"/>
                                      <xsd:element name="Description"
type="imdi:DescriptionType" minOccurs="0" maxOccurs="unbounded">
                                            <xsd:annotation>

      <xsd:documentation>Description for the content of this
session</xsd:documentation>
                                            </xsd:annotation>
                                      </xsd:element>
                               </xsd:sequence>
                         </xsd:complexType>
                   </xsd:element>
                   <xsd:element name="Participants">
                         <xsd:annotation>
                               <xsd:documentation>Groups information
about all participants in the session</xsd:documentation>
                         </xsd:annotation>
                         <xsd:complexType>
                               <xsd:sequence>
                                      <xsd:element name="Description"
type="imdi:DescriptionType" minOccurs="0" maxOccurs="unbounded">
                                            <xsd:annotation>

      <xsd:documentation>Description pertaining to all the
participants together</xsd:documentation>
                                            </xsd:annotation>
                                      </xsd:element>
```

```
                                    <xsd:element name="Participant"
minOccurs="0" maxOccurs="unbounded">
                                          <xsd:complexType>
                                                <xsd:sequence>
                                                      <xsd:element
name="Type" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Type or functional role of participant e.g.
consultant, contributor, interviewer, researcher.
OV</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Name" type="xsd:string" maxOccurs="unbounded">

      <xsd:annotation>

      <xsd:documentation>Name of the participant as used by others in
the transcription</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="FullName" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>Official name of the
participant</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Code" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>Short unique code to identify the
participant as used in the transcriptor in the transcription
</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Role" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>The role of the participant in the session
within the context of informant. OVList</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
ref="imdi:Languages"/>
                                                      <xsd:element
name="EthnicGroup" type="imdi:CVSstring">
```

```
      <xsd:annotation>

      <xsd:documentation>List of ethnic groups of participant.
OVList</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="Age" type="imdi:AgeType">

      <xsd:annotation>

      <xsd:documentation>Age of the participant in CHAT format
yy;mm.dd</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="Sex">

      <xsd:annotation>

      <xsd:documentation>Sex of the participant.
CV</xsd:documentation>

      </xsd:annotation>

      <xsd:simpleType>

      <xsd:restriction base="xsd:string">

      <xsd:enumeration value="Male"/>

      <xsd:enumeration value="Female"/>

      <xsd:enumeration value="Unknown"/>

      <xsd:enumeration value="Unspecified"/>

      </xsd:restriction>

      </xsd:simpleType>
                                                          </xsd:element>
                                                          <xsd:element
name="Education" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>The eduucation of the participant. Can slo
be used to specify litteracy</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="Anonymous" type="xsd:boolean">

      <xsd:annotation>

      <xsd:documentation>Indicated if true names were used or that
codes were employed. CV  boolean</xsd:documentation>
```

```
      </xsd:annotation>
                                                            </xsd:element>
                                                            <xsd:element
ref="imdi:Keys"/>
                                                            <xsd:element
name="Description" type="imdi:DescriptionType" maxOccurs="unbounded">

      <xsd:annotation>

      <xsd:documentation>Description for this individual
participant</xsd:documentation>

      </xsd:annotation>
                                                            </xsd:element>
                                                      </xsd:sequence>
                                                </xsd:complexType>
                                          </xsd:element>
                                    </xsd:sequence>
                              </xsd:complexType>
                        </xsd:element>
                        <xsd:element name="References" minOccurs="0">
                              <xsd:annotation>
                                    <xsd:documentation>Groups information
about external documentation associated with this
session</xsd:documentation>
                              </xsd:annotation>
                              <xsd:complexType>
                                    <xsd:sequence>
                                          <xsd:element name="Description"
type="imdi:DescriptionType" minOccurs="0" maxOccurs="unbounded">
                                                <xsd:annotation>

      <xsd:documentation>Every description is a
reference</xsd:documentation>
                                                </xsd:annotation>
                                          </xsd:element>
                                    </xsd:sequence>
                              </xsd:complexType>
                        </xsd:element>
            </xsd:sequence>
      </xsd:complexType>
      <xsd:complexType name="CorpusType">
            <xsd:annotation>
                  <xsd:documentation>Type for a corpus that points to
either other corpora or sessions</xsd:documentation>
            </xsd:annotation>
            <xsd:sequence>
                  <xsd:element name="Name" type="xsd:string">
                        <xsd:annotation>
                              <xsd:documentation>name of the (sub-
)corpus</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Title" type="xsd:string">
                        <xsd:annotation>
                              <xsd:documentation>Title for the (sub-
)corpus</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
```

```
                  <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded"/>
                  <xsd:element name="MDGroup" type="imdi:MDGroupType"
minOccurs="0"/>
                  <xsd:element name="CorpusLink"
type="imdi:NamedLinkType" minOccurs="0" maxOccurs="unbounded"/>
            </xsd:sequence>
            <xsd:attribute name="SearchService" type="xsd:anyURI"
use="optional"/>
            <xsd:attribute name="CorpusStructureService"
type="xsd:anyURI" use="optional"/>
            <xsd:attribute name="CatalogueLink" type="xsd:anyURI"
use="optional"/>
      </xsd:complexType>
      <xsd:complexType name="CatalogueType">
            <xsd:annotation>
                  <xsd:documentation>Type for group metadata
pertaining to published corpora</xsd:documentation>
            </xsd:annotation>
            <xsd:sequence>
                  <xsd:element name="Name" type="xsd:string"/>
                  <xsd:element name="Title" type="xsd:string"/>
                  <xsd:element name="Id" type="xsd:string"
maxOccurs="unbounded"/>
                  <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded"/>
                  <xsd:element name="Language"
type="imdi:LanguageIdType">
                        <xsd:annotation>
                              <xsd:documentation>Language the corpus
is described/annotated in</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="SubjectLanguage"
type="imdi:LanguageIdType">
                        <xsd:annotation>
                              <xsd:documentation>Language the corpus
is about</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Location"
type="imdi:LocationType"/>
                  <xsd:element name="ContentType"
type="imdi:Vocabulary"/>
                  <xsd:element name="Format">
                        <xsd:complexType>
                              <xsd:all>
                                    <xsd:element name="Text">
                                          <xsd:complexType>
                                                <xsd:all>
                                                      <xsd:element
ref="imdi:Keys"/>
                                                </xsd:all>
                                          </xsd:complexType>
                                    </xsd:element>
                                    <xsd:element name="Audio">
                                          <xsd:complexType>
                                                <xsd:all>
                                                      <xsd:element
ref="imdi:Keys"/>
                                                </xsd:all>
```

```
                                        </xsd:complexType>
                                  </xsd:element>
                                  <xsd:element name="Video">
                                        <xsd:complexType>
                                              <xsd:all>
                                                    <xsd:element
ref="imdi:Keys"/>
                                              </xsd:all>
                                        </xsd:complexType>
                                  </xsd:element>
                            </xsd:all>
                      </xsd:complexType>
                </xsd:element>
                <xsd:element name="Quality">
                      <xsd:complexType>
                            <xsd:all>
                                  <xsd:element name="Audio"
type="imdi:QualityType"/>
                                  <xsd:element name="Video"
type="imdi:QualityType"/>
                            </xsd:all>
                      </xsd:complexType>
                </xsd:element>
                <xsd:element name="SmallestAnnotationUnit"
type="imdi:Vocabulary"/>
                <xsd:element name="Applications"
type="imdi:Vocabulary"/>
                <xsd:element name="Date" type="imdi:DateType"/>
                <xsd:element name="Project"
type="imdi:ProjectType"/>
                <xsd:element name="Publisher" type="xsd:string"/>
                <xsd:element name="Authors" type="imdi:CVSstring"/>
                <xsd:element name="Size" type="xsd:string">
                      <xsd:annotation>
                            <xsd:documentation>Human readabusle
string that indicates total size of corpus</xsd:documentation>
                      </xsd:annotation>
                </xsd:element>
                <xsd:element name="DistributionForm"
type="imdi:Vocabulary"/>
                <xsd:element name="Access" type="imdi:AccessType"/>
                <xsd:element name="Pricing" type="xsd:string"/>
            </xsd:sequence>
      </xsd:complexType>
      <xsd:simpleType name="DateType">
            <xsd:annotation>
                  <xsd:documentation>Defines  a date that can also be
empty</xsd:documentation>
            </xsd:annotation>
            <xsd:union memberTypes="xsd:date imdi:emptyString
imdi:EmptyType"/>
      </xsd:simpleType>
      <xsd:simpleType name="MetatranscriptType">
            <xsd:annotation>
                  <xsd:documentation>Defines the different types of
metatadata descriptions</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:enumeration value="SESSION"/>
                  <xsd:enumeration value="CATALOGUE"/>
                  <xsd:enumeration value="CORPUS"/>
```

```
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:simpleType name="QualityType">
            <xsd:annotation>
                  <xsd:documentation>Quality specification scale [1-
5] or Unknown or Unspecified</xsd:documentation>
            </xsd:annotation>
            <xsd:restriction base="xsd:string">
                  <xsd:enumeration value="1"/>
                  <xsd:enumeration value="2"/>
                  <xsd:enumeration value="3"/>
                  <xsd:enumeration value="4"/>
                  <xsd:enumeration value="5"/>
                  <xsd:enumeration value="Unknown"/>
                  <xsd:enumeration value="Unspecified"/>
            </xsd:restriction>
      </xsd:simpleType>
      <xsd:complexType name="SessionType">
            <xsd:sequence>
                  <xsd:element name="Name" type="xsd:string">
                        <xsd:annotation>
                              <xsd:documentation>A short name to
identify the session</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Title" type="xsd:string">
                        <xsd:annotation>
                              <xsd:documentation>The complete title
of the session without abbrevations</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Date" type="imdi:DateType">
                        <xsd:annotation>
                              <xsd:documentation>The date when the
primary  data of the session was created in ISO8601
format</xsd:documentation>
                        </xsd:annotation>
                  </xsd:element>
                  <xsd:element name="Description"
type="imdi:DescriptionType" maxOccurs="unbounded"/>
                  <xsd:element name="MDGroup"
type="imdi:MDGroupType"/>
                  <xsd:element name="Resources">
                        <xsd:annotation>
                              <xsd:documentation>Groups information
of language resources connected to the session</xsd:documentation>
                        </xsd:annotation>
                        <xsd:complexType>
                              <xsd:sequence>
                                    <xsd:element name="MediaFile"
minOccurs="0" maxOccurs="unbounded">
                                          <xsd:annotation>

      <xsd:documentation>Groups information about the media
file</xsd:documentation>
                                          </xsd:annotation>
                                          <xsd:complexType>
                                                <xsd:sequence>
                                                      <xsd:element
name="ResourceLink" type="xsd:anyURI">
```

```
      <xsd:annotation>

      <xsd:documentation>URL to media file</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="Size" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>Size of media file </xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="Type" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Major part of mime-type</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="Format" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Minor part of mime-type</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="Quality" type="imdi:QualityType">

      <xsd:annotation>

      <xsd:documentation>Quality of the recording scale [1-
5]</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
name="RecordingConditions" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>describes technical conditions of
recording</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:element
ref="imdi:TimePosition"/>
                                                        <xsd:element
name="Access" type="imdi:AccessType"/>
                                                        <xsd:element
name="Description" type="imdi:DescriptionType"
maxOccurs="unbounded"/>
```

```
                                                    </xsd:sequence>
                                              </xsd:complexType>
                                        </xsd:element>
                                        <xsd:element
name="AnnotationUnit" minOccurs="0" maxOccurs="unbounded">
                                              <xsd:annotation>

      <xsd:documentation>Groups information about an annotation unit.
An annotation unit is refers to one layer of annotation or
transcription. This is independent of wheter these are contained in
one or more files.</xsd:documentation>
                                              </xsd:annotation>
                                              <xsd:complexType>
                                                    <xsd:sequence>
                                                          <xsd:element
name="ResourceLink" type="xsd:anyURI">

      <xsd:annotation>

      <xsd:documentation>URL to file containing the
annotations/transcription</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="MediaResourceLink" type="xsd:anyURI">

      <xsd:annotation>

      <xsd:documentation>URL to media file from which the
annotations/transcriptions originate</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="Annotator" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>Name of the person who did the annotation or
transcription</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="Date" type="imdi:DateType">

      <xsd:annotation>

      <xsd:documentation>Date when unit was
created</xsd:documentation>

      </xsd:annotation>
                                                          </xsd:element>
                                                          <xsd:element
name="Type" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>The type of the annotation unit.
OV</xsd:documentation>
```

```
      </xsd:annotation>
                                                              </xsd:element>
                                                              <xsd:element
name="Format" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>File format used for
annotation/transcription </xsd:documentation>

      </xsd:annotation>
                                                              </xsd:element>
                                                              <xsd:element
name="ContentEncoding">

      <xsd:annotation>

      <xsd:documentation>Name of encoding scheme used for the
annotation purpose</xsd:documentation>

      </xsd:annotation>
                                                              </xsd:element>
                                                              <xsd:element
name="CharacterEncoding">

      <xsd:annotation>

      <xsd:documentation>Character encoding used in the annotation
unit </xsd:documentation>

      </xsd:annotation>
                                                              </xsd:element>
                                                              <xsd:element
name="Access" type="imdi:AccessType"/>
                                                              <xsd:element
name="LanguageId" type="imdi:LanguageIdType">

      <xsd:annotation>

      <xsd:documentation>Language used for
annotating/transcribing</xsd:documentation>

      </xsd:annotation>
                                                              </xsd:element>
                                                              <xsd:element
name="Anonymous" type="xsd:boolean">

      <xsd:annotation>

      <xsd:documentation>Indicates if data has been anonymised. CV
boolean</xsd:documentation>

      </xsd:annotation>
                                                              </xsd:element>
                                                              <xsd:element
name="Description" type="imdi:DescriptionType"
maxOccurs="unbounded"/>
                                                      </xsd:sequence>
                                              </xsd:complexType>
                                      </xsd:element>
```

```
                                    <xsd:element name="AnonymousInfo"
minOccurs="0">
                                          <xsd:annotation>

      <xsd:documentation>Groups data about name conversions for
persons who are anonymised </xsd:documentation>
                                          </xsd:annotation>
                                          <xsd:complexType>
                                                <xsd:sequence>
                                                      <xsd:element
name="ResourceLink" type="xsd:anyURI">

      <xsd:annotation>

      <xsd:documentation>URL to information to convert pseudo named
to real-names</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Access" type="imdi:AccessType"/>
                                                <xsd:sequence>
                                          </xsd:complexType>
                                    </xsd:element>
                                    <xsd:element name="Source"
minOccurs="0" maxOccurs="unbounded">
                                          <xsd:annotation>

      <xsd:documentation>Groups information about the source; e.g.
media-carrier, book, newspaper archive etc.</xsd:documentation>
                                          </xsd:annotation>
                                          <xsd:complexType>
                                                <xsd:sequence>
                                                      <xsd:element
name="Id" type="xsd:string">

      <xsd:annotation>

      <xsd:documentation>Short unique code to identify
source</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Format" type="imdi:Vocabulary">

      <xsd:annotation>

      <xsd:documentation>Physical storage format.
OV</xsd:documentation>

      </xsd:annotation>
                                                      </xsd:element>
                                                      <xsd:element
name="Quality" type="imdi:QualityType">

      <xsd:annotation>

      <xsd:documentation>Quality of recorded data. Scale [1-
5]</xsd:documentation>
```

```
      </xsd:annotation>
                                                        </xsd:element>
                                                        <xsd:choice>

      <xsd:element ref="imdi:CounterPosition"/>

      <xsd:element ref="imdi:TimePosition"/>
                                                        </xsd:choice>
                                                        <xsd:element
name="Access" type="imdi:AccessType"/>
                                                        <xsd:element
name="Description" type="imdi:DescriptionType" maxOccurs="unbounded">

      <xsd:annotation>

      <xsd:documentation>Description for this
source</xsd:documentation>

      </xsd:annotation>
                                                        </xsd:element>
                                                  </xsd:sequence>
                                            </xsd:complexType>
                                      </xsd:element>
                                </xsd:sequence>
                          </xsd:complexType>
                    </xsd:element>
                    <xsd:element name="References" minOccurs="0">
                          <xsd:annotation>
                                <xsd:documentation>Groups information
about external documentation associated with this
session</xsd:documentation>
                          </xsd:annotation>
                          <xsd:complexType>
                                <xsd:sequence>
                                      <xsd:element name="Description"
type="imdi:DescriptionType" minOccurs="0" maxOccurs="unbounded">
                                            <xsd:annotation>

      <xsd:documentation>Every description is a
reference</xsd:documentation>
                                            </xsd:annotation>
                                      </xsd:element>
                                </xsd:sequence>
                          </xsd:complexType>
                    </xsd:element>
            </xsd:sequence>
      </xsd:complexType>
</xsd:schema>
```

Appendix 7

IMDI Flyer

Mission

If you have ever tried to find language resources and were unable to find them quickly and efficiently, then the EAGLES/ISLE Meta Data initiative may be the answer to your problem.

The goal of this initiative is to make the immense universe of existing and emerging language resources easy to locate by introducing a standard for metadata descriptions of Multi-media/Multi-modal Language Resources. This standard will enable interested people such as linguists, psychologists, anthropologists, speech and language engineers and many others to efficiently locate suitable resources and thus increase their reusability.

IMDI technology can be applied within an institution and/or across institutions. By simple means it is possible to hook up local metadata descriptions into a distributed and web-accessible universe of such descriptions. The IMDI Metadata standard developed within the EAGLES/ISLE project provides metadata elements and controlled vocabularies with which one can describe the main characteristics of the resources, such as the name of the language spoken, the speakers age, the speakers background etc.

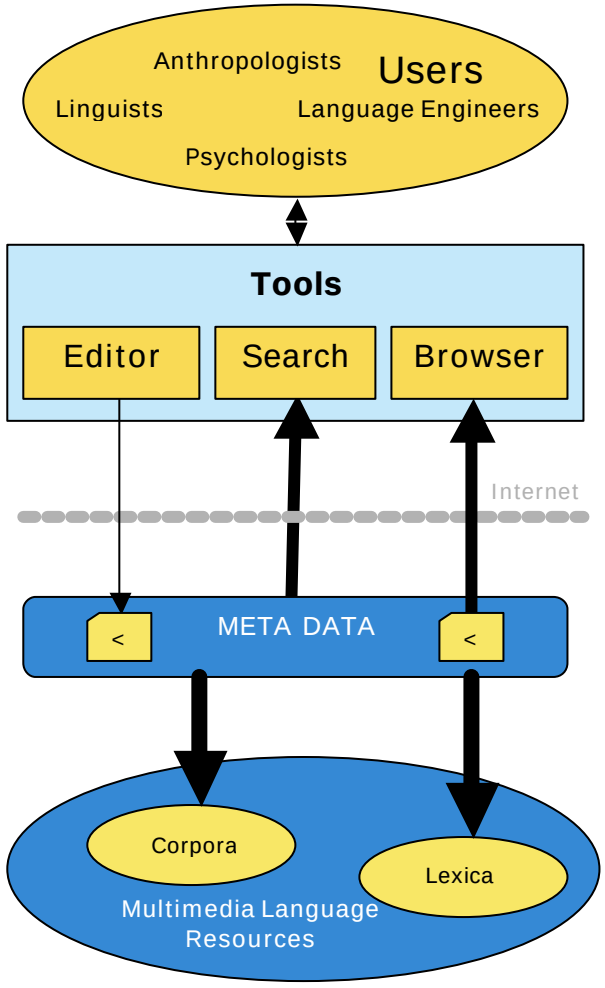
Tools are available to browse and search in the universe of metadata and to easily create your own metadata descriptions. The metadata editor is user-friendly and automatically creates the XML-based files containing the metadata descriptions. The browser can combine local and distributed files and it is easy to make the local files available on the Internet.

Access to the resources themselves is done by directly starting appropriate tools within the browser.

Support

The MPI team will continue support for all IMDI results, tools and repositories. All IMDI results and products are free for usage.

For more information, please, visit the following web-sites:
http://www.mpi.nl/ISLE
http://www.ilc.cnr.it/EAGLES



<?xml version="1.0" ?>
<!DOCTYPE METATRANSCRIPT (View Source for full doctype...)>
<METATRANSCRIPT DATE="20000830" VERSION="0" GEN="" LEVEL="0" VERSION-ID="0">
<SESSION NAME="ltdmo12a.1" LEVEL="0" DATE="20000830">

Keep yourself informed!

If you are interested and you want to know more about the **EAGLES / ISLE Metadata** project, then we invite you to subscribe to the *EAGLES/ISLE Meta Data Initiative Information Network*! To subscribe, write an email to isle@mpi.nl with the subject heading *subscribe* and we will inform you about new developments.

What is on the **web site?**

- **Project information**
- **Links** to related projects
- **Tools**
- **Documents** about meta data
- **Who's who**
- **News**

http://www.mpi.nl/ISLE

Do you want to participate?

If you would like to create metadata descriptions for yourself and organize your resource chaos for better retrieval, as well as make your descriptions open to the international community, then let us know. In the coming period we might integrate your data. In the future a resource agency will take over this function.

Documents

The following **IMDI** documents and tools are freely available or in preparation (blue marked are ready, those in black are in preparation):

- General:
  - o White Paper
  - o Metadata Overview
- Metadata Elements for:
  - o Session Descriptions
  - o Catalogue Descriptions
  - o Lexicon Descriptions
- Mapping with Dublin-Core of:
  - o Session Descriptions
  - o Catalogue Descriptions
  - o Lexicon Descriptions
- Schema Implementation for:
  - o Session Descriptions
  - o Catalogue Descriptions
  - o Lexicon Descriptions
- Metadata Tools:
  - o Session MD Editor
  - o Catalogue MD Editor
  - o Lexicon MD Editor
  - o Metadata Browser (incl. Search)
  - o support for controlled vocabularies in tools
- Demonstration
- Infrastructure / Repositories
- Public Relation Documents:
  - o This Flyer
  - o MD Reference Cards

International Standards for Language Engineering (ISLE)

IMDI

ISLE Meta Data Initiative

Improves the accessibility/availability of language resources

State: 7.12.2001

E-mail: isle@mpi.nl
Web site: http://www.mpi.nl/ISLE

Metadata Overview and the Semantic Web

P. Wittenburg, D. Broeder

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

Abstract

The increasing quantity and complexity of language resources leads to new management problems for those that collect and those that need to preserve them. At the same time the desire to make these resources available on the Internet demands an efficient way characterizing their properties to allow discovery and re-use. The use of metadata is seen as a solution for both these problems. However, the question is what specific requirements there are for the specific domain and if these are met by existing frameworks. Any possible solution should be evaluated with respect to its merit for solving the domain specific problems but also with respect to its future embedding in "global" metadata frameworks as part of the Semantic Web activities.

1. Introduction

At the LREC conference 2000 a first workshop was held which was dedicated to the issue of metadata descriptions for Language Resources [1]. It was also the official birth of the ISLE project (International Standards for Language Engineering) that has a European and an American branch. The workshop was also the moment where the European branch presented the White Paper [2] describing the goals of the corresponding ISLE Metadata Initiative (IMDI). At another workshop held in Philadelphia in December 2000 the American branch presented the OLAC (Open Language Archives Community) initiative [3].

Somewhat earlier the Dublin Core initiative mainly driven by librarians and archivists completed its work on the Dublin Core Metadata Element Set (DCMES) [4] and the MPEG community driven by the film and media industry started their MPEG7 initiative [5]. All these initiatives are closely related since they build upon each other.

After two years of hard work and dynamic developments it seems appropriate to describe the current situation, put the initiatives into a broader framework and discuss the future perspectives.

2. Concept of Metadata

2.1. Early Work

The concept of metadata is not a new concept. In general terms "metadata is data about data" which can have many different realizations. In the context of the mentioned initiatives the term "metadata" refers to a set of descriptors that allows for easily discovering and managing language resources in the distributed environment of the World-Wide-Web.

Metadata of this sort was used, for example, by librarians for many years in the form of cards and later to exchange format descriptions to describe the holdings of libraries and inform each other about them. The scope was limited to authored documents and the purpose was easy discovery and management.

Metadata has also been used for many years in some language resource archives. An example is the header information in the CHILDES database [6]. These early project specific definitions were the basis for the important work about header information within the TEI initiative (Text Encoding Initiative) [7] which was later taken over by the Corpus Encoding Standard (CES) [8] to describe the specific needs of textual corpora. The TEI initiative worked out an exhaustive scheme of descriptors to describe text documents. This header information was seen as a integral part of the described SGML structured documents themselves. It still can serve as a highly valuable point of reference and orientation for other initiatives. Some corpus projects still refer to the TEI/CES descriptors and use part of them. This approach was followed by the Dutch Spoken Corpus project [9].

Despite some projects and initiatives the concept of uniform metadata descriptions following the TEI standard was not widely accepted for different reasons. Many found the TEI/CES descriptions too difficult to understand and too costly to apply. Others took the view that their resources did not match the TEI type of categorization. Many appear not to have taken the time to investigate the extensive set of TEI suggestions.

It should not be forgotten that some companies storing language resources for various language engineering purposes such as training statistical algorithms or building up translation memories are using specifically designed databases for discovery and management purposes. These databases normally allow a shared access so that each employee can easily identify whether useful resources are available. For example Lernard&Hauspie used such a database internally[25]. The large data centers such as LDC

[25] It was not possible to get a blue-print of the structure of this database.

[10] and ELRA [11] have developed an online catalogue suitable to their needs that allows easy discovery of the resources they are housing. Other resource centers such as the Helsinki University resource server [12] use an open common web-site approach where they describe their holdings without using a formal framework such as metadata.

2.2. Classification Aspects

The creation of a metadata description for a resource is a classification process. The metadata elements define the dimensions and the values they can take define the axes along which classifications can be done. However, metadata classification of language resources is a classification in a space where the dimensions are not orthogonal, i.e. they are not independent from each other. A choice for a value in one dimension may have consequences for the choices in others. Certain properties can appear along several different dimensions. Further, we cannot always define metrics along the axes.

Therefore, a classification has to be based on a comparison with predefined vocabularies. Figure 1 shows how such classification can be done. The user may assume that the location indicated by the cross would best describe his resource. Since there is no perfect match with values along the two dimensions indicated by black and white dots, he may decide to choose the dots indicated with rectangles as the best matching ones.

Of course, this raises many problematic questions especially in communities such as the linguistic one. There does not exist yet a widely agreed ontology for language resources. Linguistic theories lead to different types of categorization systems. So who can decide about the usage of such encoding schemes and since it can be expected that sub-communities do not agree about one single scheme, the question is: how can interoperability be achieved, i.e. how can different categorizations be mapped onto each other? These questions are not simple to solve.

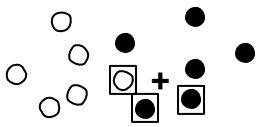
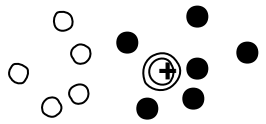


Figure 1 shows two categories represented by black and light dots. Each dot denotes a possible value of the respective category in some non-Euclidian space. The cross may indicate the “location” of the resource and the rectangles as the optimal choice for describing that resource.

A solution chosen by the IMDI initiative is to allow for flexibility, i.e. allow the addition of elements (dimensions of description/categorization) and to make the corresponding vocabularies user extendable where there is no set established yet. At first glance this solution appears acceptable but it is somewhat dangerous as can be inferred from classification literature [13]. We would like to indicate one of the possible problems with an example (fig 2). Individual users could decide to add a value to a dimension that does not seem to be characteristic for the point in space and thereby breaks the semantic homogeneity distorting the dimensions and creating problems for proper discovery.



In figure 2 an additional value is created (double circle) for one of the two categories (light circles) in an area where another dimension (black circles) is dominant. This leads to a distortion of the semantic homogeneity.

Also users could just add particular values to a vocabulary to suit their direct needs. Such a process would lead to an over-specification. The result would be a long list of specific and non-generalized terms and again problems with resource discovery are predictable.

On the other hand completely prescribing a vocabulary for a dimension not yet fully understood would mean that important areas might not be represented so that people will not make use of the categorization system at all. In the IMDI initiative a middle position was taken. A pre-defined vocabulary is proposed and at regular instances the actually used vocabulary will be evaluated to detect omissions in the proposed vocabulary. Dependent on the outcome the pre-defined vocabulary will be extended. It can of course also occur that existing values will be removed, since they are not used and are seen as obsolete by the community. One question remains: who is responsible for making decisions on such matters? This is a social and organizational issue to be solved by the whole community.

3. Reasons for Metadata

3.1. General Aspects

A re-vitalization of the metadata concept occurred with the appearance of the Web. A few figures may illustrate the problem we are all faced with. According to an analysis of IDC the amount of relevant data in companies exceeded 3.200 Petabyte in 2000 and will increase to 54.000 Petabyte in 2004[26]. The stored documents include information relevant for the success of the companies and form part of the company’s knowledge base. These documents are of various natures - partly the texts themselves explain what they are about and partly the documents need a classification to easily understand their relevance. Open questions are how to manage this knowledge base and how to make efficient use of it.

Well-known is the gigantic increase in the amount of resources available on the Web. Here, the focus is certainly on the aspect of efficient methods to find useful resources. It is often argued that the search engines that are based on information retrieval techniques have lost the game at least for the professional user who is not looking for adventures. The typical search engines use the

[26] It is not the amount of data that counts, but the number and variety of resources that increases in parallel.

occurrence and co-occurrence of words in the titles or in the texts of web documents to find what are thought to be the most suitable resources and calculate a suitability rating. Automatic clustering techniques also based on statistical algorithms are used to group information and also automatic categorization is carried out to help the user in his discovery task. Still the precision (the number of correct results compared to the number false results) and the recall (the number of hits found compared to the total number of suitable documents) are not satisfying especially if the user is looking for a specific type of information. Narrowing down the semantic scope of the queries to discover interesting documents often is a very time-consuming and tedious enterprise. Therefore, IR-based search engines will not be the only choice for professional users.

The PICS initiative [14] showed that even for general web-based information there is a need for additional type of descriptors that cannot be reliably extracted from the texts. So, metadata descriptions, i.e. characterizations of the resources with the help of a limited set of descriptive elements, were seen as a useful addition to the texts themselves. In this paper we will not deal with the aspects of how to come to valuable descriptor sets for arbitrary content, but focus on the language resource domain.

3.2. Language Resource Domain

All the content based information retrieval (IR) techniques are based on the assumption that the texts themselves, in particular the words used and their collocations, describe the topic the text is about in sufficient detail. In the domain of language resources there are a number of data types where we can assume that this may be true. Grammar descriptions or field notes in general include broad prose descriptions about the intentions and the content in addition to special explanations of linguistic or ethnographic details. IR techniques may lead to successful discovery results. Still, would professionals who are looking for "field notes about trips in Australia that lead to a lexicon about the Yaminyung language" want to rely on such statistical engines? They would prefer to operate in a structured space obviously organized by resource type, location and languages to discover the resources they are looking for. It is almost impossible to automatically derive metadata descriptions from the content of language resources such as corpora and lexica.

Also in the language resource domain we are faced with a gigantic increase in the amount of resources. An impression about this explosion of resources can be given by the example of the multimedia/multimodal corpus at the Max-Planck-Institute for Psycholinguistics where every year around 40 researchers carry out field trips, do extensive recording of communicative acts and later annotate the digitized audio and video material on many interrelated tiers. The institute now has almost 10000 sessions - the basic linguistic unit of analysis - in an online database and we foresee a continuous increase. One researcher at the institute has about 350 GB of video recordings (about 350 hours) online that are transcribed by several people in parallel. Thus the individual researchers as well as the institute as a whole are faced with a serious resource management and discovery problem.

The increase of the amount of resources was paralleled by an increase in the variety and complexity of formats and description methods. This was caused by moving from purely textual to multimedia resources with multimodal annotations. It was understood early that the traditional methods of management and discovery mostly on purely individual account led increasingly often to problems. Scientists could no longer easily find relevant data and problems arose when a researcher left the institute. Similar situations occur in other research centers, universities and also in industry.

Unified type of metadata descriptions where everyone in the domain intuitively understood the descriptors and a process where each individual researcher can easily integrate his resources and resource descriptions were seen as the solutions for the institute. These descriptions should include enough information so that a linguist can directly see whether the material is relevant for his research question at that moment. Also given an interesting resource it should be possible to immediately start relevant tools on them. Queries such as "give me all resources which contain Yaminyung spoken by 6 year old female speakers" should lead to appropriate hits.

It was clear that most of these descriptions had to be created manually since only in a few cases it may be possible to automatically extract them from directory path names, Excel sheets or other sorts of systematic descriptions. As mentioned before the great majority of the language resources are of a sort where the descriptors cannot be anticipated from the content.

3.3. New Metadata Aspects

The trend of a continuously growing number of language resources will continue. Another apparent trend is that researchers are increasingly often willing to share them online via the Internet or at least to share knowledge about their existence with others fom the community. Metadata descriptions, as previously explained, have a great potential to help researchers to manage these resources and simplify their discovery.

While the designers of the aforementioned TEI focused on text documents, current collected language resources mostly have multimedia extensions (sound and/or video). This adds new requirements on what descriptor set to use. Furthermore, it is generally agreed that the purpose of a metadata set is not so much to create a very complete description of a resource, but to support easy resource discovery and resource management. This way of looking at metadata certainly fits with the important work in the Dublin Core initiative (DC).

At the moment no-one can say with absolute authority which type of descriptor set is necessary to facilitate discovery and management, since for the domain of language resources the metadata concept (with respect to the above purposes) is very new and has hardly been applied by a greater number of linguists. We are confronted with different type of users all having different requirements that we do not know in detail. There are

- the researchers and developers who are experts and want to quickly find exactly those resources which fit to their research or development tasks[27];
- the resource manager who wants to check whether he/she wants to define a new layer of abstraction in the corpus hierarchy to facilitate browsing[28];
- the teacher who is teaching a class about syntax and wants to know whether there are resources with syntactic annotations commented in a language he/she can understand;
- the journalist who is interested in getting a quick overview about resources with video recordings about wedding ceremonies;
- the casual web-user who is interested to see whether there is material about a certain tribe he just heard about;
- many other types of users could be mentioned here whose requirements we often do not yet know.

An important point is that many of the language resource archives currently set up have a long-term perspective. So the question of their typical usage becomes an even more problematic one, since we cannot anticipate what future generations will need to discover resources. A widely used statement in such situation of uncertainty is to make the descriptor set exhaustive. But the fact is that very exhaustive sets are problematic because they are labor intensive and the inherent danger of over-specification. The IMDI team expects that a more dynamic scenario will occur where descriptor elements and even element values are seen as abstract labels which can be refined when more detail is needed. Sub-structures can also be needed to make properties more specific.

Given these uncertainties about future user needs, it makes sense to start now with a non-exhaustive element set. Also, language resource creators are reluctant to invest time in information that will primarily help others. Too much labor required will lead to a negative attitude.

Another phenomenon is that individual researchers have to participate in person in the creation and integration of metadata descriptions. There is no time to read lengthy documents about the usage of elements. Therefore everything has to be simple and straightforward, otherwise he/she will not participate. Metadata descriptions also should facilitate international collaboration. In many disciplines international collaboration with researchers located at different places is normal. Contributions from one of them must be directly visible by the others. This requires a metadata description framework that allows for regular update of the descriptions.

3.4. Resource Management Aspects

The primary task of metadata is resource discovery. However, resource management is an equally important aspect for the resource creator and manager. Metadata can help in managing resources. Linguistic data centers or companies storing language resources are used to manage large amounts of resources. Beyond discovery, management includes operations such as grouping related resources, copying valuable resources together with their context, handling different versions of resources, distributing and removing resources and maintain access lists and design copying strategies. Until a few years ago resource management was done by individual researchers using physical structuring schemes such as directory structures. This was also made possible by the relatively small size of the resources.

However, for the modern multimedia based archives of institutions and individual researchers files and corpora are becoming so huge that the physical manipulation of these resources becomes more and more a domain of the system manager. The conceptual domain defined by metadata can become the operational layer for the corpus manager. Grouping is no longer done on a physical layer that often implies copying large media files, but on the level of metadata. This means the definition of useful metadata hierarchies and to set the pointers to the resources wherever the system management may have stored them.

Resource management has acquired another dimension with the distributed nature of resources in the Internet scenario. It will become a normal scenario in the future that a video file is hosted on a certain server while two collaborators work simultaneously on that same media file. Using the Dutch scientific network this kind of collaboration is already possible. One, for example, may be annotating gestures and the other annotating semantics where speech and gesture information is needed. Annotations are generated on different tiers and are visible to both collaborators, but the place of storage could be arbitrary especially as long as the annotations have a preliminary character. The metadata description can be used to point to the location and to allow management operations as if the resources were all bundled on a single server.

4. Language Resource Data Types

Before introducing the different metadata initiatives that deal with language resources it is necessary to analyze the characteristics of the objects that have to be described. As already indicated not all objects that we find in the language resource domain are well understood. The most important ones are

- complex structured text collections
- multimedia corpora
- lexica in their different realizations
- notes and documents of various sort

The nature of text collections is very well described by the TEI initiative. The particular aspects of textual corpora were then analyzed and described by CES. Multimedia resources (MMLR) that either include multimedia material or are based on media recordings add new

[27] For a speech engineer for example it may be relevant to find resources where short-range microphones were used.

[28] For a resource manager it might be relevant to find all resources with speakers of a certain age.

requirements. MMLR can combine several resources which are tightly linked such as several tracks of video, several tracks of audio, eye tracking signals, data glove signals, laryngograph signals, several different tiers with annotations, cross-references of various sorts, comments, links to lexical entries and many others. In many MMLR it is relevant to describe that a certain annotation tier has special links with a certain media track. For speech engineers it could be relevant to know the exact relation between a specific transcription or transliteration to one specific audio track (close range microphone). On a certain level of abstraction the different sub-resources have to be seen as one or relating to one "virtual" meta resource. Metadata has to describe this macro-level complexity and has to inform the user about the type of information contained in such a bundled resource.

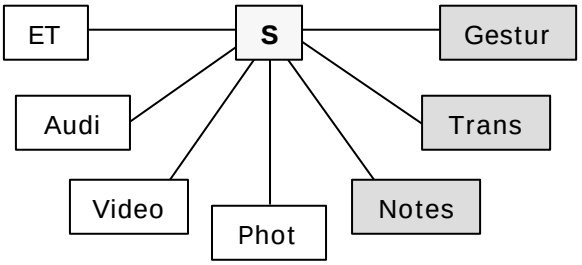


Figure 3 shows the various types of information tightly related by a common time axis.

Lexica where concepts and words are in the center of the encoding can appear in various forms such as dictionaries, wordlists, thesauri, ontologies, concordances and many others. Until now they are mostly monolithic resources with a complicated internal structure bearing the linguistic information. Metadata that wants to describe such a resource to allow useful retrieval has to indicate which type of information is available and in what format.

Linguistic notes can be of various sorts as well such as field notes, sketch grammars and sound system descriptions. Normally they appear as prose texts with no special structural properties that can be indicated by metadata. They can be treated as normal documents except that their functional type has to be indicated.

5. Metadata Goals and Concepts

In this chapter we want to briefly review the goals and concepts of the metadata initiatives that follow more or less the new paradigm described above and which are relevant for the language resource domain.

5.1. Dublin Core Metadata Initiative

The Dublin Core metadata initiative has as primary goal to define the semantics of a small set of descriptors (core set) which should allow us to discover all types of web-resources independent whether they are about steam engines or languages spoken on the Australian continent. All the experience of librarians and archivists was invested in the definition of the core set. One explicit goal was to create a significantly lighter set than defined for example within the librarians MARC standard [15]. The discussions that started seriously around 1995 ended up in the definition of 15 elements as listed in the following table.

Title	name given to the resource
Creator	entity primarily responsible for making the content of the resource
Subject	topic of the content of the resource
Description	account of the content of the resource
Publisher	entity responsible for making the resource available
Contributor	entity responsible for making a contribution to the content of the resource
Date	date associated with an event in the life-cycle of the resource
Type	nature or genre of the content of the resource
Format	physical or digital manifestation of the resource
Identifier	unambiguous reference to the resource within a given context
Source	reference to a resource from which the present resource is derived
Language	language of the intellectual content of the resource
Relation	reference to a related resource
Coverage	extent or scope of the content of the resource
Rights	information about the rights held in or over the resource

DC wanted to define a foundation for a broadly interoperable semantic network based upon a basic element set that can be widely used. This broad scope was achieved by often vague definitions of several of the DC elements. This is its strength and at the same times its weakness.

The designers well understood the limitations and problems of this approach. The Dublin Core initiative anticipated the need for other element sets and the Warwick Framework [16] was described as a way to accommodate parallel modular sets of metadata using domain specific element sets. Many initiatives work along the DC suggestions by modifying the element set in a number of dimensions, others started from scratch, however, accepting the underlying principle of simplicity. The modifications of the DC core set are done in 3 dimensions partially sanctioned by the DC initiative: (1) Qualifiers are used to refine the broad semantic scope of the DC elements. The underlying request is that qualification may not extend the semantic scope of an element. (2) Constraints may be defined to limit the possible values of an element (Example: date specification according to the W3C recommendations). (3) The usage of new elements, which of course challenges DC compatibility.

The DC initiative itself defined qualifiers and constraints for a number of elements [17]. They also foresaw a problem with uncontrolled qualification: "The greater degree of non-standard qualification, the greater the potential loss of interoperability". For long time it seemed that at least two views were disputing about the way to go forward. The ones that are in favor of a controlled extension would control the semantic scope, and thus force communities with their own semantic needs

away from adopting the DCMES. In the other view there should be loose control on the semantics of the elements, so that other communities could join easily. In the latter case DCMES would become a container for all sorts of information where querying could lead to unsatisfying results.

DCMI did not formulate any syntactic specifications. The DC Usage Group described how DC definitions could be expressed within HTML. The Architecture Working Group within DC made more extensive statements about syntactic possibilities and the inclusion of various extensions [18]. They discuss the following extensions that are common in the community applying DC:

- the usage of a scheme qualifier to put constraints on element values;
- the usage of qualifiers to narrow down the broad semantic scope of the elements such as DC:Creator.Illustrator;
- the subdivision of elements such as DC:Creator.PersonalName.Surname;
- the usage of class type relationships identifying that for example persons not only appear as values of the element creator but also belong to the class person.

There are reports about much confusion in the DC community through the usage of these uncontrolled extensions. In a proposed recommendation from April 2002 of how to implement DC with XML [19] the notion of "dcterms" is introduced which are "other elements recommended by DCMI". The proposed recommendation states that "refinements of elements are elements in their own" and give concrete examples:

use of
<dcterms:available> 2002 </dcterms:available>
instead of
<dc:date refinement="available">2002 </dc:date>
or
<dc:date type="available"> 2002 </dc:date>

These examples show that according to the recommendation refinements should be treated the same as other properties. There is no official statement yet whether this view is accepted by DCMI.

Very recently the Architecture Working Group produced another very interesting proposed recommendation about the implementation of DC with RDF[29]/XML [20]. It is argued that the situation with the simple unqualified DC is very unsatisfactory in various respects. In particular, there is no way to provide structure supporting the discovery process. It is suggested to implement a refinement of an element by applying the "subPropertyOf" relation defined within RDF Schema. A qualifier such as "dcterms:abstract" refines "dc:description" by means of the "subPropertyOf" feature. Also in this paper a replacement of the "subelement" construct (dot notation in the HTML implementation) by the "refinement" attribute is proposed.

[29] RDF = Resource Description Framework worked out by W3C. RDF will be discussed later in this paper.

With respect to language resources DC itself does not provide any special support. To describe the complex structure of MMLR DC offers the relation concept. However, the qualifiers offered do not represent the tight resource bundling very well. Since DC itself does not offer structure, dependencies as indicated in 4 cannot be represented. Also for describing lexica in more detail it does not have the necessary elements.

There is no doubt that DC is currently the most important standard for the simple description of electronically available information sources. It seems to be also clear that DC will be the standard for the casual user to look for easy discovery of simply structured resoruces. DC may form the widely agreed set. The evolution of the DC metadata set and extensions are depicted in the following graph, which is taken from Lagoze [21] and shows the "pidginization versus creolization trends" analogy from Baker.

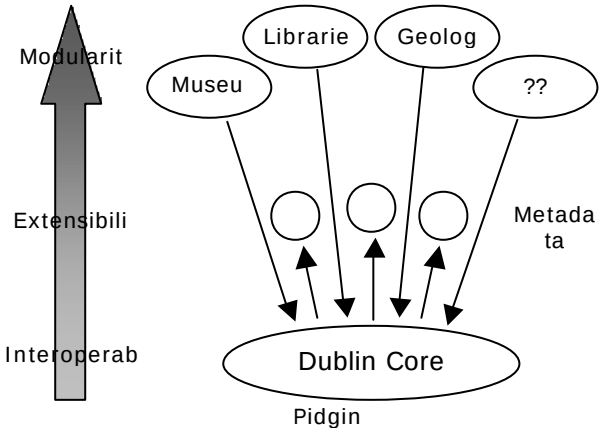


Figure 4 shows the principal problem with which DC had to cope. Interoperability leads to a pidginized form of metadata that is simple enough for the casual web user. The need for Domain specificity then leads to different specialisations of the DC set, the creoles. Dependent on the amount of extensions needed one may end up with a new metadata set.

5.2. OLAC Metadata Initiative

The OLAC metadata initiative wanted to start from the DC set and be compliant with it as far as possible, but overcome its major limitations. Therefore DC was extended in four dimensions:

- 3 attributes were defined to support OLAC specific qualifications (***refine*** to refine element semantics including controlled vocabularies; ***scheme*** to refer to an externally controlled vocabulary; ***lang*** to specify the language a description is in).
- ***Code*** attributes refer to element specific encoding schemes.
- 8 new sub-elements were created which narrow down the semantics, but need a separate controlled vocabulary (Format.cpu, Format.encoding, Format.markup, Format.os,

Format.sourcecode, Subject.language, Type.functionality, Type.linguistics).

- A special ***langs*** attribute as a list of languages which appear in a metadata description.

For various refined elements and sub-elements[30] controlled vocabularies are under preparation and their definition is part of the schema defining the metadata set [22].

The ***refine*** attribute allows OLAC to associate language resource specific semantic descriptions for DC elements that are specified too broadly and imprecisely. It is the association of a controlled vocabulary (CV) that narrows down the semantic scope even more precisely as was described in 2. OLAC wants to keep control of the CV, i.e. there is no user definable area, but there is a description of a development ***process*** that defines how definitions can be successively adapted [23].

The ***code*** attribute acts as a scheme specifier to assure that for example dates are stored in the same way (yyyy-mm-dd).

The OLAC metadata set was constructed such that it can describe all linguistic data types without creating type specific elements and software used in the area of Natural Language Processing. Also advice about and the usage of NLP software is seen as a relevant type of linguistic information.

OLAC has created a search environment that is based on the simple harvesting protocol of the Open Archives Initiative (OAI) [24] and on the standard DC set. Since OAI accepts the DC default set the OLAC designers take care to discuss how the special OLAC information is dumbed down to service providers.

OLAC's intention is to act as a domain specific umbrella for the retrieval of all resources stored in Open Language Archives. Its intent is to establish broad coalitions such that the OLAC metadata standard, i.e. the specifically extended DC set, is accepted as a standard by the whole domain.

5.3. IMDI Metadata Initiative

IMDI started its work without any bias towards any existing metadata vocabulary and wanted to first analyze how typical metadata was used in the field. A broad analysis about header information as used in various projects and existing metadata initiatives at that moment in time was the basis of the first IMDI proposal [25].

Decisive for the design of a metadata set is the question about the granularity of the user queries to be supported. From many discussions with members of the discipline, from the existing header specifications and from the 2 years of experience with a first prototypical test version, it was clear that field linguists for example wanted to input queries such as "give me all resources where Yaminyung[31] is spoken by 6 year old female speakers". Language engineers working with multimodal corpora expressed their wish to retrieve resources where "subjects were asked to give route descriptions, where speech and gestures were recorded and which allow a comparison between the Italian and Swedish way of behavior". Therefore, professional users requested much more detail than DC can offer. Furthermore the semantics of some of the DC element names did not agree with the intuition of many in the user community (e.g. Creator & Contributor). A presentation of the requirements and the needed elements in the European DC Usage Committee revealed that it did not seem advisable to use DC as a basis.

Due to the necessary detail IMDI needed modular sets with specializations for different linguistic data types. The two most prominent data types are (multimedia/multimodal) corpora and lexica. Other linguistic data types are much less common and not so well understood. Consequently two metadata sets were designed which differ in the way content and structure is described. In contrast to DC which only deals with semantics, IMDI also introduced structure and format. Structure makes it possible to associate for example a role, an age and spoken languages with every participant.

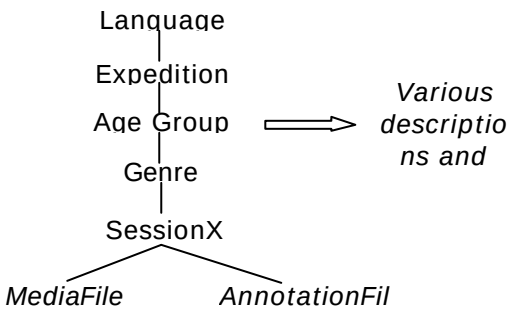


Figure 5 shows a typical metadata hierarchy with nodes representing abstraction layers. Each layer can contain references to various descriptions and notes and thereby integrating them into the corpus. All components of such a hierarchy can reside on different servers. The session nodes are the leafs in the hierarchy, since they point to the recordings and annotations.

The corpus metadata descriptions come in three flavors: (1) The metadata set for sessions is the major type, since it describes the bundle of resources which tightly belong together as described in 4. (2) Since IMDI not only created a metadata set, but also an operational environment, it allows to integrate resources into a browsable domain made up by abstraction nodes and the sessions as the leafs (see figure 5). The metadata descriptions used for the sessions and the higher nodes are basicaly the same. (3) For published corpora that appear as a whole the catalogue metadata set was designed. It contains some additional elements such as ISBN number that are typical for resources that are hosted for example by resource agencies.

The IMDI metadata set for sessions tries to describe sessions in a structured way with sufficient rich

[30] The distinction between qualifiers and sub-elements is not fully clear, especially when looking at the discussions within DC.

[31] Yaminyung is a language spoken by Australian aborigines.

information using domain specific element names [26]. It covers elements for

- administrative aspects (Date, Tool, Version, ...)
- general resource aspects (Title, DataType, Collector, Project, Location, ...)
- content description (Language, Genre, Modality, Task, ...)
- participant descriptions (Role, Age, Languages, other biographic data, ...)
- resource descriptions where a distinction is made between media resources, annotation resources, source data (URL, Type, Format, Access, Size, ...)

The IMDI set was chosen so that most elements are suitable for automatic searching, but there are also those that are filled with prose text and are meant to support browsing. The exact recording conditions can be described, but the variability is so great that it does not make sense in general to search on them. IMDI also offers flexibility on the level of metadata elements in so far that users can define their own keys and associate values with them. This can be done on the top "Session" level as well as on several substructures such as Participant and Content. This feature can be of great use especially for projects that feel that their specific wishes are not completely addressed by the IMDI set. This feature was used for example when incorporating the Dutch Spoken Corpus project within IMDI since they wanted to add a few descriptors defined by TEI. Of course, the metadata environment has to support these features also for example when searching.

For many of the elements, controlled vocabularies (CV) are introduced. Some CV's are closed such as those for continents, since the set of values is well defined. For others such as Genre, IMDI makes suggestions, but allows the user to add new values. The reason is that there is no agreement yet in the community about the exact definition of the term "genre" and how genre information can best be encoded.

For the metadata set and for the controlled vocabularies schema definitions are available at the IMDI web site. All IMDI tools apply them. In contrast to OLAC the definitions of CV are kept separate to allow for the necessary flexibility. According to the IMDI view there will be several different controlled vocabularies as is true for example for language names (ISO definitions and the long Ethnographic list) which should be stored in open repositories such that they can easily be linked.

The recent proposal for lexicon metadata [27] covers elements for

- administrative aspects (Date, Tool, Version, ...)
- general resource aspects (Title, Collector, Project, LexiconType, ...)
- object languages (MultilingualityType, Language, ...)
- metalanguages (Language)
- lexical entry (Modality, Headword type, Orthography, Morphology, ...)
- lexicon unit (Format, AccessTool, Media, Schema, Character Encoding, Size, Access, ...)
- source

Since the microstructure can be very different for the many languages and since linguistic theories also differ, it was decided not to describe structural phenomena of lexica, but only to mention which kind of information is included in the lexicon along the main linguistic dimensions such as orthography, morphology, syntax and semantics. To allow maximum re-usability of the schemas and tools the overlap between lexicon and session metadata was as large as possible.

It was felt that data types such as field notes, sketch grammars and others are resources which are in general prose texts with added semi formal notations and should not be objects which have their own specific metadata set, but they should be integrated into the metadata hierarchies at appropriate places. However, users might want to search for grammar descriptions of Finno-Ugric languages. This problem has not yet been satisfactorily solved within IMDI.

IMDI has been creating a metadata environment consisting of the following components:

- a metadata editor
- a metadata browser
- a search engine
- efficiency tools

All tools have to support the last version of the IMDI definitions of the metadata element sets and the controlled vocabularies. Since the tools are described elsewhere in greater detail [28,29], only a few special features will be described here. The editor supports isolated and connected work, i.e. in case of the PC being connected to the network new definitions of the CV etc can be downloaded and cached. A fieldworker, however, could operate independently on the basis of the cached versions. The browser can operate on local or remote distributed hierarchies allowing each user to create his own resource domain, but easily hooking it up to a larger domain. The browser is also intended to allow for the creation of nodes to form browsable hierarchies, so that a user can easily create his own preferred view on a resource domain. It also allows the user to add configuration information so that local tools of his choice can be easily started from the browser once suitable resources are found.

To increase the possibilities of resource discovery the search component is made an integral part of the browser. The current version operates on one metadata repository only and searching in a distributed domain has to be finished yet. It will make use of a simple query protocol based on HTTP to search sites with IMDI records. The macro infrastructural aspects have to be solved yet, i.e. how to gather metadata information residing at different locations in an efficient way. It is thought that the OAI harvesting protocol is suitable. Efficiency tools are of greatest importance to simplify the creation and management of large metadata repositories. For example, it has to be possible to adapt certain values of a large set of metadata descriptions with one operation. The tools currently available for this type of operation have yet to be integrated in the existing browser and editor.

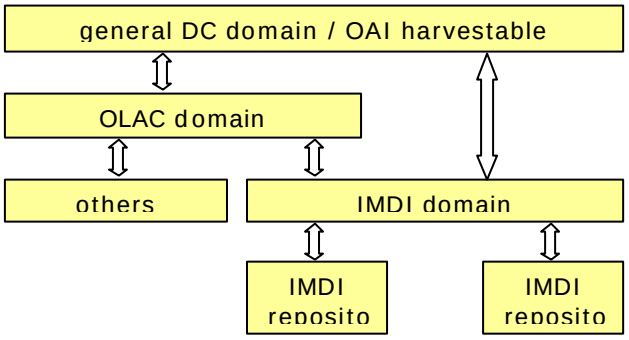


Figure 6 shows IMDI's vision about metadata services users should be able to use. It is not indicated that the general DC domain covers many more domains than just the domain of language resources.

IMDI has accepted that there are different types of users. The casual web user wishing to use a simple perhaps widely known query language based on DC encodings and the professional user interested in easily finding the correct resources. Therefore, IMDI created a document describing the mapping between IMDI and OLAC [30]. Of course, such a mapping cannot be done without losing information and such documents need updates dependent on the dynamics of the two included standards. IMDI envisages the scenario as depicted in figure 6 and will comply with it.

The way IMDI repository connectivity is done is different from how OLAC connectivity is achieved. Since OLAC is focused on metadata harvesting for search support all OLAC metadata providers have to install a script providing the OAI protocol. In IMDI it is just the URL of a local top node that has to be added to an existing IMDI portal to become member of it.

5.4. MPEG7 Initiative

In contrast to the initiatives discussed earlier MPEG7 does not just focus on metadata as the term was defined in this paper. MPEG7 is an integral part of the MPEG initiative. While the other MPEG standards are about audio and video decoding, MPEG7 is a standard for describing multimedia content. It is based on the experiences with earlier standards such as SMPTE [31]. The future MPEG4 scenario includes the definition of media objects and the user controlled assembly of several objects and streams to compose the final display in a distributed environment. The role of MPEG7 in the decoding and assembly interface is to allow the user to search for segments of multimedia content, to support browsing in some browsable space and to support filtering of specific content.

It is meant to support real-time and non-real-time scenarios. Filtering will typically operate in a real-time scenario where media streams are received and parts are not processed any further. Search and browsing typically operate before media content is actually accessed. For the real-time tasks media annotations are used to identify segments that are not appropriate with the user profile.

Due to this wide range of intended applications for the future the MPEG7 description standard is exhaustive and the metadata is just a small part of it. MPEG7 has information categories about

- the creation and production process supporting an event model (i.e. aspects of workflow)
- the usage of the content (copyright, usage history, ...)
- storage features (format, encoding, ...)
- structural information about the composition of a media resource (temporal and spatial)
- low level features (color index, texture, ...)
- conceptual information of the captured content (visual objects, ...)
- collections of objects
- user interaction (user profiles, user history, ...)

MPEG7 has adopted XML Schema as its Descriptor Definition Language (DDL)[32]. It distinguishes between the definition of Descriptors where the syntax and semantics of elements are defined and Description Schemes that define the structural relations between the descriptors. Instead of defining one huge Description Scheme, it was decided to manage the complexity of the task by forming description classes (content, management, organization, navigation and access, user interaction) and let sub-groups define suitable DS. For the description of multimedia content there seem to exist already more than 100 different schemes. Complex internal structures are possible. Summary descriptions about a film for example can contain a hierarchy of summaries.

The MPEG7 community recognized the need to be able to map to Dublin Core to facilitate simple resource discovery of atomic web resources of different media types. DC is made for such type for simple resources. In the Harmony project [32] a mapping of suitable MPEG7 elements was worked out. Finally, it was decided to apply a very restrictive mapping to not extend the semantic scope of the DC elements.

Similar to IMDI but with a much wider scope the MPEG community is working on a sophisticated environment to allow the intended broad spectrum of operations inclusive management. To create for example all the low level features describing video content one is experimenting with smart cameras.

When dealing with multimedia resources MPEG7 could be an option for the language resource community. Currently, there is no special effort within the MPEG7 community to design special DS that are suited for linguistic purposes; however, the language resource community could decide to do so. No obvious limitations can be seen. It seems that MPEG7 has still some time to go to be widely applicable.

6. Mapping Metadata

As mentioned previously DC is widely accepted as a simple metadata set for the casual web-user to search for

[32] Only two additional primitive data types (time and duration) and array and matrix data types were added to cope with the needs.

simply structured resources. To achieve interoperability on that level it is important to map between the metadata sets. We would like to use the mapping between OLAC[33] and IMDI to demonstrate a few aspects that have to be solved.

In the first example two elements are semantically similar. "dc:creator" contains at least two aspects: (1) It refers to the name of a person who created the content. (2) Creation in the sense of DC also has a Intellectual Property Rights aspect. Creators are persons who have rights about the resource. IMDI wanted to separate these two aspects to make clear that there is a responsible researcher on the one hand and participants during the recordings on the other hand, both can claim rights with respect to the resource. So, "imdi:collector" takes care of the wishes of the researchers involved. The mapping rule from IMDI to DC is very simple for this example: All collectors in IMDI descriptions are creators in DC descriptions. The mapping from DC to IMDI is not as clear, since consultants which have a formal right in the DC sense and may appear as creators should be listed under "imdi:participants".

names, all ages etc. One cannot infer which person has a certain age. To solve this problem one has to embed DC in a structure definition or use an identifier of the person and use it in all tags. Also for this example the mapping from IMDI to OLAC is simple: At first instance just the names are passed over. In second instance one could add the content of (part of) the other attributes to a description field and add it to the OLAC tag. The question is whether search engines will be able to use the information. Search engines would interpret description fields as prose text and would not use the advantages typical for structured metadata. The mapping from OLAC to IMDI is simple, since only names are expected. OLAC descriptions would be passed over to IMDI descriptions.

The third example discusses the problems inherent to resource bundling as we are used to in language resources (see figure 3). A good mapping with DC is not possible in a simple way. In IMDI the resources belonging to one session all share a large amount of metadata information and are therefore bundled in one description (if the user decides to do so). In DC one would have to describe every

	DC	OLAC	IMDI	MPEG7
addressed community	world	linguists language engineers	linguists language engineers	film & media community
scope	all web resources	all language resources	focus on (MM) corpora and lexica	all film & media documents
approach	experience of librarians and archivists	compliance to DC	based on overview about earlier work	based on earlier standards
set size	small	small	more detail	exhaustive
user extensibility	no	no	yes	?
formal definitions for	element semantics	element semantics controlled vocabularies constraints	element semantics structural embedding controlled vocabularies constraints	basic descriptor definition language Description Schemes
interoperability	-	DC compliant	mapping to OLAC/DC	mapping to DC
operations	search	search	browse, search, management, immediate execution	browse, search, filtering
tools	-	search environment	editor, browser, search tool, efficiency tools	?
connectivity by	-	OAI harvesting protocol	simple URL registration, OAI harvesting protocol	?
domain specific use of element names	no	no	yes	yes

Table 1 gives a quick overview about the goals and major characteristics of the relevant metadata proposals.

The second example implies structure. The IMDI set has a substructure for the concept "participant". Participants are those persons that are participating in interviews or other typical recording sessions. Each participant has attributes such as name, age, sex, role and languages spoken. The IMDI substructure allows one to group these attributes and therefore support questions such as "all 4 years old females speaking Yaminyung". In DC we just have the possibility to define a set, i.e. list all

atomic resource separately and use "dc:relation" to establish the links. This means that each of the atomic resources has to refer to all the others with for example the qualifier "dc:relation.isPartOf". First, such reference structure is complex and not adequate and second, nobody will actually use it. Another possibility in DC is to define a "virtual root resource" which links to the descriptions of the atomic resources to create a simple hierarchy. For the IMDI to OLAC mapping a simple solution was chosen: all atomic resources get separate descriptions. The OLAC to IMDI mapping is also very simple: since there is no structural information every atomic resource becomes an

[33] The mapping document was based on a previous OLAC version.

atomic resource in IMDI. If there would be a relation specification it would be added to the list of references. Any other scheme would be too dangerous and prone to error.

Basically, we follow the advice of the Harmony project to be very restrictive with mappings, since the semantic homogeneity of the elements can easily be distorted and conversion could lead to errors.

7. Summarizing the Metadata State

Web-accessible metadata descriptions to facilitate the discovery of language resources are a comparatively new concept. Four initiatives (DC, OLAC, IMDI, MPEG7) worked out proposals that are of more or less relevance for the linguistic domain. They differ in a number of aspects, but there is also overlap as indicated in table 1.

The concept is so new that we cannot yet draw relevant conclusions. OLAC states that they have harvested about 18.000 metadata records from their partners. From IMDI it is known that more than 10.000 metadata descriptions were created and integrated into a browsable domain. These numbers alone, however, do not answer a number of important questions such as:

- Are the creators and users convinced that metadata create an added value which is worth the additional effort? By most community members metadata is still seen as an additional effort which is not justified. Awareness is growing, however.
- Do we have a critical mass of new and relevant resources in our repositories such that users make use of the infrastructures for professional purposes? It is clear that we are being far away from such a situation.
- Which approach is the most suitable one (if there is any answer to this question at all)? We still require years to find out and have to address the question whether we have good criteria.
- What are the typical queries the different user groups are asking? We don't know yet, we need a critical mass and interesting environments to be able to answer this question.
- At which level do we need to establish interoperability? Is interoperability on DC level a useful goal? The question of interoperability cannot be seen independent from the usage scenario. Different user groups will have different requirements. The DC pidgin will not satisfy professionals. But the casual web-user may not be interested in looking for resources containing speech from 4 year old speakers.
- Which kind of tools do we need to support the resource creators and managers? Some initiatives have just started working on these issues, but it is too early to make statements.
- Upon which elements and controlled vocabularies will the community agree widely? Again, we have just started, so any answer at this moment may turn out to be wrong.

We do not know the answers to many questions yet or can only make speculations. What we know is that the number of individuals and institutions who create interesting resources is growing fast and that we need an infrastructure to allow their discovery. We also know that individuals and institutions have a management problem to solve and that traditional methods are no longer suitable. So the step to introduce metadata descriptions seems an obvious one, but we do not yet fully understand the potential of web-based metadata.

Resource discovery cannot be the only goal. Resource exploitation and management are equally important. Most important for the users is the view to step away from all sorts of details involving hardware, operating systems and runtime environments. When they have found a resource in a conceptual domain that is their domain of thinking, then they want to start a program that will help them to carry out their job. This program start should be seamless and not as it is today where users have to be computer experts. This is the dream that is still true, but not yet achieved.

Carl Lagoze pointed out that every community has different views about real entities and that these multiple views should not be integrated to one complex description, but that modular packages should emerge [33]. According to him, DC has to be seen as one simple view on certain types of objects. Consequently, he and his colleagues foresaw a scenario with many different metadata approaches where the way interoperability is achieved is not yet solved. The emergence of the Resource Description Framework [34] and the elaborations about an ABC model for metadata interoperability [35] indicate the problems we will be faced with.

Given all the uncertainties with respect to a number of relevant questions we can expect that within the next decade completely new methods will be invented based on the experiences with the methods we start applying now. Given this situation it seems to be very important to test different approaches and in so doing explore the new metadata landscape. A close network of collaboration, interaction and evaluation seems to be necessary to discuss the experiences. Probably an organization as ISO might be a good forum to start a broad discussion about the directions the language resource community should take.

Those who propose metadata infrastructures and ask persons to contribute take a high amount of responsibility. Given that our assumption is true that we will have an ongoing dynamic development[34] the designers of the metadata sets have to be sure that they can and will transform the created descriptions to new standards that will emerge by not losing the valuable information that has been gathered so far.

8. Metadata and the Semantic Web

[34] The IMDI-OLAC mapping document was created in August 2001 and has to be updated completely, since the included metadata sets have changed drastically within a year. It can happen that definitions will change again due to the uncertainty with respect to qualifiers in the DC discussion.

Some years ago Tim Berners-Lee introduced the term "Semantic Web" foreseeing that we are creating a web which can only be managed well when we apply intelligent software agents. Humans will not be able to process the gigantic amount of knowledge available. After receiving concrete tasks from users or after signaling the usefulness of own activities such agents could use the web available information about terms and their relations to find answers or to prepare such answers. Central to the idea of the Semantic Web are the ideas of seamless operation for the user and screening him from all the underlying matching and inferring processes.

Metadata as defined in this paper can play an enormous role in such a scenario, since in metadata sets the elements are more or less accurately defined and their structural relations will become increasingly often explicit as well when technologies such as RDF are used. Metadata is comparatively reliable data[35]. The current lingua franca "DC" will, if it is to be successful, be extended by structure proposals such as being worked out by the architecture group. Sets such as IMDI that include implicit structure from the beginning have to make their structure definitions explicit to make them available for use by smart agents.

Currently, especially created scripts do the mapping between metadata sets (such as IMDI to OLAC) to achieve interoperability on metadata level. These scripts contain all the reasoning implicitly which is necessary to do a useful mapping. We foresee, however, a completely different mapping scheme where the semantics behind it are explicitly formulated. To achieve this we need open repositories (referred to by XML name-spacing) that contain the definitions of elements and vocabularies and those that contain the description of relations presumed that we all could agree on the same syntax[36].

The Resource Description Framework (RDF) seems to be a promising candidate to realize some of the dreams. RDF was developed at the intersection of metadata and knowledge representation experts. From the view of knowledge management it is a decentralized scheme for representing knowledge. It is built on XML to create complex descriptions of resources. It offers a set of rules for creating semantic relations and RDF Schema can be used to define elements and vocabularies. The relations are defined with a very simple mechanism that can also be processed by machines.

In an RDF environment every resource has to have a unique identifier (URI). It can have properties and properties can have values. The simplest assertion is "the web-site http://www.mpi.nl has as author personX" (see also figure 7) where personX can be a literal or for example another web-site. The corresponding RDF code is described in example 1 in the appendix.

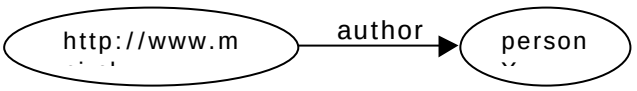


Figure 7 indicates the simple assertion mechanism of RDF where an object is characterized by the property "author" which takes the value "personX".

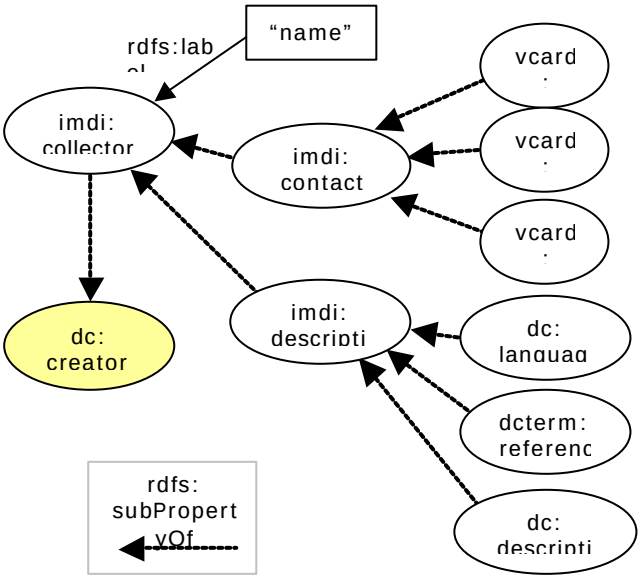


Figure 8 shows a metadata scenario where metadata sets re-use elements and relations which are defined in open repositories.

Using the RDF assertion formalism complex schemes can be realized. Example 2 in the appendix shows how Dublin Core compliant specifications could be embedded in RDF. Example 3 gives one example where Dublin Core and VCard elements are used to create one description. Example 4 shows how RDF could be used to describe the mapping between IMDI and OLAC. Especially the last two examples indicate the direction of development that we expect: A new metadata set to be defined by a (sub) community will make use of existing terminology defined in some open repositories (referred to by XML name-spacing) and write an RDF schema which puts the terms into structure/relation. This scenario is depicted in figure 8.

Smart agents that provide services can interpret these definitions. Another major assumption to make this scenario workable is that communities agree on at least a limited set of terms. When a new term is created it has to be put into an open repository and it's mapping to related terms have to be defined where feasible. This is a complicated social process and can best be guided by an organization such as ISO. Under the guidance of ISO

[35] There is the problem of how to create metadata descriptions for the huge amount of existing documents. It is clear that manual methods will not work. Automatic methods based on Information Retrieval and hopefully Information Extraction will not work on all types of resources as was explained and they would introduce unreliability.

[36] XML has got wide acceptance so that this assumption seems to be valid for the next decades.

TC37/SC4 it would make sense to create such a namespace for the language community.

Carefully designed metadata sets based on open repositories can be seen as representing parts of the ontology of the domain of language resources. It will include the commonalities as well as the differences between sub-communities. Therefore, the discussions about the metadata sets we have right now are very important contributions towards such an ontology. Shortcomings of RDF especially in its power to express semantic details have been identified and therefore initiatives such as DAML/OIL [36] suggest extensions of the framework.

Therefore, one can say that the current metadata initiatives are important steps towards the realization of the Semantic Web.

9. References

[1] LREC 2000 Workshop:
http://www.mpi.nl/ISLE/events/events_frame.html
[2] IMDI White Paper:
http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf
[3] OLAC white:
http://www.language-archives.org.docs.white-paper.html
[4] DCMES
http://dublincore.org/documents/dces
[5] MPEG7
http://mpeg.telecomitalialab.com/standards/mpeg-7.mpeg-7.htm
[6] childes
http://childes.psy.cmu.edu
[7] TEI: http://www.tei-c.org
[8] CES: http://www.cs.vassar.edu/CES
[9] CGN: http://lands.let.kun.nl/cgn/home.htm
[10] LDC: http://www.ldc.upenn.edu
[11] ELRA: http://www.icp.grenet.fr/ELRA/home.html
[12] Helsinki Linguistic Server: http://www.ling.helsinki.fi/uhlcs
[13] H. Niemann (1974) Methoden der Mustererkennung. Akademische Verlagsgesellschaft, Frankfurt
[14] PICS: http://www.w3.org/PICS
[15] MARC: http://www.loc.gov/marc
[16] Warwick:
http://www.dlib.org/dlib/july96/lagoze/07/lagoze.html
[17] DC qualifiers
http://dublincore.org/documents/2000/07/11/dmes-qualifiers
[18] DC Architecture Group
http://dublincore.org/groups/architecture
[19] DC with XML
http://www.ukoln.ac.uk/metadata/dcmi/dc-xml-guidelines
[20] DC with RDF:
http://dublincore.org/documents/2002/04/14/dcq-rdf-xml
[21] S. Weibel, C. Lagoze (1996) WWW-7 Tutorial Track
[22] OLAC MD Set:
http://www.language-archives.org/OLAC/olacms.html
[23] OLAC Process:
http://www.language-archives.org/OLAC/process.html
[24] OAI:
http://www.openarchives.org/OAI/openarchivesprotocol.htm
[25] IMDI Overview:
http://www.mpi.nl/ISLE/overview/overview_frame.html
[26] IMDI Session Set:
http://www.mpi.nl/ISLE/documents/docs_frame.html
[27] IMDI Lexicon Set:
http://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf
[28] tool paper
[29] MPI Tools: http://www.mpi.nl/tools
[30] IMDI-OLAC Mapping: http://www.mpi.nl/ISLE/documents/draft/
[31] SMPTE: http://www.smpte.org
[32] Harmony: http://metadata.net/harmony/video_appln_profile.html
[33] Carl Lagoze (2000) Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience. Invited Talk at the Archiefschool, The Hague, Netherlands, June 2000
[34] RDF: http://www.w3.org/RDF
[35] ABC: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze
[36] DAML/OIL: http://www.w3.org/TR/daml+oil-reference

10. Appendix

Example 1

The first example shows how the assertion included in figure 7 is described by using the RDF formalism and using the Dublin Core metadata element “Creator”.

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description
rdf:about="http://www.mpi.nl/OurDocument.html">
        <dc:creator> personX </dc:creator>
    </rdf:Description>
```

The first line simply indicates that XML version 1.0 is the syntax basis. The next tag indicates that we enter an RDF description. Line 2 and 3 refer to namespaces, so that machines know which elements were used. So here it is refered to the RDF syntax and the Dublin Core element set. The tag in line 5 states that an RDF-based description follows about some characteristics of the web-site ”http://www.mpi.nl”. The next line then states that we add a property “dc:creator” with the value “personX” to the description.

Example 2

In example 2 it is shown how a Dublin Core metadata description could be embedded in RDF. In doing so DC-based description could make use of the structure defining capabilities of RDF.

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description
rdf:about="http://www.mpi.nl/ISLE/whitepaper.html">
        <dc:title> IMDI White Paper </dc:title>
        <dc:creator> Daan Broeder </dc:creator>
        <dc:creator> Peter Wittenburg </dc:creator>
        <dc:creator> Freddy Offenga </dc:creator>
        <dc:subject> Metadata Initiative; XML; Metadata Environment
<dc:subject>
        <dc:lang> en </dc:lang>
        <dc:publisher> ISLE Metadata Initiative </dc:publisher>
        <dc:date> 2000-04-01 </dc:date>
        <dc:format> text/html </dc:format>
```

This description simply adds the normal attributes such as creator, subject as a list of keywords, the language it is written in, publisher, date and format to the document “IMDI White Paper” by using Dublin Core elements.

Example 3

The third example is taken from the DC-RDF proposed recommendation paper [20]. It shows how RDF allows the metadata designer to combine elements from various metadata sets.

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:rdfs:="http://www.w3.org/2000/01/rdf-schema"
        xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0">
<rdf:Description>
<dc:creator>
     <rdf:Description rdf:about="http://qqqfoo.com/staff/corky">
        <rdfs:label> Corky Crystal </rdfs:label>
        <vCard:FN> Corky Crystal </vCard:FN>
        <vCard:N> rdf:parseType="Resource">
                <vCard:Family> Crystal </vCard:Family>
                <vCard:Given> Corky </vCard:Given>
                <vCard:Other> Jacky </vCard:Other>
                <vCard:Prefix> Dr. </vCard:Prefix>
        </vCard:N>
        <vCard:BDAY> 1980-01-01 </vCard:BDAY>
     </rdf:Description>
</dc:creator>
</rdf:Description>
```

This time there are 4 namespaces mentioned, since we also have to borrow terms from RDF Schema and the vCard initiative. The RDF description is now a complex "dc:creator" structure where at first it is mentioned where it is about. Then we associate an abstract label to the attribute by using the "rdfs:label" element. Then we use a whole set of terms borrowed from vCard to describe the creator in detail.

Example 4

The fourth example shows how a formal and machine-readable relation can be established between Dublin Core "creator" and the IMDI "collector". If such descriptions are available in open repositories any engine providing some service could make use of it.

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
        xmlns:imdi="http://www.mpi.nl/ISLE/session-elements/2.5/">
     <rdf:Description rdf:about="http://www.mpi.nl/ISLE/IMDI/3.0/imdi-
schema">
        <rdfs:subPropertyOf
rdf:resource="http://purl.org/dc/elements/1.1/creator"/>
```

The description part makes an assertion which adds the "rdfs:subPropertyOf" attribute to "imdi:collector". According to this assertion "dc:creator" is the superclass, i.e. all IMDI-collectors are also DC-creators.

International Organization for Standardization

ISO/TC 37/SC4 **Language Resource Management**

WG 1: Descriptors and Mechanisms for Language Resources

This version does not include references. The reader is supposed to know the initiatives and frameworks cited. As soon as possible references will be added.

WI-3 Task Description and Scope

- SC4-internal draft version -

1. Task

ISO TC37/SC4 is dedicated to improving the management of language resources in a distributed and interlinked scenario on the World-Wide-Web.

The tasks were recently discussed and defined during a Constituent Meeting of TC37/SC4. The official report from the TC37/SC4 meeting specifies

- as focus:
  - produce an overview about existing projects/initiatives and monitor its usage
  - link with activities of the emerging Semantic Web
- as tasks:
  - provide a clear picture of the needs of the other WGs
  - identify the experts
  - draft a requirement document until the end of 2002

The TC37/SC4 resolution document adds another point explicitly: create a basic paper on goals and views. To create the requirement document a task force was installed.

To be able to achieve the goals WI-2 has to

- determine the scope of language resources
- determine the needs of the community and in the realm of TC37/SC4 of the other working groups
- determine the existing initiatives relevant to the language resource domain,
- develop a scenario how metadata will be used in the Semantic Web
- determine the set of descriptors and their vocabularies useful to describe language resources
- define all relevant terminological units (concepts and terms in major languages) and their relations
- define suitable frameworks for the definitions

A number of mandatory requirements need to be fulfilled to establish a manageable domain of language resources:

- All resources have a unique identifier conforming to the web standards (URI).

- Due to the inherent complexity and large spectrum of different types of language resources, metadata descriptions are created which describe their major characteristics to facilitate management. It is assumed that only structured descriptions will provide the necessary precision for managing language resources efficiently.
- Although the resources themselves will not always be openly accessible due to commercial, ethical or legal reasons, the metadata descriptions have to be accessible and must have the potential to be integrated into virtual management domains.
- All items in such metadata domains must adhere to interoperability mechanisms, i.e. syntax, structure, semantics (data categories, vocabularies, relations) have to be described and stored so that humans and programs can use them.
- The description level should not neglect the need for quick inspection by human readers by providing the possibility of entering prose text.
- The term "management" covers the complete workflow cycle, i.e. WI-2 has to consider the processes of resource creation, enhancement, integration, discovery, exploitation, archiving and deletion.
- The Web is international therefore multilinguality is an inherent characteristic and requirement of the language resource domain.

2. Scope of Language Resources

Only electronically available resources are included. Further, the term "Language Resource" covers at first instance all resources that contain written, spoken and non-verbal (gesture, sign, facial expression and other modalities) material. This definition includes, for example, all websites which contain language in one form or another, publications such as books, recordings of sign language whether annotated or not and lexica. Language resources can be mono- or multilingual and increasingly often language resources are based on multimedia recordings or include multimedia extensions.

There are many types of linguistic resources that contain metadata about other language resources in the wider sense. Lexica, grammar notes and many other types of language resources contain abstract linguistic material and refer to more basic types of resources such as annotated recordings. Also these derived data types are language resources.

Important for the discussions in this note is the view people have on language resources. Technical documentation of cars, material that is part of learning objects and annotated film movies are all language resources in the above-mentioned broad sense. Nevertheless, completely different communities are involved to search for them. In the case of technical documentation engineers may need certain descriptors to easily retrieve a relevant document. A linguist could look at the same document from a different perspective - a research perspective for example, i.e. he will need other type of descriptors than the engineer. We can assume that especially the description of the content will differ, but until now there is not enough knowledge to make final conclusions.

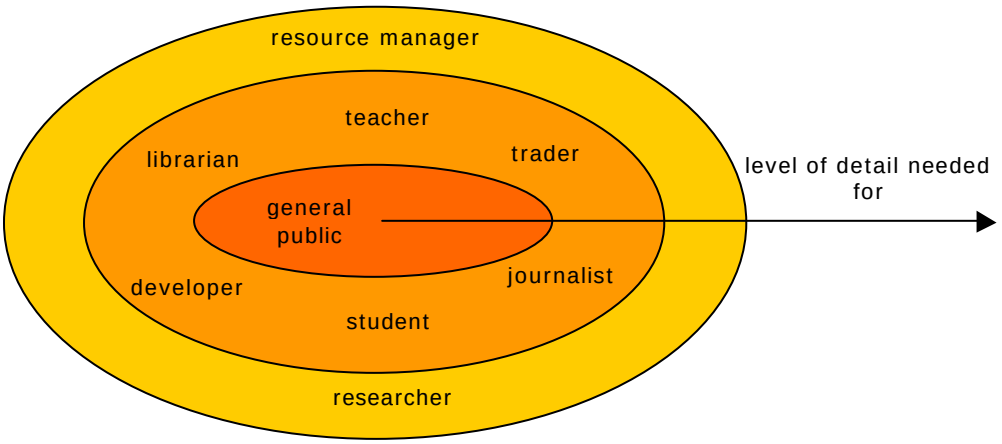
Fact is that when comparing different sets such as Dublin Core and CEI/IEC 82045-1 from the field of document management for example, that the differences are relatively large although both speak about documents with texts covered in them. Therefore, we need to restrict the scope of "Language Resources" essentially to those directed towards the study of language. Resources not providing this are not part of the domain. Metadata descriptions have to support this view.

The metadata descriptions describing language resources with a set of typical categories are not meant to be language resources themselves in the context of this note. They are solely used for resource management and discovery purposes. This, however, does not exclude that they will be viewed as LRs within the context of other work.

3. LR Community

Language resources in the above sense are of interest to many different groups and they fulfill different functions for these groups. Given this variety, it is impossible to define a complete inventory of usage scenarios for LR. We can only identify a few key communities with typical usage interests. On the one hand there is the general public, which is interested in general information on many subjects. On the other hand there are resource traders, researchers or language engineers interested in selling a particular type of language resource, deriving a new grammar or calculating the parameters of statistical recognition algorithms. While the former may be content with general information, the latter will require finer details. Consequently, we can define levels of abstractions as indicated in the following figure. The second level includes many different groups of people, since their usage profile is not at all clear at this moment.

Figure 1



4. Relevant Initiatives

There are a number of standards and initiatives that are relevant to the task of WI-2 that will be briefly mentioned in this chapter.

Header Information/Legacy Metadata

In recent years, many language resource projects and tools supported typical metadata information describing some main characteristics of the resources. An example is the ***CHAT*** header of the CHILDES project. Much experience has been assembled when the well-known ***TEI*** header standard was worked out. Most of the initiatives such as CHILDES dealt with language corpora covering only texts. Multimedia recordings as integral part or as extensions were not used when these formats were defined. Other data types such as lexica were insufficiently described. Also TEI was defined before integral media components became normal. For linguistic data types such as lexica or field notes no particular suggestions were made. Usually the header information was not used for automatic discovery, but for data management purposes. Each project defined its own individual set, since there was no plan to integrate those. Nevertheless, the experiences from these initiatives and efforts have to be considered when deriving metadata standards for language resources. An overview of such

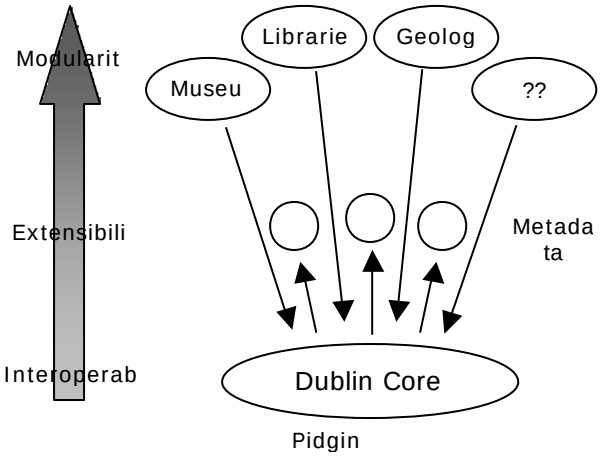
initiatives and efforts can be found in the overview document that was the basis for the IMDI metadata set.

Dublin Core Metadata Element Set

A large number of metadata initiatives were started in the last decade to describe resources of various domains. The ***Dublin Core Metadata Element Set (DCMES)*** is the most important one, as it claims to be useful to describe all type of web-resources using 15 elements so that they can be easily discovered. DCMES was primarily designed by the librarian community that had a lot of experience with categorical descriptions of all sorts of publications. Furthermore, the ***PICs*** metadata initiative influenced the DCMES development. There are many domain-driven extensions of DCMES that have produced refinements to the vaguely defined DCMES categories. However, when these initiatives do not add additional elements, they are limited to a certain extent due to the strong requirement that the semantics of DCMES may not be extended.

The Dublin Core metadata initiative did not deal with structural embedding and implementation until very recently. The Architecture Working Group recently came up with suggestions of how to implement DCMES elements with the help of XML and RDF. Their suggestions indicate the confusion in the community that arose through the uncontrolled extension of the DCMES. For matters of easiness it is suggested that "refinements of elements are elements in their own" which would indicate a change in attitude in the DC community. So, implementation considerations could lead to another phase of changes.

It is widely accepted that the DCMES designers made a wise decision in defining a simple, flat set where the semantics of the elements are vaguely specified. This will allow the data manager to describe a large variety of resources on a shallow level and the general web-user to improve the discovery of these resources compared to what would be possible at present with the usual search engines. There is no doubt that DCMES is currently the most important standard for the simple description of electronically available, simply structured resources. However, the metaphor of "pidginization" on the one hand and "creolization" on the other hand as used by Baker indicates the inherent problems of a metadata set such as DCMES.



Metadata Sets for Language Resources

On a low level of detail the ***Dublin Core*** metadata element set could be seen as a set with help of which language resources can be described. However, DCMES has

severe limitations if more granularity or detail is required. For example there is no distinction between the language a document is written in and the language a document is about. In a typical linguistic document containing for example annotations, one will have much information written in English although the language under investigation might be a Maya language such as Tzeltal. DCMES has an element "DC:Language" which is meant to codify only the language something is written in. Another element "DC:Subject" is used to describe the document is about. However, DC does not make further clarifying statements what this can mean. So one could use this element with an appropriate sub-element or refinement to describe the language the document is about.

This deficit with respect to language resources was correctly identified by the ***OLAC*** (Open Language Archives Community) motivators from LDC and SIL. In the DC community currently two mechanisms are used for extensions: refinements and sub-elements, both having highly overlapping semantics. Based on an inquiry about user needs, OLAC came to the conclusion to add a number of sub-elements. One is for example the language a document is about and another is the linguistic data type the metadata description refers to. Further, to narrow down the vague semantics of the DCMES elements OLAC defined a number of refinements of the element semantics. This makes the OLAC set which was derived from DCMES much more usable for describing language resources. OLAC makes statements that its metadata set can be used for all type of language resources, even to describe tools for language resources and best practice advice. OLAC started to define a number of controlled vocabularies that should be used within their set. Due to its approach, OLAC wants to provide a metadata set that is useful for the whole language resource community. They also see their metadata set as an umbrella to achieve interoperability on metadata level for the language resource community.

While OLAC started from the existing DCMES set to define a metadata set for language resources, the ***ISLE Metadata Initiative*** (IMDI) voted for another way. It started with an inventory of existing metadata practice in the language resource domain and a discussion about the actual needs of the users, especially when dealing with multimedia/multimodal resources. IMDI decided to only describe the major types of language resources (corpora, lexica) by metadata, since their structure and content has been analyzed in sufficient detail. Other data types such as grammar descriptions and field notes are less well described. Therefore, it is not yet understood what the requirements are to describe them to facilitate discovery. The IMDI set contains much more detail compared to DCMES, allows to enter descriptions at many levels and to specify the language these descriptions are in. In contrast to the flat DCMES specifications (just a list of elements) the IMDI set comes along with structure to be able to express that for example different participants have different attributes such as age and sex. It also allows the user to maintain the strong relation between related resources such as media files, annotation files and where available the sources (for example the original tape) implicitly. The IMDI initiative had the task to deliver a complete environment; therefore also several controlled vocabularies are supported. Knowing that we miss experience much degree of flexibility is built in into the IMDI set. Since the IMDI initiative accepts the DC/OLAC role for interoperability with the more general OAI community, a mapping was implemented which is not without information loss.

Strongly Related Metadata Initiatives

The media and film community seems to broadly support the development of the ***MPEG7*** standard which is also adopted by ISO and IEC. MPEG7 was started by looking at existing standards such as SMPTE and the emerging requirements of the community. The result was an exhaustive element set combining both,

suggestions for annotating film productions but also for creating metadata descriptions. The focus in film industry is clearly to support the production process, i.e. also annotate movies with low-level features such as for example "scene change". In the object oriented MPEG4 decoding scenario MPEG7 is intended to support the query, selection and filtering process of the user such that he can easily assemble clips and other information to new personalized presentations. MPEG7 has defined its own Description Definition Language to define Descriptors and Description Schemes that is based on XML. Many descriptors and description schemes have already been defined including a "linguistic" one defining how linguistic phenomena can be encoded. MPEG7 contains many descriptors that match with the definition of metadata in this note. In the Harmony project MPEG7 defined a very restrictive mapping of its metadata elements to Dublin Core specifically to not extend the semantics of DCMES. Although MPEG7 is not particularly designed for the view on multimedia language resources people from the linguistic community have, MPEG7 will have strong impacts on this community, since it gives the possibility to define suitable schemes for sub-communities.

Under the umbrella of IEEE the ***LOM*** (Learning Object Metadata) standard is in the process of being defined. It makes use on the knowledge gathered in DCMI, but proposes an exhaustive set of elements that is necessary for the sufficient description of learning objects. LOM not only specifies the set of elements together with constraints for order, value range and basic data type, but also groups the elements into categories. Similar to IMDI implicit structure is defined by including aggregate and simple elements. Also the LOM initiative started to define controlled vocabularies for a number of data elements.

In the area of Content Management the ***CEI/IEC 82045-1*** standard has been established. It is a joint effort (JWG15) for a metadata element set from ISO TC10 and IEC SC3B. The proposal speaks about management data as data about the content of an electronic or paper document, necessary to manage it in an Electronic Document Management System. Based on a broad analysis of possible documents and document collections including their various types of relationships during their life-time, an exhaustive metadata set is developed. Also here the many elements are grouped together in a hierarchy for intelligibility reasons.

Repository Initiatives

Beyond those initiatives that defined standards such as described above, there are also initiatives that are building up relevant repositories in the linguistic domain covering metadata. Here, we want to refer to initiatives such as the ***DFKI Tool Repository***, the ***COLLATE*** Project at DFKI and the ***TELRI*** initiative.

The first two repositories will be merged, since ***COLLATE*** is intended to become a large collection of information about people, projects, initiatives, publications and tools in the area of especially language engineering. The tool repository is built upon a well-documented taxonomy that is ready to be formulated in terms of a metadata element set for tools. Other information in COLLATE is not yet so much subject of specific metadata description standards, although information about people can be formulated according to some norm which is also used by other initiatives such as IMDI. The ***TELRI*** initiative is intended to build up a resource database, but wants to combine this also with information about tools. Until now there is no intention to use formal methods, and also the advice about tools will be given by personal information.

Terminology Initiatives

Initiatives for standardized and open terminology repositories (monolingual and multilingual) are of extreme relevance since they will contain the definitions of

terminology units, i.e. concepts and the corresponding terms in the various languages. With respect to language resources a number of initiatives and standards have to be mentioned. ***ISO 12620*** has defined a number of types of data items - where each type is called a data category. The types are grouped into 4 major categories and serve to describe terms: (1) The first category covers the data category "term" itself; (2) The second covers all term-related information which have to be associated with a term such as usage and etymological information; (3) The third covers descriptive information which describes the meaning and relates the corresponding concept entry to other concepts; (4) The fourth covers the typical administrative information. ISO 12620 does not specify the structure of term entries, i.e. nothing is said about the structural embedding of the different data items. Similar statements can be made for the ***OLIF2*** proposal that lists a number of data categories occurring in lexica.

ISO (FDIS) 12200 is a concept-based interchange format for terminology databases in SGML described with the help of a DTD. MARTIF makes use of the data categories defined in 12620 without restricting data categories to the about 150 defined within 12620. MARTIF documents have global information (typical header information), a set of concept entries (body) and a set of references to shared documents. MARTIF separates into negotiated and blind MARTIF. While the first describes a flexible interchange format that two or more partners can agree upon, the latter refers to a pre-defined standard that everyone participating will accept blindly.

CLS is a framework to define the structure and content of terminology databases and the references that can occur. CLS is very much inline with ISO 12620 and 12200. ***SALT*** can be seen as a very recent initiative to extend the terminology work to lexica, in particular those used in the translation business. This step seems to be a natural one since term information is very much related to linguistic information in lexica. SALT therefore wants mainly to test and define an XML-based lexicon/terminology interchange format called XLT and to provide tools such as an editor that allows the user to create RDF (Resource Description Framework) descriptions for complex terminology units.

Metadata Integration Initiatives

In this chapter we want to briefly discuss integration mechanisms for metadata proposals as far as they are known to us. We can distinguish a few different approaches where integration is an objective but on various unrelated levels:

- The ***OAI*** (Open Archives Initiative) approach is focusing on defining a searchable domain for metadata such that services can be built on top of search engines using structured metadata information. A simple protocol for metadata harvesting was defined. OAI is offering the means to harvest metadata records by service providers and requires that the data providers at least offer the records with DC elements. The implementation of any mapping between a domain specific element set to DC is left to the data provider.
- The ***IMDI*** approach is focusing on the management of language resources that includes both aspects: (1) creating a searchable domain as indicated above and (2) creating a linked domain of metadata descriptions which is browsable and which allows to include many different data types. IMDI therefore makes use of simple mechanisms to create an integrated metadata domain for browsing. For search it makes use of distributed databases accessed through the http protocol.
- The ***COLLATE*** initiative wants to establish linked html pages with various types of meta information in the area of language resources and use

advanced IE technology to automatically create relational links between elements of the documents included.

- The ***INTERA*** project wants to integrate the two complementary types of repositories (resource and tool repository) to facilitate the selection and execution of tools on chosen resources by interacting agents.
- The ***E-MeId*** initiative wants to act as service provider for metadata in the area of language resources. In accordance to the OLAC standard it intends to combine metadata about resources, tools and advice and allows the user to combine the three types of resources by manual intervention. Also OLAC makes use of the OAI harvesting protocol.

RDF

Due to its assumed relevance for the future ***RDF*** (Resource Description Framework) is mentioned separately. The RDF initiative by W3C is focusing on defining a common framework for complex metadata scenarios. It allows the designers to define data categories (simple or complex categories), their constraints and controlled vocabularies, and in particular the relations within a metadata set or between different metadata sets. Due to its goals RDF will open the gate to define metadata element sets such that they can play a role in the emerging Semantic Web. In the Semantic Web agents will operate on the element definitions and relations as defined in open, machine-readable repositories.

5. State and Challenges for the Future

The creation of web-accessible metadata descriptions to facilitate the management and the discovery of language resources is a comparatively new concept. All initiatives discussed are relatively young and have undergone more or less a highly dynamic phase. Therefore, it is too early to draw conclusions. A number of relevant questions such as “What are the typical queries the different user groups will input?” or “What are the widely agreed elements and controlled vocabularies?” cannot be answered yet. Therefore, it is good that the mentioned multiple initiatives were started. Also it seems that we are approaching a situation with a critical mass of resources described by metadata. It may motivate other groups and individuals to join. The community can gather experiences with metadata and different approaches such that in a few years conclusions can be drawn.

The variety of approaches and initiatives can at first glance be seen as a disadvantage. However, it was already foreseen by the driving forces behind Dublin Core that a scenario with many different metadata approaches will emerge, since every community and even sub-community has different views about real entities and that the multiple views should not be integrated per se to one complex description standard. Only the experience will show which concepts are flexible and stable enough. Having foreseen this scenario with multiple approaches and initiatives they also started a discussion about how to achieve interoperability. As already mentioned Dublin Core seems to be the accepted pidgin set by most of the initiatives such that mappings of their element sets to DC categories are provided. But these mappings imply the loss of information or bear the danger of an increased creolization of the DC set.

The emergence of the Resource Description Framework and the elaborations about an ABC model indicate possible more advanced solutions for the interoperability problem. RDF imposes formal and machine-readable structure on top of XML to support consistent representation of semantic relations. Even more exhaustive standards such as DAML/OIL will add further possibilities to express semantics. It can be expected that we end up in modular and highly structured metadata sets that refer to open repositories with specifications of terms (elements and values of controlled vocabularies) and relations between them.

Such a network of interrelated machine-readable metadata components opens the view to the Semantic Web introduced recently by T. Berners-Lee. Here intelligent agents use the term definitions and relations stored in the web to execute smart searches and other tasks for the user.

Another view to metadata for future scenarios is that it will allow to blindly executing operations as they are used for example in Information Extraction. The well-known GATE system defines a framework where different IE components can be executed in a chain reusing management information. Each component can add the necessary information to the metadata description. Amongst other purposes this information indicates which NLP components can be executed next and where these components can find the relevant information in a distributed scenario.

Summarizing we can say that a number of metadata initiatives are maturing and that the infrastructures they propose are more and more accepted by the communities. Often they propose a mapping to the simple Dublin Core set as a first basis for interoperability. This will allow us to gather broad experience with various approaches. This experience will be necessary to discuss and design the framework we need to realize the metadata infrastructure for the Semantic Web.

6. Construction of WI-2

The ISO working item on metadata descriptions about language resources has to be constructed to include all initiatives and projects which are clearly devoted to dealing with formal metadata and which have an interest in using metadata descriptions. The following initiatives are relevant in this respect and have to be in the list of official partners:

- TEI: as initiative having worked extensively in defining structures of textual resources
- DC: as most important metadata initiative world-wide with a claim for general coverage and interoperability
- OLAC: as the DC-based initiative in the domain of language resources
- IMDI: as the initiative in the domain of language resources covering more detailed descriptions
- MPEG7: as a highly relevant initiative in a closely related domain
- IMS/LOM: also as a highly relevant initiative in a closely related domain
- IEC 82045-1: also as a highly relevant initiative in a closely related domain
- COLLATE: as an initiative gathering much data in the domain of language resources almost ready to define formal metadata sets
- Terminology Initiatives: as initiatives which know much about the definition of data categories
- RDF/W3C: as an initiative which has the experts to show the way from metadata to the Semantic Web

The following contacts have already been established at an official level: OLAC, IMDI, MPEG7, LOM, IEC 82045-1 and COLLATE. Contacts to W3C, TEI and DC have to be renewed.

Further, we have to include expert users from the different linguistic sub disciplines working with metadata or having insights in metadata requirements:

- corpus and field linguistics
- language engineering
    - text-based work
    - multimodal work

- artificial intelligence
- phonetics
- psycholinguistics

Amongst these there should be experts for corpora, lexica and other data types being used in the domain of language resources.

7. Concrete Steps

First, the official statements are repeated. The TC37/SC4 meeting initiated a task force and specified the following:

- as focus:
  - produce an overview about existing projects/initiatives and monitor its usage
  - link with activities of the emerging Semantic Web
- as tasks:
  - provide a clear picture of the needs of the other WGs
  - identify the experts
  - draft a requirement document until the end of 2002

The TC37/SC4 resolution document adds another point explicitly: create a basic paper on goals and views.

This paper is seen as a basic paper on goals and views. It also includes an overview about relevant initiatives, makes first statements about the directions the Semantic Web may take and describes from which initiatives and fields experts should be invited to build the task-force. The task of WI-2 can be split into two major phases: (1) In the current phase metadata initiatives worked out excellent proposals that are in operation right now. ISO should gather the experiences made with these approaches. (2) Based on the experiences and a requirement analysis WI-2 should work out proposals that meet future needs.

In phase 1 WI-2 should carry out the following concrete steps:

- define simple schemas that are used to define elements and vocabularies
- enforce agreements on a number of controlled vocabularies
- take care that all elements and vocabularies used in IMDI and OLAC are well-defined in accordance with the schemas
- take care that these elements are available via open repositories
- work out the Semantic Web scenario.

It will require more preparation and discussions to work out the tasks for the future.