ELSNET-3

Report of the meeting for the description of an European initiative in the field of metadata for LRs

Deliverable SD.7a.2.1

Antonio Zampolli	Università di Pisa & ILC-CNR
Nicoletta Calzolari	ILC-CNR
Alessandro Lenci	Università di Pisa & ILC-CNR
Peter Wittenburg	Max-Planck-Institute for Psycholinguistics
Daan Broeder	Max-Planck-Institute for Psycholinguistics

Following a preliminary analysis concerning the ongoing activities on metadata for LRs in Europe and in the US, ELSNET representatives have participated into various meetings with researchers active in the field of metadata:

1. Meeting with the Isle MetaData Initiative group (IMDI) (autumn of 2001) – discussion of the ongoing situation on metadata in Europe and prospects of development.

2. In the context of the Workshop of the ISLE Computational Lexicon Working Group (Pisa, March 2001), representatives of the ISLE IMDI group have discussed the problem of metadata with specific reference to computational lexical resources.

3. A representative of the metadata community in Europe (Romary) has taken part to the ISLE Panel during the LREC 2002 Conference, Las Palmas, Canary islands. Many synergies have emerged with the current activities of the Computational Lexicon Working Group, aimed at establish a general standardized infrastructure for the development of multilingual lexical resources.

4. During the same LREC 2002 Conference, members of the OLAC consortium and of the ISLE IMDI group have met and discussed the possible synergies between the activities of the two groups.

As a result of these meetings, a complex and articulated situation has emerged in the field of metadata. Actually, various differences exist between the OLAC activities in the US and the current work of the IMDI group in Europe. Nevertheless, it is evident that both groups are looking for new ways to enhance the possible synergies among the two groups. With respect to this issue, further effort will be necessary on the ELSNET side to promote a more fruitful encounter between OLAC and IMDI, sponsoring new joint meetings.

It is important to remark that thanks to these meetings, the relation between the LRs field and the activities concerning metadata has increased. This has had a particularly crucial influence on the recent preparation and submission of an Expression of Interest for a project on the development of an Open Infrastructure for lexical resources, in which metadata have a key role.

The present report includes three documents attached below, prepared by Peter Wittenburg and Daan Broeder:

1. an overview paper presented at LREC 2002 in Las Palmas containing a broad overview of the current initiatives on metadata for LRs (section 1);
2. a document illustrating the plan for metadata for LRs approved by the ISO/TC 37/SC4 Committee (section 2);
3. a document illustrating the results and decisions taken at the OLAC – IMDI meeting held at LREC 2002 (section 3).

Section 1

Metadata for LRs: an Overview

P. Wittenburg, Daan Broeder

1. Introduction

At the LREC conference 2000 a first workshop was held which was dedicated to the issue of metadata descriptions for Language Resources. It was also the official birth of the ISLE project (International Standards for Language Engineering) which has a European and an American branch. The workshop was also the moment where the European branch presented the White Paper 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.

Around the same time frame the Dublin Core initiative mainly driven by librarians and archivists finished their work on the Dublin Core Metadata Element Set (DCMES) and the MPEG community driven by the film and media industry started their MPEG7 initiative. All these initiatives are closely related since they claim to address the community that deals about Language Resources partly focusing on those based on multimedia extensions.

After two years of hard work and utterly dynamic developments it seems to be a good time to describe the 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 realm of the mentioned initiatives the term "metadata" refers to a set of descriptors that allows the community members to easily discover and manage language resources in the distributed environment given by the World-Wide-Web.

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

Also in some language resource archives metadata in this sense was used. A famous example is the header information in the CHILDES database. These early project specific definitions were the basis for the important work about header information within the TEI initiative (Text Encoding Initiative) which was later taken over by the Corpus Encoding Standard (CES) 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 structural part of SGML structured documents. It still can serve as 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 for example followed by the Dutch Spoken Corpus project.

Despite some projects and initiatives the concept of uniform metadata descriptions following the TEI standard was not widely accepted from different reasons. Many found the TEI/CES descriptions too difficult to understand and too costly to apply. Others had another view of their resources not matching with the TEI type of categorization. Mostly, people seem to not have taken the time to look through the TEI suggestions.

It may not be forgotten to mention 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 such that each employee can easily identify whether useful resources are available. For example Lernard&Houtspie used such a database internally[1]. Similar statements are true for the great data centers such as LDC and ELRA. They 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 use the common web-site approach where they describe their holdings without using a formal framework such as metadata.

2.2. Classification Aspects

The assignment of a metadata description to a resource means classifying it. 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 non-Euclidian space. 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 values can appear along different dimensions. Further, we can't define metrics along the axes.

A classification has to be based on a similarity 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

[1] It was not possible to get a blue-print of the structure of that database.

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.

This raises many problematic questions, of course, especially in communities such as the linguistic one. There is no settled and widely agreed ontology for language resource yet. Linguistic theories will 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 scheme, the question raises: how can interoperability be achieved, i.e. how can different categorizations be mapped onto each other? These questions are not solved and they are not simple to be solved.

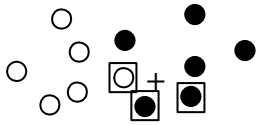


Figure 1 - Two elements are represented by black and light dots. Each dot denotes a possible value of the respective element in some non-Euclidian space. The cross may indicate the "location" of the resource and the rectangles as the optimal choice for describing that resource.

A solution chosen by the IMDI initiative is to allow for flexibility, i.e. allow to add elements (dimensions of description/categorization) and to make those controlled vocabularies user extendable where there is no set established yet. At first glance this solution seems to sound sympathetic, but it bears dangers as is known from classification literature. We would like to indicate the problems with one example (fig 2). Individual users could decide to add an element to a dimension that does not seem to be characteristic for the point in space and thereby distorting the "orthogonality" of the dimensions and creating problems for proper discovery.

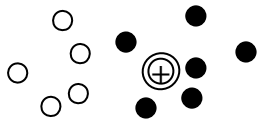


Figure 2 - An additional value is created (double circle) for one dimension (light circles) in an area where another dimension (black circles) is dominant.

Users could just use single resources to 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 during resource discovery can be predicted.

On the other hand closing a vocabulary too early 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 medium position was taken in those cases. A pre-defined vocabulary is proposed and at regular instances the actually used vocabulary will be screened to detect problem areas. 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 seen as obsolete by the community. One question remains: who is responsible for deciding about such matters. This is a social 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 numbers 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. 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 of the amount of information available in 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 occurrence and co-occurrence of words in the titles or in the texts of web documents to find most suitable resources and calculate a 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 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 are will not be the only choice for professional users.

The PICS initiative showed that even for general web-based information there is a need for additional type of descriptors that cannot reliably extracted from the texts. So, metadata descriptions, i.e. descriptions about the resource 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.

3.2. Language Resource Domain

All these information retrieval 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 resource 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 statistical engines? They would prefer to operate in a structured space obviously organized by continents and languages to discover the resources they are looking for.

Also in the language resource domain we are faced with a gigantic increase of the amount of resources and their inherent complexity. 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 were faced with a 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. Moving from purely textual to multimedia resources with multimodal annotations caused this. It was understood early that the traditional methods of management and discovery mostly on purely individual account led increasingly often to problems. Scientists could not find relevant data anymore in an easy way and when a researcher left the institute the chaos was complete. It is known that in other research centers, universities and also in industry similar situations occur.

Unified type of metadata descriptions where the descriptors were intuitively understood by everyone and where each individual researcher can easily integrate his resources and resource descriptions were seen as the solution in the institute. These descriptions should include enough information such that it can directly be seen for a linguist whether the material is relevant for his research question at that moment and given an interest that it is possible to immediately start tools on them. Queries such as "give me all resources which contain Yaminyung spoken from 6 year old females" should lead to appropriate hits.

It was clear that most of these descriptions had to be created manually since only in few cases they could be automatically extracted from directory path names, Excel sheets or other sort of systematic descriptions. As mentioned before the great majority of the resources are of a sort that the descriptors can't be anticipated from the content.

3.3. New Metadata Aspects

The trend to an extremely 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 the knowledge about their existence with others from the community. Metadata descriptions as explained seem to have a great potential to help the researchers to manage these resources and to simplify their discovery.

While the TEI designers focused on text documents, currently language resources mostly have multimedia extensions (sound and/or video). This puts new requirements on how to choose the descriptor set. Further, it is general agreement that the purpose of metadata set is not so much a complete description of a resource, but to easily support the resource discovery and the resource management. This way of looking at metadata is certainly inspired by the very important work in the Dublin Core initiative.

At this moment no one can say which type of descriptor set is necessary to facilitate discovery and management, since for the domain of language resources the metadata concept is very new and hardly used yet. We are confronted with different type of users all having different requirements which we don't know in all detail:

- the researcher and developer who is an expert and wants to quickly find exactly those resources which fit to his/her research or development task[2];
- 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[3];
- 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;

[2] For a speech engineer it may be relevant to find resources where short-range microphones were used.

[3] For a resource manager it might be relevant to find all resources with speakers of a certain age.

- the journalist who is interested to get 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 wishes we don't know yet.

An utterly important point is that many of the language resource archives currently set up have a long-term perspective. The question of their typical usage becomes an even more mystic one, since we cannot anticipate what future generations will need to discover resources. A widely used statement in such cases is to make the descriptor set exhaustive. Despite the fact that exhaustive sets will not be created due to the labor needed and due to its 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 by more detail.

Given these uncertainties about the future user needs it makes sense to start now with a non-exhaustive element set. Language resource creators are reluctant to invest time for information that will help primarily others. To high requirements will lead to a negative attitude. However, we can anticipate that the metadata that we are gathering now will not be sufficient in the long run.

Another new phenomenon is that individual researchers have to participate in the creation and integration of metadata descriptions. There is no time to read lengthy documents about the usage of elements. Everything has to be simple and straightforward, otherwise he/she will not participate. Metadata descriptions also should facilitate international collaboration. In many disciplines international collaborations with researchers located at different places are normal. Contributions from one of them must be directly visible by the others. This requires metadata descriptions which are dynamic to a certain extent and which support an event model.

3.4. Resource Management Aspects

The primary task of metadata is resource discovery. However, resources 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 were used to manage large amount of resources. Beyond the discovery management includes operations such as grouping related resources, copying valuable resources together with their context, handling versions of resources, removing resources and maintain access lists. Until a few years ago resource management was facilitated by the small size of the files. Using physical structuring schemes such as directories mostly management is done.

However, for the modern multimedia based archives of institutions and of individual researchers files and corpora are becoming so huge that the physical 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 not done anymore on physical layer that often implies copying large media files, but on the level of metadata that only means to define useful metadata hierarchies and to set the pointers to the resources wherever the system management may have stored them.

Resource management has become another dimension due to the distributed nature of resources in the Internet scenario. It will become a normal scenario for the future that a video file is hosted on a certain server and that two collaborators work on the same media file simultaneously, for example one by annotating gestures and the other by 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 could be used to point to the location and to allow management operations as if the resources where all bundled on a single server.

4. Language Resource Data Types

Before introducing the different metadata initiatives it is necessary to describe the characteristics of the objects that have to be described. As already indicated not all objects that we come across 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 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 can include multimedia material or are based on media recordings put new requirements that could not be addressed beforehand. 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 other sort of aspects. In many MMLR it is relevant to memorize that a certain annotation tier has special links with a certain media track. For speech engineers it could be of relevance 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. 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.

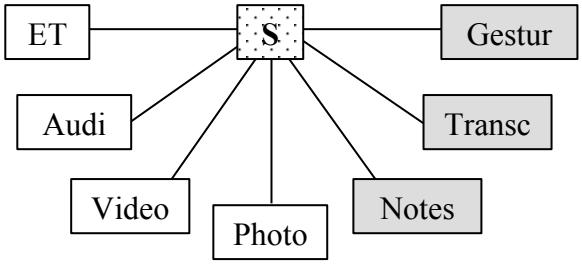


Figure 3 - Various types of information tightly related by a common time axis.

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.

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 have to be indicated by metadata. They can be treated as normal documents except that the 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 which follow more or less the new paradigm 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 librarians and archivists had was invested in the definition of the core set. One explicit goal was to end up with a significantly lighter set than defined within the librarians MARC standard. 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 utterly sloppy definitions of several of the DC elements. This strength is at the same time its weakness.

The designers understood the limitations and problems of this approach. The Dublin Core initiative anticipated the need for other elements and the Warwick Framework was described as a modular set of metadata standards using domain specific element sets. Many initiatives work along the DC suggestions by modifying the elements set in a number of dimensions, others started from scratch, however, accepting the underlying principle of simplicity. The modifications were 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 qualifycation 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 only according to the W3C recommendations). (3) The usage of new elements.

The DC initiative itself defined qualifiers and constraints for a number of elements. 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 camps were fighting about the way to go. The ones that are in favor of a controlled extension would control the semantic scope, but force communities away from starting with the DCMES. The other camp would allow communities to join, but loose control on the semantics of the elements. In the latter case DCMES would become a container for all sort of information where querying would not result in satisfying results.

DCMI did not formulate any syntactic specification. 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. 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 subelements 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;

They report about much confusion in the DC community through the usage of these uncontrolled extensions. In a note of how to implement DC with XML they introduce the notion of "dcterms" which are "other elements recommended by DCMI". The note 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 refinements should be treated in 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 note about the implementation of DC with RDF[4]/XML. 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 an equality is made between "qualification", "refinement" and "subelement" (dot notation in the HTML implementation). The statement made in the XML implementation note that refinements are elements in their own is not made again. This view would lead to a complete openness of the element semantics.

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 interoperability between different metadata sets that are currently being defined within many disciplines. DC may form the widely accepted minimal consent. The inherent problems with the DC metadata set are depicted in the following graph.

[4] RDF = Resource Description Framework worked out by W3C. RDF will be discussed later in this paper.

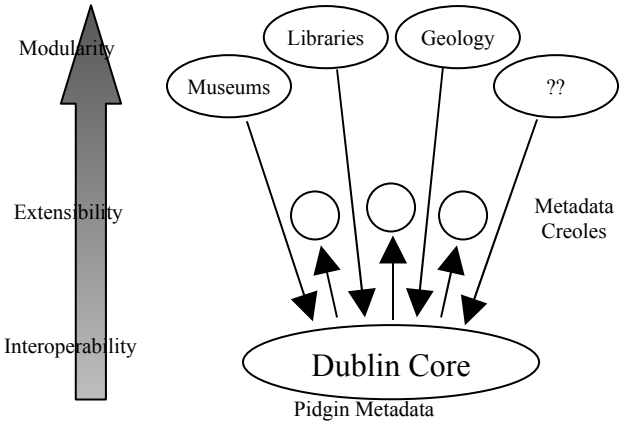


Figure 4 - The principal problem with which DC had to cope. Interoperability leads to a pidginize form of metadata that is simple enough for the casual web user. Domain specific and specialized metadata sets are the other extreme with less potential for interoperability. In between there are many creolized forms that try to operate with extensions of DC.

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:

- It defined 3 attributes to support OLAC specific qualifications (***refine*** to refine element semantics including controlled vocabularies; ***scheme*** to refer to an external controlled vocabulary; ***lang*** to specify the language a description is in).
- ***Code*** attributes refer to element specific encoding schemes.
- 8 new ***subelements*** 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 subelements controlled vocabularies are under preparation and their definition is part of the schema defining the metadata set.

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. From a functional point of view it seems that the ***refine*** attribute is similar to ***subelements***. Both narrow down the scope of the element definition. The refine attribute obviously works on the element as a whole. OLAC wants to keep control of the CV, i.e. there is no user definable area, but there is a description of a ***process*** that defines how definitions can be adapted.

The ***code*** attribute acts as a constraint 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 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) 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 passed through 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 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.

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[5] 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. A presentation of the requirements and the needed elements in the European DC Usage Committee revealed that it did not seem to be advisable to use DC as basis.

Due to the needed 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 that only dealt with semantics, IMDI also introduced structure. Structure makes it possible to associate for example a role, an age and spoken languages with an informant.

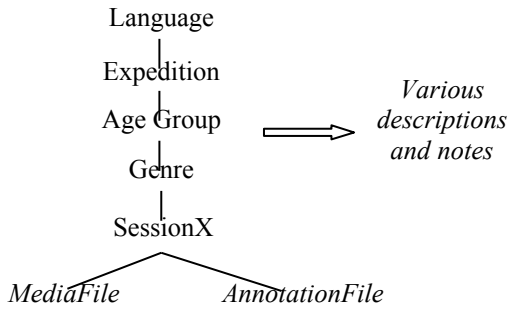


Figure 5 - A typical metadata hierarchy with nodes representing abstraction layers. Each layer can contain references to various descriptions and notes and thereby integrating them into the corpus. All components of such a hierarchy can reside on different servers. The session nodes are the leafs in 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 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 information using domain specific element names. 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 dimensions currently describing resource contents in IMDI are not seen as the final proposal. They will have to undergo a major update to achieve a higher degree of coverage and acceptance.

The IMDI set was chosen such that most elements are suitable for searching, but some are for quick inspection. 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 feature can be of great use especially for projects that feel that their specific wishes are not addressed by the IMDI set. This feature was used for example by the Dutch Spoken Corpus project 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 are closed such as 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.

[5] Yaminyung is a language spoken by Australian aborigines.

The recent proposal for lexicon metadata 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 differ as well, it was decided to not describe structural phenomena, but to only 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 should not be objects which have their own metadata description, 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 is not satisfyingly solved within IMDI yet.

IMDI is been creating a metadata environment existing 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 vocabulary. Since the tools are described elsewhere in greater detail, 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 distributed hierarchies allowing each user to create his own resource domain, but easily hooking it up to a larger domain. The browser also allows creating nodes to form browsable hierarchies, such that a user can easily create his own preferred view on a resource domain. It also allows the user to add configuration specifications such that tools of his choice can be easily started from the browser once found suitable resources.

The search component is an integral part of the browser to increase the comfort for easy navigation. The version operating on one metadata repository is ready. However, searching in a distributed domain has to be finished. It will make use of the OAI metadata harvesting protocol that can harvest metadata records of all sorts. The infrastructural aspects have to be solved yet, i.e. how to gather metadata information residing at different locations in an efficient way. 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 scripts available for this type of operation at this moment have to be integrated to the existing browser and editor.

IMDI has accepted that there are different types of users. The casual web user wanting 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. Of course, such a mapping cannot be done without loosing 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 comply with it.

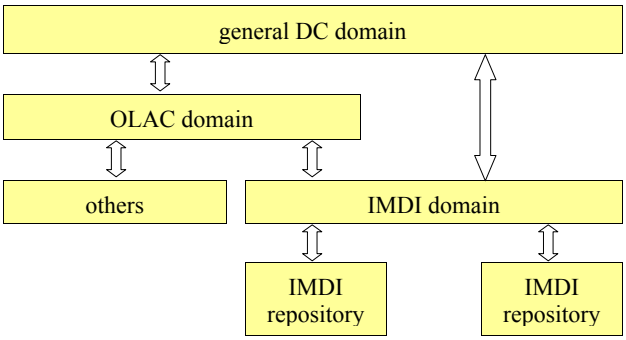


Figure 6 - IMDI's vision about metadata services users should be able to use. It is not indicated that the general DC domain covers many more domains than just the domain of language resources.

The way IMDI repository connectivity is done is different from how OLAC connectivity is achieved. While 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 domain to become member of it.

5.4. MPEG7 Initiative

In contrast to the initiatives discussed beforehand 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 standards, MPEG7 is a standard for describing multimedia content. It is based on the experiences with earlier standards such as SMPTE. 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 multimedia content (can be also segments), 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 sent. 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 part 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)[6]. 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 can contain a hierarchy of summaries.

Also the MPEG7 community accepted the role of DC as an umbrella for interoperability during searching. In the Harmony project a mapping of suitable MPEG7 elements was worked out. Finally, it was decided to suggest 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. 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, 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.

[6] Only two additional primitive data types (time and duration) and array and matrix data types were added to cope with the needs.

- Which approach is the most suitable one (if there will be any answer to this question at all)? We need many years still to find out.
- What are the typical queries the different user groups are inputting? We don't know yet, we need a critical mass and interesting environments to be able to answer this question.
- Is DC compliance a relevant goal and which information has to be passed over? Again this question cannot be answered yet, since we miss a relevant usage scenario. Will the casual web user ever be interested to find out which resources are available in Yaminyung?
- 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.
- Which kind of tools do we need to support the resource creators and managers? Some initiatives just start 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 just started, so any answer at this moment may turn out to be wrong.

	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	?

Table 1 - A quick overview about the goals and major characteristics of the relevant metadata proposals.

For many questions we don't know the answers 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 be able to discover what is available. We also know that the individuals and institutions have got a management problem to solve and that the traditional methods are not suitable any longer. So the step to introduce metadata descriptions seemed to be an obvious one. But we don't yet fully understand the potential of web-based metadata.

Resource discovery can't be the only goal. Resource management is equally important. Most important for the users is the view to step away from all sorts of details that have to do with 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 with help of which they can 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 which still is 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. According to him, DC has to be seen as one simple view on certain type 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 and the elaborations about an ABC model for metadata interoperability 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 doing so explore the new metadata landscape. A close network of collaboration and interaction seems to be necessary to discuss the experiences. Probably 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 asking persons to contribute take over a high amount of responsibility. Given that our assumption is true that we will have an ongoing dynamic development[7] 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 loosing the valuable information that has been gathered.

[7] The IMDI-OLAC mapping document was created at ?? 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.

International Organization for Standardization

ISO/TC 37/SC4 **Language Resource Management**

WG 1: Descriptors and Mechanisms for Language Resources
WI-2: Multimodal and Multilingual Information Documentation

Section 2

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.

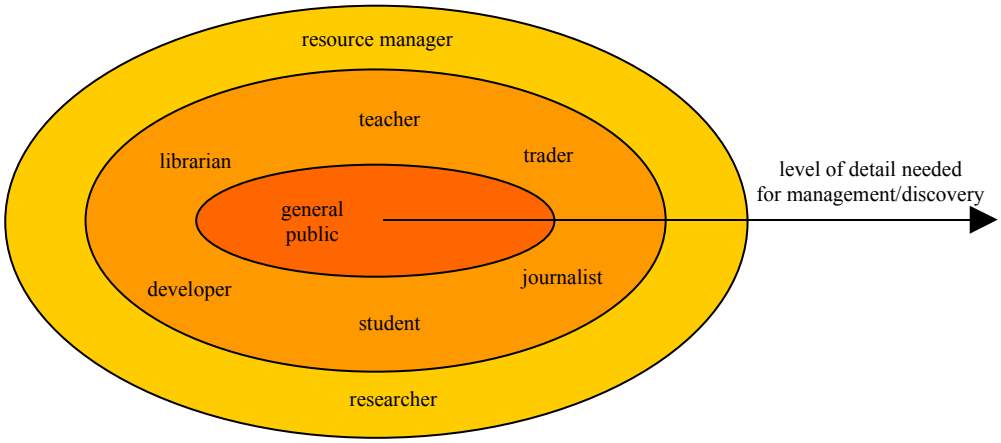
Fact is that when comparing different sets such as Dublin Core and CEI/IEC 82045-1 from the field of document management for example, that the differences are relatively large although both speak about documents with texts covered in them. Therefore, we need to restrict the scope of "Language Resources" 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

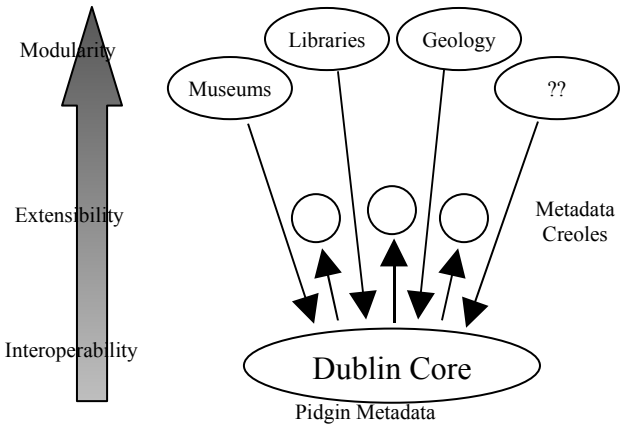
In recent years, many language resource projects and tools supported typical metadata information describing some main characteristics of the resources. An example is the ***CHAT*** header of the CHILDES project. Much experience has been assembled when the well-known ***TEI*** header standard was worked 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-Meld*** 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.

Section 3

OLAC-IMDI Meeting: Results and Decisions

LREC 2002, Las Palmas

In addition to the enclosed ISO document in section 2, it seems necessary to add some comments about the current situation with respect to metadata for language resources.

1. We have two metadata sets which are relevant for Language Resources: OLAC and IMDI. MPEG7 might be an alternative, but they are far away from being implemented and from offering specific schemas for the LR community.
2. Both initiatives have undergone a dynamic development. While IMDI started from the beginning as a set offering detail for the community, OLAC started as a lightweight metadata set with simply one new sub-element: Subject.Language. In the meantime OLAC added more linguistically relevant classifiers such as to describe linguistic data type and necessary refinements of the vaguely defined DC set.
3. IMDI will continue to offer specificity where necessary. For example for lexica a specific metadata set was designed by 4 experts as a consequence of the MILE discussions.
4. At DFKI the experts will design a metadata set to describe language resource tools, since also here OLAC does not offer enough detail.
5. While OLAC basically offers a search mechanism for resources, tools and advice, IMDI offers infrastructure for resource management, browsing facilities, searching and a direct, user configurable option to start tools.
6. Both metadata sets will have to develop to be useful for Semantic Web type of activities. While IMDI maps the structural aspects of the language resource objects, OLAC would have to adapt substantially to be useful for automatic procedures.
7. While OLAC is very active on a political level and makes agreements etc, IMDI simply offers infrastructure and tools for the interested community.

Summarizing one can say that OLAC has another more global focus while IMDI offers more detail for the interested community and a potential for automatic services of the future. Given the comments during the LREC meeting we expect that OLAC will extend their set stepwise to fulfill the needs of their current community, i.e. they will move away from the DC set more and more. In a Semantic Web scenario we expect that both sets will merge to a new more flexible RDF-based framework. Therefore, it is good for the community to test out two different approaches.

Based on these facts, the OLAC and the IMDI motivators sat together during the LREC conference to discuss interoperability issues. These are the agreements:

- The IMDI domain will be searchable from the OLAC domain. Only a restricted set of IMDI-elements can be queried.
- The OLAC domain will be searchable from the IMDI domain. Not all queries can be extended in all detail on OLAC compliant metadata records.
- We will agree on a simple XML-based schema for controlled vocabularies to be able to use each other's vocabularies.
- We will agree as extensively as possible to use the same controlled vocabularies.
- We will further collaborate under the ISO umbrella where Steven Bird and Peter Wittenburg are both members of WI-2, which is about metadata descriptions.