Metadata Infrastructure for Language Resources and Technology

2009-02-04 - Version 5

Editors: Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Erhard Hinrichs, Lothar Lemnitzer, Dieter Van Uytvanck, Andreas Witt, Peter Wittenburg

The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.

Metadata Infrastructure for Language Resources and Technology

CLARIN-2008-5

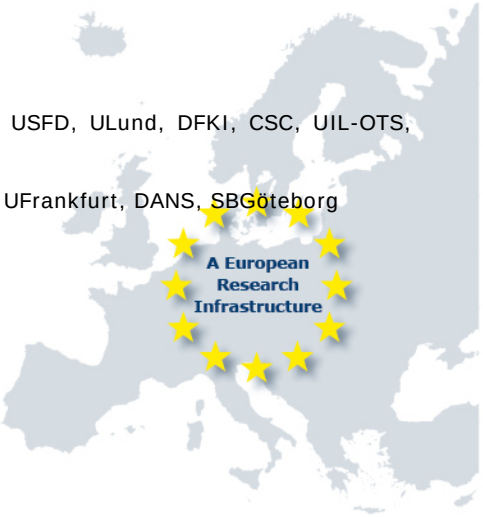
EC FP7 project no. 212230

Deliverable: D2.4 - Deadline: 1.7.2008 (postponed to 1.12.2008 due to late start)

Responsible: Peter Wittenburg

Contributing Partners: MPI, INL, OTA, RACAI, WROCUT, UPF, ELDA, ILSP, ILC, USFD, ULund, DFKI, CSC, UIL-OTS, ULeuven, AKSIS, ATILF, UTuebingen, HASRIL, CST, UTartu

Contributing Members: ULeipzig, UMasaryk, CELTA, TILDE, Meertens, IDS, BBAW, UFrankfurt, DANS, SBGöteborg



Scope of the Document

This document gives an overview about how metadata descriptions are used until now, what the deficits of the current infrastructures are and which lessons we as community learned from about a decade of experience. Based on this the requirements for a new CLARIN approach are being worked out.

This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

CLARIN References

• CLARIN Centers Types	CLARIN-2008-1	May 2008
• CLARIN Centers	CLARIN-2008-3	August 2008
• CLARIN Persistent Identifiers	CLARIN-2008-2	October 2008
• CLARIN LRT Federation	CLARIN-2008-4	October 2008

Contents

1 Introduction

Descriptive metadata that is characterizing a resource with the help of keyword-value pairs is becoming increasingly important to manage and find electronic resources in a time where the sheer amount of resources and the complexity of the relations between them are increasing in an unforeseen way. Many communities, in particular the library community, recognized about two decades ago the need for new sets of descriptive elements and a new electronic infrastructure. These are necessary to create smoothly operating environments for users to manage and find research data which goes beyond the traditional publications. About a decade ago in particular two initiatives (IMDI, OLAC) started to use descriptive metadata systematically in the field of language resources. At DFKI an attempt was made to come to a central registry of tools and the Text Encoding Initiative worked out suggestions for header elements to be used. Little later the ENABLER initiative came out with a broad overview about which metadata concepts were being introduced by the various initiatives.

While the major initiative from the librarians (Dublin Core) was started in the nineties, the major work in the LRT domain started in 2000. In particular IMDI and OLAC can look back on a broad experience in the usage of metadata in the LRT field, the hesitations of linguists to produce high quality metadata, the ongoing debates about the usefulness of suggested elements and the increasing awareness about the importance and benefits of descriptive metadata in the LRT field. This awareness is paralleled by the increasing understanding of the importance to take care of the preservation of the scientific data.

Three events that took place in 2008 confirmed the growing awareness about the need to pay more attention to preservation and management of research data. The special ESFRI working group on repositories made a clear recommendation that each research infrastructure initiative needs to work out a solution for a proper repository infrastructure. Both the e-IRG meeting in Paris as well as the meeting of the Alliance for Permanent Access in Budapest discussed aspects of data preservation from different points of views - nevertheless indicating the importance of high quality metadata descriptions. At LREC 2004 a new ISO group, ISO TC37/SC4, was founded that has as central topic the "management of language resources". Also in this group descriptive metadata is one important dimension. The central dimension of the work of this group is to come to a central registry of relevant linguistic concepts that can be used as a point of reference to achieve semantic interoperability.

The main lesson learned in the past decade is that there is no "one-schema" approach that will satisfy the needs of all researchers. There are too many different resource types; there are so many different sub-disciplines with their own requirements for elements and terminology and there are resources where only little information is available. The introduction of a central concept registry is the basis for changing the focus from syntax to semantics when talking about interoperability. This is exactly where CLARIN is aiming at and it is fully in line with initiatives such as Dublin Core.

This document wants to give an overview about metadata infrastructures and their usage, describe the linguistic requirements and work out a procedure for the CLARIN work.

2 Definitions

This section intends to define some key terms that will be frequently used in the course of this document. As some of these notions clearly have ambiguous meanings it should be stressed that this definition list should only be considered as a mean to prevent further on that semantic discussions overtake the real substance of this paper.

Application profile

An application profile is a set of elements chosen from different existing metadata sets that is supposed to optimally suited to describe resources in a specific field. The new set is defined in an XML-schema that refers to the elements in their original schemas via a namespace identifier. An application profile may also define a subset of a single existing metadata set.

Metadata [Understanding Metadata]

structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource.

Types of metadata [Understanding Metadata]

Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

Structural metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.

Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; three that sometimes are listed as separate metadata types are:

− Rights management metadata, which deals with intellectual property rights,

− Preservation metadata, which contains information needed to archive and preserve a resource.

− Technical metadata that tells how the files were created and stored (what can be used as specific information for workflow frameworks).

Resource [RFC 3986]

The term "resource" is used in a general sense for whatever might be identified by a URI. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., "today's weather report for Los Angeles"), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., "parent" or "employee"), or numeric values (e.g., zero, one, and infinity).

Metadata registry (short: registry) [Understanding Metadata]

a formal system for the documentation of the element sets, descriptions, semantics, and syntax of one or more metadata schemes

Repository [CiTER]

facility that provides reliable access to managed digital resources

Archive [CiTER]

repository dedicated to the long-term preservation of the associated data

Scheme [Understanding Metadata]

a metadata element set and rules for using it

Metadata harvesting [Understanding Metadata]

a technique for extracting metadata from individual repositories and collecting it in a central catalog

Catalog (or **catalogue)** [WP:catalog]

an organized, detailed, descriptive list of items arranged systematically.

Profile [Understanding Metadata]

(see application profile)

Resource provider [CiTER]

organization that makes a resource available on-line

Bundle (of resources)

Bundle of resources. A collection of tightly related resources that all relate to the same (linguistic) event. Usually all are recordings or descriptions of such an event. The concept is used in IMDI and the AILLA metadata schemas. The bundle is represented by a metadata description, but has no separate URI addressing it.

Complex resource [CiTER]

resource consisting of multiple constituent parts each of which can be accessed individually.

Web service [Brown 2007]

A web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format.

Reference [CiTER]

link to resource stored elsewhere

Crosswalk [WP:crosswalk]

A crosswalk is a knowledge base that shows equivalent elements (or "fields") in more than one schema. It maps the elements in one metadata scheme to the equivalent elements in another scheme.

3 Metadata

3.1 History

For the origins of the concept of metadata, its usage and terminology we have to look at the library world where the problem of tagging and retrieving of large amounts of resources already existed from an early age. The emergence of standards to describe (at the start non-electronic) written resources like the [MARC] standards (MARC, USMARC, MARC 21) (1970 -1998) enabled the exchange of data between different libraries and the creation of general applicable catalogue instruments and software. But it is only with the emergence of interconnected libraries and other repositories that need for standardization of metadata sets becomes essential. Because the technology and experience of the librarians was already advanced compared to other disciplines it was natural for them to take the lead in trying to develop metadata description systems such as DCMI [DC] (1995), that aim to also incorporate other domains. However with this attempt that originally advocated describing all objects with a system of 15 classifiers (although qualifiers for more specificity were allowed), too much focus was put on librarian's terminology and interests (IPR etc.) to enable wide acceptance in other domains. For interoperability between domains that require mutually intelligible resource descriptions, it is still a solution of choice even if much information is lost in translating domain specific metadata into DCMI. Problematic with DCMI is also that it is a flat list of descriptors lacking any structure and making it difficult to describe complex objects.

The tension between the need of adequate and sufficiently rich (domain specific) terminology in order to correctly describe resources and otherwise the need for interoperability where terms have to be understood by people from different disciplines has lead to a kind of rebound effect of description systems moving from small sets with descriptors with broad significance to big complicated sets with highly specific descriptors and back again. Some [Baker 1998] have compared this to the linguistic theory of pidginization and creolization where pidgins arise when mutual intelligibility is needed and pidgins are creolized to achieve richer semantics. How this tension is resolved is often a matter of purpose or pragmatism.

In the linguistic domain that is our focus, metadata has been used already for a long time but not seen as a separate data category. Usually it was present in the headers of text or annotation files as for instance in [CHAT] or [ESF]. But the encoding and semantics of the metadata were all corpus specific and no attempts were made to cover a wider field of resources. The inclusion of the metadata in the resource, although useful from the view of data management, meant also that the resource format was fixed but it would be more accurate to say that a choice particular resource format implied also a choice for a metadata set.

One initiative, the [TEI] (1990) has been very successful in next to establishing a widely accepted system for text annotation also specifying a metadata set and format for types of text resources, the TEI header. Although the metadata is not independent from the TEI annotation format itself, it achieved wide support within the linguistic community that transcends the use for a particular corpus or project. TEI allows extracting the header from a TEI document and storing it as an independent entity in metadata repositories, just as IMDI and OLAC.

To address the need for metadata descriptions for the linguistic domain also suitable for multi-media resources, using a domain specific terminology and supporting complex resources [IMDI] (2000) was created and used in several international projects [ISLE], [INTERA]. Its development has been influenced by other initiatives [CGN], [ECHO], [ENABLER], [MILE] and several years of interaction with various communities including the Sign Language Community. Although applicable to text corpora and lexica, its main focus is the deep description of multi-media/multimodal corpora and endangered language documentation [DOBES]. Although IMDI also has a special set of elements to describe complete corpora with a single metadata record, it is primarily used to describe bundles of tightly related resources. Profiles were created as community specific extensions such as for CGN and the Sign Language community.

As an application of DCMI for the linguistic domain, [OLAC] (2000) was created that adds a (extensible) set of descriptors to the DC set. Although it started with the addition of only a single element, this set has been extended through the years. OLAC also caters for language technology like ACL Tool repository metadata [DFKI]. OLAC is accepted as a metadata exchange format between language resource archives.

The need to describe complex objects independent from the objects themselves has been tried by developing metadata sets especially for this purpose. With IMDI there is an implicit relation between resources being

described by the same descriptive metadata record. A standard like [METS] (2001) focuses especially on the so called structural metadata that describes the relation between the complex object's constituents and is neatly separated from the descriptive metadata for which it allows different descriptive metadata systems being included. It is of course also possible to describe complex objects also by providing information about an object's relations with other objects in the metadata records for the individual objects, but this is more complicated. DCMI provides for this the "RELATION" element, this also practiced by OLAC.

Depending on the purpose and available resources, the level of granularity addressed by the metadata description usually varies. Top-level corpus description usually requires different a set of elements than are needed for individual resource descriptions. The ELDA universal catalogue entries [ELDA UC] is an example of this as is [EAD] (1993), although EAD is not specific for the linguistic domain. IMDI has a different set for each purpose [IMDI].

The need to exchange metadata between repositories to gather the metadata at a single central point has led to the development of special protocols and formats for the transport of metadata records. Some of these protocols require the use of a specific metadata set or the use of a specific set next to other possibilities. Metadata exchange can also take the form of transporting a metadata query to other repositories; the result to the user is similar: a single (virtual) metadata repository. An exchange protocol that plays an important role is OAI-PMH, which relies on HTTP to harvest metadata records by so-called OAI service providers from OAI data providers. The OAI protocol requires at least DCMI metadata in the records that are exchanged. Although maintaining an OAI data provider infrastructure was designed to be simple, it has been noted that in the linguistic domain many repositories cannot cope with it [Hochstenbach 2003]. A very simple way of harvesting metadata is to use a WebCrawler and extract the metadata from web pages or special XML schema based metadata records, the IMDI framework uses the last option to gather metadata in catalogues.

The option of transferring the metadata query instead of the metadata records themselves was standardized for the library world already in pre-web times as [Z39.50] (1970). A replacement using the HTTP protocol was provided by the [SRU]/[SRW] protocols (2005). Z39.50 is widely used in library environments and often incorporated into integrated library systems; the semantics of its queries are Z39.50-specific, defined by the Bib-1 attribute set. In general this strategy of propagating a query to other repositories is called "federated search" and when applied in a cascaded form, can search a very large number of information sources.

When gathering metadata records from repositories using sets with different semantics for the purpose of offering a single catalogue, it is required to perform a translation or mapping of the different elements. A common technique is the metadata crosswalk (see sections 2 and 4.1.1.6) that has for instance been applied in the [ECHO] project (2002). Usually different sets are translated to a pivot set of limited specificity like DC. The loss of information should be balanced against cost of maintaining a limited set of mappings.

Attempts to reuse existing metadata sets when designing metadata descriptors for a certain application led to the emergence of "application profiles". This is a method to have a metadata (XML) schema refer to elements from existing metadata schemas by drawing on the different existing namespaces. Application profiles can also constrain existing element definitions, but only to make the semantics more specific or narrower. The dangers of semantic overlap are real but can be intentional by the profile creator and no real solution is provided. The application profile approach to reuse metadata schemas has been promoted by the [SCHEMAS] project.

Current developments that represent the state of the art of metadata descriptions in the linguistic domain are:

1) the integration of metadata with a PID system like in the [ARK] proposal, [DOI] and the TC37/SC4 attempts and providing a standard for citation of language resources [CiTER] (2008)

2) Attempts to create frameworks where metadata sets or schemas are selected or combined in a controlled way for a particular need. The concept is similar to that of application profiles but the implementation can be different e.g. not using XML schemas. TEI can serve as an example, although the resulting element set is meant for annotation purposes instead of metadata description. The big challenge with this type of work is to avoid the problems of application profiles. This can be achieved by e.g. limiting semantic overlap and improving coherence by using terminology databases or data category registries.

3.2 Role and Parts of Metadata

Keyword-value pairs of descriptions of resources stored have a very long tradition. They are necessary whenever the volume stored becomes big or when the resources have an anonymous state as is the case when different people deposit books or data at a central place for example. In both cases it is important to classify the deposits along a few clever chosen dimensions allowing the professionals to manage and the users to find the data. In the digital age where the sheer amount of data is increasing in unprecedented dimensions and where the complexity of relations between the stored objects are beyond what humans can easily handle, metadata descriptions are the anchor point of any operation. [NISO] states that metadata is key to ensuring that resources will survive and continue to be accessible into the future. Actually metadata replace the resources they stand for in a number of usage scenarios and they even become themselves object of research if they are informative enough.

Recently at two meetings where the role of digital repositories was emphasized [eIRG] and where the costs of preserving data [APA] were analyzed, metadata was very a prominent topic. Almost all speakers referred to the great challenge for data management and preservation posed by the increase of volume and complexity and that metadata plays a crucial role in these efforts. Huc [HUC] argued that the extreme increase of data stored in repositories will even ask for more detailed metadata in the future. [Beagrie 2008] showed that the creation of good quality metadata should be generated in the process of resource creation. According to an investigation in the UK metadata creation after 10 years would be problematic and the costs would rise by a factor of 30 at least.

In the emerging eScience scenario we can expect that metadata will have more roles as just supporting management and retrieval. It will be used to build virtual collections, to carry out automatic profile matching to find appropriate processing components in workflow systems, to do quick inspection by human users, to automatically check in which environments certain operations on resources could take place etc. It will even be used for other than scientific purposes such as evaluating the visibility and quality of research work for example.

To fulfil all these criteria, metadata needs to have various types of classifications. In [Understanding metadata] the following distinctions are made:

- descriptive metadata that describes a resource for purposes such as discovery and identification
- structural metadata that describe how compound objects are put together
- administrative metadata that provides information to help manage a resource - here they distinguish two sub-types:
    - rights management metadata
    - preservation metadata necessary for long-term survival

These dimensions are rather general which is why we would like to add a few topics. We need elements that describe (1) where and when a resource was created, (2) who created it and who participated in it, (3) how and from whom the resource can be obtained, (4) how the content can be classified, (5) which language the resource contains and/or which culture the resource comes from, (6) which format the resources has and how it was created, (7) in case of tools what its runtime and import/export requirements are, (8) what the quality of the resource is and (9) what its name and unique identifier is. While many of the elements need to be formalized (controlled vocabulary) there will be some fields that allow prose descriptions for quick inspection purposes.

In addition it is widely agreed that metadata needs to be open to create visibility and enable re-usage of resources (data and tools). Protection of personality rights needs to be addressed by encrypting descriptions where necessary.

4 Contemporary Metadata Infrastructures

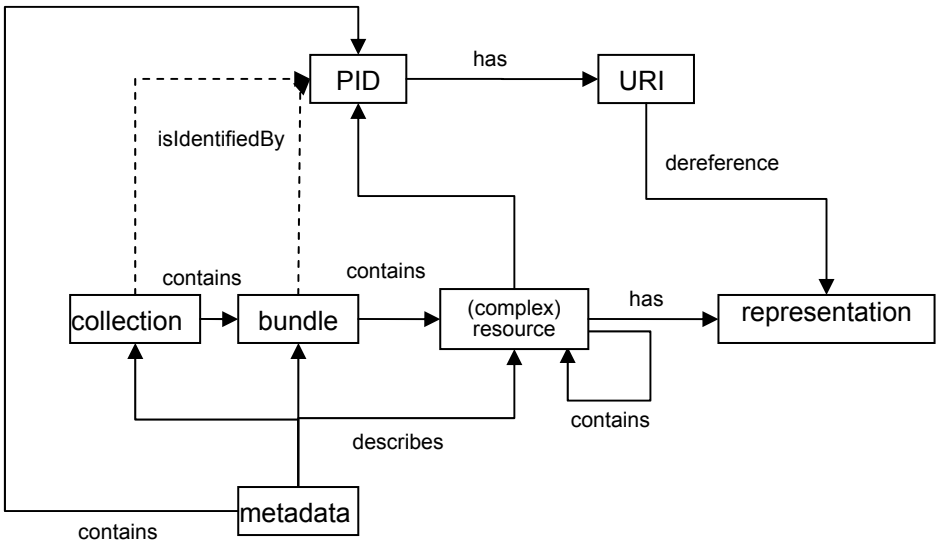
4.1 Technical Architecture

4.1.1 Introduction

With a technical infrastructure we mean actual existing implementation for a metadata infrastructure. This includes aspects as (1) how the metadata infrastructure integrates with the data model for the resources; (2) the encoding of the metadata, the way flexibility of the set is provided and how constraints are put on the metadata values; (3) harvesting of metadata and (4) the way the metadata is registered or put in a catalog.

4.1.1.1 A data model for describing resources

Important elements of the metadata and resource world that should be named and whereof the mutual relations need to be described are: resource, resource bundle, collection, URI, PID, representation, dereferencing. The model is meant to cover relevant practices e.g. it needs to take into account the use of collections and resource bundles as used in our domain.



The model in this figure depicts the different relations between (complex) resources[1], resource aggregations (collections and bundles[2]) and the metadata. The model presupposes that every resource has a PID as an identifier. Collections and resource bundles can be issued a proper PID themselves that will refer to the metadata, but that is not obligatory. The question of granularity: what is a complex resource, what are the constituents and what are bundles and collections depend on the resource type and also the way a specific repository wishes to handle them. The model also assumes that a constituent part of a resource can be referenced by a PID also, since in reality this will be accomplished by adding a fragment identifier to the PID of the containing resource, the implication is that such a combination of PID and fragment identifier is again a PID.

[1] According to definitions used in ISO documents a complex resource is a resource such as a PDF document for example which has a set of photos in it as well which are themselves identifiable resources.

[2] The term "bundle" is used by some initiatives to specify groups of resources that share a special narrow relationship such as videos, audios and annotations sharing the same time axis for example. In essence bundles are collections with special relations amongst them.

The definition what a resources is will get rather complex when for example a database management system is used to store metadata and data together in a table structure. It is then the access software of the database system that will implement the solution based on its internal logical structure.

4.1.1.2 Aspects of metadata encoding and storage formats

Probably the only relevant practice for encoding metadata for interoperability and interchange is to use XML with a suitable constraining schema. Used schemas are of the XML schema and RNG types, DTDs are now of less relevance. As an example of plain text metadata we name the ARK PID framework [ARK] that supports retrieving some simple text encoded metadata (ECS) from the PID resolver. Although for exchange XML is considered the standard, for the storage of metadata different ways are used. It can be argued that metadata is a resource in itself and should be stored in an archivable format, implying the use of XML files. This is the practice of archives using LAT/IMDI archiving software. Others store the metadata in a database and rely on the database tools often not addressing the persistency issue.

The type of database system that is used to store the metadata varies depending on the repository size and needs. Some repositories that house also texts and annotations in XML format store the XML metadata and the texts together in a (native) XML database. Others use an object relational mapping to store the XML records in a RDBMS for better performance.

So most relevant metadata frameworks use XML encoding that is constraint by suitable schemas [IMDI], [OLAC], [TEI]. For constraining the values of the by the schema defined metadata descriptors there exist different approaches. The constraints including controlled vocabularies may be also defined by the schema or the controlled vocabularies might be defined outside the schema allowing for the possibility to adapt them for use by specific projects without changing the overall schema (OLAC vs. IMDI). The relevant vocabularies must then be made available to such tools as metadata editors via a separate service.

The adaptation of a metadata schema for specific sub domains or projects is an important property of the metadata framework that determines to a high degree its flexibility. Either the framework works with a flexible schema that allows different specializations [IMDI] or there exists a family of schemas where each specialization has a proper schema [OLAC].

ODD [ODD] is an interesting feature within the TEI framework and although TEI is mainly directed towards schemas for text encoding, the technology is usable for creating xml schemas for metadata too. An ODD (One Document Does it all) file is an XML resource that is the source of descriptions, examples and formal declarations and schemas.

4.1.1.3 Metadata for web services

Web Services Description Language [WSDL] is an XML format that is used to describe web service interfaces. It describes the operations (network end points or ports) and data formats (messages) in an abstract manner. This results in a set of reusable bindings which are subsequently bound to concrete network protocols and message formats. WSDL is often used in combination with SOAP (Simple Object Access Protocol) and XML schema to define web services over the web. WSDL version 2.0 can be used to describe both REST and SOAP web services. WSDL is maintained by W3C.

The Universal Description, Discovery and Integration [UDDI] project was advocated as a universal method for dynamic discovery and invocation of web services. It was initiated by Arriba, Microsoft and IBM in recognition of the need for a global registry for discovering web services. Microsoft, IBM and SAP all have operated public implementations of UDDI and Universal Business Registry that represents a master directory of publicly available e-commerce services. By the end of 2006 all of these had been shut down. Private and community UDDI nodes remain in operation.

UDDI conceptually consists of three types of sections where information is stored. White pages contain basic business details information such as name, address, and business identifiers such as tax id. This makes it possible to locate services using organizational characteristics. Yellow pages contain information on web services using taxonomic classifications. These classification criteria, such as ‘annotation tools’, may also be used to locate web services of interest. Green pages finally describe technical aspects of a web service such as location and service bindings.

UDDI also provides as web service API for publishing, searching, retrieving and replicating this information.

Electronic business XML [ebXML] is joint effort from United Nations/CEFACT and OASIS to create a single global XML framework solution. It is intended to facilitate trade by providing a specification that allows organizations to express their business processes in a manner that is understandable by other organizations thereby allowing process integration. The primary focus of ebXML is therefore on e-business. It describes a data model for e-business objects (including services), messaging for transactions and a registry for e-business objects.

Both UDDI and ebXML are maintained by [OASIS] (Organization for the Advancement of Structured Information Standards).

4.1.1.4 Metadata Tools

Creation

The process of creating new metadata is something that can make or break an infrastructure. The usability of the tools to generate metadata is therefore from the utmost importance. We will further on give an overview of contemporary metadata creation software. As mentioned in [NISO] the metadata creation tools can largely be divided into three classes:

1. **Template tools**, that basically fills in the the user-entered data into the element set. An example of this approach is the IMDI Metadata Editor [IMDI] or DCdot[3]

2. **Mark-up tools**, helping the user to create a file that correctly corresponds to the metadata schema. In most cases this means the creation of a (valid) XML-file. A program often used for this goal (especially for the creation of TEI-documents) is Oxygen[4].

3. **Extraction tools** aim at the automatic creation of metadata based on the contents of the resource to be described. More than as a complete solution should they be seen as a kick-start for the sometimes time-consuming process of metadata creation. The results should in general be reviewed by a human editor afterwards. Scorpion[5] is an example of this approach as it uses automatic classification to create a DC subject element.

One should keep in mind that during the creation of new metadata the end user is often confronted with the limits of a fixed metadata schema.

To overcome the limits of existing metadata schemes, 2 approaches can be taken:

First, there might be no support for certain information which the user needs to store. In this case an extension of the scheme could bring an outcome. However, this could result in semantic incompatibility and should therefore only be considered in very specific cases.

Second, in case there are some elements which play a more prominent role than others, these elements can be specified into a metadata profile. Such a profile generally results in a user interface prompting the end user to fill in the important fields, in turn resulting in a more complete and apt metadata descriptions.

Usage

Obviously the whole process of creating metadata should result in an improved accessibility of the data it describes. Roughly speaking this can be achieved in the realm of the following fields:

1. **Searching**. Ranging from string matching with all metadata fields to the more sophisticated approaches (e.g. natural language queries and faceted classification) – metadata is an important key in the process of guiding the user to the data she is looking for.

---

[3] http://www.ukoln.ac.uk/cgi-bin/dcdot.pl

[4] http://www.oxygenxml.com/

[5] http://purl.org/scorpion/

2. **Resource discovery.** In many cases one might just be looking around in a data collection without knowing exactly what to search for. These kind of exploratory activities too can be well supported by suitable metadata, preferably enriched with prose descriptions.

3. **Administration.** The administrative metadata can of course also offer help with regards to the management of the resources: determining access rights, keeping track of version information, processing history, etc.

Modification

Whether it is related to the availability of additional information or to curation efforts, in some cases there is the need for tools that allow the modification of metadata. There are mainly two approaches to achieve this: either the modifications are applied directly to the repository or there is a workflow mechanism that provides the means to successively checkout the current version, edit it and commit the adapted metadata. Some metadata frameworks keep track of the changes in the metadata description itself. as does TEI.

Transformation

A reasonable number of tools do exist that perform conversions between different metadata formats. Their success depends largely on the feasibility of defining a semantic common denominator. Often some human post-processing is required.

For more information on metadata transformations, see section 4.1.1.6.

4.1.1.5 Interoperability

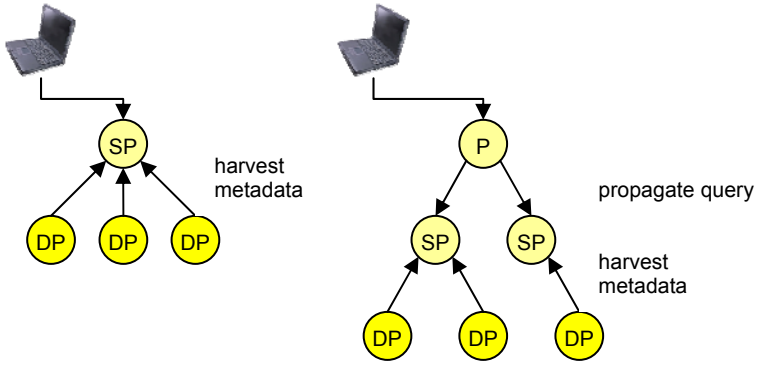
The wish to have a central catalog covering metadata from different repositories has led to the emergence of the OAI-PMH protocol [OAI-PMH] as a de-facto standard for gathering metadata into a single catalog - a process that is called metadata harvesting. In the OAI-PMH model the world is divided in data providers, that offer metadata for harvesting, and service providers that harvest the metadata and offer a service to the world, for instance a central metadata catalog. The OAI-PMH protocol is fairly efficient and simple to implement although it was reported that for some organizations implementing and maintaining an OAI-PMH data provider is difficult. The OAI-PMH protocol requires that the metadata is in all cases also offered in DC format next to any other format. This allows that metadata from different disciplines can be harvested and put into one catalog although presumably much information was lost by mapping all metadata to the DC set. When all harvested data providers have agreed to also provide metadata of another set than DC, it is of course possible to create a more useful catalog.

Although the use of OAI-PMH is widespread, alternatives exist. When metadata is present in web-pages, a web-crawler can gather the metadata for discovery purposes [METATAG]. The IMDI/LAT archives use a similar crawling strategy where linked IMDI XML metadata records are harvested. This form of harvesting only requires a web-server; however, format compliance with an accessible schema needs to be checked.

The strategy of gathering all metadata into a single catalog can be augmented by having the service providers transform the offered metadata into a standard set other than DC. The service provider would need to know different mappings for all the different offered sets. See next section for more details on this.

In all the above scenarios the catalog is located at one site and metadata searching is done within that catalog. A different scenario is that of "federative search", where a special web portal offers a web interface for metadata search and has knowledge about all relevant repositories that offer a search service. The portal also knows how to access the search service and how to formulate queries and how to map to the specific metadata sets used at the different repositories. The portal then translates a by the user specified query into a separate query for every relevant search service and propagates it. Of course a search service may be provided by a service provider that harvests multiple data providers. See the figure. The portal is also responsible to solve the ranking problem by merging the results of the different service providers in a meaningful way, which can be difficult. When a resource has metadata descriptions in multiple repositories using different schemas, the chance of finding it increase. Federative search in a coherent metadata environment was standardized with Z93.50 as mentioned in "Metadata

History".



Single catalog Federative search

4.1.1.6 Metadata Crosswalk

Much work is done currently about supporting metadata crosswalks, i.e. to allow users navigating in several domains. In particular in the library world joint catalogues are very well known of course. The vocabularies for describing publications and their semantic scope has been much more standardized and the descriptions are in general comparatively simple. But also in the research and cultural heritage area some work is known to join the metadata domains.

In the ECHO project a sub-project was carried out to join 10 different metadata sets from five different domains (Philosophy, History of Arts, History of Science, Ethnology, and Linguistics). Two of these used complex thesauri to classify content resulting in a search engine supporting cross-walk based on an underlying pragmatic ontology. In the CATCH-STITCH[6] project different description systems using also content classification thesauri of different types automatic mapping techniques are being tested. Despite a conversion to one semantic representation framework (SKOS) automatic mapping techniques in general do not deliver positive results if no contextual information can be used. A few other crosswalk projects can be found on Wikipedia[7].

4.1.2 Existing Frameworks and Practices

4.1.2.1 Using Semantic Representation Frameworks

Descriptive Metadata Descriptions (DMD) in general are relatively simple structures specified by a schema. The semantics of structural relations can in general be described by "associations" and "properties". However, this is different for classification systems that are used to define the element values in some cases such as "subject" where in history of art the Iconclass and ATT thesauri are very well known. In the area of language resources except for some obvious cases such as for "languages" and "geographic locations" no classification systems are used widely. As one can see for the Iconclass thesaurus for example they contain more complex semantic relation types.

As indicated, the core metadata descriptions don't offer rich semantics. Of course, there can be implicit relations embedded in the DMD. There could be relations between the interviewed persons and it could be useful for some research to make these relations explicit and to do some reasoning on it. Again, this requires rich metadata which in general is not the case yet. This is the reason that some colleagues tend to transform DMD into RDF representations. The new CLARIN infrastructure should offer the generation of suitable RDF as one output format to support such work, but this certainly does not yet have high priority.

---

[6] http://www.cs.vu.nl/STITCH/

[7] http://en.wikipedia.org/wiki/Crosswalk_(metadata)#Examples }

4.1.2.2 Data Categories for Semantic Interoperability

An increasing amount of experts in the metadata field is shifting its focus away from schemas, since it is widely understood that schemas are important to be able to correctly parse and interpret the content of an XML file, but not to achieve interoperability. For achieving interoperability the proper definition of linguistic concepts and their registration in machine readable registries is of greatest relevance. T. Baker (DC) summarized it as "there will be many schemas that hopefully will use registered semantics that can be referred to by using persistent identifiers".

This is the reason why ISO TC37/SC4 saw it as its central activity to create a "data category registry" and to define a process guided by domain experts to integrate concept definitions. The state of this work can be described in the following way: (1) The DCR data model has been stabilized. (2) The ISOcat distributed database system will become usable in January 2009. (3) Several colleagues have already created many concept definitions that are waiting to be processed by community experts. Also the Dublin Core and TEI concepts can be referenced via the web although one cannot yet speak in all cases about persistent identifiers. On purpose ISO TC37/SC4 did not want to include relations in the DCR, since relations are frequently very much dependent on practical issues such as the purpose of a search. The idea is now that schemas reference such categories (concept definitions), that users create and register relations between such categories where necessary and that search engines make use of the definitions and relations to find useful resources.

For CLARIN this development means that when we want to build a new infrastructure we need to step away from a fixed schema. Instead we will rely on a vocabulary which is registered in open and accepted registries and will include existing schemas as possible components to be upwards compliant and to support living communities. Here we should primarily think of the existing OLAC, IMDI and ELDA schemas.

4.2 Relevant Initiatives

4.2.1 METS

METS (Metadata Encoding and Transmission Standard) is an XML-schema[8] that was developed as a standard data structure describing complex digital library objects. Such a METS file can contain 7 sections[9]:

- **METS header**: It contains metadata describing the METS document itself, including such information as creator, editor, etc.
- **Descriptive Metadata**: This section may point to descriptive metadata external to the METS document or contain internally embedded descriptive metadata, or both. Multiple instances of both external and internal descriptive metadata may be included in the descriptive metadata section.
- **Administrative Metadata**: This section provides information regarding how the files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object (i.e., master/derivative file relationships, and migration/transformation information). As with descriptive metadata, administrative metadata may be either external to the METS document or encoded internally.
- **File Section**: The file section lists all files containing content which comprise the electronic versions of the digital object. **<file>** elements may be grouped within **<fileGrp>** elements, to provide for subdividing the files by object version.

[8] http://www.loc.gov/standards/mets/mets.xsd

[9] Source: [WP:METS]

- **Structural Map**: This section is the heart of a METS document. It outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element.

- **Structural Links**: This section allows METS creators to record the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map. This is of particular value in using METS to archive Websites.

- **Behavioral**: A behavior section can be used to associate executable behaviors with content in the METS object. Each behavior also has a mechanism element which identifies a module of executable code that implements and runs the behaviors defined abstractly by the interface definition.

Depending on its use, a METS can be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System [OAIS] Reference Model. Sometimes METS records are exchanged between different centers using the OAI-PMH protocol (which is described below) [Tansley 2006].

Earlier experiences have shown that METS is a suitable format for the meta-description of linguistic resources, as indicated in the report of the *Fieldwork Data Sustainability Project* [FIDAS].

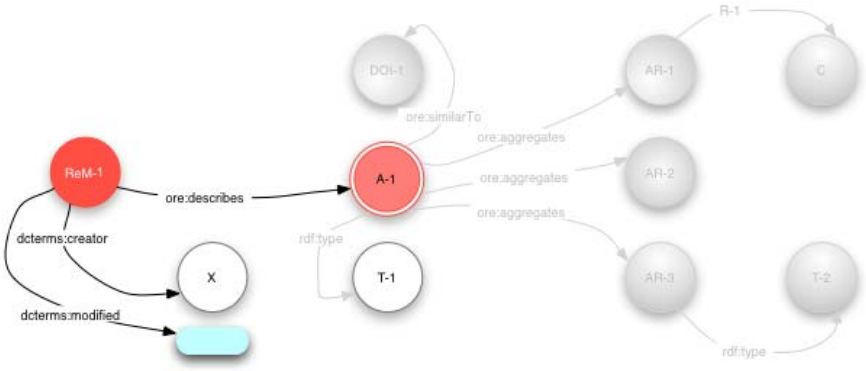
4.2.2 OAI-PMH and OAI-ORE

The Open Archives Initiative, which has its roots in the open access and institutional repository movements, has been the initiator of two standards that aim to provide interoperability between heterogeneous repositories.

The first one, OAI-PMH (Protocol for Metadata Harvesting) provides an XML message format for the exchange of XML records (typically metadata). It supports selective or incremental harvesting which allows a client repository to maintain an up-to-date copy of records in all or part of a source repository.

For the details about this process, see section 4.1.1.5.

OAI-ORE[10] (Object Reuse and Exchange) on the other hand defines a data model for Resource Maps that describe aggregations of web resources, and recommends serialization formats for these Resource Maps. ORE is based on the Web Architecture where every information object is made available via a URI. No new protocol is defined. Exchange of Resource Maps is possible individually by direct web access, and via batch discovery mechanisms. OAI-PMH is one protocol that may be used to implement batch discovery.



The figure above shows a resource map[11] that describes an aggregated resource. The metadata part is highlighted. It is taken from the ORE User Guide[12].

[10] source: http://www.openarchives.org/ore/0.9/primer.html#RelationToPMH

[11] The ORE-concept of a resource map is a named graph that indicates aggregation relations between web resources.

4.2.3 Dublin Core[13]

The Dublin Core Metadata Initiative started by defining a restrictive set of 15 elements with semantically broad categories. Later it extended this by the qualified Dublin Core set which are elements with rather precise semantics.

The Dublin Core Metadata Element Set arose from discussions at a 1995 workshop sponsored by OCLC and the National Center for Supercomputing Applications (NCSA). As the workshop was held in Dublin, Ohio, the element set was named the Dublin Core. The continuing development of the Dublin Core and related specifications is managed by the Dublin Core Metadata Initiative (DCMI). The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. Faced with a proliferation of electronic resources and the inability of the library profession to catalog all these resources, the goal was to define a few elements and some simple rules that could be applied by non-catalogers. The original 13 core elements were later increased to 15: *Title*, *Creator*, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Identifier*, *Source*, *Language*, *Relation*, *Coverage*, and *Rights*.

Finally the discussions between supporters of the "simple view" offered by un-qualified Dublin Core and those who requested fined semantic distinctions resulted in what is called qualified Dublin Core elements. These elements can help to refine the broad DC elements.

Still Dublin Core does not prescribe any syntax, i.e. the elements can be included in any kind of schema. Further details on Dublin Core (including more recent developments) can be found in Appendix **Error! Reference source not found.**.

4.2.4 TEI

The Text Encoding Initiative [TEI] has developed a widely used standard for the markup of electronic texts (ranging from corpora like the BNC to poetry). The encoding relies on SGML (for the older versions of the TEI specifications) or XML. Metadata can be embedded in the header of a TEI-file, which generally contains fields that correspond to a bibliographic record (e.g. title, distributor). As from the most recent version of TEI (P5) its schema can be customized by creating on [ODD] file which contains prose descriptions and a formal specification of the newly added elements.

TEI header elements are widely known also in the LRT domain and are used in a number of projects to characterize resources. It seems that TEI has received a new momentum and therefore it is of great relevance for CLARIN to look at the elements introduced by TEI for example to describe lexica and integrate them into the component model to be established.

4.2.5 IMDI

The IMDI Framework [IMDI] offers next to a suitable set of XML-based metadata descriptors for language resources, a set of tools and an infrastructure to use these. It is characterized by that:

(1) It allows related resources to be bundled by metadata descriptions. This bundling feature was introduced by the desire to be able for example to group all resources that pertain to a recording: One or more video or audio files, images, the annotations, possible loosely connected texts and references to the original tape that may be stored in a tape archive.

(2) Metadata descriptions can be linked to form structured virtual organizations facilitating browsing and management. This linking feature allows users to create well-designed hierarchically structured sub-corpora as well as to create relations of all sorts. Links are realized by embedding pointers in the XML format metadata description files.

IMDI focuses on the description of annotated multimedia/multimodal resources and the element set for this type of resources emerged over 8 years of discussions between linguists from various sub-disciplines, in particular field linguists and multimodality specialists. Despite this focus it covers a number of other aspects.

---

[12] http://www.openarchives.org/ore/primer

[13] Source: [Understanding Metadata]

(1) It has a schema for lexica with multimedia extensions which resulted from the MILE project. (2) It has a special schema for corpora to cater for descriptions of for example the Dutch Spoken Corpus (CGN). (3) It has a special profile for CGN to allow the integration of the TEI header elements this project was using. (4) It has a special profile for Sign Language which has been worked out by European sign language researchers. (5) It is harvesting a number of European sites on its portal.

Currently there are about 60.000 IMDI metadata descriptions and CLARIN needs to make sure that upwards compliance is guaranteed. A deep analysis has been presented on 27.000 descriptions showing that the quality of metadata descriptions in general is not satisfying [Klassmann 2006].

4.2.6 Universal catalogue

In addition to the ELRA catalogue of language resources, the Universal Catalogue is a service offered to ELRA members who will be given an early access to browse through all resources identified by ELRA before they become part of the catalogue. Since the universal catalogue has become available only during the last weeks still more analysis needs to be done.

4.2.7 OLAC[14]

The [OLAC] Metadata Set is the set of metadata elements that members of the Open Language Archiving Community have agreed to use for discovering language resources that come from various archives. OAI PMH is used to harvest metadata of all registered data providers. Some archives use the OLAC set to describe their resources. Uniform description across archives is ensured by limiting the values of certain metadata elements to the use of terms from agreed-upon controlled vocabularies.

The OLACMS is equally applicable whether the resources are available online or not. The metadata set consists of all the elements of the Dublin Core Metadata Set. To this core set, OLACMS adds a set of refinements and qualifications that are designed for describing fundamental properties of language resources, such as subject language, language data type, and software functionality. The OLACMS Standard uses XML to represent metadata descriptions.

Discussions about the granularity offered by data providers emerged when one provider offered a large number of resource descriptions which dominated the hit lists then. This experience indicates that CLARIN needs to think of smart algorithms in the future that can automatically create hierarchies of descriptions according to some criteria.

OLAC has currently about 35.000 records and many gateways have been defined that map internal metadata sets to the DC/OLAC set including one from IMDI to OLAC. Also here the installed base is impressive and any attempts from CLARIN to move ahead needs to come up with upwards compatibility.

4.2.8 MPEG7

[MPEG-7] is a multimedia content description standard. This description will be associated with the content itself, to allow fast and efficient searching for material that is of interest to the user. MPEG-7 is formally called Multimedia Content Description Interface. Thus, it is not a standard which deals with the actual encoding of moving pictures and audio, like MPEG-1, MPEG-2 and MPEG-4. It uses XML to store metadata, and can be attached to time code in order to tag particular events, or synchronize lyrics to a song, for example.

In general it was seen as being too complex to find broad acceptance.

4.2.9 ISOcat DCR

[ISOcat] is the Data Category Registry for ISO TC 37. The mission of the DCR is to overcome problems associated with semantic interoperability among language-related data resources. The DCR provides carefully defined linguistic concepts together with relevant modeling constraints. Interoperability can be achieved between different schemas, e.g., if they reference the same ISOcat entries. Pointing to an entry without further

[14] Source: http://www.language-archives.org/documents/faq.html

constraints indicates that the "concept" used is identical to the one found in ISOcat. Further constraints can of course be expressed by using another relation type, such as "is subclass of".

Right now the DCR contains metadata elements from the IMDI and part of the OLAC set. More details are given in section 4.1.2.2.

4.2.10 Natural Language Software Registry

The Natural Language Software Registry [NLSR] at DFKI is a concise summary of the capabilities and sources of a large amount of natural language processing software available to the NLP community. It comprises academic, commercial and proprietary software with specifications and terms on which it can be acquired clearly indicated. Yet the registry is restricted mainly to NLP tools. Its content is integrated into the LT World information system that can be used to ask about various sorts of information in the NLP domain. It makes use of an underlying ontology.

Since it is one of the very few software registries in the domain of LRT, it is important for CLARIN to analyze the used elements and vocabularies. Also the extension of LT World to cover LRT in the broader sense would be a very interesting option.

4.2.11 ACL Data and Code Repository [15]

The *ACL Data and Code Repository* is a repository of data (e.g., hand-labeled text, hand-parsed text, feature vectors for machine learning, etc.) and source code (e.g., taggers, parsers, chunkers, etc.) for computational linguistics and natural language processing. The goal of the repository is to make it easier for researchers to replicate each other's work and to compare different approaches using the same benchmarks. Resources in the ACL are described with a minimal metadata set (ID, name, Contributor, Copyright, License, Citation, Description).

4.2.12 LOM[16]

The IEEE Learning Technology Standards Committee (LTSC) developed the Learning Object Metadata standard to enable the use and re-use of technology-supported learning resources such as computer-based training and distance learning. The LOM defines the minimal set of attributes to manage, locate, and evaluate learning objects. Generally an XML encoding is used.

Its main difference compared to other sets is the inclusion of educational elements like *Typical AgeRange* (of the intended user), *Difficulty*, *Typical Learning Time* and *Interactivity Level*.

5 Metadata Usage

This chapter is devoted to describe the current state of metadata usage by the different groups of users, the state of standardization work, missing functionality and the lessons learned. We will restrict ourselves to the LRT domain knowing that in other disciplines the situation may differ. Also we will not consider the many useful web-sites that offer resources, but where human interpretation of prose text is required to find them. Excellent examples are the Helsinki Language Corpus Server[17] where for example even formal metadata was created, but not in a way that it can be harvested easily by machines, and the Phonogramm Archive Vienna that has detailed prose descriptions about the various collections. There are many resource providers of this type where we cannot speak about a machine readable keyword type of description that can be aggregated to new virtual collections for example without applying information extraction methods.

5.1 Coverage

In 2000 overviews were created within the ENABLER project about the types of metadata sets used and about the coverage and visibility. According to these overviews the following initiatives were busy to define

[15] Source: http://aclweb.org/aclwiki/index.php?title=ACL_Data_and_Code_Repository

[16] Source: [Understanding Metadata]

[17] (http://www.ling.helsinki.fi/uhlcs/)

metadata sets or tags: EAGLES, ISLE Meta Data Initiative (IMDI), Open Lexicon Interchange Archives Format (OLIF2), Open Language Archives Community (OLAC), Browsable Corpus (BC), Corpus Encoding Standard (CES), Codes for the Human Analysis of Transcripts (CHAT), Dublin Core (DC), European Language Resources Association Catalogue (ELRA), Gesture Databank (GDB), International Corpus of English (ICE), Linguistic Data Consortium Catalogue (LDC) and Multimedia Content Description Interface (MPEG-7). As far as is known only two initiatives were successful in building an infrastructure that went beyond the purposes of the project in focus and that were meant to support a distributed domain of resources: OLAC and IMDI - both with different and complementary focus. When discussing the coverage in the LRT domain we therefore would like to refer to the results of the ENABLER study and on the coverage of the two mentioned initiatives.

The ENABLER study is based on 31 replies from 134 distributed questionnaires to major players that included lexica, multimodal resources, speech resources, written resources and tools. From the statistics we can conclude that in 2000 the visibility of resources was extremely bad. The situation has improved; still the coverage in terms of machine readable harvested metadata records is very poor. OLAC is a typical service provider in the LRT domain that is harvesting metadata that adheres to the OAI PMH protocol. It currently contains material from 37 archives worldwide and offers 36161 metadata records. The IMDI repository is not very active as being a service provider, nevertheless the IMDI service provider offers in total 93.000 metadata records including those of invited deposits from researchers without proper archiving facilities. The metadata records are harvested from 6 different archives and are made available for harvesting by other service providers via a OAI PMH gateway. The metadata for some collections is partly harvested by the OLAC harvester and some by service providers outside of the LRT domain.

The Talkbank project has a different concept in so far as it harvests content from various sources and then offers metadata records of its 17.000 resources for searching purpose. Parts of these descriptions are harvested by OLAC. The catalogs of LDC and ELDA are growing, but these organizations focus on resources which they can offer to consumers, i.e. they are not acting as metadata service provider. In total the ELDA catalogue covers 1000 resources (469 spoken, 242 written, 283 terminologies, 6 multimodal) and the LDC catalogue 411 resource collections.

With respect to tools the situation is even worse. Currently we know of three registries that contain tool information. The DFKI registry contains 287 records about available tools.

In total we can summarize the state in the following words:

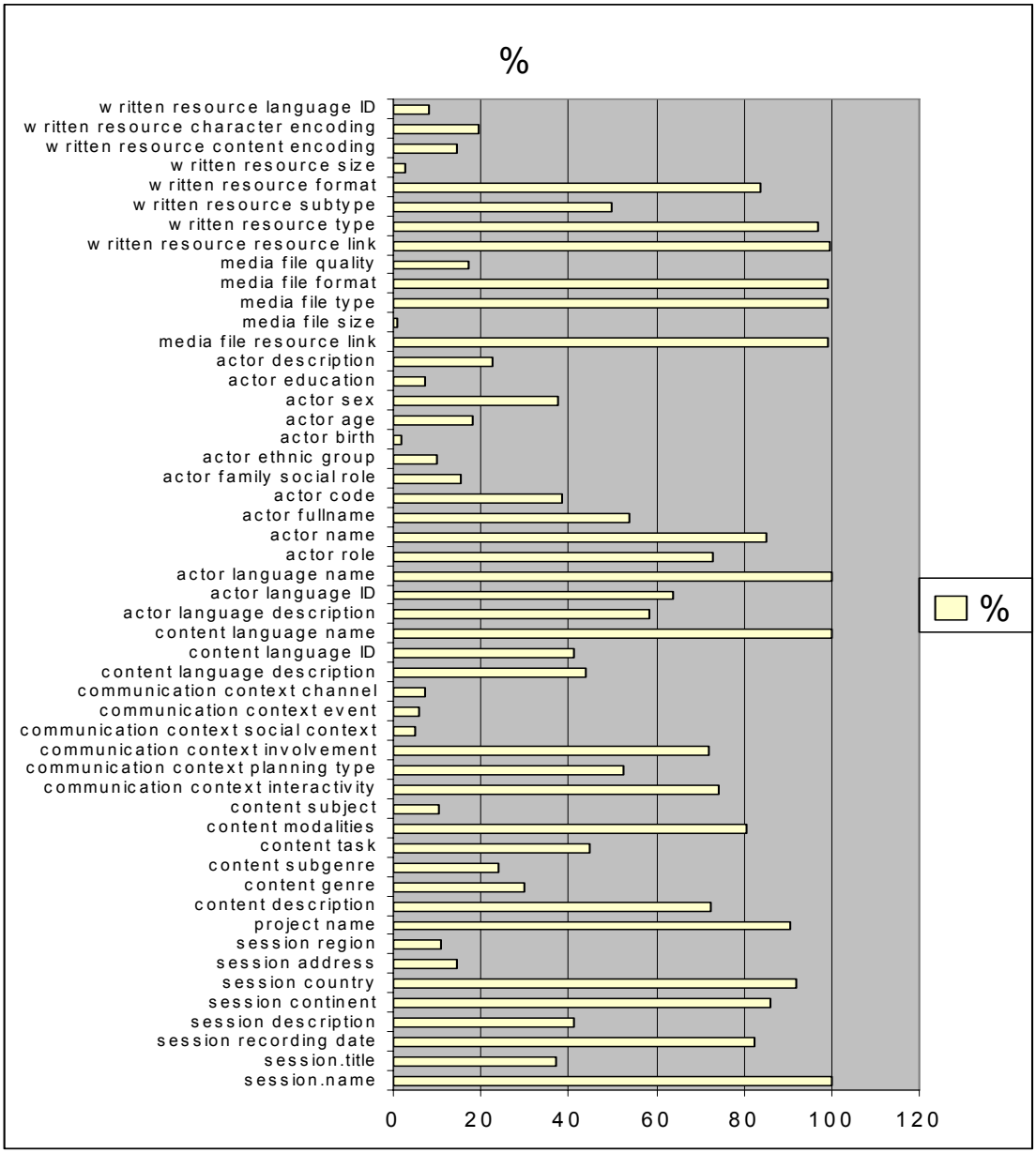
- Despite all efforts the total coverage of metadata descriptions is still far away from having a critical mass or exhibiting a satisfying coverage for various reasons.
- Until now mainly those resources were registered that are of relevance for the field of Language Technology (Computational Linguistics, Speech Recognition).
- The level of granularity of visible resources is very inconsistent. It ranges from a single metadata record for example for the Dutch Spoken Corpus which in its own has about 12.000 resource bundles to single small lexicons.
- Yet there is no systematic approach to harvest all sorts of metadata descriptions that are available on the web and offer various types of filtering, automatic organization and selection mechanisms.
- It probably still holds that many resources or collections are in a bad state hampering progress.
- Also the quality of the metadata descriptions is varying considerably in terms of level and correctness.

In various national roadmap reports it was stated that the funding policies need to change so that sufficient time and funds are reserved to create proper metadata descriptions of resources and tools. The awareness is growing that metadata descriptions are one of the most important factors to improve re-usage and to make resources and tools part of the eScience scenario. Therefore we can expect that the coverage will improve in all respects, however, the LRT community itself needs to intensify its efforts to make all existing resources and tools visible. This is one of the missions of CLARIN. There should be descriptions independent of the state of the resources themselves.

5.2 Quality of Metadata

The evaluation of the quality of metadata descriptions can be made according to a number of criteria such as validity according to a formal schema specification, degree of usage of the available elements and semantic coherence of the usage of the elements. As already indicated most of the resource and tool descriptions currently are available in prose texts as web-sites etc., i.e. any useful statement about quality can be made.

For schema based machine readable metadata descriptions a quality assessment can be done as far as it concerns a formal validation against a schema. Any proper OAI PMH service provider such as the OLAC service provider will check the correctness of the offered metadata records. Also the IMDI service provider will check the schema conformance of the



Only one broader analysis about the quality of metadata is known to us. This analysis was carried out on 23.710 IMDI metadata descriptions and the degree of filling is indicated in the following table. All computer generated metadata descriptions were excluded from this analysis to be able how users make use of the offered metadata frameworks. It should be noted here that some formal descriptions of the resources (links, types, formats) are always checked to be correct. From this statistics and discussions at various meetings we can derive a few phenomena:

- researchers hesitate to invest time for the generation of good quality metadata, since in general there is no funding reserved, there is no duty to create proper metadata and there is not always an immediate benefit for themselves
- researchers hesitate to carry out classifications such as would be required in the "Genre" element where there is no widely agreed vocabulary

- even for elements with controlled variables such as for language codes (content language ID) only 40% were filled in although language names were filled in for 100% which points in particular to two aspects: the increased effort to look up the exact codes is not taken, the suggested classification is not accepted

- often information is simply missing, i.e. elements can't be filled in

- there is no understanding yet that other researchers may look at a resource in a different way which requires additional elements to be filled in; the age of speakers for example may only be interesting for developmental linguistic studies, but not for language technologists, i.e. researchers tend to only look at their own interests

- the tools are not always so simple to use that researchers are motives to invest more time

In general the experience is that for those fields that are not governed by a closed vocabulary and where the tools only support this vocabulary much curation effort is necessary to achieve proper metadata descriptions.

5.3 Types of Usage

Metadata act as fingerprints for publications for all sorts of retrieval, referencing and management activities. In particular in the After-Gutenberg area when the number of books and copies increased considerably the usage of library cards became indispensable. People needed to make a difference between a book as described by a few of its typical attributes such as author, title and year of publication and the various physical copies that were distributed. Library cards became abstract incarnations of publications meant for human (librarians, users) consumption. Needless to say that also museums and archives started using the concept of DMD[18] to describe and manage their large holdings of physical objects. In parallel large content classification systems such as Iconclass [ICONCLASS] were developed to allow experts to classify the content of paintings and sculptures and to add such classifications to the DMD. To distinguish the object as such and to register every publication in a unique way the ISBN numbers were invented.

In the digital era the number of resources increased again by factors and also their structure is getting more complex. Single resources such as a text, a sound or a video file are the basic objects, however meaningful for research purposes are complex resources of various types such as annotated media recordings where several resources share the same underlying axis and exhibit an extensive internal relation structure, a lexicon with a large amount of multimedia extensions, a collection of resources that is meaningful for a community and a virtual collection that was created by a single researcher as basis for a scientific analysis. It is increasingly better understood that metadata descriptions are excellent representations of all these simple and complex objects. In addition they can take over similar function as library cards and the world wide community is ready to introduce unique identifiers for each object, therefore naturally associated with metadata descriptions.

During the last decades it became much more obvious what the possibilities for DMD in future scientific information management and usage will be. First of all DMD in general adds valuable additional information to resources, information that is mostly not encoded in the resource itself. This information can have different character such as administrative, describing the content, describing technical and usage details, etc. Basis for such added value are classification steps since resources need to be associated with a certain language, a certain genre, a certain group of creators, a certain subject etc.

Of primary relevance is DMD to help researchers and other users to find useful data to answer a given research question. Dependent on the researcher's background the queries will be more or less detailed. Linguists will be interested in combining metadata and content queries such as for example for a longitudinal study: *Give me the frequencies of correct usage of the 3rd person plural inflectional forms for 3, 4 and 5 years old children and allow me to compare between boys and girls.* Such queries have a direct research impact. Non expert users may ask more general queries such as *where can I find resources about the Kuikuru language* to then have a quick look whether they are relevant for the work. In both cases the user expects that the DMD will provide a direct link to the resources themselves to be able to utilize them. Increasingly often

[18] In the following we call the keyword type of metadata descriptions of resources Descriptive Metadata (DMD).

researchers will want to create their own virtual collection probably by virtually combining resources from different repositories. Actually, the collection building is carried out by creating, re-grouping and linking DMD descriptions, the resources in general will not be moved for many reasons. Such virtual collections will then form the basis for ongoing scientific work and the context needs to be preserved for documentation purposes. In all these cases DMD are research tools.

For other users metadata could be used for general discovery purposes or to advertise resources and facilitate access. DMD could be linked with geographical locations or used to group resources dynamically for exhibition purposes like web presentations and portals. For large online repositories such as Flickr and YouTube[19] the principle of social tagging became very popular where users associate values with certain fields. Also in these cases DMDs are created for later discovery and grouping purpose based on individual and non constrained classification steps.

The other important pillar for DMD creation and usage is information management. In scientific repositories or archives with a long-term preservation intention authenticity, proper classification and grouping are essential. DMD can be the basis of various management operations such as copying, moving, associating access rights, migration of formats, checking consistence, etc. Relating various resources with each other has a completely different function, since for management tasks it is crucial to treat for example a sound recording and a video recording created at the same time and describing the same event and various annotations of them as one unit. Here metadata becomes the function of the glue that allows managers to bundle resources and therefore to facilitate sensitive operations.

For agencies like LDC and ELDA metadata is aggregated to catalogues to show interested users what the agencies are offering and under which terms they do so. While in the research world metadata descriptions are rather living objects in itself that can be subject of queries, here the metadata is more static in nature, but nevertheless a very important information source.

Finally (1) metadata is very important for long-term preservation, since it can for example store information about the creation and transformation process a resource undertook over time; (2) it can be used to store user experience which is very important for example in the case of tools; (3) it can be used to automatically match various descriptions for example to determine which tools can be applied to a certain selection of resources.

So far DMD are mostly used for interpretation by humans either facilitated by a search engine or by visually browsing. In future eScience scenarios increasingly often we will have to support machine driven operations such as the creation of (semi) automatic abstractions, i.e. hierarchically grouping resources in new data driven manners, or as automatic profile matching to suggest alternatives for certain tasks. The latter can lead to a (semi) automatic extraction of processing chains which would enormously facilitate the work in particular of laymen.

5.3.1 Metadata Assertions

Based on the usages we can describe a few basic assertions about metadata which have been stabilized during the last decades for electronic resources.

DMD represent Resources
Metadata descriptions represent (collections of) resources in all possible contexts where it is not useful or not possible to use the resource itself. To act this way metadata need to refer to the resource so that it is possible to access the resource if necessary.

DMD are Open
Metadata descriptions need to be open so that they can be used for all kinds of purposes in the research world and beyond. It needs to be possible to harvest metadata without restrictions to combine them to new create new services.

[19] See http://www.flickr.com/ and http://www.youtube.com/

DMD are Rich
Ideally metadata are rich objects that contain a lot of information that describes a given resource or collection in various dimensions as has been indicated. Incrementally DMD should be enriched dependent on the usage and life cycle transformations of a resource. So DMD should include a version history or at least point to it.

DMD serve many Functions
DMD can serve many different functions that range from pure management aspects to scientific aspects. Since CLARIN is an infrastructure dedicated to research it must be possible to enrich DMDs to make them suitable to use them for typical research questions.

DMD are Living Objects
Metadata descriptions are living objects in the domain of research that can be collected, manipulated, combined, etc, i.e. they are much more than mere catalogues. In the CLARIN research domain metadata are ideally rich objects that can be used for research purposes. The metadata descriptions are also volatile because they can be incrementally enriched (especially so with metadata for growing collections) and might contain workflow history information.

DMD are Registered
Due to the described research needs where DMD are subject of manipulations we need to separate registered from processed metadata. Registered DMD are metadata descriptions created and maintained by experts such as depositors and archive managers. These DMD are registered in official and well-maintained registries that carry out various quality and consistency checks.

DMD Gathering is Systematic
Yet the process of creation, harvesting and offering is lacking a systematic approach in many respects such as quality, granularity, filtering and views. A CLARIN solution needs to be based on a more systematic approach that needs to be worked out in form of guidelines for the community.

5.4 Types of Users

With respect to the usage of Descriptive Metadata (DMD) we can distinguish a large variety of users types. We will characterize a few of them excluding the creators of DMD:

- **Archive/Repository managers** that are using metadata to organize and structure their holding, that want to create selections for copying purposes, that want to define domains of responsibility and ownership etc
- **Occasional Users** who are curious about languages and just want to know a bit more and who are probably satisfied with broad Dublin Core like semantics or even with a simple query in Google
- **Researchers from the linguistic domain** who want to know what kind of resources are available for a certain language, what kind of genre it is, what quality it is in etc (typically these users would prefer portals from service providers who harvest metadata from various archives)
- **Researchers or developers from industry** who want to know the price to test for example the quality of a speech recognizer or a computer linguistic tool (tokenizer, parser etc) and want to see a catalogue of a resource distributing company
- **Researchers from the linguistic domain** who want to use metadata to answer research questions such as in a longitudinal study where they want to compare linguistic features between speakers of different age and se, where they want to compare linguistic features between certain dialect background of speakers etc (typically these users would first build their virtual collections)
- **Researchers from other disciplines** who for example want to establish a network of persons in a certain area where metadata is one of the sources of input and where they finally link metadata fragments with other online data (here one can imagine a huge number of different and unpredictable applications)
- **Language community members** (minority languages, sign languages etc) who want to know where they can find material for their specific language without having the knowledge of the field
- **Policy makers** who want to see how good a national language is represented on the web.
- **Computer systems/web services**, i.e. software crawlers that will read and interpret metadata descriptions

This list is not comprehensive, but can serve as an indicator of how different the types of users are that will make use of metadata given that there is a critical amount of data available from a certain portal. The major

point will be the general acceptance or visibility of the portal. Of course we need to accept that the Google portal is the one that is used most widely, i.e. we need to make sure that metadata records can be harvested by Google and other general purpose search engines. The list also indicates that the types of user interface and the types of views a portal needs to provide will be very different.

CLARIN is a research infrastructure primarily mentioned for researchers in the humanities and social sciences, i.e. we need to focus on the users from the research domain. However, when designing a new metadata infrastructure we should consider other user types as well.

5.5 State of Standardization

First, we need to discuss what we can call a standard. Obviously a wide usage of a specified rule does not mean that this rule can be called a "standard". In addition the rule needs to have passed the process specified by an internationally accepted standardization organization such as ISO which has formed a special sub-committee TC37/SC4 devoted to standards in language resource management. There are other classical standardization organizations such as IETF. In addition a number of widely respected initiatives have got a normative role such as W3C, OASIS, Dublin Core and TEI and companies that set de facto standards due to their large coverage such as Microsoft. Important for the acceptance of a certain rule that has been specified is the shared belief of users that investments will not be wasted and that organizations or initiatives that launch the rule can guarantee some stability, support and wide usage over a longer period in time.

With respect to metadata we can hardly speak about standards in the traditional sense of the word, but there are some serious recommendations for the field of LRT. All relevant initiatives were mentioned in chapter 3. Here we want to make short statements about the "standardization" process.

Dublin Core
The DCMI elements are offered under a PURL construction, i.e. there is an indirection mechanism which will provide stable identifiers as long as the PURL registration record will be maintained by the Dublin Core community. Dublin Core offers the standard element set of 15 elements and semantically more specific qualified elements. The PURL mechanism allows users to define an appropriate namespace and refer to each element. Various variants of embedding (XML, text, RDF) are offered. Since DC is widely supported community users can expect that the references and definitions will be maintained for quite a while.

OLAC
OLAC offers a mechanism for refinements of standard DCMI vocabulary and 4 elements (language, linguistic-field, linguistic-type, discourse-type) specialized for the LRT domain that can be used in these refinements. The refinement mechanism is defined on an earlier DCMI embedding specification and DC itself admits that concepts used for refinements should be usable as elements in a schema. This is the solution for Qualified DC. We propose to include the definitions of the 4 additional elements in the ISOcat reference to ensure referencing stability and long-term persistence.

IMDI
The IMDI element set is described on the web and has been entered into the preliminary version of the the ISOcat concept registry. The description itself will not be sufficient to refer to an element and the IMDI community is certainly not sufficient to guarantee long life times. The usage of ISOcat for all elements which are not already covered by DCMI or TEI will ensure referencing stability and long-term persistence. The problem of semantic context in the ISO DCR is not yet fully solved i.e. the difference between language used in the context of a participant description (the language somebody can speak) and language in the context of the content of a recording (the language actually spoken)..

TEI
TEI is a worldwide initiative with a broad support from many disciplines, so we can assume that TEI will be one of the initiatives which will exist for quite some time in the future as well. TEI until now follows a slightly different policy compared to ISO DCR, since currently it offers for every concept defined web-pages with a semi-formal way of description. Each such page has a URL, yet there was no attention for long-term reference stability. Within the TEI name space every label is unique. So references are unique already now and they are persistent in so far that the URI of TEI will be supported for long time. The difference with ISOcat is that the definitions are not machine readable yet.

CHAT
Here we use the CHAT format from the CHILDES project as an example of a very well designed format where also header tags were defined. Yet the tags are only defined in a document (manual), i.e. they cannot be used by others and the status is very much dependent on the projects lifetime.

Summarizing, we can say that there are three initiatives where we can assume a long-term stability and proper standardization of definitions and references: DC, TEI and ISOcat. For all other initiatives such as IMDI it is necessary to include the element definitions in the ISOcat concept registry for example.

5.6 Lessons Learned

Since in particular the IMDI and OLAC infrastructures are around for about 8 years now, we can draw a number of conclusions from the experience so far.

- Both sets, IMDI and OLAC, have stabilized over the years and offer solid infrastructures. OLAC is focusing on cross-repository services. Although it offers an editor to create descriptions, its main focus is on acting as service provider, i.e. harvest DMD via the OAI PMH, requiring a low granularity[20] and offering a search engine to look for interesting resources applying the OLAC/DC vocabulary. IMDI offers a structured schema allowing users describing resources and resource bundles in greater linguistic detail. This enables users to formulate queries that are directly relevant for research questions. It comes with an editor, allowing to create DMD and to embed them in browsable hierarchies, it allows depositors to create canonical hierarchies that can be used for management purposes and users to create their own private hierarchies, it offers native XML and HTML browsing facilities by applying on the fly XSLT transformations, it offers a gateway to act as full OAI PMH data provider, it harvests data from other registered data providers and it offers structured and unstructured search options. REST interfaces allow users to embed for example metadata queries in web portals. With respect to its foci OLAC and IMDI are fairly complementary.

- Their coverage has grown during the last years and is impressive in spite of the limited funds that were available. However in total the amount of language resources that have been registered and that are accessible via the portals is very small compared to the amount of language resources that have been created. Therefore we cannot speak about a satisfying solution. The reasons for this are very heterogeneous. Here we can only mention the most important ones: (1) Metadata creation is expensive and extra work, which is in general not budgeted for. (2) Researchers still lack convincing arguments to invest in efforts for the benefit of other users. (3) The available metadata sets were not always useful since their schema and terminology are not appropriate for the resources to be integrated. Users often want to be able to tailor their sets to their needs. (4) Available knowledge about existing language resources is often such that even the responsible researchers don't know exactly how to classify them and where they are exactly physically stored. (5) Some researchers still see their resources as their private capital which they don't want to share. (6) In some cases there are ethical considerations or privacy reasons that forbid users to even publish metadata about resources.

- From broad discussions in our discipline we know that terminology and localization issues are crucial for researchers to create DMD. Sub discipline terminology is different from what is used in sets such as IMDI and people hesitate to use non-familiar vocabularies to classify their resources. Missing support for a working language is also a point of many uncertainties.

- Even for professional frameworks such as IMDI with ample technical support, we can see that the willingness to create richly filled in metadata descriptions is rather low and that adherence to standards is not guaranteed. Statistics carried out on 27.000 metadata records in the MPI and DOBES archive show that some fields such as for example "genre" are not used since categorization is seen as problematic or since it costs too much time to decide about it and that other fields such as language ID are not used properly since it would cost too much time to look up in the integrated Ethnologue list what the exact ID is.

- Right now we see the first real applications where depositors themselves see a benefit from investing time for metadata creation. Archives such as the one at the MPI with about 60 Mio annotations for resources from many different teams and a large variety of languages are of a size and richness that it makes sense

[20] From the IMDI domain only DMD are accepted that represent language-oriented collections.

for a researcher to formulate scientific queries that contain metadata constraints to restrict the collection on which content queries can be formulated. Another application where metadata is required are dynamic portals that exhibit the richness according metadata categories such as "genre". On the fly metadata queries can present those resources that contain for example stories about certain subjects etc. A third increasingly accepted argument for the usefulness of metadata is its importance for building virtual collections suitable to work on a certain research question. Without the need for copying the real resources users can simply copy and recombine metadata descriptions for this purpose.

- The general pressure from funding agencies is growing to produce well-organized and well-described collections at the end of funding periods. Also the insight of the disciplines is growing that accessibility and re-usage of digital resources is ultimately dependent on proper metadata descriptions and proper digital archiving.

The eight years of experience resulted in a much deeper understanding of a number of "technical" problems that need to be solved such as metadata granularity, resource bundling, mapping between different metadata vocabularies, irrelevance of schemas for semantic interoperability in most cases of discovery, usefulness of registering concepts in open registries as basis for semantic interoperability, granularity of concept descriptions in such registries etc.

5.7 Missing Functions

While the ordinary functions such as browsing and (structured/unstructured) searching are well supported, we can see many useful functions that are not well supported, but which would make the metadata domain much more attractive for users.

- Yet the granularities presented in metadata registries are too different which gives very unbalanced results. We cannot assume that the resource providers create different levels of granularity since it is a matter of usage. In the research domain we increasingly often see that individual resources are used in different combinations, i.e. the notion of "fixed published" corpora needs to be augmented by the visibility of the individual resource. Highest priority for resource providers is therefore to offer a high granularity. However, we need smarter technologies that allow to automatically creating abstractions.
- Closely related is the function of automatically creating user-defined hierarchies for selected collections. Given arbitrary collections it must be possible to specify an ordered list of metadata elements according to which hierarchies can be built. These are useful for various operations such as selection, filtering etc. However, these techniques can only be usefully applied when the metadata descriptions are rich.
- Also very related is the function of creating views on sets of metadata descriptions. Views are based on selections and filtering combined with visualization techniques such as facetted searching. These techniques allow users to create customized views and are complementary to creating user defined hierarchies.
- Tools for creating virtual collections are not yet sophisticated enough and don't support to operate across various repositories. It must be easy for users to gather collections and then to sort them etc.
- Yet we miss tools to allow users to create their own schemas from prefabricated components or by re-using elements from open registries such as DC, TEI, ISOcat, etc. Such a flexible schema approach requires new types of editors and navigation tools. This all needs to be developed from scratch.
- Arbitrary users increasingly often like to add their own tags. Our infrastructures yet do not allow "social tagging" as it is called.
- Metadata is often badly filled in. To create a rich environment we need to apply semantic web technologies to automatically enrich metadata descriptions.

6 Linguistic Requirements

6.1 Introduction

This chapter is devoted to formulating the linguistic requirements for a list of metadata concepts that can be suggested to establish the CLARIN virtual observatory. This needs to be done on the basis of the experiences made in the linguistic domain and beyond during the last few decades as described in the previous chapters. The main lesson learned is that a single schema or profile will not be satisfying to cope with the variety of resource types, the variety of linguistic sub-domains[21], the variety of linguistic needs and the variety of conditions in which resources are created. Therefore we will choose the following approach:

- we need to work out a taxonomy of data and tool resources[22]
- we need to work out an inventory of basic components to describe such resources and identify which components can be used for the various resource types
- we need to determine the concepts that are not yet in the existing metadata sets, but required to describe the various resource types
- we need to address the issue of how to describe collections of resources[23]
- to guide users who are looking for advice we will add typical metadata descriptions for a number of linguistic resource types[24]
- finally the CLARIN community needs to work out suggestions for elements that should be submitted to the ISOcat standardization process

For this process we will start analyzing which element sets the various initiatives used and whether these are relevant for our domain. In particular we will look at DC, ENABLER[25], IMDI, OLAC and TEI for resources and ENABLER and DFKI Tool Registry for tools. Before starting the analysis we should rehearse on the functions of keyword type of metadata. It is meant for discovery, management and quick inspection. It is not meant to give complete descriptions of resources, although some experts expect that metadata descriptions have to become increasingly rich to be able to find useful resources in the ever growing repositories. Rich metadata will also facilitate automatic operations such as profile matching to find appropriate tools which will become increasingly important in the eScience scenario. The richer the metadata are the more advanced and specific access methods can be defined.

As is known from other initiatives such as LMF that also specify a core model that can be extended by components and elements taken from component and concept registries, many people like to have guidance. Therefore CLARIN will design a number of typical components and profiles that are suggested to be used.

---

[21] An excellent overview can be found under http://www.language-archives.org/REC/field.html

[22] It should be noted here that we need to start in the preparatory phase with a limited set of resources, since the taxonomy will be developed in parallel.

[23] Aggregated objects are differentiated from complex objects according to CLARIN -2/2008. A complex resource is for example a PDF object that contains texts and embedded in texts some images where each of the images itself can be identified. Collections of resources have external relations amongst them, but they exist separately.

[24] Many users will look for best practice of how to describe for example a lexicon. Since there are different views per sub-discipline CLARIN should make suggestions for some sub-disciplines such as NLP, field linguistics etc.

[25] ENABLER emerged from a extensive comparison taking into account various sets.

Nevertheless users can modify them according to their needs. These typical profiles will be worked out in subsequent documents to come and will be available via the web.

6.2 Resource and Technology Taxonomy

A taxonomy of linguistic resources and technology components should adhere to the following principles:

- It should encompass the whole domain of language resources and technology components, i.e. every language resource or component should belong to one class.
- It should exactly subdivide this domain, i.e. a language resource or component should belong to exactly one class.
- It should be possible to suggest a distinct set of metadata for each type of resource or tool.

Ideally, a taxonomy would be also mono-criterial, i.e. the set of classes is based on only on criterion. This is not the case for our taxonomy.

An additional motivation for the preliminary taxonomy which we present in the following is to help our target group in finding the resource which they are looking for. The taxonomy should therefore use names and descriptors which are familiar to the target group.

Before we set out with defining a fine-grained and detailed taxonomy we need to start with the major ones in order to describe the corresponding metadata requirements. In this document we distinguish the following major resource types:

- **Text Resources**: These are all types of resources that are composed of linear texts such as books, papers, articles etc. Q: is a transcript of an audio file a text resource?
- **Speech Resources**[26]: These are audio recordings that can contain speech, singing and other events that can be seen as language material.
- **Multimedia/Multimodal Resources**: These are moving picture recordings mostly with integrated sound channel.
- **Time Series Resources**[27]: These are other types of linear recordings over time created by some device such as data gloves, EEG, FMRI, articulator measurements, eye trackers, motion trackers etc. The object of these recordings are persons who perform some linguistically relevant action (e.g. Articulating a word or sentence) or perform an action on a linguistically relevant object (e.g. Reading a text)
- **Images**: These are data structures encoding for example photos, drawings, video extractions, scanned texts etc.
- **Annotations**: These are structured linguistic encodings that refer to events in time (e.g. Audio or video resources) or to sequences of characters.
- **Lexica**: These are complex data structures where lexical units of different types are described with the help of structured attributes. Lexica can include multimedia elements.
- **Concept Registries/Terminologies:** These are flat lists of terms with their related concepts or normative definitions.
- **Ontologies***:* An ontology is a specification of a conceptualization*.* As such, the term refers to a wide variety of structured objects that in general will include concepts, their definitions and relations between them. Ontologies have to be formalised to a degree that formal inferences can be performed on these data.
- **Objects**: An object is an entity in the real world from which linguistic data are derived. Such an object can be a book or a copy of a newspaper on which a text resource is based.
- **Situations**: A situation is a real-world event from which linguistic data are derived. Such a situation can be an interview which is recorded (by audio and / or video devices) or an experiment.

[26] Within some LRT communities speech resources are expected to be accompanied with the necessary annotations.

[27] In this document we made the difference between speech, multimodal and time series resources to indicate that we include a large variety of resource types. Actually the way to describe these resources will be very similar.

- **Tools/Services**: This is a cover term to describe all sorts of tools, performing NLP speech etc. tasks, web applications and web services that are registered. Time will show whether further sub-classifications will be necessary.

Some additions to this list which we assume to be necessary:

- **Typological databases**: Typological databases contain sets of samples which illustrate one or more linguistic phenomena across a wide range of typologically distinct languages.
- **Grammars**: these are more or less formalised accounts of the structure of natural languages
- **Rule sets:** Rules describe a set of well-formed constructions of a natural of formal language or recurring patterns in linguistic data.
- **Concordances**: Concordances are extracts mainly from textual resources which present a keyword with a user-defined proportion of its immediate context
- **Wordlists**: Wordlists are derived from corpora and present the word form types together with some (often quantitative) information of class information (e.g. gazetters)
- **Transcripts**: transcripts are written records of audio recordings or the audio part of video recordings
- **Training data sets**: Training sets typically include small sets of representative data together with their classification. Training sets are needed for learning tasks and competitions, e.g. CLEF.

Beyond these classes of basic language resources, we have to deal with aggregate resources. In principle, a large number of aggregate resources can be formed from the basic resources mentioned above. However, there are some types of aggregate resources which are well-known in the community and which therefore should be searchable by their own class names:

- **Text Corpus:** are carefully compiled collections of textual data.
- **Speech Corpus:** are carefully compiled collections of audio data in combination with symbolic information (e.g. annotations) and metadata (like speaker descriptions).
- **Multimodal Corpus:** are carefully compiled collections of data which are produced and / or received through different channels or modes
- **Treebank**: these are aggregates of textual (or audio) resources and some levels of (linguistic) annotation. Treebanks are used for linguistic and lexical studies.
- **Session**: This resource type is central to the IMDI metadata, where it is defined as follows: 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.

On another, higher level, we need components to facilitate working with components in various areas such as in the area of metadata and lexica. It will therefore be a requirement to register components and profiles to make them re-usable.

WP5 will further elaborate on the issue of an LRT taxonomy and come up with a document at a later stage in the project.

6.3 Metadata Components

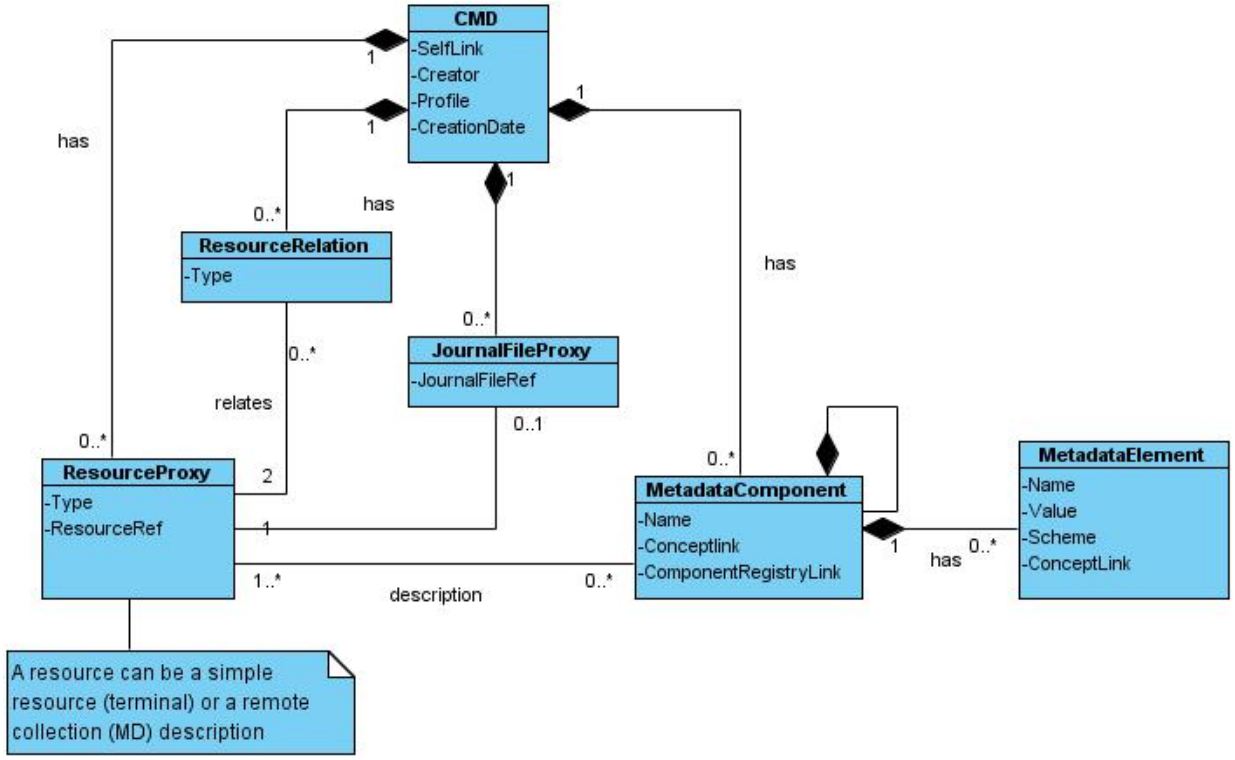
6.3.1 Introduction

Core of the new flexible metadata setup in CLARIN is a component approach as it was inspired by ISO TC37/SC4 when working out [LMF]. Components are the building blocks that can be used to describe different aspects or dimensions of a resource. CLARIN will suggest a number of components, that will also be made available in a component registry, but users can use and create their own components as long as these make use of registered concepts.

A component can exist of components from an accepted concept registry and elements taken from the ISO Data Category Registry (ISOcat), thus components are defined as a recursive structure[28]. The essentials of

[28] It needs to be checked whether we want to define restrictions.

this model are depicted in the following UML diagram taken from the tentative component framework design plans. Next to enabling recursive descriptive metadata components for resources the model is able to describe collections at different levels of granularity and also relations between member resources.

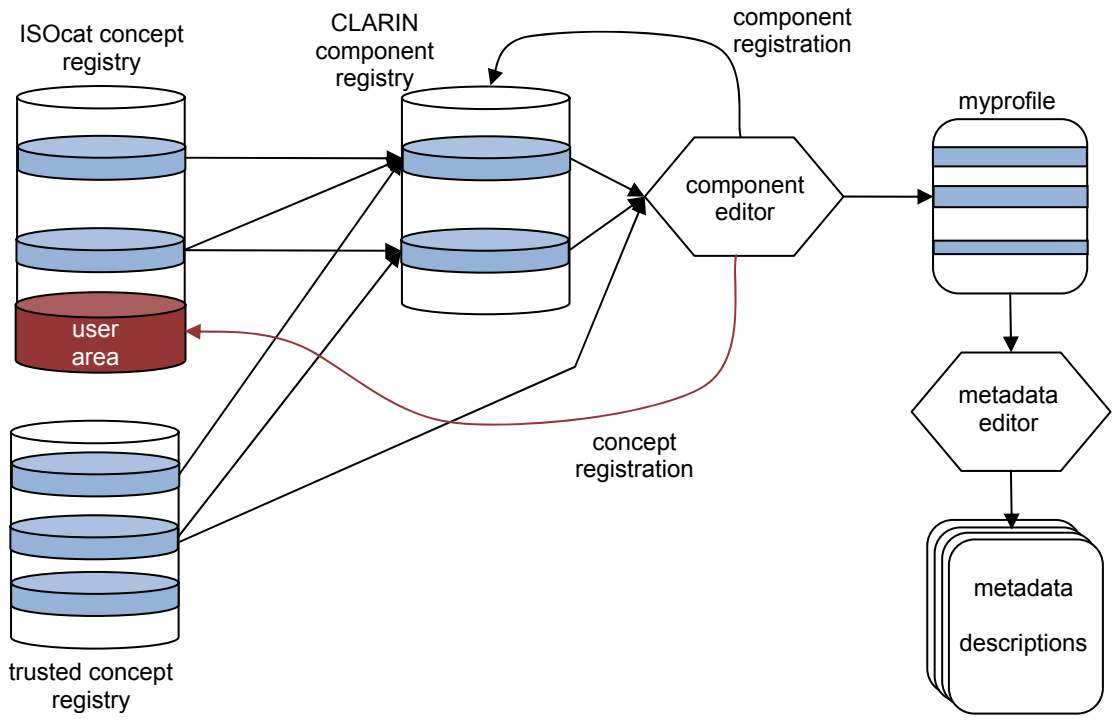


This UML diagram of a Clarin Metadata Description (CMD) expresses (1) the recursive nature of the metadata components: every component can have any number of (sub) components. (2) The possibility of describing collections by having a CMD not only refer to resource(s) but also to other metadata descriptions enabling a description at any level of granularity in a collection. Relations between resources or collections described by a CMD can be modeled by specifying a ResourceRelation. (3) The leafs in the metadata component hierarchy are metadata elements and have next to the elements name and value a link to the concept registry and a value scheme. All administrative data is modeled by attributes of the top CMD datastructure itself. Each resource description can optionally hold a reference to a journal file that contains information about its creation history.

Basic requirement for all components is that the component's metadata elements reference ISOcat registered concepts or concepts in another trusted registry. If a user wants to include an element which is not yet registered since he feels that it is necessary for the proper description of the resource, he would need to register his concept at least in the so-called "user space" in ISOcat first. The ISOcat process will decide if the new category will be integrated in the official part of the registry. CLARIN will be strict and only accept categories that are registered in ISOcat, since otherwise no semantic interoperability can be established.

The diagram below indicates the general architecture CLARIN will implement. All concepts used must be taken from either the ISOcat or other trusted registries such as from Dublin Core. There will be many components in CLARIN certified component registries that can be re-used, but the user can also integrate own concepts or create own components. After selecting and building the user comes to a final ensemble of components that can be transformed into a metadata schema that is used then to create the real metadata resource descriptions. Since there will be various schemas for the various resource types dependent for example on the sub-community (sign-language researchers will need a different type of description of a video signal than a multimodality expert) we better speak about profiles. This terminology is in line with what initiatives such as Dublin Core defined [Baker 2008].

CLARIN metadata components can be created by users, it will be checked by the component editor whether all elements used are indeed taken from ISOcat or another trusted registry. The component editor needs to have a facility to store final component ensembles or profiles to make them re-usable.



This diagram shows the metadata scenario when moving towards a flexible component model where the trusted concept registries such as ISOcat form the basis of interoperability and not schemas anymore. It indicates that users can create their own components and profiles based on existing profiles and by integrating elements that come from trusted concept registries. Finally the user will create his profile that is tailored to his needs and then create the metadata descriptions. To allow some degree of flexibility the framework should accept also the re-usage of well-defined categories that are in the open user area in the ISOcat registry.

6.3.2 Methodology

The component-based model CLARIN is suggesting knows about three layers:

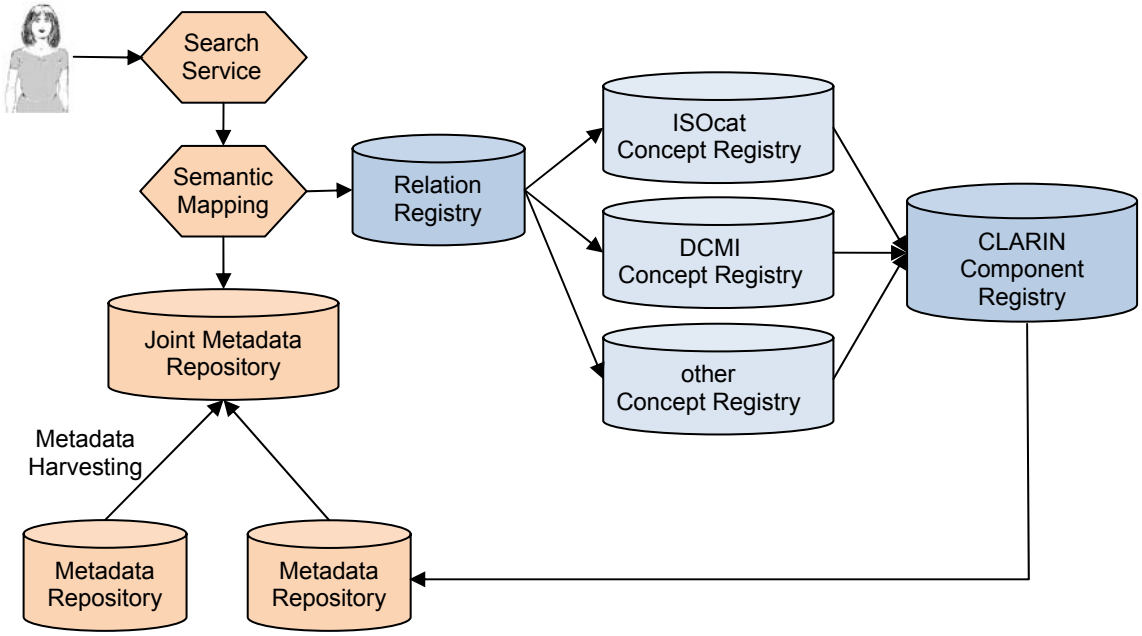
- Concepts that are stored in accepted concept registries such as the ISOcat concept registry to cater for semantic interoperability.
- Relation registries that store simple relations between some concepts from the concept registries to allow semantic mappings.
- Components that can be extended with new elements or used recursively to create new components. They can be aggregated in component profiles that are used by communities, projects or individuals to describe specific resource types.

Components can be viewed as aggregations of useful concepts embedded in a recursive structure while profiles are aggregations of components. The difference between a component and profile is only in the intention of its creator and usage by others, a profile is meant to be reused as a standard way for metadata descriptions while a component is a building block for such profiles. Theoretically however every individual component can function as a profile.

In the foreseen CLARIN metadata infrastructure every user can:

- introduce new metadata elements as long as they are linked to registered concepts in the ISOcat concept registry or other concept registries
- add new components by creating new ones or by modifying existing ones and register them in the CLARIN component registry
- add new profiles (templates) by creating new ones or modifying existing ones and register them in the CLARIN profile registry

CLARIN will provide a fair number of basic components to motivate re-usage and to prevent a proliferation of too many new components at the start of the project. The basis of these initial components will be a deconstruction of existing metadata sets to provide compatibility with the installed base. CLARIN will also provide a number of profiles largely based on existing practices of using metadata sets as IMDI/OLAC/TEI, so that many current projects and archival practices may continue with only limited changes[29]. Also best practice advice about which components and profiles to use for specific resource types and domains will be published[30]. The ENABLER overview will serve as a source of inspiration. The profiles and advice will be restricted to a selection of the most used resource types and a few sub-disciplines such as NLP, field linguistics and the speech and multimodality domain.



All metadata descriptions are based on components from accepted registries, i.e. they are the blueprints for all metadata descriptions that have been created within the CLARIN infrastructure and that are stored at the various metadata repositories. Components are made up of concepts that are taken from accepted concept registries and relations describe the semantic relations between the various concepts. Thus the blue colored registries contain the accepted concepts in CLARIN and the semantic relations between them and the possible components that can be composed. Metadata descriptions created based on these concepts and components and stored in local repositories will be harvested to form centralized CLARIN repositories. These can be subject of searches and other types of operations. Since there will be semantic overlap between some concepts characterized by different terms a search operation on the joint repository for example needs to make use of the relations to find useful hits.

Metadata descriptions based on these profiles can of course be used by individual projects and archives to administrate their resources using whatever tools and infrastructure they see fit. However the added value of the "CLARIN compliant" profiles is that the descriptions can be harvested by CLARIN metadata service providers and represented in metadata catalogs. These service providers will need using the possibilities given by the links to the concept and relation registries to offer users a single unified virtual metadata domain (see figure above).

Later in this chapter we will make an analysis of the existing metadata sets as DC/OLAC, IMDI and the DFKI Tool registry as also of the ENABLER overview to see what basic components and profiles can be derived from them. Since users can create their own components, issues such as granularity and semantic scope are not crucial. This also implies that semantic overlap between suggested components and profiles is not a problem either.

---

[29] The current existing metadata schemas and the instantiated descriptions will have to be adapted slightly with CLARIN specific administrative information. XSLT style sheets will be provided for easy conversion.

[30] Suggestions for proper components and profiles are handy for those users who like guidance.

The suggested CLARIN methodology therefore is as follows:

- Analyzing the currently widely used metadata sets and overviews, we will identify a first set of components taking into account the need for backwards compatibility.
- By an analysis per resource type we will identify remaining gaps with respect to components and add them and suggest profiles.
- By a comparison of element sets we will identify which elements should be included in the ISOcat registry and also identify the semantic relations between them in case of semantic overlap.

6.3.3 Dublin Core / OLAC

Dublin Core and OLAC are widely used in a number of institutions to describe their resources. Dublin Core only describes elements that can be included in any component or profile. Since DCMI elements have been defined properly and are available via PURLs CLARIN will accept this as an "accepted registry", i.e. DCMI elements can be used to semantically anchor metadata elements in CLARIN components. The well-known unqualified set covers 15 elements[31].

> Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type

OLAC is compliant with this approach and defined a few refinements that can be used to refine DCMI semantics. In the following the suggested OLAC extensions are listed:

> Discourse-type, Subject-Language, Linguistic-field, Linguistic-type, Creator-Role

However, OLAC also defines a schema. Due to the important role OLAC has in the community we will turn it into a component and profile so that upwards compatibility is guaranteed.

6.3.4 TEI

The Text Encoding Initiative [TEI] has developed over more than twenty years a widely used standard for the markup of electronic texts. The standard defines several schemata for the annotation of various text types which are used primarily in the Humanities. It covers a wide range of documents and document collections ranging from text corpora like the British National Corpus (BNC) to terminological databases. The encoding is based on SGML for the older versions of the TEI specifications and XML for the more recent versions. Metadata can be placed in the header of a TEI document. As from the most recent version of TEI (P5) its schema can be customized by creating on [ODD] file which contains prose descriptions and a formal specification of elements added by the user. Each TEI conformant document consists of a 'teiHeader' and a 'body' element. In the case of collections, for example linguistic corpora, the body might contain several TEI documents together with their respective headers. In addition, collections must have a separate header for the whole aggregated object.

TEI header elements are widely known in the LRT domain and are used in many projects which have to characterize resources. It seems that TEI has received a new momentum lately and is therefore of relevance for CLARIN. It is at least worth to investigate the elements introduced by TEI. One example is the description of lexica and their integration into the component model which CLARIN is going to specify.

In the teiHeader the metadata are presented in a highly structured way. The outermost TEI header element ‘teiHeader’ encloses up to four elements: fileDesc, encodingDesc, profileDesc, revisionDesc.

- fileDesc: The first component of a TEI header is the file description. It contains three required elements and four optional elements that might be used to specify information about the electronic resource. This element has been modelled in analogy to cataloguing standards used in libraries. Its most important sub-elements are:

[31] Qualified DC covers many more elements which are refinements of the unqualified set of elements.

    - titleStmt: contains information about the title, the authors etc. Sub elements are used to specify, amongst others, information of the title and the author(s) of a resource, to indicate the funder and the principal investigator responsible for the creation of the resource
    - editionStmt: contains information about the edition of a text.
    - extent: contains information about the size of the resource, specified in a unit appropriate for the resource type, e.g. bytes, characters, sentences.
    - publicationStmt: contains information about the publication or distribution of the resource
    - sourceDesc: is used when the electronic resource was derived from an original resource. In this case this sub-element is used to specify a description of the source
- encodingDesc: The encoding description specifies the methods and principles applied for the creation of the electronic resource. The encoding might be described with a free text be or by using pre-defined sub-elements. In the context of CLARIN, the most relevant of these elements are:
    - projectDesc: contains information about the purpose for which an electronic file was created and about the methodology according to that it was constructed.
    - samplingDecl: if applicable, it is used to give information about the methods used to create a sampling in the creation of a resource. It is used, for example, to describe so-called balanced corpora.
    - tagsDecl: provides detailed information about the tagging applied to an XML document. This element might be used, for example, to describe a vocabulary used to annotate linguistic concepts.
- profileDesc: The profile description contains information about 'non-bibliographic aspects' of a resource. Within this element the following optional elements might be used:
    - creation: might be used to describe the origin of the text.
    - langUsage: might be used to describe the language and/or dialects used in the text. The specification of the language(s) used within a text should make use of a language identifier that "should be constructed as in Best Current Practice" (TEI Guidelines). According to the TEI Guidelines P5 current best practice is defined in the IETF documents RFC 4646 and RFC 4647.
    - textClass: allows the user to include a text classification according to classification scheme.

  For some language resources, the profileDesc is also the appropriate place for information about the persons involved in a communication or information about the context of a linguistic interaction. The following elements might be used to present this kind of information:

    - textdesc: might be used to describe the situation within which a language resource was produced or experienced. To do this, several predefined sub-elements can be used, e.g. channel, derivation, domain, and preparedness
    - particDesc: the participants of a language interaction are described within this element. The description might be done with a free text be or by making use of the pre-defined element 'person', along with its sub-elements, e.g. affiliation, age, birth, death, education, sex and socecStatus (social status)
    - settingDesc: setting description provides information about the situation the communication took place. Besides a prose description the element 'setting' with its sub-elements, e.g. date, name, time, and locale, might be used.
- revisionDesc: the revision description is used to provide versioning information

TEI in total offers a self-contained comprehensive descriptive system which cannot be transformed into the CLARIN metadata infrastructure due to its inherent complexity. However, similar to various projects such as the Spoken Dutch Corpus we will make use of header elements that were defined by TEI and that can be used to describe the typical CLARIN resources. Often used TEI elements will be bundled into usable profiles for re-usage by the CLARIN community.

6.3.5 ENABLER Components[32]

ENABLER suggested a variety of components which are partly true for all resource types identified and partly specific for the different resource types. There are no readymade ENABLER schemas and the elements were both taken from other sets and integrated based on bottom-up discussions with LRT providers and users. Due to its overview character ENABLER components can serve as an excellent source of inspiration.

As ENABLER was partly inspired by the ELRA catalogue, we consider its components as sufficient to include the ELRA metadata, i.e. we will not explicitly refer anymore to the ELDA catalogue.

All resources in Enabler share the same "external" description components which are those components that are independent of the resource type:

Resource Identification:
Name, Short Name, ID Number, Version, Type, Description
Creation:
Organization, RelevantProjects, **Dates**
Function:
Application purposes
Validation:
Validation, Type, Methodology, Level
Distribution:
AvailabilityStatus, **Organization**, DistributionFormat, Medium, Price, Documentation
Copyright:
Organization
Dates:
StartingYear, CompletionYear,UpdateFrequency, LastUpdate
Organization:
Name, LegalStatus, URL, ftp, **ContactPerson**
ContactPerson:
Name, Position, PostalAddress, Tel, FAX, E-Mail

In addition ENABLER defined the following components for the various resource types.

Lexicon Description

Macrostructure:
LexiconType, **Languages**, **Size**, **Coverage**, LexicalUnit, **EncodingFormat**, RepresentationLevels, **Development**
Microstructure:
Orthography, Etymology, Frequency, Phonology, Morphology, Morphosyntax, Syntax, Semantics, Definition, Comment, Usage, OtherInformation
Languages:
NrOfLanguages, Languages
Size:
SizePerLanguage, SizePerRepLevel
Coverage:
Type, Domain
EncodingFormat:
Type, CharacterSet

[32] Components are indicated in bold face.

Development
    Sources, Mode

Text Corpora Description

ResourceData:
    Languages, **Size**, **Coverage**, **EncodingFormat**, **Annotation**, **Development**
DocumentData:
    Title, Size, Topic, Genre, Medium, Creator, CreationDate, Other
TextData:
    TextAuthor, PublicationDate, Publisher, TranslatedText, Other
Coverage:
    Type, Domain, TextProductionDates

Speech Resource Description

ResourceData:
    Languages, **Size**, **Coverage**, **EncodingFormat**, **Segmentation**, **Annotation**, **Development**, Content
SpeechData:
    Title, Size, Medium, Waveform, SAmpleRate, SamplingFormat, NoSamples, Topic, Language, Type, Content, Date of Recording, Region of recording, **SpeakerData**
EncodingFormat:
    Type, CharacterSet, Waveform
Segmentation:
    Segmentation, Methodology
Annotation:
    AnnotationType, Annotation, Schema
Development:
    Medium, Type
SpeakerData:
    NoSpeakers, Sex, Age, Origin, Education, Profession

Multimodal Resource Description

ResourceData:
    Size, **Languages**, **Coverage**, **Format**, Type, **Annotation**
Coverage:
    CaptureRegion, ProductionDates
Format:
    Medium, CompressionFormat
MultimodalData:
    Title, Size, Medium, CompressionFormat, CreationLocation, CreationDate, DeviceInstrumentType, DeviceInstrumentSettings, Type, Topic,Language, **ParticipantData**
ParticipantData:
    NoParticipants, Sex, Age, Origin, Education, Profession

Tool Description

Here the "external" descriptors and components are slightly different.

Resource Identification:
    Name, Short Name, ID Number, Version, Type, Description
Creation:
    Organization, RelevantProjects, **Dates**
Function:
    Application purposes
Validation:
    Testing, Evaluation, Performance, Accuracy
Distribution:

AvailabilityStatus, **Organization**, Type, Medium, Price, Documentation
Copyright:
Organization
TechnicalRequirements:
SystemRequirements, OperatingSystem, OtherRequirements
Technical Description:
ImplementationLanguage, ExecutionEnv, InputFormat, OutputFormat, **LanguageDependency**
LanguageDependency:
LanguageDependency, Languages

6.3.6 IMDI Components

IMDI is a structured set offering components at different granularity and it has a high granularity with respect to components which came out of the discussions amongst the involved researcher communities. IMDI distinguishes 3 different schemas: metadata for published corpora (IMDI catalog schema), metadata for collections and metadata for resources and resource bundles. For the latter it has a similar distinction as ENABLER: it has elements that describe non-linguistic resource specific characteristics and elements that describe the specifics of the included linguistic resource type. Within these two main parts it has additional components that can be distinguished:

- The non-specific descriptions dimensions are: session/lexicon, location, project, actor, access, contact, description
- The type-specific description dimensions are: content, language and resource, technical metadata

In almost all sets users can make semantic refinements. An actor in IMDI as well as a creator in DC can have a date of birth or speak certain languages. The latter is a "linguistic" attribute, which is directly related with a person and only indirectly related with the resource to be described.

Here we only will list those components that determine the major description dimensions. For each dimension we indicate the included elements.

Session Description

Session:
Name, Title, Recording Date, LocationAddress, LocationCountry, LocationContinent, Regions, **Description**, **Key**
Project:
Name, Title, ID, **Contact**, **Description**
Content:
Genre, Subgenre, Task, Modalities, Subject, Interactivity, PlanningType, Involvement, Social Context, EventStructure, Channel, **Description**, **ContentLanguages, Key**
ContentLanguages:
Description, **ContentLanguage**
ContentLanguage:
DominantLanguage, SourceLanguage, TargetLanguage, **Language**
Actor:
Role, Name, FullName, Code, FamilySocialRole, EthicGroup, Birthdate, Age, Sex, Education, **Contact**, Source, **Description**, **ActorLanguages**
ActorLanguages:
Description, **ActorLanguage**
ActorLanguage:
MotherTongue, PrimaryLanguage, **Language**
Resources:
MediaFile, **WrittenResource**, **Source**, **Contact**, **Description**, ResourceLink, **Access**, **Description**, **Key**
Generic Resources: ResourceLink, Access, Description, Key
MediaFile:
Size, Type, Format, Quality, Recording Conditions**, TimePosition, GenericResource, Validation**, **Description**

WrittenResource:
Date, MediaResourceLink, Type, Subtype, Format, Size, Derivation, Content Encoding, Character encoding, Language ID, **Validation**
Validation:
Type, Methodlogy, Level, **Description**
Source:
SourceID, Format, Quality, TimePosition, CounterPosition, **Generic Resource**
Contact:
Name, Address, Email, Organization
Description:
Language, Text, Reference

Lexicon Description

The IMDI lexica can have multimedia extensions.

Session:
Name, Title, FinishingDate, LocationAddress, LocationCountry, LocationContinent, Regions, **Description**, **Key**
Project:
Name, Title, ID, **Contact**, **Description**
Content:
Genre, Subgenre, Task, Modalities, Subject, Interactivity, Planning Type, Involvement, Social Context, Event Structure, Channel, **Description**, **ContentLanguages, Key**
ContentLanguages:
Description, **ContentLanguage**
ContentLanguage:
DominantLanguage, SourceLanguage, TargetLanguage, **Language**
Actor:
Role, Name, FullName, Code, FamilySocialRole, EthicGroup, Birthdate, Age, Sex, Education, **Contact**, Source, **Description**, **ActorLanguages**
ActorLanguages:
Description, **ActorLanguage**
ActorLanguage:
MotherTongue, PrimaryLanguage, **Language**
Lexicon:
ID, Datum, Source, MediaRef, Type, Format, NoEntries, NoSubentries, Size, SchemaRef, CharacterEnc, LexicalInfo, **MetaLanguages**, Access, Descriptions, References
MetaLanguages:
Description, **MetaLanguage**

Nodes

IMDI allows generating hierarchies of metadata descriptions by defining nodes. These mainly have a set of links to subsequent nodes.

CorpusNode:
Name, Title, CorpusLinks, **Description**, **Services**
Services:
SearchService, CorpusStructure

Catalogue Descriptions

IMDI provides a separate profile for catalogue collections, i.e. published corpora. These need different types of metadata elements as they also appear in ELDA and LDC element sets.

Catalogue:
Name, Title, ID, Description, Publisher, Authors, Size, DistributionForm, **DocumentLanguages, SubjectLanguages, Location, Format, Project, Access**
DocumentLanguages:
ID, Name
SubjectLanguages:

Description, **Language**
Language:
ID, Name, Dominant, SourceLanguage, TargetLanguage
Location:
Continent, Country, Region, Address
Format:
TextRef, AudioRef, VideoRef, Quality, SmallestAnnotationUnit, Applications, Date
Project:
Name, Title, ID, Contact, Description
Access:
Availability, Date, Owner, Publisher, **Contact**, Description, Pricing
Contact:
Name, Address, Email, Organization

Sign Language Profile

IMDI was used by a large group of European Sign Language Researchers to describe resources; however they needed special extensions to the IMDI schema. Here we will list the extensions that were introduced in addition to IMDI components. It is suggested to include these as registered components so that all SL resources can be represented. The elements are so different that there is no semantic overlap with other elements.

SignLanguageContent:
ElicitationMethod, **InterpretingType**
InterpretingType:
Source, Target, Visibility, Audience
Actor:
DeafnessType, SignLanguageExperience, Family, Education
DeafnessType:
Status, AidType
SignLanguageExperience:
AcquisitionLocation, SignTeaching
Family:
MotherDeafness, MotherPrimaryCommunicationForm, FatherDeafness, FatherPrimaryCommu nicationForm, PartnerDeafness, PartnerPrimaryCommunicationForm
Education:
Age, SchoolType, ClassKind, EducationModel, Location, BoardingSchool

CGN

The Dutch Spoken Corpus project chose to also add specific extensions taken from TEI to describe their resources and these were implemented as a specific CGN profile in IMDI. There is obviously quite some overlap so that it needs to be checked which elements should be reused.

CGNSession:
WordCount, RecCount, ByteCount, TempoAV, RecDate, LocName, Locale, Segmentation, Availability
CGNActor:
Age, BirthYear, BirthPlace, BirthRegion, FirstLang, HomeLang, WorkLang, ResidencePlace, ResidentSize, EducationPlace, EducationRegion, EducationDegree, EducationLevel, Occupation, OccupationLevel

DBD

The Dutch Bilingualism Database project also added special elements as a specialized profile to the IMDI schema. Also these are semantically very special.

DBDContent:
LanguageMode
DBDActor:
CountryOfBirth, AgeAtImmigration, LevelOfBilingualism

6.3.7 DFKI Tool registry

The DFKI tool registry does is presented as a classical web-site; however, it supports a number of metadata elements which are presented here in form of components.

Tool
Name, Description, Abstract, Price, Languages, DistributionMedium, Operating System, Documentation, **Type, Author, IOAspects, ExecutionAspects**
Author
Affiliation, Location, URL, Email
IOAspects
MimeTypeIn, MimeTypesOut, TagSetsIn, TagSetsOut, LanguagesIn, LanguagesOut
ExecutionAspects
ExecutionLocation, RequiredSoftware
Type
SpokenLanguage, WrittenLanguage, Multimodality, Language Resource, NLPDevelopmentAid, Multimedia, Evaluation, Annotation, Others.

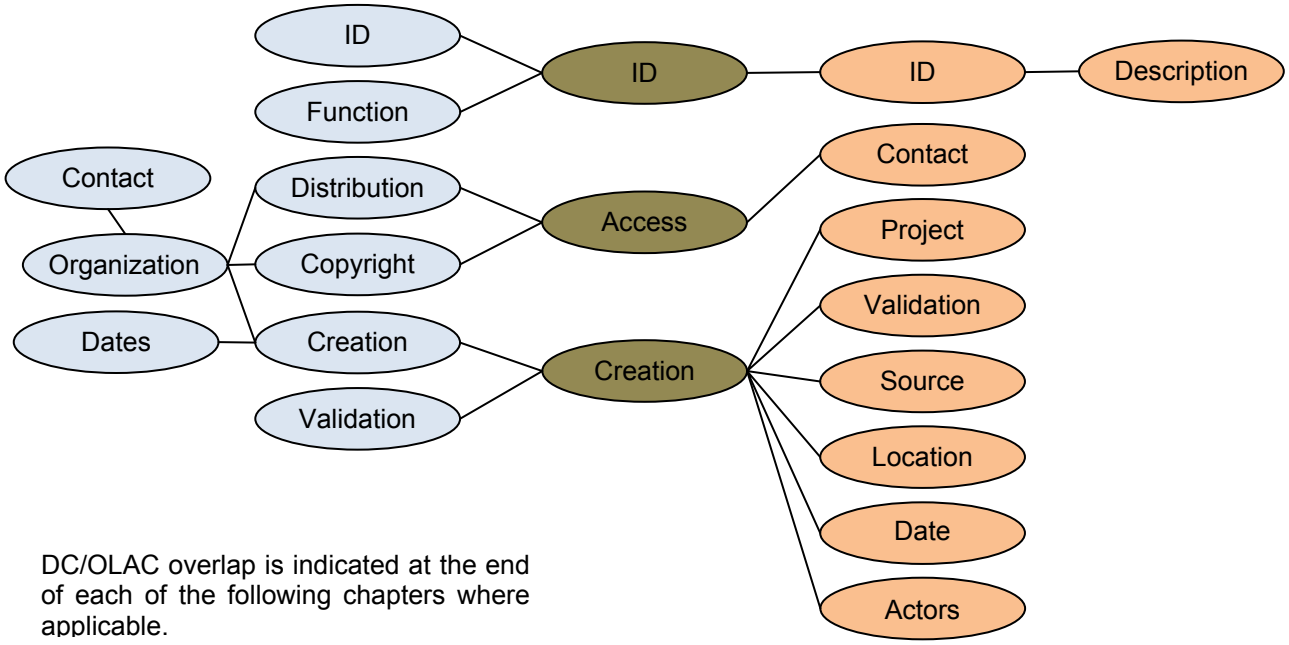
6.4 Comparison of Components

In the following, CLARIN analyses per selected resource type the set of components that were identified to describe them for discovery, management and inspection purposes and indicate where possible gaps in terms of elements. We will start the process for a number of well-understood resource types and work out element sets for other ones in the near future. These highly relevant resource types are: Text Resources, Annotations, Audio Resources, Images, Video/Multimedia Resources and Lexica.

The purpose of this section is to have a basis for the discussions about the concepts that need to be entered into ISOcat and that need to be used. This selection of elements to be put forward to the ISO process will be worked out in subsequent documents and will be available on the web.

With respect to tools we need to synchronize with the discussions in the working group dealing with web services.

6.4.1 Components with General Information



DC/OLAC overlap is indicated at the end of each of the following chapters where applicable.

An analysis of ENABLER and IMDI clearly reveals that one can derive three abstract dimensions that are at the root of the descriptions. These are then further specified. The Actor, Content and Resources components of IMDI mainly describe the content of the resources, i.e. they are not listed here. The bundle nature of IMDI makes the analysis slightly more difficult since some general information is associated with the various

resources in a bundle. In the diagrams in this chapter the ENABLER "components" are in blue, the IMDI components are in red and the suggested main components are in brown.

From this semantic analysis we can derive the need for the following general components:

- ID needs to offer components that identify the resource and give information about it:
    - Identification information: identifying a resource[33], referencing to the resource
    - Function Information: describing the function or purpose of a resource
    - Description Information: describing the resource in general terms
- Creation needs to offer components that describe the complete creation process without describing the content of the resource itself:
    - Project: describing the formal context in which the resource was created
    - Organization: describing the organization which is responsible for its creation
    - Location: describing the geographic location in which the resource was created
    - Dates: describes all dates that are relevant for the creation process
    - Source: describing the sources that were used to create the new resource
    - Validation: describing the quality controls that were carried out on the control
    - Creating Persons: describe the people that were involved in the creation
- Access needs to offer components that describe the way users can access the resource:
    - Distribution Information: describes the media and costs that are involved in receiving information
    - Technical Access Information: describes the way technical way a resource can be accessed
    - Copyright Information: describes the legal and ethical aspects that are related with the usage of the resource
    - Contact Information: describes the organization or person to be contacted to get access or information about the resource

These dimensions can nicely be linked with the semantic dimensions of DC. The identity is described by the elements: Identifier, Description and Title. The creation is described by the elements: Creator, Contributor, Coverage, Date, Type and Source. The access is described by the elements: Publisher, Rights and Format

The DC elements Subject, Language and Relation as well as the OLAC extensions Discourse-Type, Language, Linguistic-Field, Linguistic-Type and Role all describe content aspects of the resources.

Special sub-discipline profiles as listed in chapter 6.3.6 don't require extensions at this level.

6.4.2 Metadata Components for Lexica

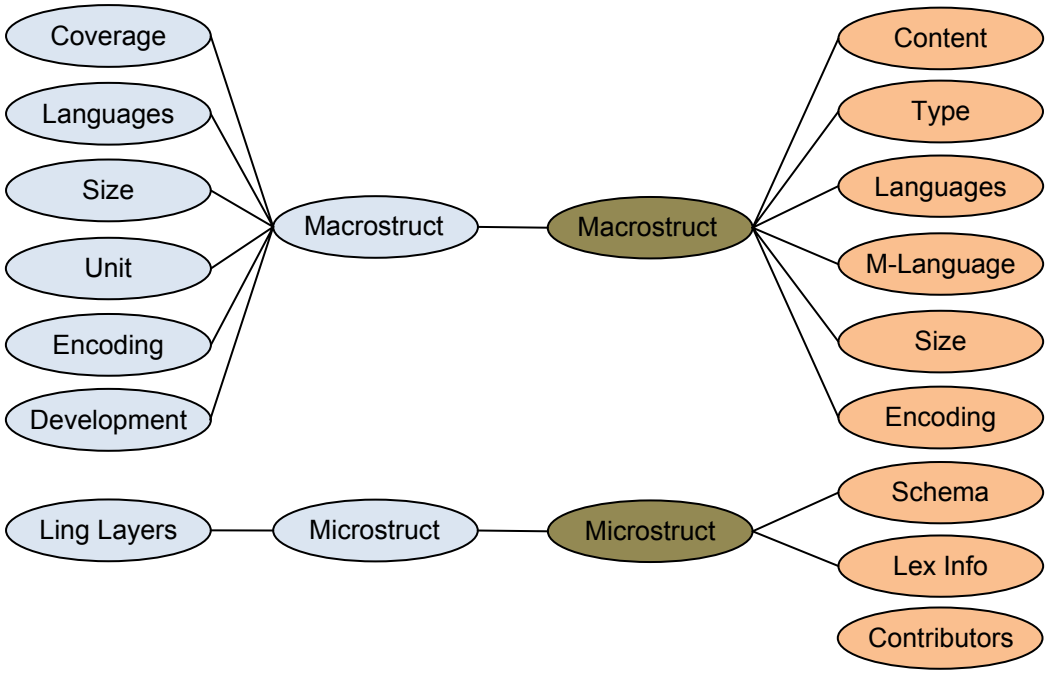
Next we will analyze the components that are necessary for describing lexica as indicated by ENABLER and IMDI.

It seems to be wise to use the distinction between macrostructure and microstructure ENABLER is making. For both sets there is much agreement between the kinds of components that need to be available which is not surprising since both are derived from the intensive MILE discussions.

- Macrostructure information needs to cover a wide range of topics:
    - Coverage/Content information: describes the type of language and/or the type of semantic domain a lexicon is about
    - Languages information: describes the languages that are involved; IMDI differentiates between object and working languages which is necessary

[33] In this chapter the term "resource" denotes a single resource as well as an aggregation of resources.

- Size information: describes the size of a lexicon in different terms; a variety of elements needs to be provided to support the different ways of specifying the size of a lexicon; in case of lexica with multimedia extensions the sheer size in terms of storage capacity is essential as well
- Unit information: characterizes mainly the type of headword that is used in the lexicon
- Encoding information: describes the type of internal encoding which has two dimensions: the character encoding dimension, the structuring mechanism (XML, rDB etc)[34]
- Development information: describes the way the lexicon is derived from source material



- Microstructure information needs to describe which kinds of information types can be found in the lexicon:
  - Linguistic Layer information: describes a qualified list of linguistic layers that are included in the lexicon
  - Schema information: actually a reference to the exact schema for further information of the microstructure
  - Contributor information: describes the persons or machines that participated in the creation of the various lexical attributes

The DC elements Subject and Language can be mapped easily to Content and the Language information. OLAC as well makes the distinction between object and working languages.

6.4.3 Metadata Components for Audio Resources

Although audio resources are dealt with for quite some time they becoming increasingly important in the area of linguistics due to the increasing amount of accessible resources. They are used for a large variety of purposes ranging from automatic speech recognition with more detailed expectations with respect to quality description to field linguistics where elements characterizing the linguistic environment are very much important. Audio recordings contain a variety of language material ranging from speech to songs.

Both ENABLER and IMDI apply a description scheme that bundles transcriptions with annotations which should not be maintained. We need to describe atomic objects as the basis and from there describe

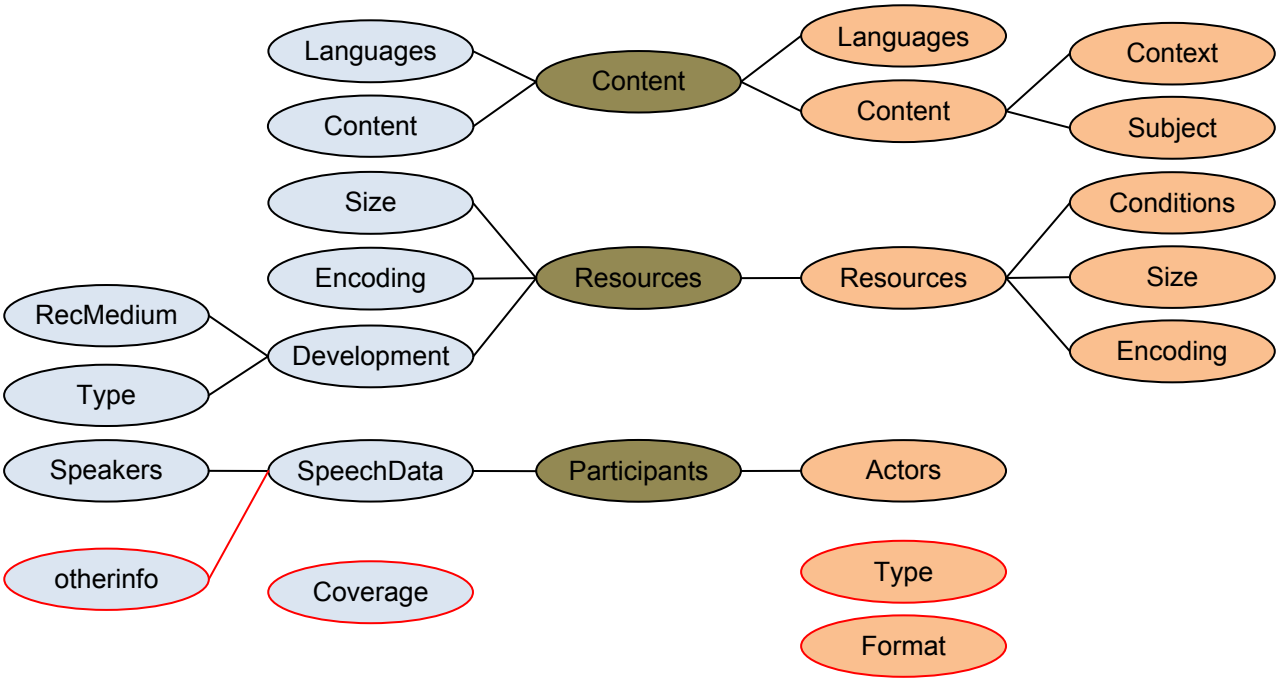
[34] This information probably should be covered under general information.

aggregations. In the following we need therefore to disentangle the element sets and only mention the ones that are required to describe an audio resource. Also both use some elements that are already covered under general information which will be left out in this analysis. These are indicated with a red line.

A further analysis will be required to see whether audio resources can't be included under the heading of multimedia resources as well, since they obviously share many elements.

We can distinguish three main components all having sub-components:

- Content information that describes the content aspects of the resource
  - Language information: describes the languages involved and their setting
  - Content information: describes a whole variety of issues that is relevant to interpret the content in the right way such as elements describing the subject and elements describing the linguistic context, task etc
- Participant information covers all relevant information that is important to describe those persons (and computers) that are participating in the recording (thus not the creators). On purpose we left out all elements and sub-components that may be necessary to describe the participants, since the information required is very much dependent on the sub-discipline. An impression can be gained by looking at the special IMDI profiles.
- Resources information covers all characteristics that describe the recording circumstances. Be aware that an audio recording could include several channels.
  - Conditions information: description of all technical context information of the recording
  - Encoding information: description of all information that describes the way the content is encoded



The DC elements Subject and Language can be mapped easily to Content and the Language information. OLAC as well makes the distinction between object and working languages.

Special sub-discipline profiles such as mentioned in 6.3.6 need to have extensions in the description of the three main dimensions. Since this is also true for the following resource types we will not further mention this.

6.4.4 Metadata Components for Multimedia/Multimodal Resources

Since we miss a good cover term for all types of multimedia (here combined audio and video) and other typical time series recordings as they become increasingly popular in linguistic research labs working on multimodality, virtual reality or even brain research studies, we have summarized this kind of resources under

the title "multimedia/multimodality". All these recordings actually share the same main components as were mentioned under 6.4.3. Only some of the elements and their values that are needed to describe an EEG or Eye Tracking recording in detail for example may differ from those used for describing audio recordings.

Thus we expect changes with respect to the sub-components, elements and values with respect to the "Resources" component. Dependent on where to put it we certainly would need a classifier for determining the type of resource.

Special cases we need to deal with are images. We could interpret them as time series with a length of one time unit. Yet we need to analyze what the habitudes are. In IMDI the "sessions" component is used in this way to describe thousands of photos and drawings covering language material in some form.

DC and OLAC don't add special dimensions in this respect.

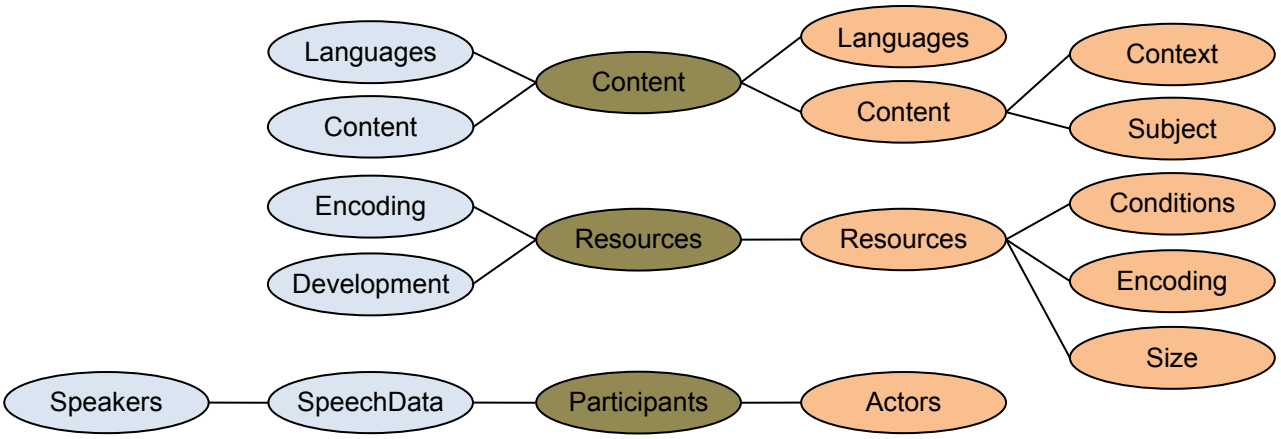
6.4.5 Metadata Components for Text Resources

Text resources as basic language material are very common in particular in all kinds of humanities disciplines. They may come from newspapers, books, singular documents and many other sources. We want to separate these basic material from all sorts of annotations that are adding layered information to other basic resources such as source texts, audio and video recordings etc.

ENABLER makes a distinction between ResourceData which is a cover term, DocumentData which refers to the digitized version and TextData which refers to the original text. Again the distinction between ResourceData and DocumentData will not be maintained due to the atomic approach in CLARIN. Similarly IMDI includes "WrittenResources" in a bundle which will also be amended by an atomic approach.

Collections of textual resources will be dealt with under aggregations; however, it is up to the resource provider to determine the granularity of what a resource is. When a resource provider is harvesting newspaper issues every day for example, the resource could be such an issue. If done so many genres will be included in one single resource and based on metadata the user could not do a selection with the help of a query. If every article of such an issue would be seen as a resource then the metadata description would be more specific, but the amount of metadata would increase extremely.

A comparison of the ENABLER and IMDI suggestions indicates that the main components are the same. Therefore we can re-use the main dimensions.



Dependent on the granularity choice there may be additional sub-components, elements and vocabularies required to describe textual resources. In specific cases the components describing participants do not make sense or can better be extracted from the content by named entity recognition modules. But these could be used to automatically enrich the metadata descriptions.

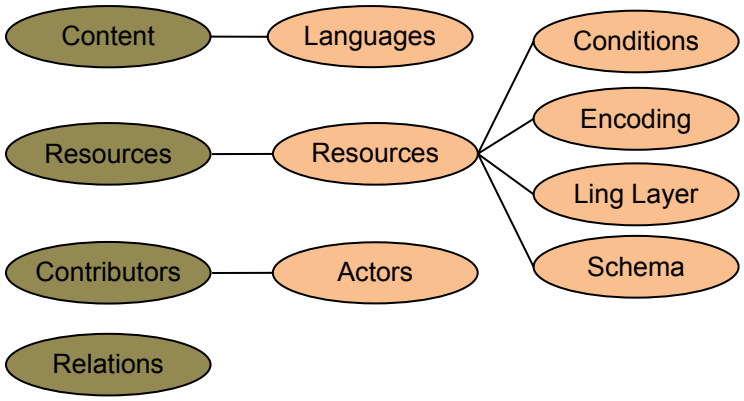
In the past most discussions were about terminology, i.e. different "labels" were required, since typically those researchers who work with texts are not the same than those working with multimedia/multimodality resources. It has to be analyzed in how far these terminology differences can be handled at the concept level in the data category registry or at the level of component registries or whether different components need to be provided.

6.4.6 Metadata Components for Annotations

Annotations can occur in different forms. They can occur as single unstructured comments on specific text or media fragments. They can also occur as structured linguistic extensions - a type of metadata in its general meaning - on texts or media data. In the latter case in general we speak about annotation structures which can have multiple linguistic layers, i.e. starting with a transcription of what is being said up to a complex semantic or pragmatic description. In general we can speak about tiers of annotations where each tier contains linguistic encodings of a certain type and where each annotation is associated with a sequence of characters in another set of tiers or with a period of time. Therefore, annotations are always associated with another resource which needs to be specified by a relation element. This relation must be qualified since it could for example refer to one of the channels of a stereo recording.

ENABLER has some provisions for transcriptions as part of media resources and IMDI design was based on annotation layers which are part of bundles. Not all annotation tiers will be in separate resources. Many tools are allowing the users to create one complex annotation structure as one XML file or as a table structure in a database for reason of simplicity. In (semi) automatic annotation systems such as in NLP operation chains it is often a matter of principle to apply standoff methods, i.e. for every annotation tier a separate resource is created. In reality a mix will mostly be found. Thus similar to lexica we can speak about the difference between macro and micro structure.

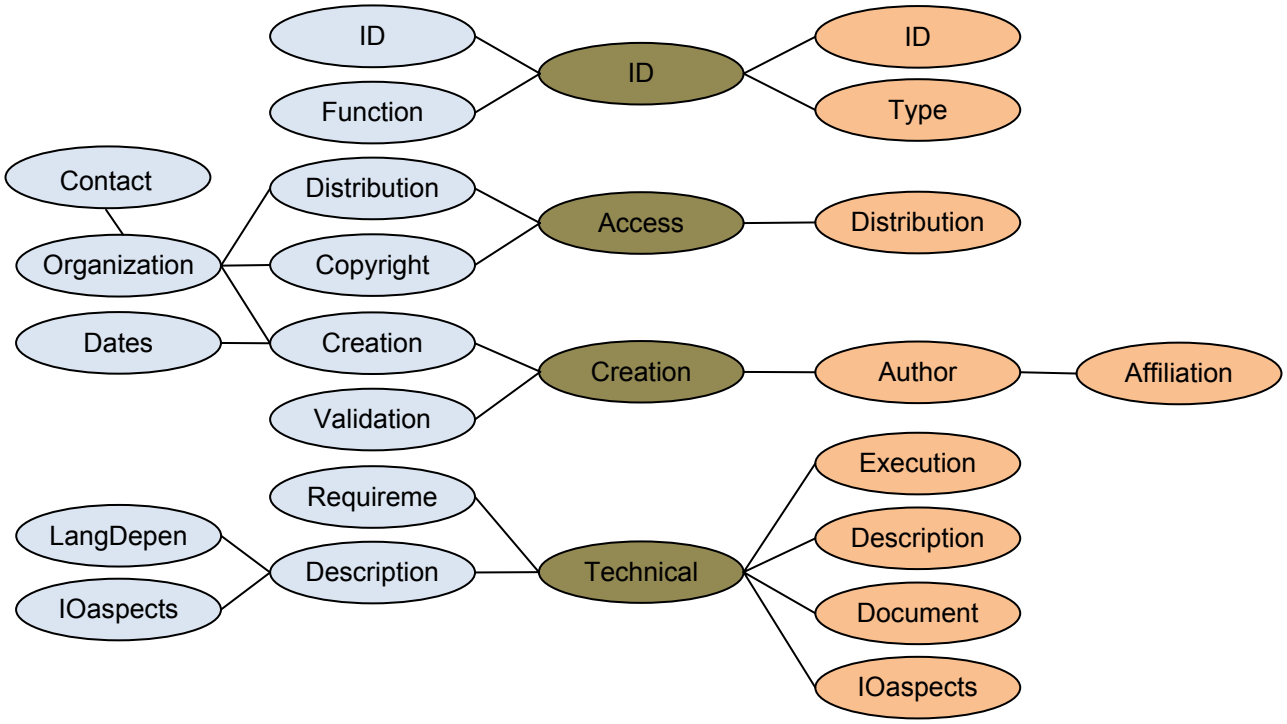
Due to the association with another resource much information such as content is already stored in the primary resource. The disadvantage is that when studying the metadata of an annotation one would need to find the root resource in certain discovery tasks. However, we need to provide a few content description elements such as languages since a tier could cover a translation to another language for example.



Due to the similarity with other resource descriptions the major components don't need to be described in more detail. It should be mentioned here that these diagrams only describe components but not their embedding in structure. Contributors would be related to the linguistic annotation layers. The component "Encoding" in this context will point to tag sets used etc.

6.4.7 Metadata for Tools

With respect to tools the statements here can only be of preliminary nature, since a careful analysis about the needs has been started in WG 2.6. A first workshop about web services and their description took place in November and another one will follow in February 2009. In this analysis we are comparing the ENABLER specifications with those of DFKI. Since it is suggested to use slightly different general information we list the possible components.



DC and OLAC don't add special dimensions.

6.4.8 Provisions for Relations

It is obvious that the choice for a atomic approach requires new ways of how to handle bundles of closely linked resources such as annotations on texts or audio recordings. Both IMDI and ENABLER include some form of bundling which is very handy for certain tools, but difficult to deal with in a component based framework. DC has a separate element to include relations of all sorts which CLARIN needs to adopt. This needs to be done in collaboration with ISO TC37/SC4.

6.5 Aggregated Resources

As already indicated we need to distinguish four forms of aggregations: (1) complex resources, (2) bundles of resources, (3) collections of resources, (4) published collections (corpora).

We can speak of ***complex resources*** when we can speak about an integration of several resources into one new resource. Let's take a PDF document as an example that includes a photo where the photo and the PDF document both can have identifiers.

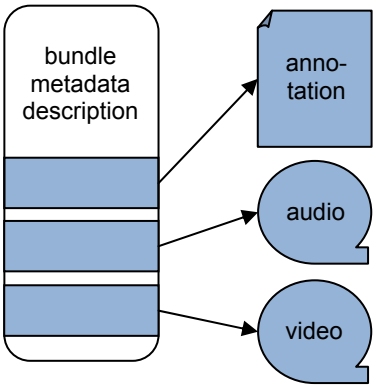
Some speak about ***bundles of resources*** if the resources share a specific semantic relation. Let's take an annotated media file that covers video and audio recordings and several layers of annotations. All these resources share the same time line or sequence of characters. Actually in sets such as IMDI this special and implicit relation is exploited by the multimedia tools. DCMI did not make such a step, but allows users to explicitly add any type of relation to other resources.

Collections of resources are user or depositor defined aggregations that fulfil a certain purpose. They can be created explicitly by the depositors to come to a browsable and therefore manageable hierarchy, they can be created by a user who published a paper based on a selection of resources by using either explicit or implicit relations, they can be created by machines by gathering resources that adhere to specific criteria etc. Bundles are therefore specific types of collections.

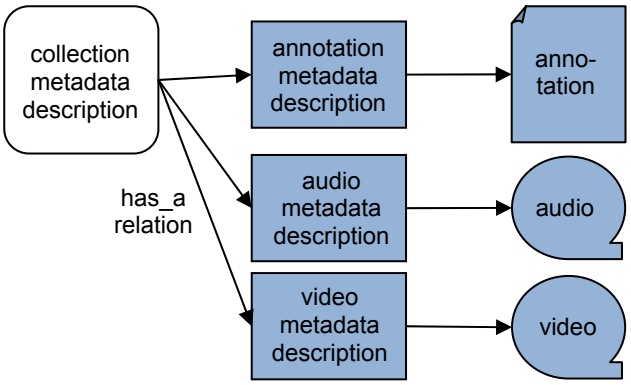
Published collections (synonyms: published corpora and reference collections) are collections that have a certain status, a wider relevance, mostly a specific name and special attributes associated with it. Examples for such collections are Brown Corpus, British National Corpus, and Dutch Spoken Corpus. Published collections are collections, however, they need to be treated especially since often they were evaluated according to special criteria, they are associated with special access restrictions etc.

Collections are generally defined in a recursive way, i.e. (virtual) collections can be built from collections and resources.

The metadata description of complex resources follows exactly the same rules as for resources except that links to the parts may be included. Bundles were introduced for example by IMDI to indicate the close relationship due to the same time axis. The concept of implicit relations embedded in a single structure is strong for further exploitation by tools since all relevant information can be found in one object. However, it has its limitations that become apparent for example in cumulative annotation scenarios. Here a more generic approach with explicit relations is required.

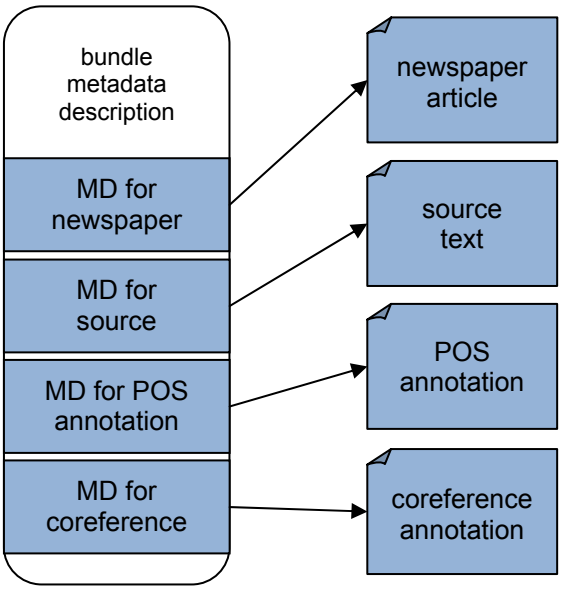


Bundle of resources with a metadata representation and implicit structure as can be used in IMDI for example. Also ENABLER suggested this concept.



Generic collection metadata description with separate descriptions for all resource types, a metadata description representing the collection and explicit relations.

For any form of collection it is important that they need to be identified by separate metadata descriptions as indicated by the above figure. This is the only way to make them referencable, associate a PID and access policy information with them etc.



This example shows an implementation of the bundle concept for a treebank example. This example indicates the disadvantage of the bundle concept. In an accumulative annotation scenario different creators are active creating new annotation layers. A bundle metadata description cannot be used as a container for contributions from different sides. A collection approach with separate metadata descriptions is more appropriate since every creator can create his own collection and refer to the shared resources.

A concrete example can also be given for a Treebank which can be seen as a bundle, i.e. as a special form of a collection of resources. Nevertheless, it is important to be able to refer to it as one resource. Such a resource typically consists of a text resource that is based on some issues of a newspaper. The text content of these newspaper issues may have been harvested from the CD-Rom version for example. Later on, the text was annotated by different annotators in different time spans, resulting in several annotation layers. As a

result such a Treebank is a collection with metadata records for the newspaper issues, the digitized text and currently six different annotations.

6.5.1 Metadata Principles for Workflows

This chapter will be written at a later phase when the working group on web services and workflows (WG2.6) has a deeper insight into the requirements,

6.6 Views and Filters

Experience has shown that large unstructured repositories of metadata as they will be gathered by the CLARIN service provider will be difficult to navigate. There are a number of reasons such as

- the sheer amount of resources the metadata of which will be harvested
- the high granularity which we expect to handle where every annotation tier to an interview could be a resource in itself
- the wide spread of responsibilities between various countries and institutions

Without structuring mechanisms at the side of the service provider we will not be able to create an attractive domain for finding material. The OLAC service provider decided to only harvest descriptions of corpora to prevent situations where people are looking for Dutch resources and get a huge amount of hits from one data provider with a high granularity and only one hit from a data provider with a low granularity. Finally such a singular hit would not be visible anymore. On the other hand we can only influence the policy of the individual data providers to a certain extent. In the research world the notion of corpora is blurring, since individual resources tend to become part of different "virtual" collections dependent on the goal of the scientific analysis. In general researchers are not creating resources to create a specific corpus, but they create a collection of resources from which they believe they can give answers to certain research questions. Older resources or resources from colleagues may be as interesting for new research questions as the newly created ones.

CLARIN therefore is looking for innovative solutions to cope with the mass of metadata descriptions. All must be based on the principle of delegated responsibility, i.e. national coordinators are responsible to check the quality and availability of metadata descriptions. It is obvious that high quality metadata, i.e. correctly used elements and a high degree of filling, will be crucial for any type of intelligent operation.

The following operations need to be foreseen:

- It must be easy for users to specify hierarchies of elements to create browsable trees that will help in navigation. Such hierarchy specification will result in queries that will be executed on the resources and in virtual nodes characterizing the resources under the node. The user must be free to choose the dimensions of ordering, however, a few typical views should be given as standard options.
- It must be easy for users to filter out resources according to certain characteristics. If the researcher is only interested in German and Dutch resources he should be able to specify that only these will be visible.

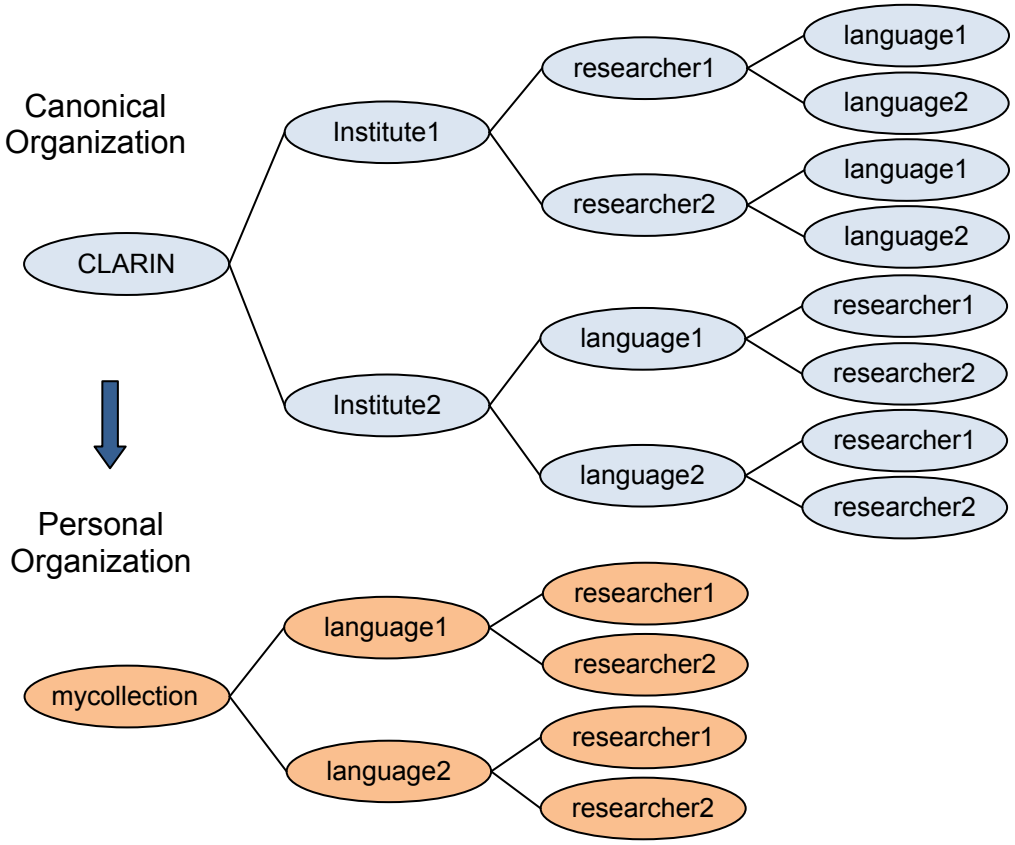
The following figure indicates schematically the two options.

WP5 will work out a document that describes the requirements for these options in more detail. An automatic abstraction process can overcome the problem mentioned above that different granularities lead to unbalanced hit lists, since a virtual node would be automatically created to represent all resources that adhere to a certain criterion. The problem that certain fields are not filled in in many cases needs to be dealt with.

In general we can distinguish a few layers of representation that are relevant to implement these principles:

- Institutions offer metadata descriptions for harvesting, these metadata descriptions are not organized, i.e. they appear as list of records (probably offered via the OAI PMH protocol).
- At two levels (national and European) these records will be harvested and stored in an internal representation format which may be in the form of a relational database to implement fast searching for example.

- Portal tools are available for the user to specify hierarchies and filters to create a special personalized view on these stored representations.



This diagram indicates the process of abstraction and filtering to generate a new view. Here we assume that metadata descriptions are already organized according to some criteria, but this could also be generated from the description by an abstraction process. It can happen that the organization in different institutes is done according to different criteria, i.e. in one scheme researchers are differentiated at the highest layer and at another one languages are the highest organization criteria. A user does not want to see the institute nodes, but would like to see an organization based first on languages and second on researchers. By specifying this hierarchy "Language->Researcher" he should get the personal organization. Of course any element could be used to define the organization. For example a researcher may want to organize is collection according to two languages and a number of ages etc.

Yet there are no suggestions except in IMDI of how to represent virtual nodes. IMDI nodes don't include linguistic information except a description. On the other hand IMDI as well as the ELDA and LDC catalogues have information about "published corpora", which represent one level of a collection description. In CLARIN we need to work out how we want to represent nodes and what kind of information we want to store in them.

7 Procedure

The work in CLARIN with respect to the metadata infrastructure will occur in 6 almost parallel tracks:

1. preparation
2. profiles, components and elements
3. architecture and portals
4. centers network
5. infrastructure
6. goals for 2009 and 2010

7.1 Preparation Work

There are a number of issues that need to be taken up by other work packages and working groups. We want to briefly mention them again:

1. WP5 needs to come up with a comprehensive taxonomy of resource types which will serve as basis for adding other profiles, components and elements.
2. WP5 will deliver a paper on views and filters which will help to specify the nature of the tools which will become part of the infrastructure.
3. The Joint WP2/5/7 effort to get a comprehensive and complete LRT inventory needs to be carried on to get an overview about the resources and resource types. This could be used to increase the mass of metadata descriptions that CLARIN will offer.
4. In WG2.1 we need to define the centers that are powerful enough to offer metadata so that they can act as metadata provider in the CLARIN scenario.
5. Based on the analysis work in WG2.6 we need to come to an improved specification of how to describe tools, web applications and in particular web services.

This work will move on as described in other documents.

7.2 Profiles, Components and Elements

In this track quite some work needs to be done which finally will result in the set of elements to be proposed to become part of the ISOcat concept registry and in a set of registered components and profiles. In detail the steps to be done are:

1. Based on the input from WP5 and by further analysis steps we need to extend the set of suggested components so that all relevant resource types are covered. In particular we need to look for ways how CLARIN is going to represent virtual nodes and collections.
2. Based on a refined analysis of components we need to come to the set of elements that a) will take care of upwards compliance with the existing installed base and b) will close the existing gaps.
3. For this purpose it is necessary to come to a final decision about the semantic granularity of categories in the ISOcat registry which is a joint effort together with ISO experts. A document needs to specify which is left to the context in which a concept will be used.
4. Specifications about elements need to be distributed to the whole CLARIN community for comments before submitting them to the ISO process. To do this we need to set up a suitable commenting platform.
5. All accepted elements need to be entered into ISOcat and a standardization process needs to be started. CLARIN can already start working when data categories will appear in the CLARIN workspace.
6. From various partners in the various countries the language dedications within ISOcat need to be filled in.
7. A number of prototypical profiles need to be established that take care of the installed base, a number of typical sub-communities and resource types.

The next step along this track will be an expert workshop in January.

7.3 Architecture and Portals

Based on the requirements specified in this document and others forthcoming and the experience that is available in the CLARIN community we now need to come from individual architectural sketches as they can be found in this document for example to a full architectural description. This description needs to describe the core of the infrastructure which are

- the registries for concepts, components and profiles,
- the metadata repositories,
- the harvesting principles based on hierarchical domains of responsibility,
- the mechanisms of referencing used,
- the embedding of tools such as searching, browsing, filtering and viewing and
- the specification of the component mechanism.

In addition we need to specify what the portals will do and how they should be designed. One of these portals is based on the LT World maintained at DFKI which needs to be extended to a CLARIN LTR World.

The plan is that the architectural drawing and the description of the portal will become ready in January 2009.

7.4 Centres Network

In relation to the progress of WG2.1 we will specify which of the centres will classify as class A, B or C centres. All are potential metadata providers. We will interact with each of the centres to understand in which way they can offer metadata for harvesting. Preferably harvesting will be done according to the OAI PMH principle, however, since we know that this can form a serious obstacle we would also accept XML files that can be harvested and checked against a schema such as the one from IMDI.

In case of OAI PMH support we need to help centres setting up this port and carry out functional and quality checks.

In addition we need to clarify which institutes will run the national portals and what kinds of responsibilities these hubs will have in CLARIN to provide a smoothly functioning infrastructure.

In a few documents we will describe all aspects. The main work will be done in January and February. Setting up harvesting pipes and testing the quality will be done step by step in 2009.

7.5 Metadata Infrastructure

Based on the architectural drawing and the descriptions of the portal tasks we will start specifying and designing tools that will be part of the new component based metadata infrastructure. As indicated we need to guarantee upwards compatibility for the installed base and we have to balance between short term and long term goals. This specification and design work needs to cover all aspects of the infrastructure. We will not create very detailed specifications for all components; however, sufficient detailed specification documents will be necessary where the development work will be carried out by several different participants.

One of the major tasks will be to understand which of the various CLARIN members has national funds to participate in the development work. The work in setting up registries and building tools should start in February. It is intended to discuss these issues in the February workshop.

7.6 Goals for 2009 and 2010

It will be necessary to work out a development roadmap that will include all the above mentioned aspects. It will be necessary to find a balance between short term requirements to present early results and long-term requirements without risking too much inefficiency. A future document will be drafted which will describe a development and implementation roadmap as a basis for further discussion and decisions by the Executive Board at the next meeting in March.

At present, we recommend that short term activities should make as much use as possible of the infrastructures that already exist and have proven to work smoothly, i.e.

- It is recommended to currently use the LRT inventory to describe resources and tools in an easy and handy way; CLARIN will take care that the entered information will be re-used for the metadata creation process.
- Where institutes are under a high pressure to create proper metadata on a short term, it is recommended to describe LRT with the help of the existing IMDI or OLAC frameworks and to use the currently available extension mechanisms of these frameworks where necessary. Extensions should be chosen carefully and where possible they should make use of available elements described by TEI or DC. Institutes should provide XML or OAI-PMH based harvesting options, so that the CLARIN service provider can harvest these metadata descriptions already now. CLARIN will take care that these will be harvested and that the descriptions can be transformed later on so that they can be integrated into the emerging component-based infrastructure.
- CLARIN already started the necessary work to be able to operate as a metadata service provider, i.e. currently efforts and investments are made to implement the OAI-PMH harvesting protocol and a suitable gateway between OLAC and IMDI[35].

[35] The Gateway from IMDI to OLAC is already operating since years and has been defined in a collaborative effort between IMDI and OLAC specialists.

8 Appendices

8.1 Dublin Core Element Set

The Dublin Core metadata set was developed to be simple and orthogonal, and to describe all types of Web-based documents. However, Dublin Core has been used with other types of materials and in applications demanding some complexity. There has historically been some tension between supporters of a minimalist view, who emphasize the need to keep the elements to a minimum and the semantics and syntax simple, and supporters of a structuralist view who argue for finer semantic distinctions and more extensibility for particular communities.

These discussions have led to a distinction between qualified and unqualified (or simple) Dublin Core. Qualifiers can be used to refine (narrow the scope of) an element, or to identify the encoding scheme used in representing an element value. The element *Date*, for example, can be used with the refinement qualifier *created* to narrow the meaning of the element to the date the object was created. *Date* can also be used with an encoding scheme qualifier to identify the format in which the date is recorded, for example, following the ISO 8601 standard for representing date and time. All Dublin Core elements are optional and all are repeatable. The elements may be presented in any order. While the Dublin Core description recommends the use of controlled values for fields where they are appropriate (for example, controlled vocabularies for the Subject field), this is not required.

A more recent evolution within the DC realm is the introduction of the Dublin Core Abstract Model [DCAM]. This RDF-based object model provides a framework to describe resources in terms of binary relations (cfr. RDF triples), like *Resource X has-title Y* or *Resource X has-publisher Z.* These relations are expressed by assigning a value to a property, like in the following example:

```
<dcds:statement dcds:propertyURI="http://purl.org/dc/terms/title">

   <dcds:literalValueString>DCMI Home Page</dcds:literalValueString>

</dcds:statement>
```

Here the property *title* – described as an RDF file which can be found by looking to the URI specified in the **dcds:propertyURI** attribute – gets a value which corresponds to a string ("DCMI Home Page").

It should be noted that this model does not necessarily rely on a specific syntax: it can be expressed using XML [DC-DS-XML] or [DC-TEXT] which resembles in large extent to the RELAX-NG (compact) syntax. Another remarkable difference when comparing DCAM to the traditional DC set is the fact that there is no longer the obligation to map all metadata elements to one of the 15 core elements. As such it provides a higher degree of flexibility.

8.2 OLAC Extensions

The OLAC extensions can be found in the specification document: http://www.language-archives.org/REC/olac-extensions.html.

8.3 The ENABLER Overview

These overview lists resulted from a broad survey of metadata elements used by different initiatives: EAGLES, ISLE Meta Data Initiative (IMDI), Open Lexicon Interchange Archives Format (OLIF2), Open Language Archives Community (OLAC), Browsable Corpus (BC), Corpus Encoding Standard (CES), Codes for the Human Analysis of Transcripts (CHAT), Dublin Core (DC), European Language Resources Association Catalogue (ELRA), Gesture Databank (GDB), International Corpus of English (ICE), Linguistic Data Consortium Catalogue (LDC), Multimedia Content Description Interface (MPEG-7).

The ENABLER overview also describes the type of vocabulary which is left away here. For a more detailed view we refer to the ENABLER document.

8.3.1 External Metadata for Language Resources

Resource identification*
- Name
- Short name
- ID number
- Version
- Type
- Description

Creation
- Organization
  - Name
  - Legal status
  - URL
  - ftp
- Contact person
  - Name
  - Position
  - Postal address
  - Telephone
  - Fax
  - E-mail
- Relevant project(s)

Name
- Dates
  - Starting year
  - Completion year
  - Update frequency
  - Last update

Function
- Application purposes

Validation
- Validation
- Type
- Methodology
- Level

Distribution*
- Availability status
- Organization
  - Name
  - Legal status
  - URL
  - ftp
- Contact person
  - Name
  - Position
  - Postal address
  - Telephone
  - Fax
  - E-mail
- Distribution format
- Medium
- Price
- Documentation
- Documentation Language(s)

Copyright
- Organization
  - Name
  - Legal status
  - URL
  - ftp
- Contact person
  - Name
  - Position
  - Postal address
  - Telephone
  - Fax
  - E-mail

8.3.2 Lexicon Metadata

8.3.2.1 Macrostructure

Lexicon Type*
- Language(s)*
- Number of languages
- Language(s)
- Size*
- Size per language
- Size per level of representation
- Coverage*
- Type
- Domain(s)
- Lexical unit / Headword
  - Type
- Encoding format*
  - Type
  - Character set
- Level(s) of representation*

Type
Development
Sources
Mode

8.3.2.2 Microstructure

Orthography
Etymology
Frequency
Phonology
Morphology
Morphosyntax
Syntax
Semantics
Definition
Comment
Usage
Other information

8.3.3 Metadata set for Text Corpora

Language(s)*
Number of languages
Language(s)
Size*
Size per language
Size per level of annotation
Coverage*
Type
Domain(s)
Text production dates
Encoding format*
Type
Character set
Annotation*
Structural annotation
Structural annotation
Level of annotation
Schema / Tagset
Linguistic annotation
Linguistic annotation
Level of annotation
Schema / Tagset
Development
Sources
Mode

Document data (refers to digitized data / file)
Document data encoded
Title
Size
Topic
Genre
Medium
Creator
Creation date
Other
Text data encoded
Text author
Publication date
Publisher
Translated text
Other

8.3.4 Metadata set for Speech Resources

Language(s)*
Number of languages
Language(s)
Size*
Size per language
Size per level of annotation
Coverage*
Region(s) where captured
Production dates
Encoding format
Type
Character set
Waveform
Segmentation*
Segmentation
Methodology
Annotation*
Transcription
Transcription
Tagset
Other Annotation
Type of annotation
Annotation
Schema / Tagset
Development

Medium
Type
Content

Speech file data
Title
Size
Medium
Waveform
Sample rate (Hz)
Sampling format
Number of samples
Topic

Language(s)
Type
Content
Date of recording
Region of recording
Speaker data
Number of speakers
Sex (per speaker)
Age (per speaker)
Origin (per speaker)
Education (per speaker)
Profession (per speaker)

8.3.5 Metadata set for Multimodal Resources

Size
Language(s)
Coverage
Region(s) where captured
Production dates
Format
Medium
Compression format
Type
Annotation
Type of annotation
Annotation
Schema / Tagset
Multimodal file data
Title
Size
Medium
Compression format
Creation location
Creation date
Device
Instrument type
Device
Instrument settings
Type
Topic
Language
Participants' data
Number of Participants
Sex (per Participant)
Age (per Participant)
Origin (per Participant)
Education (per Participant)
Profession (per Participant)

8.3.6 Metadata set for Tools

Technical requirements*
System requirements
Operating system(s)
Other s/w applications required
Technical description
Implementation language
Execution environment
Input format
Output format
Language dependency*
Language dependency
Language(s)

8.4 IMDI Schemas

The various IMDI schemas for annotated corpora, published corpora, corpus nodes and for the various profiles can be found at the IMDI web site, via http://www.mpi.nl/imdi/schemas/schemas.html

9 Bibliography

Projects and abbreviations

Reference	Abbreviation of	Link
[APA]	Alliance for Permanent Access	http://www.alliancepermanentaccess.eu
[ARK]	Archival Resource Key	http://www.cdlib.org/inside/diglib/ark/
[CGN]	Corpus Gesproken Nederlands	http://lands.let.kun.nl/cgn/
[DAM-LR]	Distributed Access Management for Language Resources	http://www.dam-lr.eu/
[DC]	Dublin Core	http://dublincore.org/
[DCAM]		http://dublincore.org/documents/abstract-model/
[DC-DS-XML]		http://dublincore.org/documents/dc-ds-xml/
[DC-TEXT]		http://dublincore.org/documents/dc-text/
[DEISA]	Distributed European Infrastructure for Supercomputing Applications	http://www.deisa.eu/
[DFKI]		http://www.language-archives.org/archive/dfki.de
[DOBES]	Dokumentation Bedrohter Sprachen	http://www.mpi.nl/dobes
[DOI]	Digital Object Identifier	http://www.doi.org/
[EAD]	Encoded Archival Description,	http://en.wikipedia.org/w/index.php?title=Encoded_Archival_Description&oldid=250469911
[ebXML]	e-business XML	http://en.wikipedia.org/wiki/Ebxml
[EGEE]	Enabling Grids for E-sciencE	http://www.eu-egee.org/
[e-IRG]	e-Infrastructure Reflection Group	http://www.e-irg.eu/
[ELDA UC]	Universal Catalogue	http://universal.elra.info/
[ENABLER]		http://www.ilsp.gr/enabler/
[ESF]	European Science Foundation Second Learner Study	http://books.google.de/books?id=g292tXMX4tgC&pg=PA1&lpg=PA1&dq=esf+Second+learner+perdue&source=bl&ots=WKi3GUQQP6&sig=n7QSWy3StXvD06nMfAzY7GBbm9w&hl=de&sa=X&oi=book_result&resnum=3&ct=result#PPP1,M1
[FIDAS]	Fieldwork Data Sustainability Project	http://www.apsr.edu.au/fidas/fidas_report.pdf
[HS]	Handle System	http://www.handle.net/
[ICONCLASS]		http://en.wikipedia.org/wiki/Iconclass
[INTERA]	Integrated European language	http://www.mpi.nl/intera/

	data Repository Area	
[ISOcat]		http://www.isocat.org
[LMF]	Lexical Markup Framework	http://en.wikipedia.org/w/index.php?title=Lexical_Markup_Framework&oldid=255448197
[METATAG]		http://en.wikipedia.org/w/index.php?title=Meta_element&oldid=256779491
[METS]	Metadata Encoding and Transmission Standard	http://en.wikipedia.org/wiki/METS
[MILE]		http://www.mileproject.eu/
[MPEG7]		http://en.wikipedia.org/w/index.php?title=MPEG-7&oldid=241494600
[NLSR]		http://registry.dfki.de/
[OAIS]	Open Archival Information System	http://en.wikipedia.org/wiki/Open_Archival_Information_System
[OASIS]	Organization for the Advancement of Structured Information Standards	http://www.oasis-open.org/
[ODD]	One Document Does all	http://www.tei-c.org/wiki/index.php/ODD
[OLAC]	Open Language Archives Community	http://www.language-archives.org/
[PMH]	Protocol for Metadata Harvesting	http://www.openarchives.org/pmh/
[SCHEMAS]		http://www.schemas-forum.org/
[SRU]	Search/Retrieve via URL	http://www.loc.gov/standards/sru/
[SRW]	Search/Retrieve Web Service	http://en.wikipedia.org/wiki/Search/Retrieve_Web_Service
[TEI]	Text Encoding Initiative	http://www.tei-c.org/
[UDDI]	Universal Description Discovery and Integration	http://en.wikipedia.org/wiki/UDDI
[WSDL]	Web Services Description Language	http://www.w3.org/TR/wsdl20
[Z39.50]		http://en.wikipedia.org/wiki/Z39.50

Literature

[Baker 1998] Baker, T. (1998). Languages for Dublin Core. D-Lib Magazine, 4:12.

[Baker 2008] Guidelines for Dublin Core Application Profiles (Working Draft), http://dublincore.org/documents/2008/11/03/profile-guidelines/

[Hochstenbach 2003] Hochstenbach, P., Jerez, H., and Van de Sompel, H. (2003). The OAI-PMH static repository and static repository gateway. In Digital Libraries, 2003. Proceedings. 2003 Joint Conference on, pages 210-217.

[Brown 2004] Brown, A. and Haas, H. (2004). Web Services Glossary. *W3C working group note*. http://www.w3.org/TR/ws-gloss/

[Beagrie 2008] Beagrie, N., Chruszcz, J., and Lavoie, B. (2008). Keeping research data safe. Technical report. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

[CiTER] Citation of Electronic Resources, ISO Draft (2008)

[IMDI] Broeder, D. and Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119-132.

[HUC] http://www.alliancepermanentaccess.eu/documenten%5Chuc.pdf

[Klassmann 2006] Klassmann, A., Offenga, F., Broeder, D., Skiba, R., and Wittenburg, P. (2006). Comparison of resource discovery methods. LREC.

[RFC 3986] Berners-Lee, T, et al., "Uniform Resource Identifier (URI): Generic Syntax", IETF RFC 3986, January 2005, http://tools.ietf.org/rfc/rfc3986.txt

[Tansley 2006] Tansley, R. (2006). Building a Distributed, Standards-based Repository Federation. D-Lib Magazine, 12(7/8):1082-9873.

[Understanding Metadata] Guenther, R. and Radebaugh, J. (2004). Understanding Metadata. National Information Standard Organization (NISO) Press, Bethesda, USA.

[WP:catalog] Catalog. (2008, August 3). In Wikipedia, The Free Encyclopedia. Retrieved 09:05, August 22, 2008, from http://en.wikipedia.org/w/index.php?title=Catalog&oldid=229575451

[WP:crosswalk] Crosswalk (metadata). (2008, August 16). In Wikipedia, The Free Encyclopedia. Retrieved 13:46, September 25, 2008, from http://en.wikipedia.org/w/index.php?title=Crosswalk_(metadata)&oldid=232244315