Metadata Overview and the Semantic Web

P. Wittenburg, Daan Broeder

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

Abstract

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

1. Introduction

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

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

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

2. Concept of Metadata

2.1. Early Work

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

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

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

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

It should not be forgotten that some companies storing language resources for various language engineering purposes such as training statistical algorithms or building up translation memories are using specifically designed databases for discovery and management purposes. These databases normally allow a shared access so that each employee can easily identify whether useful resources are available. For example Lernard&Hauspie used such a database internally[1]. The large data centers such as LDC [10] and ELRA [11] have developed an online catalogue suitable to their needs that allows easy discovery of the resources they are housing. Other resource centers such as the Helsinki University resource server [12] use an open common web-site approach where they describe their holdings without using a formal framework such as metadata.

2.2. Classification Aspects

The creation of a metadata description for a resource is a classification process. The metadata elements define the

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

dimensions and the values they can take define the axes along which classifications can be done. However, metadata classification of language resources is a classification in a space where the dimensions are not orthogonal, i.e. they are not independent from each other. A choice for a value in one dimension may have consequences for the choices in others. Certain properties can appear along several different dimensions. Further, we cannot always define metrics along the axes.

Therefore, a classification has to be based on a comparison with predefined vocabularies. Figure 1 shows how such classification can be done. The user may assume that the location indicated by the cross would best describe his resource. Since there is no perfect match with values along the two dimensions indicated by black and white dots, he may decide to choose the dots indicated with rectangles as the best matching ones.

Of course, this raises many problematic questions especially in communities such as the linguistic one. There does not exist yet a widely agreed ontology for language resources. Linguistic theories lead to different types of categorization systems. So who can decide about the usage of such encoding schemes and since it can be expected that sub-communities do not agree about one single scheme, the question is: how can interoperability be achieved, i.e. how can different categorizations be mapped onto each other? These questions are not simple to solve.

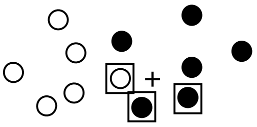
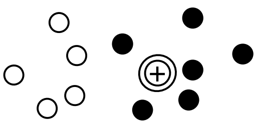


Figure 1 shows two categories represented by black and light dots. Each dot denotes a possible value of the respective category in some non-Euclidian space. The cross may indicate the “location” of the resource and the rectangles as the optimal choice for describing that resource.

A solution chosen by the IMDI initiative is to allow for flexibility, i.e. allow the addition of elements (dimensions of description/categorization) and to make the corresponding vocabularies user extendable where there is no set established yet. At first glance this solution appears acceptable but it is somewhat dangerous as can be inferred from classification literature [13]. We would like to indicate one of the possible problems with an example (fig 2). Individual users could decide to add a value to a dimension that does not seem to be characteristic for the point in space and thereby breaks the semantic homogeneity distorting the dimensions and creating problems for proper discovery.



In figure 2 an additional value is created (double circle) for one of the two categories (light circles) in an area where another dimension (black circles) is dominant. This leads to a distortion of the semantic homogeneity.

Also users could just add particular values to a vocabulary to suit their direct needs. Such a process would lead to an over-specification. The result would be a long list of specific and non-generalized terms and again problems with resource discovery are predictable.

On the other hand completely prescribing a vocabulary for a dimension not yet fully understood would mean that important areas might not be represented so that people will not make use of the categorization system at all. In the IMDI initiative a middle position was taken. A pre-defined vocabulary is proposed and at regular instances the actually used vocabulary will be evaluated to detect omissions in the proposed vocabulary. Dependent on the outcome the pre-defined vocabulary will be extended. It can of course also occur that existing values will be removed, since they are not used and are seen as obsolete by the community. One question remains: who is responsible for making decisions on such matters? This is a social and organizational issue to be solved by the whole community.

3. Reasons for Metadata

3.1. General Aspects

A re-vitalization of the metadata concept occurred with the appearance of the Web. A few figures may illustrate the problem we are all faced with. According to an analysis of IDC the amount of relevant data in companies exceeded 3.200 Petabyte in 2000 and will increase to 54.000 Petabyte in 2004[2]. The stored documents include information relevant for the success of the companies and form part of the company’s knowledge base. These documents are of various natures - partly the texts themselves explain what they are about and partly the documents need a classification to easily understand their relevance. Open questions are how to manage this knowledge base and how to make efficient use of it.

Well-known is the gigantic increase in the amount of resources available on the Web. Here, the focus is certainly on the aspect of efficient methods to find useful resources. It is often argued that the search engines that are based on information retrieval techniques have lost the game at least for the professional user who is not looking for adventures. The typical search engines use the occurrence and co-occurrence of words in the titles or in the texts of web documents to find what are thought to be the most suitable resources and calculate a suitability rating. Automatic clustering techniques also based on statistical algorithms are used to group information and also automatic categorization is carried out to help the user in his discovery task. Still the precision (the number of correct results compared to the number false results) and the recall (the number of hits found compared to the total number of suitable documents) are not satisfying especially if the user is looking for a specific type of information. Narrowing down the semantic scope of the queries to discover interesting documents often is a very time-consuming and tedious enterprise. Therefore, IR-based search engines will not be the only choice for professional users.

[2] It is not the amount of data that counts, but the number and variety of resources that increases in parallel.

The PICS initiative [14] showed that even for general web-based information there is a need for additional type of descriptors that cannot be reliably extracted from the texts. So, metadata descriptions, i.e. characterizations of the resources with the help of a limited set of descriptive elements, were seen as a useful addition to the texts themselves. In this paper we will not deal with the aspects of how to come to valuable descriptor sets for arbitrary content, but focus on the language resource domain.

3.2. Language Resource Domain

All the content based information retrieval (IR) techniques are based on the assumption that the texts themselves, in particular the words used and their collocations, describe the topic the text is about in sufficient detail. In the domain of language resources there are a number of data types where we can assume that this may be true. Grammar descriptions or field notes in general include broad prose descriptions about the intentions and the content in addition to special explanations of linguistic or ethnographic details. IR techniques may lead to successful discovery results. Still, would professionals who are looking for “field notes about trips in Australia that lead to a lexicon about the Yaminyung language” want to rely on such statistical engines? They would prefer to operate in a structured space obviously organized by resource type, location and languages to discover the resources they are looking for. It is almost impossible to automatically derive metadata descriptions from the content of language resources such as corpora and lexica.

Also in the language resource domain we are faced with a gigantic increase in the amount of resources. An impression about this explosion of resources can be given by the example of the multimedia/multimodal corpus at the Max-Planck-Institute for Psycholinguistics where every year around 40 researchers carry out field trips, do extensive recording of communicative acts and later annotate the digitized audio and video material on many interrelated tiers. The institute now has almost 10000 sessions - the basic linguistic unit of analysis - in an online database and we foresee a continuous increase. One researcher at the institute has about 350 GB of video recordings (about 350 hours) online that are transcribed by several people in parallel. Thus the individual researchers as well as the institute as a whole are faced with a serious resource management and discovery problem.

The increase of the amount of resources was paralleled by an increase in the variety and complexity of formats and description methods. This was caused by moving from purely textual to multimedia resources with multimodal annotations. It was understood early that the traditional methods of management and discovery mostly on purely individual account led increasingly often to problems. Scientists could no longer easily find relevant data and problems arose when a researcher left the institute. Similar situations occur in other research centers, universities and also in industry.

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

It was clear that most of these descriptions had to be created manually since only in a few cases it may be possible to automatically extract them from directory path names, Excel sheets or other sorts of systematic descriptions. As mentioned before the great majority of the language resources are of a sort where the descriptors cannot be anticipated from the content.

3.3. New Metadata Aspects

The trend of a continuously growing number of language resources will continue. Another apparent trend is that researchers are increasingly often willing to share them online via the Internet or at least to share knowledge about their existence with others from the community. Metadata descriptions, as previously explained, have a great potential to help researchers to manage these resources and simplify their discovery.

While the designers of the aforementioned TEI focused on text documents, current collected language resources mostly have multimedia extensions (sound and/or video). This adds new requirements on what descriptor set to use. Furthermore, it is generally agreed that the purpose of a metadata set is not so much to create a very complete description of a resource, but to support easy resource discovery and resource management. This way of looking at metadata certainly fits with the important work in the Dublin Core initiative (DC).

At the moment no-one can say with absolute authority which type of descriptor set is necessary to facilitate discovery and management, since for the domain of language resources the metadata concept (with respect to the above purposes) is very new and has hardly been applied by a greater number of linguists. We are confronted with different type of users all having different requirements that we do not know in detail. There are

- the researchers and developers who are experts and want to quickly find exactly those resources which fit to their research or development tasks[3];
- the resource manager who wants to check whether he/she wants to define a new layer of abstraction in the corpus hierarchy to facilitate browsing[4];
- the teacher who is teaching a class about syntax and wants to know whether there are resources with syntactic annotations commented in a language he/she can understand;
- the journalist who is interested in getting a quick overview about resources with video recordings about wedding ceremonies;
- the casual web-user who is interested to see whether there is material about a certain tribe he just heard about;

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

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

- many other types of users could be mentioned here whose requirements we often do not yet know.

An important point is that many of the language resource archives currently set up have a long-term perspective. So the question of their typical usage becomes an even more problematic one, since we cannot anticipate what future generations will need to discover resources. A widely used statement in such situation of uncertainty is to make the descriptor set exhaustive. But the fact is that very exhaustive sets are problematic because they are labor intensive and the inherent danger of over-specification. The IMDI team expects that a more dynamic scenario will occur where descriptor elements and even element values are seen as abstract labels which can be refined when more detail is needed. Sub-structures can also be needed to make properties more specific.

Given these uncertainties about future user needs, it makes sense to start now with a non-exhaustive element set. Also, language resource creators are reluctant to invest time in information that will primarily help others. Too much labor required will lead to a negative attitude.

Another phenomenon is that individual researchers have to participate in person in the creation and integration of metadata descriptions. There is no time to read lengthy documents about the usage of elements. Therefore everything has to be simple and straightforward, otherwise he/she will not participate. Metadata descriptions also should facilitate international collaboration. In many disciplines international collaboration with researchers located at different places is normal. Contributions from one of them must be directly visible by the others. This requires a metadata description framework that allows for regular update of the descriptions.

3.4. Resource Management Aspects

The primary task of metadata is resource discovery. However, resource management is an equally important aspect for the resource creator and manager. Metadata can help in managing resources. Linguistic data centers or companies storing language resources are used to manage large amounts of resources. Beyond discovery, management includes operations such as grouping related resources, copying valuable resources together with their context, handling different versions of resources, distributing and removing resources and maintain access lists and design copying strategies. Until a few years ago resource management was done by individual researchers using physical structuring schemes such as directory structures. This was also made possible by the relatively small size of the resources.

However, for the modern multimedia based archives of institutions and individual researchers files and corpora are becoming so huge that the physical manipulation of these resources becomes more and more a domain of the system manager. The conceptual domain defined by metadata can become the operational layer for the corpus manager. Grouping is no longer done on a physical layer that often implies copying large media files, but on the level of metadata. This means the definition of useful metadata hierarchies and to set the pointers to the resources wherever the system management may have stored them.

Resource management has acquired another dimension with the distributed nature of resources in the Internet scenario. It will become a normal scenario in the future that a video file is hosted on a certain server while two collaborators work simultaneously on that same media file. Using the Dutch scientific network this kind of collaboration is already possible. One, for example, may be annotating gestures and the other annotating semantics where speech and gesture information is needed. Annotations are generated on different tiers and are visible to both collaborators, but the place of storage could be arbitrary especially as long as the annotations have a preliminary character. The metadata description can be used to point to the location and to allow management operations as if the resources were all bundled on a single server.

4. Language Resource Data Types

Before introducing the different metadata initiatives that deal with language resources it is necessary to analyze the characteristics of the objects that have to be described. As already indicated not all objects that we find in the language resource domain are well understood. The most important ones are

- complex structured text collections
- multimedia corpora
- lexica in their different realizations
- notes and documents of various sort

The nature of text collections is very well described by the TEI initiative. The particular aspects of textual corpora were then analyzed and described by CES. Multimedia resources (MMLR) that either include multimedia material or are based on media recordings add new requirements. MMLR can combine several resources which are tightly linked such as several tracks of video, several tracks of audio, eye tracking signals, data glove signals, laryngograph signals, several different tiers with annotations, cross-references of various sorts, comments, links to lexical entries and many others. In many MMLR it is relevant to describe that a certain annotation tier has special links with a certain media track. For speech engineers it could be relevant to know the exact relation between a specific transcription or transliteration to one specific audio track (close range microphone). On a certain level of abstraction the different sub-resources have to be seen as one or relating to one “virtual” meta resource. Metadata has to describe this macro-level complexity and has to inform the user about the type of information contained in such a bundled resource.

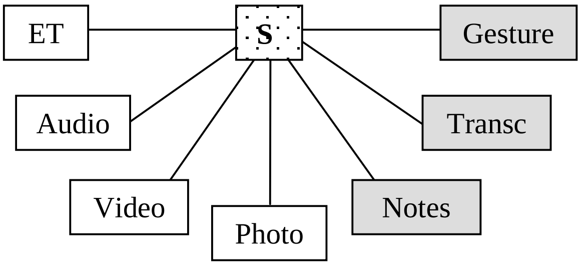


Figure 3 shows the various types of information tightly related by a common time axis.

Lexica where concepts and words are in the center of the encoding can appear in various forms such as dictionaries, wordlists, thesauri, ontologies, concordances and many others. Until now they are mostly monolithic resources with a complicated internal structure bearing the linguistic information. Metadata that wants to describe such a resource to allow useful retrieval has to indicate which type of information is available and in what format.

Linguistic notes can be of various sorts as well such as field notes, sketch grammars and sound system descriptions. Normally they appear as prose texts with no special structural properties that can be indicated by metadata. They can be treated as normal documents except that their functional type has to be indicated.

5. Metadata Goals and Concepts

In this chapter we want to briefly review the goals and concepts of the metadata initiatives that follow more or less the new paradigm described above and which are relevant for the language resource domain.

5.1. Dublin Core Metadata Initiative

The Dublin Core metadata initiative has as primary goal to define the semantics of a small set of descriptors (core set) which should allow us to discover all types of web-resources independent whether they are about steam engines or languages spoken on the Australian continent. All the experience of librarians and archivists was invested in the definition of the core set. One explicit goal was to create a significantly lighter set than defined for example within the librarians MARC standard [15]. The discussions that started seriously around 1995 ended up in the definition of 15 elements as listed in the following table.

Title	name given to the resource
Creator	entity primarily responsible for making the content of the resource
Subject	topic of the content of the resource
Description	account of the content of the resource
Publisher	entity responsible for making the resource available
Contributor	entity responsible for making a contribution to the content of the resource
Date	date associated with an event in the life-cycle of the resource
Type	nature or genre of the content of the resource
Format	physical or digital manifestation of the resource
Identifier	unambiguous reference to the resource within a given context
Source	reference to a resource from which the present resource is derived
Language	language of the intellectual content of the resource
Relation	reference to a related resource
Coverage	extent or scope of the content of the resource
Rights	information about the rights held in or over the resource

DC wanted to define a foundation for a broadly interoperable semantic network based upon a basic element set that can be widely used. This broad scope was achieved by often vague definitions of several of the DC elements. This is its strength and at the same times its weakness.

The designers well understood the limitations and problems of this approach. The Dublin Core initiative anticipated the need for other element sets and the Warwick Framework [16] was described as a way to accommodate parallel modular sets of metadata using domain specific element sets. Many initiatives work along the DC suggestions by modifying the element set in a number of dimensions, others started from scratch, however, accepting the underlying principle of simplicity. The modifications of the DC core set are done in 3 dimensions partially sanctioned by the DC initiative: (1) Qualifiers are used to refine the broad semantic scope of the DC elements. The underlying request is that qualification may not extend the semantic scope of an element. (2) Constraints may be defined to limit the possible values of an element (Example: date specification according to the W3C recommendations). (3) The usage of new elements, which of course challenges DC compatibility.

The DC initiative itself defined qualifiers and constraints for a number of elements [17]. They also foresaw a problem with uncontrolled qualification: “The greater degree of non-standard qualification, the greater the potential loss of interoperability”. For long time it seemed that at least two views were disputing about the way to go forward. The ones that are in favor of a controlled extension would control the semantic scope, and thus force communities with their own semantic needs away from adopting the DCMES. In the other view there should be loose control on the semantics of the elements, so that other communities could join easily. In the latter case DCMES would become a container for all sorts of information where querying could lead to unsatisfying results.

DCMI did not formulate any syntactic specifications. The DC Usage Group described how DC definitions could be expressed within HTML. The Architecture Working Group within DC made more extensive statements about syntactic possibilities and the inclusion of various extensions [18]. They discuss the following extensions that are common in the community applying DC:

- the usage of a scheme qualifier to put constraints on element values;
- the usage of qualifiers to narrow down the broad semantic scope of the elements such as DC:Creator.Illustrator;
- the subdivision of elements such as DC:Creator.PersonalName.Surname;
- the usage of class type relationships identifying that for example persons not only appear as values of the element creator but also belong to the class person.

There are reports about much confusion in the DC community through the usage of these uncontrolled extensions. In a proposed recommendation from April 2002 of how to implement DC with XML [19] the notion of “dcterms” is introduced which are “other elements recommended by DCMI”. The proposed recommendation states that “refinements of elements are elements in their own” and give concrete examples:

use of
<dcterms:available> 2002 </dcterms:available>

instead of
<dc:date refinement=”available”>2002 </dc:date>
or
<dc:date type=”available”> 2002 </dc:date>

These examples show that according to the recommendation refinements should be treated the same as other properties. There is no official statement yet whether this view is accepted by DCMI.

Very recently the Architecture Working Group produced another very interesting proposed recommendation about the implementation of DC with RDF[5]/XML [20]. It is argued that the situation with the simple unqualified DC is very unsatisfactory in various respects. In particular, there is no way to provide structure supporting the discovery process. It is suggested to implement a refinement of an element by applying the “subPropertyOf” relation defined within RDF Schema. A qualifier such as “dcterms:abstract” refines “dc:description” by means of the “subPropertyOf” feature. Also in this paper a replacement of the “subelement” construct (dot notation in the HTML implementation) by the “refinement” attribute is proposed.

With respect to language resources DC itself does not provide any special support. To describe the complex structure of MMLR DC offers the relation concept. However, the qualifiers offered do not represent the tight resource bundling very well. Since DC itself does not offer structure, dependencies as indicated in 4 cannot be represented. Also for describing lexica in more detail it does not have the necessary elements.

There is no doubt that DC is currently the most important standard for the simple description of electronically available information sources. It seems to be also clear that DC will be the standard for the casual user to look for easy discovery of simply structured resoruces. DC may form the widely agreed set. The evolution of the DC metadata set and extensions are depicted in the following graph, which is taken from Lagoze [21] and shows the “pidginization versus creolization trends” analogy from Baker.

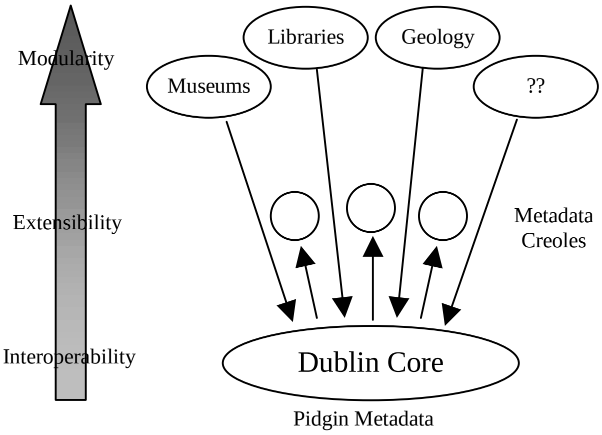


Figure 4 shows the principal problem with which DC had to cope. Interoperability leads to a pidginized form of metadata that is simple enough for the casual web user.

The need for Domain specificity then leads to different specialisations of the DC set, the creoles. Dependent on the amount of extensions needed one may end up with a new metadata set.

5.2. OLAC Metadata Initiative

The OLAC metadata initiative wanted to start from the DC set and be compliant with it as far as possible, but overcome its major limitations. Therefore DC was extended in four dimensions:

- 3 attributes were defined to support OLAC specific qualifications (***refine*** to refine element semantics including controlled vocabularies; ***scheme*** to refer to an externally controlled vocabulary; ***lang*** to specify the language a description is in).
- ***Code*** attributes refer to element specific encoding schemes.
- 8 new sub-elements were created which narrow down the semantics, but need a separate controlled vocabulary (Format.cpu, Format.encoding, Format.markup, Format.os, Format.sourcecode, Subject.language, Type.functionality, Type.linguistics).
- A special ***langs*** attribute as a list of languages which appear in a metadata description.

For various refined elements and sub-elements[6] controlled vocabularies are under preparation and their definition is part of the schema defining the metadata set [22].

The ***refine*** attribute allows OLAC to associate language resource specific semantic descriptions for DC elements that are specified too broadly and imprecisely. It is the association of a controlled vocabulary (CV) that narrows down the semantic scope even more precisely as was described in 2. OLAC wants to keep control of the CV, i.e. there is no user definable area, but there is a description of a development ***process*** that defines how definitions can be successively adapted [23].

The ***code*** attribute acts as a scheme specifier to assure that for example dates are stored in the same way (yyyy-mm-dd).

The OLAC metadata set was constructed such that it can describe all linguistic data types without creating type specific elements and software used in the area of Natural Language Processing. Also advice about and the usage of NLP software is seen as a relevant type of linguistic information.

OLAC has created a search environment that is based on the simple harvesting protocol of the Open Archives Initiative (OAI) [24] and on the standard DC set. Since OAI accepts the DC default set the OLAC designers take care to discuss how the special OLAC information is dumbed down to service providers.

OLAC’s intention is to act as a domain specific umbrella for the retrieval of all resources stored in Open Language Archives. Its intent is to establish broad coalitions such that the OLAC metadata standard, i.e. the

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

[6] The distinction between qualifiers and sub-elements is not fully clear, especially when looking at the discussions within DC.

specifically extended DC set, is accepted as a standard by the whole domain.

5.3. IMDI Metadata Initiative

IMDI started its work without any bias towards any existing metadata vocabulary and wanted to first analyze how typical metadata was used in the field. A broad analysis about header information as used in various projects and existing metadata initiatives at that moment in time was the basis of the first IMDI proposal [25].

Decisive for the design of a metadata set is the question about the granularity of the user queries to be supported. From many discussions with members of the discipline, from the existing header specifications and from the 2 years of experience with a first prototypical test version, it was clear that field linguists for example wanted to input queries such as “give me all resources where Yaminyung[7] is spoken by 6 year old female speakers”. Language engineers working with multimodal corpora expressed their wish to retrieve resources where “subjects were asked to give route descriptions, where speech and gestures were recorded and which allow a comparison between the Italian and Swedish way of behavior”. Therefore, professional users requested much more detail than DC can offer. Furthermore the semantics of some of the DC element names did not agree with the intuition of many in the user community (e.g. Creator & Contributor). A presentation of the requirements and the needed elements in the European DC Usage Committee revealed that it did not seem advisable to use DC as a basis.

Due to the necessary detail IMDI needed modular sets with specializations for different linguistic data types. The two most prominent data types are (multimedia/multimodal) corpora and lexica. Other linguistic data types are much less common and not so well understood. Consequently two metadata sets were designed which differ in the way content and structure is described. In contrast to DC which only deals with semantics, IMDI also introduced structure and format. Structure makes it possible to associate for example a role, an age and spoken languages with every participant.

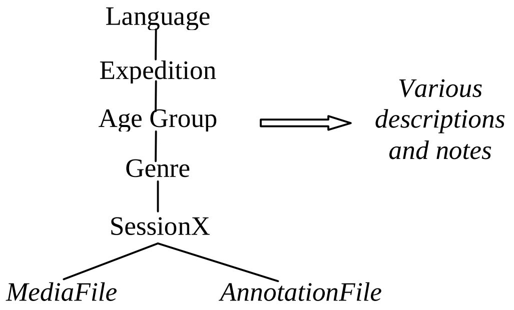


Figure 5 shows a typical metadata hierarchy with nodes representing abstraction layers. Each layer can contain references to various descriptions and notes and thereby integrating them into the corpus. All components of such a hierarchy can reside on different servers. The session nodes are the leafs in the hierarchy, since they point to the recordings and annotations.

The corpus metadata descriptions come in three flavors: (1) The metadata set for sessions is the major type, since it describes the bundle of resources which tightly belong together as described in 4. (2) Since IMDI not only created a metadata set, but also an operational environment, it allows to integrate resources into a browsable domain made up by abstraction nodes and the sessions as the leafs (see figure 5). The metadata descriptions used for the sessions and the higher nodes are basicaly the same. (3) For published corpora that appear as a whole the catalogue metadata set was designed. It contains some additional elements such as ISBN number that are typical for resources that are hosted for example by resource agencies.

The IMDI metadata set for sessions tries to describe sessions in a structured way with sufficient rich information using domain specific element names [26]. It covers elements for

- administrative aspects (Date, Tool, Version, ...)
- general resource aspects (Title, DataType, Collector, Project, Location, ...)
- content description (Language, Genre, Modality, Task, ...)
- participant descriptions (Role, Age, Languages, other biographic data, ...)
- resource descriptions where a distinction is made between media resources, annotation resources, source data (URL, Type, Format, Access, Size, ...)

The IMDI set was chosen so that most elements are suitable for automatic searching, but there are also those that are filled with prose text and are meant to support browsing. The exact recording conditions can be described, but the variability is so great that it does not make sense in general to search on them. IMDI also offers flexibility on the level of metadata elements in so far that users can define their own keys and associate values with them. This can be done on the top “Session” level as well as on several substructures such as Participant and Content. This feature can be of great use especially for projects that feel that their specific wishes are not completely addressed by the IMDI set. This feature was used for example when incorporating the Dutch Spoken Corpus project within IMDI since they wanted to add a few descriptors defined by TEI. Of course, the metadata environment has to support these features also for example when searching.

For many of the elements, controlled vocabularies (CV) are introduced. Some CV’s are closed such as those for continents, since the set of values is well defined. For others such as Genre, IMDI makes suggestions, but allows the user to add new values. The reason is that there is no agreement yet in the community about the exact definition of the term “genre” and how genre information can best be encoded.

For the metadata set and for the controlled vocabularies schema definitions are available at the IMDI web site. All IMDI tools apply them. In contrast to OLAC the definitions of CV are kept separate to allow for the necessary flexibility. According to the IMDI view there will be several different controlled vocabularies as is true for example for language names (ISO definitions and the

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

long Ethnographic list) which should be stored in open repositories such that they can easily be linked.

The recent proposal for lexicon metadata [27] covers elements for

- administrative aspects (Date, Tool, Version, ...)
- general resource aspects (Title, Collector, Project, LexiconType, ...)
- object languages (MultilingualityType, Language, ...)
- metalanguages (Language)
- lexical entry (Modality, Headword type, Orthography, Morphology, ...)
- lexicon unit (Format, AccessTool, Media, Schema, Character Encoding, Size, Access, ...)
- source

Since the microstructure can be very different for the many languages and since linguistic theories also differ, it was decided not to describe structural phenomena of lexica, but only to mention which kind of information is included in the lexicon along the main linguistic dimensions such as orthography, morphology, syntax and semantics. To allow maximum re-usability of the schemas and tools the overlap between lexicon and session metadata was as large as possible.

It was felt that data types such as field notes, sketch grammars and others are resources which are in general prose texts with added semi formal notations and should not be objects which have their own specific metadata set, but they should be integrated into the metadata hierarchies at appropriate places. However, users might want to search for grammar descriptions of Finno-Ugric languages. This problem has not yet been satisfactorily solved within IMDI.

IMDI has been creating a metadata environment consisting of the following components:

- a metadata editor
- a metadata browser
- a search engine
- efficiency tools

All tools have to support the last version of the IMDI definitions of the metadata element sets and the controlled vocabularies. Since the tools are described elsewhere in greater detail [28,29], only a few special features will be described here. The editor supports isolated and connected work, i.e. in case of the PC being connected to the network new definitions of the CV etc can be downloaded and cached. A fieldworker, however, could operate independently on the basis of the cached versions. The browser can operate on local or remote distributed hierarchies allowing each user to create his own resource domain, but easily hooking it up to a larger domain. The browser is also intended to allow for the creation of nodes to form browsable hierarchies, so that a user can easily create his own preferred view on a resource domain. It also allows the user to add configuration information so that local tools of his choice can be easily started from the browser once suitable resources are found.

To increase the possibilities of resource discovery the search component is made an integral part of the browser. The current version operates on one metadata repository only and searching in a distributed domain has to be finished yet. It will make use of a simple query protocol based on HTTP to search sites with IMDI records. The macro infrastructural aspects have to be solved yet, i.e. how to gather metadata information residing at different locations in an efficient way. It is thought that the OAI harvesting protocol is suitable. Efficiency tools are of greatest importance to simplify the creation and management of large metadata repositories. For example, it has to be possible to adapt certain values of a large set of metadata descriptions with one operation. The tools currently available for this type of operation have yet to be integrated in the existing browser and editor.

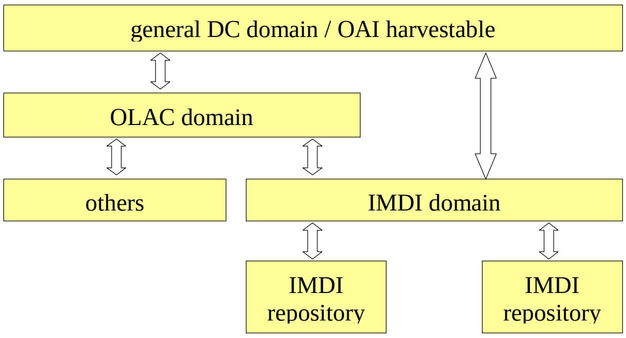


Figure 6 shows IMDI’s vision about metadata services users should be able to use. It is not indicated that the general DC domain covers many more domains than just the domain of language resources.

IMDI has accepted that there are different types of users. The casual web user wishing to use a simple perhaps widely known query language based on DC encodings and the professional user interested in easily finding the correct resources. Therefore, IMDI created a document describing the mapping between IMDI and OLAC [30]. Of course, such a mapping cannot be done without losing information and such documents need updates dependent on the dynamics of the two included standards. IMDI envisages the scenario as depicted in figure 6 and will comply with it.

The way IMDI repository connectivity is done is different from how OLAC connectivity is achieved. Since OLAC is focused on metadata harvesting for search support all OLAC metadata providers have to install a script providing the OAI protocol. In IMDI it is just the URL of a local top node that has to be added to an existing IMDI portal to become member of it.

5.4. MPEG7 Initiative

In contrast to the initiatives discussed earlier MPEG7 does not just focus on metadata as the term was defined in this paper. MPEG7 is an integral part of the MPEG initiative. While the other MPEG standards are about audio and video decoding, MPEG7 is a standard for describing multimedia content. It is based on the experiences with earlier standards such as SMPTE [31]. The future MPEG4 scenario includes the definition of media objects and the user controlled assembly of several objects and streams to compose the final display in a distributed environment. The role of MPEG7 in the decoding and assembly interface is to allow the user to search for segments of multimedia content, to support browsing in some browsable space and to support filtering of specific content.

It is meant to support real-time and non-real-time scenarios. Filtering will typically operate in a real-time scenario where media streams are received and parts are not processed any further. Search and browsing typically operate before media content is actually accessed. For the real-time tasks media annotations are used to identify segments that are not appropriate with the user profile.

Due to this wide range of intended applications for the future the MPEG7 description standard is exhaustive and the metadata is just a small part of it. MPEG7 has information categories about

- the creation and production process supporting an event model (i.e. aspects of workflow)
- the usage of the content (copyright, usage history, ...)
- storage features (format, encoding, ...)
- structural information about the composition of a media resource (temporal and spatial)
- low level features (color index, texture, ...)
- conceptual information of the captured content (visual objects, ...)
- collections of objects
- user interaction (user profiles, user history, ...)

MPEG7 has adopted XML Schema as its Descriptor Definition Language (DDL)[8]. It distinguishes between the definition of Descriptors where the syntax and semantics of elements are defined and Description Schemes that define the structural relations between the descriptors. Instead of defining one huge Description Scheme, it was decided to manage the complexity of the task by forming description classes (content, management, organization, navigation and access, user interaction) and let sub-groups define suitable DS. For the description of multimedia content there seem to exist already more than 100 different schemes. Complex internal structures are possible. Summary descriptions about a film for example can contain a hierarchy of summaries.

The MPEG7 community recognized the need to be able to map to Dublin Core to facilitate simple resource discovery of atomic web resources of different media types. DC is made for such type for simple resources. In the Harmony project [32] a mapping of suitable MPEG7 elements was worked out. Finally, it was decided to apply a very restrictive mapping to not extend the semantic scope of the DC elements.

Similar to IMDI but with a much wider scope the MPEG community is working on a sophisticated environment to allow the intended broad spectrum of operations inclusive management. To create for example all the low level features describing video content one is experimenting with smart cameras.

When dealing with multimedia resources MPEG7 could be an option for the language resource community. Currently, there is no special effort within the MPEG7 community to design special DS that are suited for linguistic purposes; however, the language resource community could decide to do so. No obvious limitations can be seen. It seems that MPEG7 has still some time to go to be widely applicable.

6. Mapping Metadata

As mentioned previously DC is widely accepted as a simple metadata set for the casual web-user to search for simply structured resources. To achieve interoperability on that level it is important to map between the metadata sets. We would like to use the mapping between OLAC[9] and IMDI to demonstrate a few aspects that have to be solved.

In the first example two elements are semantically similar. “dc:creator” contains at least two aspects: (1) It refers to the name of a person who created the content. (2) Creation in the sense of DC also has a Intellectual Property Rights aspect. Creators are persons who have rights about the resource. IMDI wanted to separate these two aspects to make clear that there is a responsible researcher on the one hand and participants during the recordings on the other hand, both can claim rights with respect to the resource. So, “imdi:collector” takes care of the wishes of the researchers involved. The mapping rule from IMDI to DC is very simple for this example: All collectors in IMDI descriptions are creators in DC descriptions. The mapping from DC to IMDI is not as clear, since consultants which have a formal right in the DC sense and may appear as creators should be listed under “imdi:participants”.

The second example implies structure. The IMDI set has a substructure for the concept “participant”. Participants are those persons that are participating in interviews or other typical recording sessions. Each participant has attributes such as name, age, sex, role and languages spoken. The IMDI substructure allows one to group these attributes and therefore support questions such as “all 4 years old females speaking Yaminyung”. In DC we just have the possibility to define a set, i.e. list all names, all ages etc. One cannot infer which person has a certain age. To solve this problem one has to embed DC in a structure definition or use an identifier of the person and use it in all tags. Also for this example the mapping from IMDI to OLAC is simple: At first instance just the names are passed over. In second instance one could add the content of (part of) the other attributes to a description field and add it to the OLAC tag. The question is whether search engines will be able to use the information. Search engines would interpret description fields as prose text and would not use the advantages typical for structured metadata. The mapping from OLAC to IMDI is simple, since only names are expected. OLAC descriptions would be passed over to IMDI descriptions.

The third example discusses the problems inherent to resource bundling as we are used to in language resources (see figure 3). A good mapping with DC is not possible in a simple way. In IMDI the resources belonging to one session all share a large amount of metadata information and are therefore bundled in one description (if the user decides to do so). In DC one would have to describe every atomic resource separately and use “dc:relation” to establish the links. This means that each of the atomic resources has to refer to all the others with for example the qualifier “dc:relation.isPartOf”. First, such reference structure is complex and not adequate and second, nobody will actually use it. Another possibility in DC is to define

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

[9] The mapping document was based on a previous OLAC version.

a “virtual root resource” which links to the descriptions of the atomic resources to create a simple hierarchy. For the IMDI to OLAC mapping a simple solution was chosen: all atomic resources get separate descriptions. The OLAC to IMDI mapping is also very simple: since there is no structural information every atomic resource becomes an atomic resource in IMDI. If there would be a relation specification it would be added to the list of references. Any other scheme would be too dangerous and prone to error.

Basically, we follow the advice of the Harmony project to be very restrictive with mappings, since the semantic homogeneity of the elements can easily be distorted and conversion could lead to errors.

7. Summarizing the Metadata State

Web-accessible metadata descriptions to facilitate the discovery of language resources are a comparatively new concept. Four initiatives (DC, OLAC, IMDI, MPEG7) worked out proposals that are of more or less relevance for the linguistic domain. They differ in a number of aspects, but there is also overlap as indicated in table 1.

The concept is so new that we cannot yet draw relevant conclusions. OLAC states that they have harvested about 18.000 metadata records from their partners. From IMDI it is known that more than 10.000 metadata descriptions were created and integrated into a browsable domain. These numbers alone, however, do not answer a number of important questions such as:

- Do we have a critical mass of new and relevant resources in our repositories such that users make use of the infrastructures for professional purposes? It is clear that we are being far away from such a situation.
- Which approach is the most suitable one (if there is any answer to this question at all)? We still require years to find out and have to address the question whether we have good criteria.
- What are the typical queries the different user groups are asking? We don’t know yet, we need a critical mass and interesting environments to be able to answer this question.
- At which level do we need to establish interoperability? Is interoperability on DC level a useful goal? The question of interoperability cannot be seen independent from the usage scenario. Different user groups will have different requirements. The DC pidgin will not satisfy professionals. But the casual web-user may not be interested in looking for resources containing speech from 4 year old speakers.
- Which kind of tools do we need to support the resource creators and managers? Some initiatives have just started working on these issues, but it is too early to make statements.
- Upon which elements and controlled vocabularies will the community agree widely? Again, we have just started, so any answer at this moment may turn out to be wrong.
- Are the creators and users convinced that metadata create an added value which is worth the additional effort? By most community members metadata is still seen as an additional effort which is not justified. Awareness is growing, however.

	DC	**OLAC**	**IMDI**	**MPEG7**
addressed community	world	linguists language engineers	linguists language engineers	film & media community
scope	all web resources	all language resources	focus on (MM) corpora and lexica	all film & media documents
approach	experience of librarians and archivists	compliance to DC	based on overview about earlier work	based on earlier standards
set size	small	small	more detail	exhaustive
user extensibility	no	no	yes	?
formal definitions for	element semantics	element semantics controlled vocabularies constraints	element semantics structural embedding controlled vocabularies constraints	basic descriptor definition language Description Schemes
interoperability	-	DC compliant	mapping to OLAC/DC	mapping to DC
operations	search	search	browse, search, management, immediate execution	browse, search, filtering
tools	-	search environment	editor, browser, search tool, efficiency tools	?
connectivity by	-	OAI harvesting protocol	simple URL registration, OAI harvesting protocol	?
domain specific use of element names	no	no	yes	yes

Table 1 gives a quick overview about the goals and major characteristics of the relevant metadata proposals.

We do not know the answers to many questions yet or can only make speculations. What we know is that the number of individuals and institutions who create interesting resources is growing fast and that we need an infrastructure to allow their discovery. We also know that individuals and institutions have a management problem to solve and that traditional methods are no longer suitable. So the step to introduce metadata descriptions seems an obvious one, but we do not yet fully understand the potential of web-based metadata.

Resource discovery cannot be the only goal. Resource exploitation and management are equally important. Most important for the users is the view to step away from all sorts of details involving hardware, operating systems and runtime environments. When they have found a resource in a conceptual domain that is their domain of thinking, then they want to start a program that will help them to carry out their job. This program start should be seamless and not as it is today where users have to be computer experts. This is the dream that is still true, but not yet achieved.

Carl Lagoze pointed out that every community has different views about real entities and that these multiple views should not be integrated to one complex description, but that modular packages should emerge [33]. According to him, DC has to be seen as one simple view on certain types of objects. Consequently, he and his colleagues foresaw a scenario with many different metadata approaches where the way interoperability is achieved is not yet solved. The emergence of the Resource Description Framework [34] and the elaborations about an ABC model for metadata interoperability [35] indicate the problems we will be faced with.

Given all the uncertainties with respect to a number of relevant questions we can expect that within the next decade completely new methods will be invented based on the experiences with the methods we start applying now. Given this situation it seems to be very important to test different approaches and in so doing explore the new metadata landscape. A close network of collaboration, interaction and evaluation seems to be necessary to discuss the experiences. Probably an organization as ISO might be a good forum to start a broad discussion about the directions the language resource community should take.

Those who propose metadata infrastructures and ask persons to contribute take a high amount of responsibility. Given that our assumption is true that we will have an ongoing dynamic development[10] the designers of the metadata sets have to be sure that they can and will transform the created descriptions to new standards that will emerge by not losing the valuable information that has been gathered so far.

8. Metadata and the Semantic Web

Some years ago Tim Berners-Lee introduced the term “Semantic Web” foreseeing that we are creating a web which can only be managed well when we apply intelligent software agents. Humans will not be able to process the gigantic amount of knowledge available. After receiving concrete tasks from users or after signaling the usefulness of own activities such agents could use the web available information about terms and their relations to find answers or to prepare such answers. Central to the idea of the Semantic Web are the ideas of seamless operation for the user and screening him from all the underlying matching and inferring processes.

Metadata as defined in this paper can play an enormous role in such a scenario, since in metadata sets the elements are more or less accurately defined and their structural relations will become increasingly often explicit as well when technologies such as RDF are used. Metadata is comparatively reliable data[11]. The current lingua franca “DC” will, if it is to be successful, be extended by structure proposals such as being worked out by the architecture group. Sets such as IMDI that include implicit structure from the beginning have to make their structure definitions explicit to make them available for use by smart agents.

Currently, especially created scripts do the mapping between metadata sets (such as IMDI to OLAC) to achieve interoperability on metadata level. These scripts contain all the reasoning implicitly which is necessary to do a useful mapping. We foresee, however, a completely different mapping scheme where the semantics behind it are explicitly formulated. To achieve this we need open repositories (referred to by XML name-spacing) that contain the definitions of elements and vocabularies and those that contain the description of relations presumed that we all could agree on the same syntax[12].

The Resource Description Framework (RDF) seems to be a promising candidate to realize some of the dreams. RDF was developed at the intersection of metadata and knowledge representation experts. From the view of knowledge management it is a decentralized scheme for representing knowledge. It is built on XML to create complex descriptions of resources. It offers a set of rules for creating semantic relations and RDF Schema can be used to define elements and vocabularies. The relations are defined with a very simple mechanism that can also be processed by machines.

In an RDF environment every resource has to have a unique identifier (URI). It can have properties and properties can have values. The simplest assertion is “the web-site http://www.mpi.nl has as author personX” (see also figure 7) where personX can be a literal or for example another web-site. The corresponding RDF code is described in example 1 in the appendix.

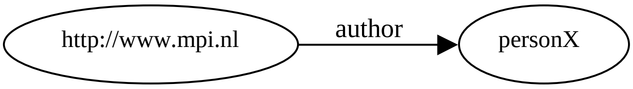


Figure 7 indicates the simple assertion mechanism of RDF where an object is characterized by the property “author” which takes the value “personX”.

[10] The IMDI-OLAC mapping document was created in August 2001 and has to be updated completely, since the included metadata sets have changed drastically within a year. It can happen that definitions will change again due to the uncertainty with respect to qualifiers in the DC discussion.

[11] There is the problem of how to create metadata descriptions for the huge amount of existing documents. It is clear that manual methods will not work. Automatic methods based on Information Retrieval and hopefully Information Extraction will not work on all types of resources as was explained and they would introduce unreliability.

[12] XML has got wide acceptance so that this assumption seems to be valid for the next decades.

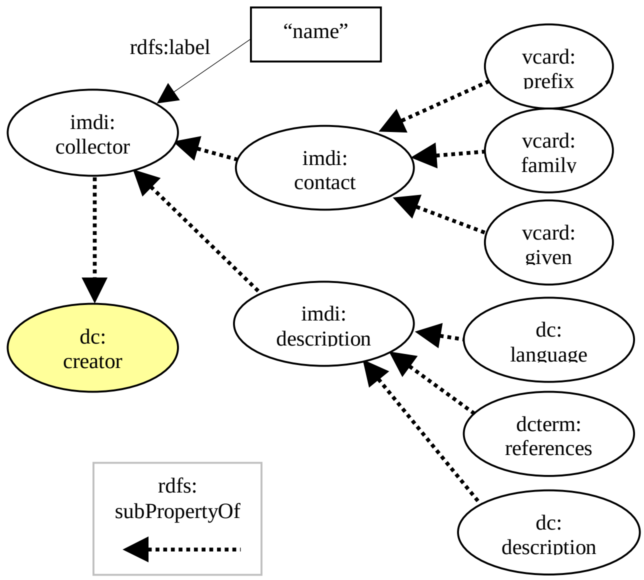


Figure 8 shows a metadata scenario where metadata sets re-use elements and relations which are defined in open repositories.

Using the RDF assertion formalism complex schemes can be realized. Example 2 in the appendix shows how Dublin Core compliant specifications could be embedded in RDF. Example 3 gives one example where Dublin Core and VCard elements are used to create one description. Example 4 shows how RDF could be used to describe the mapping between IMDI and OLAC. Especially the last two examples indicate the direction of development that we expect: A new metadata set to be defined by a (sub) community will make use of existing terminology defined in some open repositories (referred to by XML name-spacing) and write an RDF schema which puts the terms into structure/relation. This scenario is depicted in figure 8.

Smart agents that provide services can interpret these definitions. Another major assumption to make this scenario workable is that communities agree on at least a limited set of terms. When a new term is created it has to be put into an open repository and it's mapping to related terms have to be defined where feasible. This is a complicated social process and can best be guided by an organization such as ISO. Under the guidance of ISO TC37/SC4 it would make sense to create such a namespace for the language community.

Carefully designed metadata sets based on open repositories can be seen as representing parts of the ontology of the domain of language resources. It will include the commonalities as well as the differences between sub-communities. Therefore, the discussions about the metadata sets we have right now are very important contributions towards such an ontology. Shortcomings of RDF especially in its power to express semantic details have been identified and therefore initiatives such as DAML/OIL [36] suggest extensions of the framework.

Therefore, one can say that the current metadata initiatives are important steps towards the realization of the Semantic Web.

9. References

[1] LREC 2000 Workshop:
http://www.mpi.nl/ISLE/events/events_frame.html
[2] IMDI White Paper:
http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf
[3] OLAC white:
http://www.language-archives.org.docs.white-paper.html
[4] DCMES
http://dublincore.org/documents/dces
[5] MPEG7
http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
[6] childes
http://childes.psy.cmu.edu
[7] TEI: http://www.tei-c.org
[8] CES: http://www.cs.vassar.edu/CES
[9] CGN: http://lands.let.kun.nl/cgn/home.htm
[10] LDC: http://www.ldc.upenn.edu
[11] ELRA: http://www.icp.grenet.fr/ELRA/home.html
[12] Helsinki Linguistic Server: http://www.ling.helsinki.fi/uhlcs
[13] H. Niemann (1974) Methoden der Mustererkennung. Akademische Verlagsgesellschaft, Frankfurt
[14] PICS: http://www.w3.org/PICS
[15] MARC: http://www.loc.gov/marc
[16] Warwick:
http://www.dlib.org/dlib/july96/lagoze/07/lagoze.html
[17] DC qualifiers
http://dublincore.org/documents/2000/07/11/dmes-qualifiers
[18] DC Architecture Group
http://dublincore.org/groups/architecture
[19] DC with XML
http://www.ukoln.ac.uk/metadata/dcmi/dc-xml-guidelines
[20] DC with RDF:
http://dublincore.org/documents/2002/04/14/dcq-rdf-xml
[21] S. Weibel, C. Lagoze (1996) WWW-7 Tutorial Track
[22] OLAC MD Set:
http://www.language-archives.org/OLAC/olacms.html
[23] OLAC Process:
http://www.language-archives.org/OLAC/process.html
[24] OAI:
http://www.openarchives.org/OAI/openarchivesprotocol.htm
[25] IMDI Overview:
http://www.mpi.nl/ISLE/overview/overview_frame.html
[26] IMDI Session Set:
http://www.mpi.nl/ISLE/documents/docs_frame.html
[27] IMDI Lexicon Set:
http://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf
[28] tool paper
[29] MPI Tools: http://www.mpi.nl/tools
[30] IMDI-OLAC Mapping: http://www.mpi.nl/ISLE/documents/draft/
[31] SMPTE: http://www.smpte.org
[32] Harmony: http://metadata.net/harmony/video_appln_profile.html
[33] Carl Lagoze (2000) Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience. Invited Talk at the Archiefschool, The Hague, Netherlands, June 2000
[34] RDF: http://www.w3.org/RDF
[35] ABC: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze
[36] DAML/OIL: http://www.w3.org/TR/daml+oil-reference

10. Appendix

Example 1

The first example shows how the assertion included in figure 7 is described by using the RDF formalism and using the Dublin Core metadata element "Creator".

```
<?xml version=”1.0”?>
<rdf:RDF
            xmlns:rdf =”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
            xmlns:dc=”http://purl.org/dc/elements/1.1/”>
       <rdf:Description rdf:about=”http://www.mpi.nl/OurDocument.html”>
            <dc:creator> personX </dc:creator>
       </rdf:Description>
</rdf:RDF>
```

The first line simply indicates that XML version 1.0 is the syntax basis. The next tag indicates that we enter an RDF description. Line 2 and 3 refer to namespaces, so that machines know which elements were used. So here it is refered to the RDF syntax and the Dublin Core element set. The tag in line 5 states that an RDF-based description follows about some characteristics of the web-site ”http://www.mpi.nl”. The next line then states that we add a property “dc:creator” with the value “personX” to the description.

Example 2

In example 2 it is shown how a Dublin Core metadata description could be embedded in RDF. In doing so DC-based description could make use of the structure defining capabilities of RDF.

```
<?xml version=”1.0”?>
<rdf:RDF
            xmlns:rdf =”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
            xmlns:dc=”http://purl.org/dc/elements/1.1/”>
       <rdf:Description rdf:about=”http://www.mpi.nl/ISLE/whitepaper.html”>
            <dc:title> IMDI White Paper </dc:title>
            <dc:creator> Daan Broeder </dc:creator>
            <dc:creator> Peter Wittenburg </dc:creator>
            <dc:creator> Freddy Offenga </dc:creator>
            <dc:subject> Metadata Initiative; XML; Metadata Environment <dc:subject>
            <dc:lang> en </dc:lang>
            <dc:publisher> ISLE Metadata Initiative </dc:publisher>
            <dc:date> 2000-04-01 </dc:date>
            <dc:format> text/html </dc:format>
       </rdf:Description>
</rdf:RDF>
```

This description simply adds the normal attributes such as creator, subject as a list of keywords, the language it is written in, publisher, date and format to the document “IMDI White Paper” by using Dublin Core elements.

Example 3

The third example is taken from the DC-RDF proposed recommendation paper [20]. It shows how RDF allows the metadata designer to combine elements from various metadata sets.

```
<?xml version=”1.0”?>
<rdf:RDF
            xmlns:rdf =”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
            xmlns:dc=”http://purl.org/dc/elements/1.1/”
            xmlns:rdfs:=”http://www.w3.org/2000/01/rdf-schema”
            xmlns:vCard=”http://www.w3.org/2001/vcard-rdf/3.0”>
<rdf:Description>
<dc:creator>
       <rdf:Description rdf:about=”http://qqqfoo.com/staff/corky”>
            <rdfs:label> Corky Crystal </rdfs:label>
            <vCard:FN> Corky Crystal </vCard:FN>
            <vCard:N> rdf:parseType=”Resource”>
                        <vCard:Family> Crystal </vCard:Family>
                        <vCard:Given> Corky </vCard:Given>
                        <vCard:Other> Jacky </vCard:Other>
                        <vCard:Prefix> Dr. </vCard:Prefix>
            </vCard:N>
            <vCard:BDAY> 1980-01-01 </vCard:BDAY>
       </rdf:Description>
</dc:creator>
</rdf:Description>
</rdf:RDF>
```

This time there are 4 namespaces mentioned, since we also have to borrow terms from RDF Schema and the vCard initiative. The RDF description is now a complex “dc:creator” structure where at first it is mentioned where it is about. Then we associate an abstract label to the attribute by using the “rdfs:label” element. Then we use a whole set of terms borrowed from vCard to describe the creator in detail.

Example 4

The fourth example shows how a formal and machine-readable relation can be established between Dublin Core “creator” and the IMDI “collector”. If such descriptions are available in open repositories any engine providing some service could make use of it.

```
<?xml version=”1.0”?>
<rdf:RDF
          xmlns:rdf =”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
          xmlns:dc=”http://purl.org/dc/elements/1.1/”>
          xmlns:imdi=”http://www.mpi.nl/ISLE/session-elements/2.5/”>
      <rdf:Description rdf:about=”http://www.mpi.nl/ISLE/IMDI/3.0/imdi-schema”>
          <rdfs:subPropertyOf rdf:resource=”http://purl.org/dc/elements/1.1/creator”/>
      </rdf:Description>
</rdf:RDF>
```

The description part makes an assertion which adds the “rdfs:subPropertyOf” attribute to “imdi:collector”. According to this assertion “dc:creator” is the superclass, i.e. all IMDI-collectors are also DC-creators.