

Management of Language Resources using Metadata

P. Wittenburg, Daan Broeder

Max-Planck-Institute for Psycholinguistics
peter.wittenburg@mpi.nl

Abstract

Technology development allows many more researchers than before to create language resources especially with multimedia extensions. This creates a resource management problem that exceeds the boundaries of established resource centers. Metadata environments such as the one proposed by IMDI that offer a metadata set and also tools to operate on them have a strong potential to help the individual researcher to carry out his resource management tasks. In addition, it allows him to easily integrate his resources into a large distributed domain of resources. The work at the Max-Planck-Institute for Psycholinguistics to establish a large multimedia language corpus helped to understand the needs and requirements. Due to this experience the IMDI environment has reached a state of maturity, but still some important features have to be added.

1. Introduction

Researchers and developers in the area of language resources are faced with four very dominant trends in the recent years: (1) The number and complexity of language resources stored in digital archives is growing fast, (2) there is an increasing acceptance of the need to improve the availability of the resources, (3) the Internet now connects many archives storing such resources and this asks for interoperability and (4) for many language resources need to be stored in archives for a large period of time due to economical and ethical reasons.

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 already more than 7000 annotated sessions - the basic linguistic unit of analysis - and we foresee a continuous increase. It was usual that researchers managing their resources with individually designed Excel-Sheets eventually were not able to keep control of them and that the institute effectively lost all access to resources when a researcher left. Thus the individual researcher as well as the institute was both faced with a resource management problem. It is known that in other research centers, universities and also in industry similar situations occur.

The increase of the amount of resources was paralleled by an increase in the variety and complexity of formats and description methods. Moving from purely textual to multimedia resources with multimodal annotations caused this. Media can include not only several audio and video tracks, but also increasingly often other information such as for example from eye trackers, data gloves and brain image recorders.

In many areas resources were seen as the private capital of a researcher or a specific project that served only to investigate a limited number of research questions. Therefore, the need to make resources available for other research was not seen. However, researchers now understand the potential of modern technology to immediately access the raw material, which enables for example re-coding, or incremental annotation procedures that can be part of collaborations. These opportunities increase the individual researchers willingness to share his

resources and to invest time to create publicly available descriptions. We clearly recognize a trend towards making the resources themselves available via the Internet or at least indicating what resources exist by creating structured descriptions available on the Internet.

The usage of the Internet demands for interoperability on various levels. Therefore new technologies devoted to the special requirements of the Internet such as RDF (Resource Description Framework), XML and UNICODE are have been developed to improve the exchange and re-usage of data. The usage of open standards is even more important when repositories of language resources have to support long archive periods. The Internet also adds another dimension of complexity since people want to create distributed repositories where the resources of a corpus can be scattered over different locations, nevertheless requiring transparent access to them.

Summarizing we can say that a much broader group of researchers besides the experts who have always handled expensive resources are now involved. They are managing larger amounts of more complex structured resources, making them available in standardized formats and descriptions via the Internet. Now that resource creation has become much more easy many individual researchers are also coping with resource management problems pushing the management task beyond the experts at large data centers.

2. Resource Management

The increased relevance of resource management can best be seen in the document domain by the emergence of various sorts of commercial Content Management Systems. It is widely understood that only improved management concepts will allow us to prevent a chaotic situation where we will have an increasing amount of data on our storage devices, but don't know about them nor know how to access them.

We can identify at least four different groups of people involved in resource management each one with their own views: (1) the computer system specialists have to be able to manage data on a physical level. They allocate physical resources, define structures in file systems and take care of redundant copies for secure data storage. (2) The producer of resources wants to integrate his resources into the repository in an easy way and describe them easy and correctly to facilitate retrieval. (3) The user wants to deal with data on a domain-oriented level, i.e. a level where the

well-established concepts and terminology of a domain are used. He is not interested in file system details. This view includes distributed scenarios where the user wants to combine resources from different institutions without having to know where exactly the resources reside. Often the producer is himself a user. (4) The archive manager acts as an interface between system specialists and producers and also prefers to manage data at the level of domain concepts. At least he has to know how the system managers handle the resources since he has to draw the links between logical and physical structure and influence for example the policies for protecting the data. In many cases the producer/user is also the archive manager, since there is no support staff. Management has to consider all views.

The following is a non exhaustive list of points to be addressed by modern resource management (resource discovery is in general seen as being a component of resource management, but in this paper we will mention it, but not focus on it).

- How to store resources such that they can survive for many years independent from technology changes.
- How to protect resources against unauthorised access
- How to create personalized views on resource repositories to facilitate easy and optimised navigation
- How to offer easy and immediate access to resources after access is approved?
- How can descriptions of sets of resources be modified easily?
- How to easily integrate new resources into the distributed resource repository?
- How to keep track of old versions?
- How to make such a management scheme available to interested parties.
- How to easily move groups of resources to other locations transparent to the user/producer?
- How to achieve hardware and operating system independent operation within the resource domain?
- How to easily integrate different data types that belong together and allow access while hiding the complexity?
- How to inform people about the existence of a resource and its major characteristics?
- How to easily discover resources in a distributed scenario from a conceptual perspective?

In this paper we will focus on the resource manager and user views. This although many important problems such as for example the problems of long-term archiving of digital media are not at all solved.

3. Pillars of Management

As already indicated, industry delivers a wide range of software solutions that are meant to cover documents of all sorts. In this paper we will not discuss Document Management Systems although they may deliver much functionality, but focus on the key pillars of open distributed solutions aimed at our specific environment and data types.

3.1. Standards

Open standards are very important to achieve interoperability, to build up long-term archives and to produce long-term available tools. Especially in the domain of computer-based language resources, however, we are faced with an extremely dynamical situation. This means we are confronted with a multitude of standards making many people turn over to use the word “best practice guidelines” instead. For multimedia resources for example we are confronted with a long list of media compression methods (MPEG1/2/4, Cinepak, Sorensen, MP3, ATRAC etc) all emerging within the last decade. Each having its advantages and disadvantages dependent on the field of application. For an archive one has to decide about major backend standards (such as MPEG2) which allows creating other representations for specific applications on the fly.

Referring to the earlier questions we need a couple of standards. We claim that many of the management problems can be solved with the help of establishing a suitable metadata environment existing of a metadata element set and appropriate tools. Tools themselves are not subject of standardization per se, since it is good to have competing solutions. With respect to the metadata set, however, we need agreements on various levels. The metadata elements are the dimensions of how to characterize a resource and it is clear that each choice for a set of dimensions limit the expressiveness for other groups of users. Therefore, we can expect that there will be different sets of dimension to describe multimedia/multimodal language resources. Important for the community is that we have open accessible definitions of the elements such that schemes can refer to them. They should be described as Data Categories if this will be the common practice for terminology repositories.

In addition, in the case of non-orthogonal spaces as the one we need to describe, these dimensions can only be defined appropriately by specifying suitable controlled vocabularies. They are the values that a specific dimension can take. Also these controlled vocabularies have to be openly accessible and should be defined in the same way. Both elements and their controlled vocabularies, have to be known exactly to achieve interoperability. Of course, it makes sense to use just one controlled vocabulary for example for language codes, but also here we are faced with different (quasi) standards such as ISO 639-2, the Ethnologue list from SIL¹ [1,2] and the various lists handled by specific projects. Also here we must accept that different vocabularies will exist.

Consequently, we are faced with mapping problems on different levels. RDF will be the primary language to try and bring all the different pieces of the mosaic together. This problem has not been tackled yet with the exception of a few cases such as in the Harmony project and in the mapping proposal from IMDI² to DC³/OLAC⁴. MPEG7⁵ categories were mapped on Dublin Core categories in a very restricted way and the element relations are described

¹ Summer Institute of Linguistics

² ISLE Metadata Initiative

³ Dublin Core Metadata Initiative

⁴ Open Language Archives Community

⁵ MPEG7 is the standard for media annotation within the family of MPEG standards in the film and media industry

with the help of the RDF formalism. Such a formal framework has not yet described the IMDI to OLAC mapping. At the moment we don't know which expressional power the community will need to accomplish the big task to create such a mapping for the language resource domain. The emergence of DAML/OIL [3] indicates, however, that RDF itself will probably not be sufficient.

It is assumed here without further comment that XML is our common language, i.e. all definitions and frameworks to be used should be based on XML.

3.2. Metadata Descriptions

The usage of metadata descriptions for improving the management of documents is not a new concept. Librarians are used to describe their documents with cards since many years. Linguists and speech engineers were used to describe characteristics of their resources and put these in file headers - mostly project specific formats. The community learned a lot from the TEI⁶ work about standards for resource headers (later adopted by the CES⁷) and it is still used as a reference to look at. Also in some projects such as CGN⁸ the TEI recommendations were followed to a certain extent.

TEI is a comparatively exhaustive descriptor set meant to describe the characteristics and structure of a resource. Newly developed metadata sets do not want to describe the resource in a too great detail, but address the problem of easy discovery primarily, i.e. a resource would be described sufficiently well, if a user manages to find it. Metadata sets such as DC, OLAC and IMDI follow this approach. DC tries to address the discovery problem with 15 sloppily defined categories ordered in a flat structure. In doing so DC allows the user to describe resources about steam engines as well as resources about Sign Language both on a very general level. For many DC categories it is not clear how they can be applied to different domains, therefore refinements are defined as was done by the OLAC initiative. The "DC:Type" element that defines the resource type is refined by the characteristic "CPU" to describe the type of CPU a NLP tool can run on. The semantics of such an element are stretched extremely.

MPEG7 and IMDI followed another approach since they started with studying the domain specific requirements. For MPEG7 it is essentially the production process of movies that has to be covered to later be able to retrieve relevant segments that are covered by the metadata set in addition to the ordinary elements such as "Creator". The basis of IMDI was an extensive survey of the different ways in which linguistic resources in all their variety have been described. Often this was done in the form of a proprietary "file-header" that contained metadata information about the annotation as a whole such as for instance the CHAT file format [4]. CES (being TEI compliant with respect to corpora) suggestions were applied were useful for discovery, however, we have not found sufficient support for other types of linguistic data than text. TEI/CES also mixes metadata and content in the same way as MPEG7. IMDI has favored a physical separation of metadata and content allowing

uncomplicated protection schemes which is important for some groups of users. It also allows separate management of resources and metadata, usefull because the integration of legacy data formats has to be supported.

3.3. IMDI

3.3.1. Session Concept

The IMDI set was especially targeted at multimodal/multimedia resources and their inherent complexity, i.e. basis is in general the existence of media recordings. This led to the development of the "Session" concept. For linguists a session is defined as the basic unit of linguistic analysis and covers a coherent type of linguistic action or performance. From a corpus organization point a session is the leave in the tree. A session is in general associated with a bundle of tightly related resources: a video recording of a native speaker, a set of pictures of that persons house, some field notes about this scene and afterwards some multimodal annotations. The IMDI definition of the term "session" covers this bundling from an access and management point of view.

In DC one would have to use the "DC:Related" element to describe the relation between these resources that is associated with much overhead. This was described in more detail in the IMDI-OLAC mapping document [5].

From a management point of view the session concept makes sense since accessing or extracting subcorpora implies accessing resp. copying of complete sets of related information.

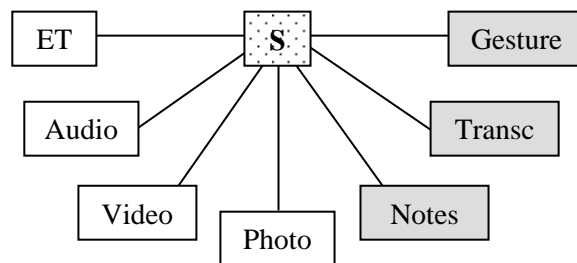


Figure 1 shows a typical session with its related resources all referring to the same linguistic event. It covers different types of recordings and different annotations.

In IMDI its the structured metadata set which describes this relation, i.e. there is only one metadata description (if the user decides to do it that way) with different sub-blocks describing the characteristics of the individual components. This way allows a user to ask questions such as "give me all resources which have eye movement recordings and a phonetic transcription of what was spoken"

3.3.2. Browsable Domain

Next to the "Session" concept, IMDI introduced the idea of structuring corpora in a conceptual space by having hierarchies of (sub-) corpora where description nodes representing a certain level of abstraction with respect to other (sub-) corpus nodes culminating eventually in pointers to session nodes (see figure 2). Each level represents a certain abstraction layer that is meaningful to the resource manager or user.

⁶ Text Encoding Initiative

⁷ Corpus Encoding Standard

⁸ Spoken Dutch Corpus Project

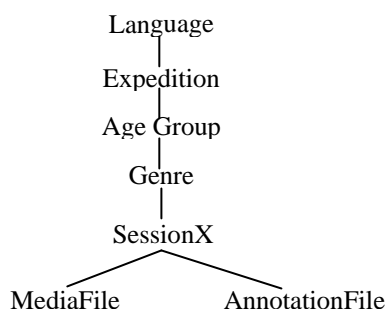


Figure 2 shows a typical hierarchy from field linguistics

Since corpus nodes create logical structures several parallel hierarchies can be created to structure the same (sub-)corpus and to express different interests of users. This allows each user to establish his own preferred view on the distributed resource domain and by also using bookmarks to create his own conceptual space (see figure 3). These parallel hierarchies can also be used to support versioning. Of course, there is no reason for the user to not create cross-references. For management purposes such cross-references are of course difficult to handle, i.e. the resource managers preferably would work with just the canonical tree.

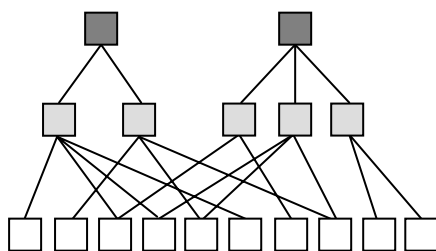


Figure 3 shows two user defined hierarchies referring to the same set of session nodes that are at the bottom level. One view could make a sex distinction, another one by age groups.

The mechanism by which the (sub-) corpora nodes refer to each other is to use URL's. This has the advantage to support distributed corpora frameworks and create a unique namespace for all resources.

3.3.3. Data Type Integration

Such a browsable domain as indicated is of course very useful for integrating various data types that we find in complete corpora. We already described the integration on the session level. For many data types however it only makes sense to associate them with higher nodes in a corpus tree. Such a node represents an abstraction with respect to a number of metadata elements (for example sharing the same language). Lexica can be related to a sub-corpus associated with a language or a set of recordings for a language (lexicon of a 3 year old child). Field notes and comments about dialect variants in general can appear on all levels of a corpus. In general many of these data types do not have any definite structure, but are just prose texts in some general format such as DOC, HTML or PDF. Corpus management has to provide mechanisms to include such descriptions in a flexible way.

IMDI allows the resource manager to do so, but of course, will exclude proprietary formats such as DOC.

3.3.4. Practical Considerations

A strong concern was and still is how one can enforce creators and managers to adhere to standards with all its consequences as described above. The stricter the rules are such as full adherence to the chosen controlled vocabulary of a certain element, the more sensitive these procedures will become. Although the IMDI type of operations are now in operation for 3 years we cannot claim that a "standard" such as IMDI for describing language resources will not undergo changes. In IMDI for example we expect changes with respect to the dimensions and vocabularies that describe the resource content.

It was found - and this experience is nothing new - that it is very important to support the creators and managers with professional tools. Within IMDI it was always tried to have a balance between the development of the metadata set and an editor that supports the creation of IMDI descriptions. The IMDI editor now supports

- All metadata elements including their controlled vocabularies in a dynamic way, i.e. if the definition in the repositories change the editor will adapt its representations
- Sub-blocks which allow the user to save and reuse reoccurring information such as participant or project information

Version changes in the metadata set can of course lead to severe problems for corpus management and metadata usage. There efficient tools are of the greatest importance to modify all whole sets of existing metadata descriptions. Currently, a script allows the resource managers to change the values of the elements for a whole set of metadata descriptions. Of course, such operations are very sensitive and such a script may not be given to the general user. The intention is to include such an option in the editor such that all changes are conforming to the actual IMDI definitions.

The browser offers the same feature as the editor in so far that it also uses the actual vocabulary definitions from the repository. Further, the browser offers the following management relevant features:

- A user can create new (private) nodes and therefore define his own view on a sub-corpus
- It is possible to start the editor from the browser environment to modify metadata descriptions
- It is possible for the users (managers) to associate tools with individual or bundles of resources such that when a (set of) useful resources was found immediately a tool can be started to operate on the resources.

Both tools will have to provide for version conversion in case they find metadata descriptions in an older format. They should not however work with old versions without forcing (if possible) an update.

In the future the editor has to be extended to be able to create formatted lists (Spreadsheet type) of the content of a range of metadata descriptions for easy check and input to for example statistic programs. This is a favorite view on metadata of many users. The user has to be able to select the elements he wants to see. One complication is

given through the fact that some elements can occur several times such as participants, i.e. the number of entries for the spreadsheet can only be computed by first reading all selected metadata descriptions.

3.3.5. Difference to Normal HTML Domains

Of course, the basic organization principles sound very familiar, since we use the same for designing web pages. Instead of creating XML based descriptions one could create HTML pages and include all information and data types as hyperlinks in the usual way. Some archives are operating this way. Metadata descriptions could be included in the headers of the HTML files to support element-based search.

The IMDI team did not choose for this way for the following major reasons:

- HTML is basically a way to describe how documents should be displayed and not to describe data structures.
- Using HTML would not have made sense without also using HTTPD servers and browsers. Otherwise HTML is just a much less powerful version of XML. The current HTML browsers however are not suited to perform all computation tasks required of a metadata browser such as making intelligent choices for tools to work on resources.
- We needed a format to transfer information. Tools should be able to interpret this information either to display parts of it or to offer the user a choice of tools to work on referenced resources.

4. Conclusions

Based on 3 years of experience with a multimedia/multimodal corpus which covers already more than 7000 metadata descriptions and a showcase application including sample corpora from 6 European institutions we can draw some conclusions.

1. All questions raised in chapter two are addressed by the IMDI environment with two exceptions: (1) Version handling of resources and metadata description schemes are not yet supported by the tools by the tools. (2) The tool for extracting complete sub-trees of a corpus is not yet available.
2. The need to apply the definitions and tools to such a big and heterogeneous corpus as for example the MPI corpus was a useful and necessary enterprise. It made us understand the underlying processes and requirements to establish an environment such as IMDI.
3. Corpus management was performed during the development phase of the IMDI environment. This meant that frequent updates of the metadata schema took place that required frequent transformation of the metadata files.
4. We now have an environment where it is comparatively easy to integrate or build up IMDI based archives that supports the creator, the user and especially the resource manager with suitable mechanisms and tools.
5. Since all definitions are open everyone can create his own set of tools to work on the metadata descriptions,

i.e. improve the search engine or write another browser.

6. Using a file oriented framework for storing metadata only appears as an advantage when distributing or integrating small (personal) archives or making extractions of sub corpora on portable media for off-line use. It does however create confidence of the linguists that they can take their metadata descriptions with them on a floppy and are not dependent on server bound DBMS's.
7. Using metadata in a uniform, controlled and structured way is a new experience for our linguists. It did and still costs a large persuasion effort to have them input their metadata. It has only been since a short time that they themselves can reap the benefits by using for instance metadata search, since a critical mass is necessary and since the improvements for resource management had to become apparent.
8. The introduction of a complete and operational metadata environment was the first experience for the development team of this sort. Often the practical experience guided us in designing and improving the tools, since we did not foresee all aspects of efficient resource management beforehand.

Finally, it seems to be appropriate to add a statement about future perspectives. We see metadata for language resources still in its beginning phase, since there are not so many resource repositories which already created the appropriate files. Especially there are only few attempts to do resource management with the help of metadata environments. We have shown their great potential but also the difficulties involved. Especially the inclusion of metadata element and vocabulary definitions in open repositories and the formulation of their relations with the help of Semantic Web compliant mechanisms such as RDF will motivate more groups to contribute and participate. Interoperability between different metadata sets will also be facilitated by applying these agreed standards.

The soon to be started INTERA project is aiming to realise and work at the above mentioned points.

- [1] [ISO639-2]
Codes for the representation of names of languages - part 2: alpha-3 code, International Organization for Standardization (ISO), 1998.
<http://lcweb.loc.gov/standards/iso639-2/langhome.html>
- [2] Ethnologue language name index
<http://www.sil.org/ethnologue/names/>
- [3] DAML/OIL: <http://www.daml.org>
- [4] Childes: <http://childes.psy.cmu.edu>
- [5] IMDI-OLAC-Mapping: <http://www.mpi.nl/ISLE>