

Towards a Standard for Meta-Descriptions of Language Resources

D. Broeder, H. Brugman, A. Russel, R. Skiba, P. Wittenburg

Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
e-mail broeder@mpi.nl

Abstract

The desire is to improve the availability of Language Resources (LR) on the Intra- and Internet. It is suggested that this can be achieved by creating a browsable & searchable universe of meta-descriptions. This asks for the development of a standard for tagging LRs with meta-data and several conventions agreed within the community.

1. Introduction

At the Max Planck Institute for Psycholinguistics (MPI) the concept of the Browsable Corpus (BC) was introduced as a means of organizing and structuring the growing mass of complex multi-media language resources. In the BC concept LRs such as annotated media files or separate transcriptions are pointed to by a collection of meta-description files. These meta-description files hold meta-data about the LRs to characterize their form and content to the user in a way that is meaningful to the user. Their own structure is well defined and the semantics of the vocabulary is based on agreements amongst the users such that they can be browsed and searched. As a consequence the user does not have to inspect the resources themselves to establish their usefulness. For many questions it is sufficient to simply scan the meta-universe.

The BC has been applied to various LRs from different researchers and research areas and although not yet all the MPI's LRs have been described and structured in this way, enough has been done to suggest that this approach is very promisingⁱ.

Several other communities are working on similar ideas to create such browsable and searchable subspaces in the Internet. We refer here especially to the work of the librarians on Dublin Coreⁱⁱ and to the Resource Description Framework proposal of the W3Cⁱⁱⁱ.

The concept of the Browsable Corpus and the ongoing activities in the internet itself suggest that this technology should be expanded in such a way that there can grow an international universe of linked meta-descriptions available on the internet.

This idea of creating a standard for LR meta-data has led to a joint NSF/EC initiative^{iv}.

By looking more closely at the way the Browsable Corpus concept was implemented we will try to identify likely problems with its generalization.

2. The Browsable Corpus

The BC concept implies that individual LRs (the leaves in a corpus) are associated with so called session meta-description files (session MDF) containing meta-data and location information of those LRs. Other meta-description files called corpus meta-description files (corpus MDF) may then describe parts of a large corpus and point to these session meta-description files and other corpus MDFs forming a hierarchical structure (see Figure 1). This hierarchical structure that forms the meta-universe may then be used by appropriate tools for browsing and searching.

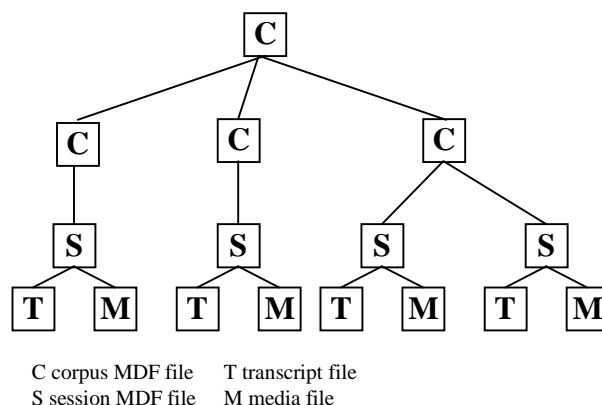


Figure 1

The internal structure of MDF files has to provide efficient encoding for the meta-data elements and LR references. XML was chosen as the format for the meta-data in the MDF's because it has the power to express the required structure, it is generally accepted for this sort of activities, and allows us to take advantage of the expanding range of XML-capable software. Thus the MDF structure is specified in a DTD-specification file. The current operational version of BC uses an inflexible DTD that defines a fixed MDF structure which is sufficient for our own purposes but is unlikely to satisfy a

wider community. Version two has more flexibility in that it will allow extensions of the DTD.

There are obvious problems in extending the BC approach beyond the well-controlled domain of a single institute. These problems are mainly concerned with the choice of an optimal set of meta data items that have to describe a disparate set of LRs.

3. Data and Meta-Data

Meta-data is data about data. To see what meta-data items are possible we first have to define the set of language resources we want to describe. Corpora based on transcriptions, audio/video material and annotations are the major type of LR. There are many types of secondary resources such as lexicons, grammar descriptions, sound system descriptions, amongst others, all of which are based on primary corpora.

A primary data file in a corpus is an observation of a subject; here it is useful to define meta-data elements referring to the subject(s) such as the subject's age, socio-linguistic background etc. The content of the linguistic action requires meta-data elements like language spoken, discourse type etc.

Secondary resources can give rise to quite other meta-data items. This often involves an interpretation of primary observation data involved. For instance in a lexicon a reference to the linguistic theory used to describe the syntactic categories could be a valuable meta-data item. Despite this diversity, there are a number of elements that may be common to all classes of LRs although there can be a small difference in semantics.

4. Organizing meta-data

A number of strategies present themselves when looking at ways to describe LRs with meta-data:

1. Define a broad set of meta-data elements that will cover any LR, ignoring that many elements will only have significance for a single class of LR.
2. Define a minimal set of meta-data elements that is common to all LRs but will probably be insufficient to describe the LR in sufficient detail. This approach formed the basis of the Dublin Core programme.
3. Define - for every subtype of LR or sub-community - specific sets of meta-data items. In addition to 2. From the experience of the Dublin Core community it seems wise to define only a core set of elements and describe their semantics. In order to make progress we should not strive to describe too broad a domain.

Another approach to achieving flexibility is to create sets and subsets of meta-data items by selecting bundles or structures of meta-data items from a predefined collection. For instance look at the meta-data items associated with

the concept of language like "name" and "dialect". We can describe this structure (as an XML element);

```
<!ELEMENT LANGUAGE EMPTY>
<!ATTLIST LANGUAGE
  TAG CDATA "LANGUAGE"
  NAME CDATA #REQUIRED
  DIALECT CDATA #REQUIRED
>
```

This structure can be used to describe part of the meta-data items associated with the "content" of the corpus, but also to describe part of the meta-data items associated with "participants" or subjects of the corpus, as you would like to specify the native language of a subject. The difference between the two contexts is being expressed in the TAG attribute of the LANGUAGE ELEMENT which in the CONTENT case has value "LANGUAGE USED" and in the PERSON case has value "NATIVE LANGUAGE". This represents some sort of structured name space mechanism.

```
<!ELEMENT CONTENT (LANGUAGE+, ...) >
  <!ATTLIST CONTENT
    TAG CDATA "CONTENT"
    DISCOURSE_TYPE CDATA
  ...
>
<!ELEMENT PARTICIPANTS (PERSON+, ...) >
<!ELEMENT PERSON (LANGUAGE+, ...) >
```

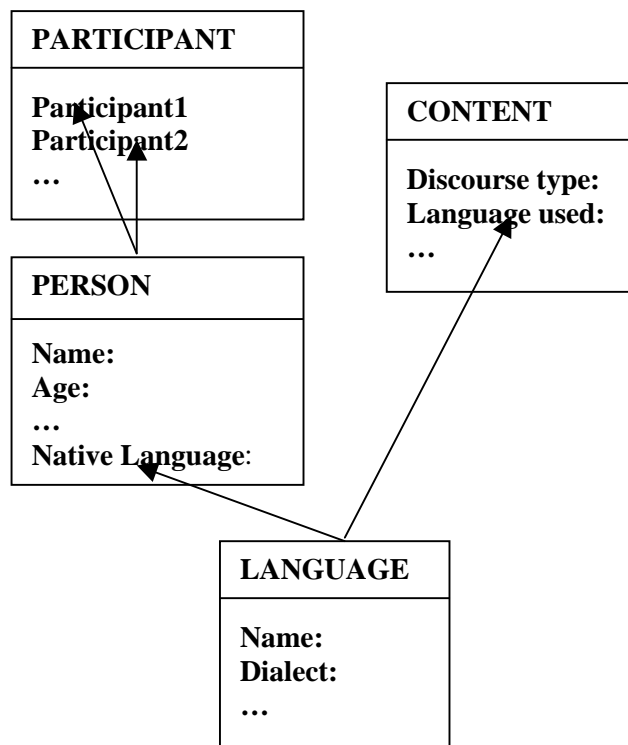


Figure 2

A possible weakness of this strategy is that the semantic difference between LANGUAGE in the context of "language spoken" or "native language" of a subject can not be expressed by the same set of meta-data items.

Clearly, this approach avoids having a flat namespace where we would have elements with names like: "name of native language participant 1" and "name of native language participant 2". We now simply define the meta-data item structure LANGUAGE and define that every PERSON meta-data item structure has a LANGUAGE type meta-data item bundle associated with it with TAG "NATIVE LANGUAGE". Also this approach enables that software developed for handling LR meta-data becomes simpler and easier to write.

A different but important other class of meta-data is the type of annotation used in a resource. This includes information on the format of the annotation layers, the kind of annotation layers available and the coding conventions applied. This would enable a statistical analysis tool to automatically select the correct set of available annotations and map similar codes found in different types of resources if this mapping were linguistically feasible.

Some users will have very specific tools they want to apply to corpora that need very specific meta-data items. It will therefore be necessary to be able to define additional meta-data elements for specific purposes. Such flexibility could be created by having each meta-data structure provide space to accommodate a place to add additional meta-data sub structures.

5. Connection to existing standards

The Dublin Core (DC) initiative of the librarians was already mentioned. What is proposed here, however, is a much more restricted approach. DC appears to claim to be a general purpose standard for "all" kind of meta-descriptions with the consequence that the standard is still not finalized and that its formalisms get increasingly complex. It is proposed here that we limit ourselves to create a browsable and searchable universe of LRs only. We have to be able to move fast to be able to test such new mechanisms and adopt them according to our needs.

However, it is expected that the Resource Description Framework initiative of the W3C (RDF) will be very useful for all meta-data initiatives. Elements specified by other communities could be re-used simply by using the name-space mechanism. Elements defined by different communities can be used together in the meta-descriptions.

A predictable problem is explaining the semantics of the elements to the user in a simple way. This is important in two areas: (1) When entering new descriptions and (2) when a search has to be specified. The user in both cases has to know the vocabulary and understand the semantics of the elements. Thus it may be necessary to redefine certain elements for the LR community, to make sure there is a standard term all members agree upon. Reusing

vocabulary from other communities is likely to be counter-intuitive and therefore bad. The use of meta-descriptions should solve problems and not create new ones.

6. Dynamic structuring

Although the structure created by mutual referring MDF files as described in 1 is not directly connected with the question of meta-data standards, the implications of this structure for the tools we use is important. A hierarchical structure seems to be essential for creating useful and meaningful browsable spaces and there should exist at least one hierarchy to get to the eventual LRs.

Multiple hierarchical structures built upon the same LRs are even better. For instance a corpus can be divided between male and female speakers or between adult and child speakers (see Figure 3).

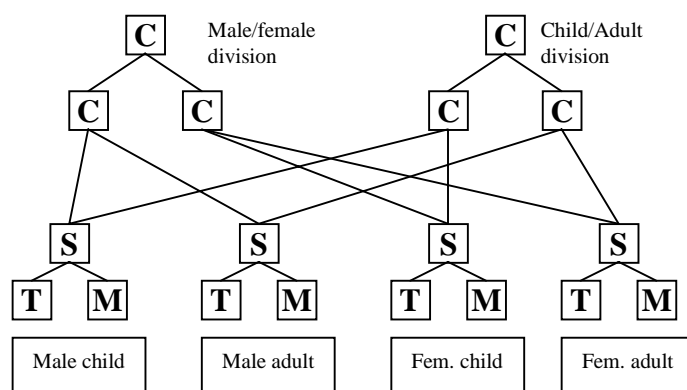


Figure 3

It would be even better to be able to compute and create new hierarchies on the fly from any point in an existing hierarchy. Suppose a user would be browsing a corpus and would have selected a tree of all LRs with male speakers, it might be appropriate for that user to want the next level be a division between age groups for all males but the user might want a division on the basis of social status for the females. This can give rise to asymmetric expansions such as in Figure 4

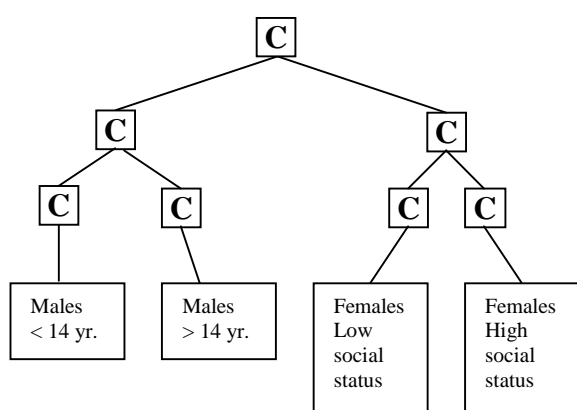


Figure 4

An unsolved question is how we can create and maintain such browsable hierarchies for reuse in an international context. The organizational effort is such that they can't be created manually by a central organization. Only small coherent sub-communities with a high degree of IT organization structure could manage this.

For instance in the area of anthropology one natural browsing hierarchy would be a geographic division at the top levels. If this hierarchy is to be created and maintained automatically the question is if this can be achieved on the basis of the available meta-descriptions. In general abstraction and classification processes have to be carried out based on the preferences of users or sub-communities.

7. Tools

For the development of a meta-description standard for the LR community, it is important to have suitable tools that help maintain that standard right from the beginning. An editor has to be available for everyone which supports the meta-description structure and vocabulary, gives a maximum of help to the individuals when creating the meta descriptions, and constrains the input possibilities. At the MPI such a meta-editor has been built, although it lacks the necessary flexibility and user guidance for general applicability. The user interface has to be dependent on the schema definitions, i.e. the editor has to be able to read them and adapt its user interface.

Browsers have to be available which support the specific meta-description file structure and the meta-data vocabulary. At the MPI a special browser was developed based on Tcl/Tk which parses the XML meta-description files and displays the corpus structure and meta-data elements.

Powerful search tools supporting the vocabularies of the name spaces used have to be available. They have to help the users by means of guiding them through the dimensions of the search space defined by the elements used in the meta-descriptions.

The facility of meta-data search has to be coupled with the browser to support incremental search, where the user increasingly narrows down the search domain after studying previous search results. The browser should show then the users their search results by marking up the found resources. This is called "search aided browsing".

So far the assumption has been that the meta-description files are physically located somewhere on the hard disk of a computer. These files are pointed to by URL's from other meta-description files or web pages. Although this form of meta-data should always be supported because it allows an individual researcher on a home PC to work with the browser and search-engine there are other forms possible:

- Meta-description files stored in a DB. They are fetched by a HTTP server.
- Meta-description files are generated real-time by a program from data in a DB.

In its simplest form the meta-data search will act just like a web-crawler going through the meta-description files looking for the right meta-data items. This will not be an acceptable approach for a site with many meta-description files where this would simply take too long even if all the meta-description files were virtual files such as described above. In that case a meta-data search would better query a DB directly. The simple "crawler" approach should always be possible.

At the MPI the possibility to directly start application tools after having found a corpus part by browsing and searching in the meta-universe turned out to be very handy. Often users don't know and understand the limitations of the exploitation and analysis tools relevant to the specialized resources being looked at. If a way could be found to directly show the user which tools can be applied to the selection made, it would be very helpful. This can only be achieved, if the meta-descriptions have links to the basic resources such as annotation and media files. Some sort of resource typing must also be achieved and a toolmaker has to indicate which type of LR the tool can handle. A schema similar to MIME-types would be necessary to achieve this.

8. Registries and Portals

Given that a user creates a meta-description of a specific language resource by using a constrained editor as described in 4. Where must these descriptions be placed and how must they be linked to form a browsable and searchable universe? It could be suggested that for a number of sub-communities meta registries are setup in such a way that the meta-editors directly create entries in such registries. But in principle this has to be a supervised activity to avoid the chaos that could be created by anonymous individuals. Only a process of constant evaluation can lead to the high-quality universe we are looking for.

It has to be mentioned that such a scheme where LRs local at a site are pointed to by browsable hierarchies

established at a central organization requires that the maintainer of the LRs guarantees their availability at a constant location.

Centers have to be established which can be used as entry points to the LR-meta universe. Here institutions such as ELRA and LDC could play an enormous role. They have to maintain the universe, define permanent browsable structures and evaluate browsable hierarchies offered for acceptance in the LR-meta universe.

9. Summary

Finding and accessing Language Resources over the Intra- and Internet is made possible if the LRs are tagged with meta-data and structured in hierarchies.

These hierarchies cross the borders of individual institutes and communities and impose special demands with respect to their creation and maintenance.

The meta-data structure and vocabulary must be standardized rigidly enough to allow general tools for LR search and resource discovery to interact with it and yet flexible enough to allow sub-communities to make their own extensions and specialized tools

10. References

ⁱ MPI's Browsable Corpus website
<http://www.mpi.nl/world/tg/lapp/browscorp/browscorp.html>

ⁱⁱ Dublin Core web page <http://purl.oclc.org/dc>

ⁱⁱⁱ RDF web page <http://www.w3.org/RDF>

^{iv} International Standards in Language Engineering- a EC 5th Framework and NSF initiative