

# XML-Based Language Archiving

Peter Wittenburg, Hennie Brugman, Daan Broeder, Albert Russel  
Max-Planck-Institute for Psycholinguistics

## 1. Introduction

At the MPI for Psycholinguistics the multimedia/multimodal archive now comprises close to 30.000 sessions that can be seen as linguistically meaningful units of analysis. Most of these sessions have a multimedia basis in so far that the primary data is either based on sound or on video recordings – in total more than 5000 hours. A large fraction of these recordings are associated with annotations of various types. The archive also contains other linguistic data types such as lexicons, sketch grammars, field notes and others. Various projects contributed to this digital collection such as

- field workers from the MPI studying language behavior of different cultures and language acquisition processes by children and adults with the help of longitudinal observations
- researchers of the MPI studying multimodal interactions in various circumstances and from various cultural backgrounds
- researchers of the MPI and within the ECHO project studying sign languages from different countries
- teams documenting endangered languages from all areas of the world
- Dutch and Belgium researchers building the Dutch National Spoken Corpus
- researchers from 5 European countries studying the language use of immigrants

Forming and maintaining this large archive that comprises various corpora from different researcher groups so that it is visible as one coherent collection and that it can be exploited with a limited set of tools has been a major effort during the last 4 years and was only feasible by relying on XML-based technologies. However, it has also to be made clear that proper data modeling is the step that has to be made first.

In the following we will outline how proper data modeling for different aspects of corpus creation, management and exploitation together with XML-based instantiation of these models helped us to cope with the challenges.

In this paper we shall use the following terminology: An **archive** denotes the full and organized collection of resources that has to be administered and offered to the user community. A **corpus** is a sub-part of this collection that was created by a person or a project. In the general sense **metadata** can be any data associated with other data, i.e. metadata can be annotations of video streams or of annotations, lexicons the entries of which refer to tokens appearing in a corpus, ontological entries that refer to concepts of the real world, keyword type descriptions of the resources in a collection or many others. For reasons of simplicity we will refer to metadata in this paper as the keyword type of description of resources that is useful for discovery and exploitation purposes.

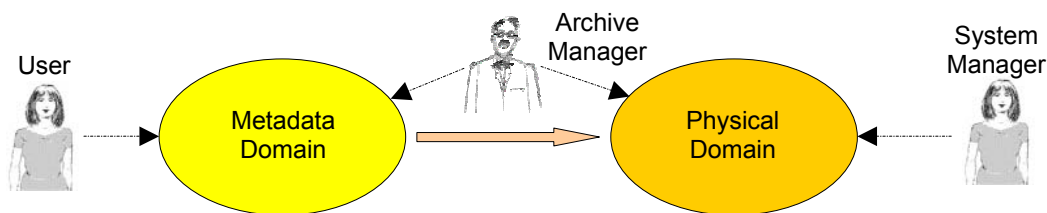
## 2. Archive Organization and Management

When we took the decision to digitize all material and provide access to multimedia recordings not any longer via traditional audio/video technology but via computers, it was clear that we would be faced with the problem of how to organize the resulting large and ever increasing collection, how to give users access to the resources it includes and how to allow managers to maintain and extend the collections without ending in chaos.

A number of fundamental and far going conclusions guided us during the design and development phase. (1) From the beginning we assumed that our archive has to be seen as just one building block in a world-wide domain of online archives of language resources that are brought together by the Internet. Users and in particular agents would need an interoperable domain of language resources. As a consequence our organization solutions should be open for easy integration and exploitation. (2) We assumed that many of our typical users would work often without connection to the Internet at least temporarily. So our tools should be able to work with local collections without the need of connecting to a central site. (3) We understood that even within a discipline such as linguistics very different types of users would like to make use of the emerging domain of archives, i.e. shells addressing the specialists and those for the computer illiterates should be available. (4) Knowing that the underlying physical structure (storage architecture) would change regularly it was clear to us that we would have to enable and convince people that they should discover and access useful resources via a virtual layer and not by using physical access paths. (5) Also from the beginning it was evident that the wishes of users and user groups in describing their data would be different so that flexibility was necessary. For more details about this work we refer to [1,2].

## 2.1 Metadata Model

The basis of all our archiving work was the design of the IMDI<sup>1</sup> metadata model in collaboration with others mainly within the ISLE<sup>2</sup>, INTERA<sup>3</sup> and DOBES<sup>4</sup> projects. It should be used by the archive managers to organize the material in a way independent of the actual physical location and to carry out typical management tasks as far going as possible. It should be used by the users to discover and access the resources. Only such a system would give the system managers the freedom to take appropriate decisions at relevant moments without affecting the usage of the resources.



When we wanted to uncouple users from the physical organization of the resources we had to understand first the way how users organize their data. Therefore the metadata model had to preserve the most relevant elements of resource organization. The major conclusions can be summarized as:

- Resources are organized in bundles of data at various levels. Recordings and their different levels of annotations are tightly coupled. They share the same time axis or the same notion of sequential order. Therefore the term “session” (later bundle) was introduced to denote the smallest manipulable unit in an archive structure and to ask users to create metadata descriptions at this level, since the individual resources belonging to one session share most of the information necessary for discovery and retrieval.
- Resources are organized according to various criteria such as field trip dates, languages, age groups and others into manageable sub-corpora. Users also wanted to have some flexibility to regroup resources at this level dependent on their actual research interests. It was decided that this level of bundling could best be described

<sup>1</sup> IMDI=ISLE Metadata Initiative, <http://www.mpi.nl/IMDI>

<sup>2</sup> ISLE=International Standards for Language Engineering

<sup>3</sup> INTERA=Integrated European language Resource Area

<sup>4</sup> DOBES= Dokumentation Bedrohter Sprachen; <http://www.mpi.nl/DOBES>

by abstraction nodes representing some concept that the sub-subsequent sessions share.

- Other structural relations between resources are for example that a lexicon is created for a specific language and at least partially derived from some of the resources in the archive. The creation of an abstract “language” node could document this relation and the lexicon would typically be associated with such a node. There are many different types of these relations.

These considerations on the one hand and the need to offer users a full-fledged alternative to the physical structuring methods led us, amongst other considerations, to a metadata model that is different from what was suggested for example by the Dublin Core (DC)<sup>5</sup> and OLAC<sup>6</sup> initiatives. Here metadata was introduced just for resource discovery via search engines. In our case we also decided to cover the organizational and management aspects. Only the possibility of for example grouping based on abstractions would allow us to treat them as units of management (copying, associating access rights, associating unique identifiers, etc).

## 2.2 Metadata Set

Another essential pillar of the IMDI metadata model was the definition of a metadata set. It had to mimic the bundling structure of sessions and be flexible enough to meet the individual needs of different projects. In addition to the mentioned criteria it was understood that only a structured set would meet the user’s needs. While a flat set such as DC would not allow to relate attributes such as “sex” and “education” to participants except by introducing refinements, but then modifying the semantic scope of the concept itself, IMDI introduced structure into the set, i.e. various elements in the IMDI set can be associated with attributes that are part of the IMDI standard.

Flexibility was achieved defining a core set of IMDI elements on the one hand, but on the other hand data providers can add own element-value pairs at various places. Such extensions could theoretically lead to a proliferation of element categories reducing the success rate in resource discovery. First, the creation of metadata is a painful effort for data providers and the experience showed that they are in general satisfied with the semantics already offered by the core elements. Second, the introduction of “profiles” for specific user groups such as Sign Language researchers or projects such as the Dutch Spoken Corpus project that used a number of additional elements taken from the TEI<sup>7</sup> give controlled extensions a quasi official status. So the current IMDI set has three layers with respect to the semantics supported: (1) the Core IMDI set, (2) special project or sub-discipline oriented profiles and (3) user specific extensions.

Further, the IMDI set needed to be accompanied by definitions of controlled vocabularies that define the range of values certain elements can take. It was decided in the metadata model to not make controlled vocabularies part of the schema controlling the IMDI set itself, since it was foreseen that they will change frequently to meet the needs of the users and to not force users to define their own elements due to a missing value. For specific elements such as “language code” it must also be possible to allow the inclusion of different vocabularies such as the ISO norms and the Ethnologue list.

## 2.3 IMDI Infrastructure

Based on this elaborated IMDI metadata model we were able to design the basics of the concrete IMDI infrastructure. In contrast to other approaches that start with a relational

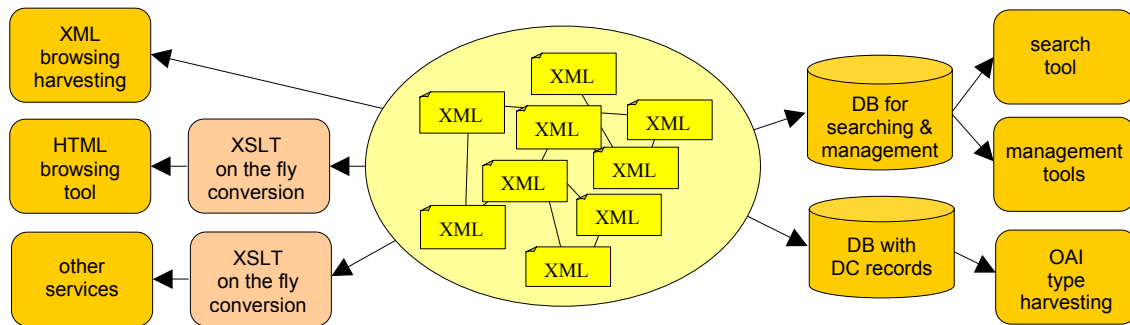
---

<sup>5</sup> DC = Dublin Core Initiative; <http://www.dublincore.org>

<sup>6</sup> OLAC = Open Language Archives Community; <http://www.language-archives.org>

<sup>7</sup> TEI = Text Encoding Initiative; <http://www.tei-c.org>

database implementation as the core, we decided to use a linked domain of XML-files as the core data structures of the IMDI infrastructures. All other data structures would be derived from these XML representations.



This approach has shown to have many advantages for us. (1) It immediately allows us to connect various emerging IMDI sub-domains by simply adding links, i.e. IMDI can operate in a distributed domain of resource providers without any additional efforts. (2) It allows everyone to crawl in this domain and create any service that can be useful to harvest and exploit the rich IMDI metadata domain. (3) It allowed us to build an XML browser exploiting the special features of the IMDI model such as the bundle concept, but also allowed us to offer HTML versions such that users can exploit the IMDI domain with normal HTML browsers. (4) From a structural perspective it is easy to generate DC records so that OAI type of harvesting is supported. (5) It is easily possible to generate all kinds of specialist databases to efficiently support searching and management. (6) Extractions of sub-corpora from the pool of IMDI metadata descriptions are easily made and copied for example to a notebook including all relevant structure information.

With respect to management a whole set of integrated analysis programs are available to check the correctness and state of the archive's structure. Further, an access management framework was developed that allows archive managers and responsible researchers to define access rights. Here a canonical metadata hierarchy is essential since the manager can select an arbitrary node and set rights for all resources that occur below the selected node.

The creation of HTML representations of the metadata descriptions can easily be achieved by style-sheet-based XSLT transformations that are carried out on the fly. However, the indication of the position in the linked metadata structure as a navigation help requires additional program execution at the server side.

Therefore central in many ways is also the availability of well-documented XML-schemas and concept definitions. The XML schemas describe the structure of the metadata descriptions and of the controlled vocabularies. Aspects such as support for multilingualism had to be considered since IMDI is used in various countries. All XML-Schemas defining the IMDI set are openly available in the Web ([www.mpi.nl/IMDI](http://www.mpi.nl/IMDI)).

## 2.4 Metadata Interoperability and Future

Metadata interoperability will become one of the essential pillars of the Semantic Web. At the encoding level interoperability is achieved by using UNICODE and at the syntactic level by using XML and by having validated the created IMDI files. We know from practical experience in the ECHO project where we created an interoperable metadata domain of 5 different humanities disciplines how important the encoding and syntactical interoperability is. We were confronted with non-validated XML repositories generated from databases of various types also including different character encodings. It was and is a very time-

consuming effort to transform these repositories into useful representations. In a dynamic environment where such repositories change continuously this is not feasible. For many holdings we are still far away from the ideal state that the OAI MHP protocol is used where well-checked data is offered by the data providers.

It is even more problematic to achieve interoperability at the semantic level. Currently, mapping relations between elements are hardwired into a wrapper to realize for example the IMDI-to-OLAC mapping. This is an unsatisfying approach since no one can influence the mapping. Within ISO TC37/SC4 we work on a framework where all concepts used are described in an XML-based and ISO compliant (ISO 11179, ISO 12620) Data Category Registry (DCR). In this way semantics are defined in a machine readable form and individual schemas will refer to entries in the DCR. While equality relations between two metadata sets can easily be implemented by referring to the same DCR entry, more complex relations can be implemented as RDF assertions referring to two different entries. The emerging ISO framework will naturally improve the semantic interoperability and open the possibility for projects to define their own metadata sets by re-using concepts that are already defined in the open DCR.

### **3. Archive Exploitation**

The archive contains a number of different linguistic data types such as lexicons, field notes and others that we do not want to discuss in this paper. We would like to focus on the complex problems associated with annotated multimedia recordings and texts and the important role of XML in this context. We will also not discuss the encoding aspects in detail, but focus on structural aspects in such annotated multimedia recordings.

Also in this respect we were guided by a number of fundamental decisions: (1) We are faced with incrementally added and updated annotations within a tier, but also on newly created annotation tiers. (2) In multimodal interaction studies one can easily have a large amount of annotation tiers (>50). (3) Annotations will exhibit all possible time relations since the multimodal streams such as speech, gesture, eye movements and others have to be seen as independent from each other. (4) Annotations will refer to periods in media time, but also to sequences in other annotations. References to points in time are seen as periods of unity length. (5) Some annotations will exhibit hierarchical relations within tiers (syntax trees) or across tiers as in the case of interlinearized morphological glossing. (6) Different types of cross-referencing have to be supported that allow to refer to many objects on different tiers in the extreme case.

Similar as in the case of metadata it was the intention that the annotated recordings should be available in well-documented formats to users directly so that they can carry out their own types of processing on them. On the other hand we wanted to provide (multimedia) tools that allow to create and exploit complex annotations. For those who have problems to even download and install tools, access options via normal web-browsers were planned.

#### **3.1 Abstract Corpus Model**

Guided by the above mentioned criteria and our experience with complex multimodal annotations for more than 6 years we were able to design and further develop an Abstract Corpus Model that was seen as the blueprint for writing programs like ELAN<sup>8</sup>. This ACM has the power to express the complexity mentioned above and it should be possible to import many of the well-known formats such as CHAT<sup>9</sup>, SHOEBBOX<sup>10</sup> and others. Therefore ACM can be seen as an attempt to define a general model for complex annotations. Our

---

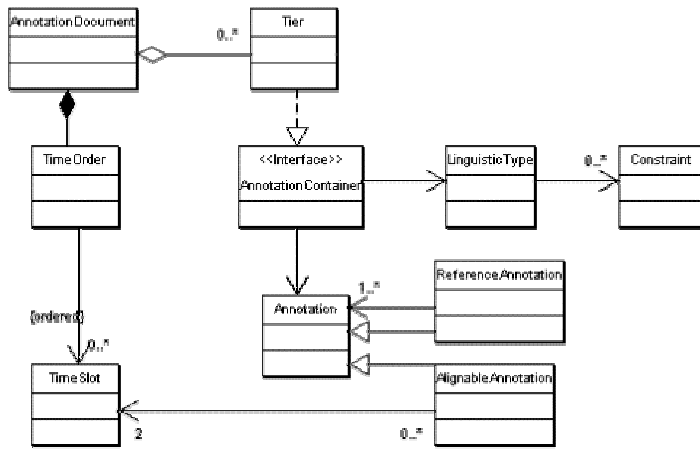
<sup>8</sup> ELAN = EUDICO Linguistic Annotator; <http://www.mpi.nl/tools>

<sup>9</sup> CHAT = Format used in the CHILDES project; <http://chilides.psy.cmu.edu>

<sup>10</sup> SHOEBBOX is a program frequently used by field linguists; <http://www.sil.org/computing/shoebbox>

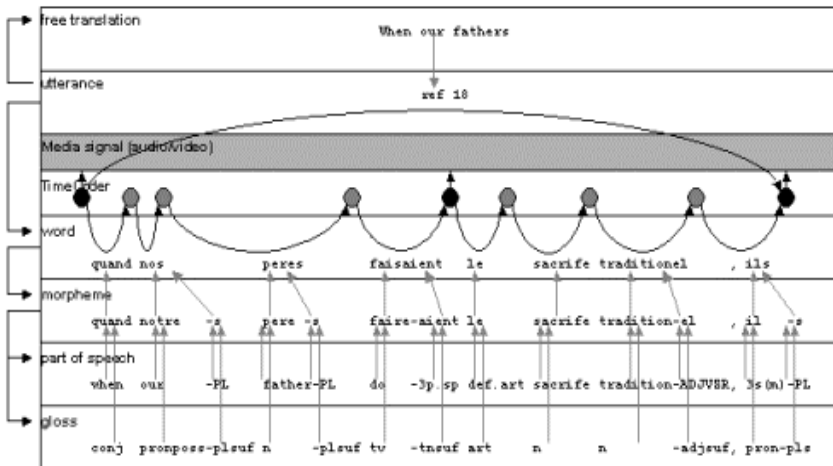
experience showed, however, that we had to refine the model several times to make it powerful enough to handle the continuously increasing demands of the scientists.

Further, the model had to account for annotation tier types and constraints that are specified for annotation types. The modeling was carried out in UML. The following figure indicates the core part of ACM.



This figure shows the core part of the UML chart that explains the Abstract Corpus Model. Tiers are seen as containers for annotations of certain linguistic type that share the same set of constraints. Annotations are split between those who are referring to periods in time, i.e. every annotation has a begin and end time and annotations can share such slots, and those that refer to one or several other annotations.

Many concrete annotation configurations were tested to see whether the model is powerful enough to handle all phenomena mentioned above. The following figure gives such a configuration, more elaborations can be found in [3,4,5].



This figure shows a typical complex annotation configuration where utterances and words are linked to time slots. All other annotations refer to ordered annotations. The annotations on the word, morpheme, POS and gloss tiers are part of an hierarchical system.

### 3.2 ELAN Annotation Format (EAF)

Given the concepts of the ACM and the ways they are associated, it was straightforward to design a format to deal with persistence and implement it using XML. The EAF is based on an open XML-schema which defines that each annotation is treated as an atomic entry referring to another annotation or to time slots. This way of formulation guarantees that parallel independent and therefore only partially overlapping streams can be represented without problems and that changes can be integrated easily.

The format is open to whether the annotations belonging to different annotations tiers are represented in one or several files. We assume that whenever a set of annotation tiers is created by one researcher or a closely collaborating team it makes sense to represent them

in one file. However, as soon as several persons work independently from each other a stand-off annotation is the natural choice. It is essential that the full complexity of annotation structures as described above can be represented in the case of a stand-off as well. The extensive use of ID/IDREFs should solve this aspect. The general requirements of stand-off annotations can be found in [6].

Several different types of research tasks were carried out using ACM and EAF as the basis. In the DOBES (Documentation of Endangered Languages) project the linguistic analysis work based on sound and video material is represented in EAF. Its obliged tiers are an orthographic or phonemic transcription and a translation into a major language. For some material a deep linguistic analysis has to be added allowing later generations to reconstruct the language. The most complex system is Advanced Glossing [7] defining 24 tiers to describe morphology and syntax. Similar annotation work is carried out by many researchers of the MPI.

Sign language studies are characterized by very complex multi-tier annotations since the movements of all articulators have to be described and analyzed in terms of their contribution to the standard linguistic layers. Such studies have been carried out within the ECHO project by 3 European SL communities for a comparative study of sign languages [8] and by a new research group bringing together signers from very different cultures.

Much work has been done in the area of gesture analysis and the relation of gestures to speech utterances. Several cross-cultural studies were carried out including annotations from the articulator to the interpretation level. Also here complex annotations covering many tiers, various types of time overlapping and cross-references can be found [9].

### **3.3 Access and Interoperability**

As mentioned above we have to understand that user groups differ in the way how they work with such richly annotated corpora. XML is an excellent basis to serve to generate the different usage types. For those who prefer using standard web-browsers to exploit annotated media we are currently testing different ways for web-presentations. One option is to generate SMIL formatted representations that produce media streams with synchronized sub-titles, another way is to generate HTML versions where annotations can be clicked on to invoke the appropriate media fragments. Starting from the XML representation it is not difficult to generate other representational forms with the help of style sheets. The difficulty is mostly to find an appropriate general layout for presenting complex annotations with the help of HTML. Because of the extensive use of ID/IDREFs it is less trivial to do style sheet based transformations that maintain the full structural complexity of annotation documents.

Since XML is now a widely agreed and powerful enough basis for structuring and tagging complex text documents, it is perfect to transform all other formats to XML. Import modules and converters were created even for structured WORD documents that are still frequently used by field linguists. However, often the differences in the underlying data models create problems for the transformation step. TRANSCRIBER<sup>11</sup> for example has the notion of events and its annotations are marked by just the begin time which is also the end time of the previous annotation. An interpretation step beyond XML is necessary before the TRANSCRIBER created XML file can be transformed into for example the EAF format. The emergence of XML has reduced the number of different formats and obviously it helped to stimulate a world-wide discussion about and convergence of suitable annotation formats, which will result in a unification. Currently, efforts are taken within ISO TC37/SC4 in this direction.

---

<sup>11</sup> TRANSCRIBER is a program used for audio transcription; <http://www.etca.fr/CTA/gip/Projets/Transcriber/>

## 4. Conclusions

Many institutions still prefer to take a relational database instantiation as the core for their holding. We have described the reasons for choosing XML files as basis in both cases where we are confronted with more complex documents that are increasingly often to be seen as objects in a distributed Internet scenario. We also indicated that we primarily see advantages in the fact that experienced users can exploit the files directly. This is of particular relevance in the Semantic Web era where we assume that intelligent agents will find their way through a domain of related and complex structured documents which ideally will be associated with schemas that refer to data category registries and ontological repositories. Using relational databases as core would always mean to introduce a web-service that exports the database contents. Therefore, we see relational databases as special containers that include optimized representations for specific purposes such for implementing fast searching.

We have shown that XML plays a fundamental role in our archiving work. It is at the center of our representations of complex information about the archive organization and its content and it helps to easily generate different types of presentation formats. We expect that XML will foster the international unification and help us to increase interoperability. However, we have also shown that (1) it is important to design a proper data model before designing an XML-based representation format and that (2) it is a good container for defining the structural elements, but that (3) it does not solve the semantic and interpretation problems. Here data category repositories also applying XML and RDF-based repositories with relational information will emerge.

## References

- [1] P. Wittenburg, W. Peters, D. Broeder (2002), *Metadata Proposals for Corpora and Lexica*. LREC 2002 Conference. Las Palma, Mai
- [2] P. Wittenburg, D. Broeder (2002), *Management of Language Resources with Metadata*. Workshop on International Standards of Terminology and Language Resources Management. Las Palmas, Mai.
- [3] H. Brugman, P. Wittenburg (2001), *The application of annotation models for the construction of databases and tools*. IRCS Workshop on Linguistic Databases, University of Pennsylvania.
- [4] H. Brugman (2003), *Annotated Recordings and Texts in the DOBES project*. EMELD Workshop, East Michigan University.
- [5] S. Bird, M. Liberman (2001), *A formal framework for linguistic annotation*. *Speech Communication* 33 (1,2), pp 23-60
- [6] H.S. Thompson (1997), *Towards a Base Architecture for Spoken Language Transcript{s,tion}*; [www.ltg.ed.ac.uk/~ht/rhodes.html](http://www.ltg.ed.ac.uk/~ht/rhodes.html)
- [7] S. Drude (2002), *Advanced Glossing – a language documentation format and its implementation with Shoebox*. *Int. Workshop on Resources and Tools in Field Linguistics, LREC 2002*
- [8] H. Brugman, O. Crasborn, A. Russel (to appear), *Collaborative Annotation of Sign Language Data with Peer-to-Peer Technology*, submitted for LREC2004
- [9] H. Brugman, P. Wittenburg, St. Levinson, S. Kita (2002), *Multimodal Annotations in Gesture and Sign Language Studies*. LREC 2002 Conference. Las Palma, Mai