

---

## Sprachdokumentation

---

# Documentation of Languages and Archiving of Language Data at the Max Planck Institute for Psycholinguistics in Nijmegen

Daan Broeder, Hennie Brugman & Gunter Senft

### Abstract

This paper illustrates the general procedures linguists and technicians at the MPI for Psycholinguistics (MPIP) have been developing for gathering (cross-cultural and cross-linguistic) data, for processing and documenting these data, and for archiving them. The paper ends with an outlook on future developments with respect to the documentation of (endangered) languages and language archiving.

### 1 Introduction<sup>1</sup>

In this paper we illustrate the basic and general procedures technicians at the Max Planck Institute for Psycholinguistics have been developing for gathering data, for processing and documenting these data, and for archiving them. We first present a tool for data elicitation developed by researchers at the MPI which combines a number of different research interests. Then we use data that were collected with this tool on the Trobriand Islands (Papua New Guinea) to illustrate the procedures just mentioned. Information with respect to the tool and the Kilivila sessions and metadata are publicly available and can be accessed on the web under the following URLs: <http://www.mpi.nl/world/data/fieldmanuals> and <http://www.mpi.nl/IMDI/tools> (for available tools to access the Kilivila sessions).

<sup>1</sup> This paper is based on a talk the authors presented at the "Ringvorlesung "Bedrohte Sprachen": Sprachenwert – Dokumentation – Revitalisierung" at the Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld in February 2003. The authors would like to thank the organizers of the Ringvorlesung and the audience for insightful questions and comments.

## 2 "Staged Events" – a tool for collecting cross-cultural and cross-linguistic data

"Staged Events" is a data elicitation tool designed by Miriam van Staden, Gunter Senft, Nick Enfield, Jürgen Bohnemeyer and Alex Dukers at the Language and Cognition Group at the MPI for Psycholinguistics. This method for data elicitation combines research interests in serial verb constructions, in event typicality, and in event complexity. It is designed to collect descriptions of complex events in order to examine how these are segmented into macro-events, what kind of information is expressed and how the information is ordered in the descriptions. It is also designed to allow comparative cross-cultural and cross-linguistic studies. The tool consists of two tasks:

- 1.) a description and recollection task, designed to elicit
  - elaborate descriptions of complex events for the description task and
  - concise equivalents for the recollection task;
- 2.) a re-enactment task of some of the scenes on the basis of descriptions given in task 1.

Task 1 consists of two sets of video-clips and stills (on DV tape and digitized on a CD). Set 1, a subset of Set 2, consists of 53 clips and 53 stills. Set 2 consists of 86 clips and 86 stills. The video clips depict various scenes with human actors and recognisable objects (for example, an actor fetches an axe and chops wood, an actor bumps into another actor who drops a plate which breaks, an actor plays guitar over his head, scenes from a football (soccer) game, etc.). The clips are arranged in a specific order. Every seven or eight video-clips for the description task are followed by seven or eight corresponding stills for the recollection task. These stills were carefully selected by Alex Dukers from the video-clips. Every still depicts a crucial moment in the event staged in the clip from which it was chosen.

The researcher elicits these data with two consultants, one acts as the addressee who has not seen the clips and stills before, and one acts as the describer who first describes the clips and then the stills. The researcher makes the addressee ask "what happened?" (using a language/culture appropriate phrasing that focusses on the action) and the describer knows that her or his description must be such that the addressee knows what happened. After seven or eight video-clips the researcher presents the stills to the describer and asks him or her to describe from memory which scene the picture belongs to, using the appropriate equivalent of the question "which clip was this?". The task is run on a laptop or on a DV-camera. It takes about 40 minutes per consultant to run set 1 and at least 60 minutes per consultant to run set 2.

The re-enactment task aims to test whether the information contained in the descriptions yielded by the first task is sufficient for a hearer to re-enact the scene correctly, but it is also designed to check which parts of a complex scenario are left to inferences based on 'stereotypicality' of events (for instance, if a

scene is described as “a man throws an apple to a woman” does this imply that the apple is caught by the woman?). This second task requires that the researcher selects one representative description from the data collected during the first task for 14 scenes depicted in the video-clips there. Moreover, the researcher needs some objects (a shawl or cloth, a fruit, a guitar, a chair, a table, a ball) that are necessary to act out the described scenes. The researcher either plays the tape recorded description or reads it out himself to a pair of consultants that are asked to re-enact what they have just heard in this description. Not all scenes require two actors. Then the actors themselves may decide who is the actor. When two actors are required, they may decide for themselves who acts which part.

This task takes about 30 minutes plus optional discussion time. Again, the elicitation session should be video-recorded. As mentioned above, the “staged events” tool, together with other elicitation devices developed for the Language and Cognition group’s 2001 field season can be found on our web-site under the following URL: <http://www.mpi.nl/world/data/fieldmanuals>.

### 3 Data gathering and data processing

In 2001 and in 2002 members of the Language and Cognition Group used the elicitation tool just described to collect data in their various field-sites. The data collection was not problematic at all – even when researchers used the complete set of staged events for their data elicitation purposes. In this paper we use data of the Austronesian language Kilivila to illustrate the procedures of data gathering, data processing, language documentation and language archiving. These data were collected by Gunter Senft on the Trobriand Islands (Papua New Guinea) in 2001. The data can be found on the web using tools from <http://www.mpi.nl/IMDI/tools> and browsing to the Oceania subcorpus from the MPI/LAC corpus.

The video- and audio-taped data were first transcribed and then morpheme interlinearized and translated. This first step in processing the data resulted in word documents that noted the place where the data were gathered (Tauwema, a village on Kaile’una Island), the date when these data were gathered (15<sup>th</sup> of June, 2001) and the names of the consultants (in the example given below this was Moyadi, a man belonging to the Malasi clan, who acted as speaker, and Toyogima, a man belonging to the Lukwasisiga clan, who acted as hearer) involved. The transcription of the Kilivila language data looked as follows (the example represents the data elicited with the staged event No. 8):

(008)

Makena turaki, esakaula ela, ekatukwevivila ema ikota beya.**ma-ke-na turaki e-sakaula e-la e-katukwevivila**

Dem-CP.wood-Dem truck3.-run 3.-go3.-turn.round

**e-ma i-kota beya**

3.-come 3.-arrive here

*This truck, it runs it goes, it turns it comes it arrives here.*

The Kilivila data corpus encompasses 860 event reports and documents more than five hours of data elicitation. On the basis of these transcriptions Markus Kramer and Marc Pippel wrote a Visual Basic programme that converts word documents into Elan readable EAF (Eudico Annotation Format) files. Why this was necessary and how this was done will be explained in the next section.

#### 4 ELAN

The case discussed so far represents a realistic scenario for collection, processing and archiving of annotation data in the domain of Endangered Languages. At the MPI we have been confronted with annotation data stored in a range of file formats. Researchers sometimes use general purpose formats like Microsoft Word (as in the case above), others use specific tools like Shoebox or Transcriber, with their own proprietary formats.

During the process of developing software tools for handling these various forms of data annotations we came to the conclusion that it was necessary to insert a software layer between all these different file formats with their own implicit or explicit models of annotations and the tools we developed for handling them. This software layer made it possible to use the same tool for a range of different annotation documents. We have been working on this project since 1997. The resulting software framework that incorporates a number of tools, is called EUDICO (EUropean DIstributed COrpora); one of the main tools within EUDICO is ELAN, the 'Eudico Linguistic ANnotator'.

ELAN is a tool developed for creation, editing, inspecting and searching complex linguistic annotations of video and/or audio signals. It allows the user to define multiple annotation layers or tiers that are – in principle – independent, but can have complex dependencies. Annotations can be visualized by means of multiple time-synchronized views. ELAN has full support for entry and display of Unicode characters. For a detailed description of ELAN and its manual, look at <http://www.mpi.nl/tools>.



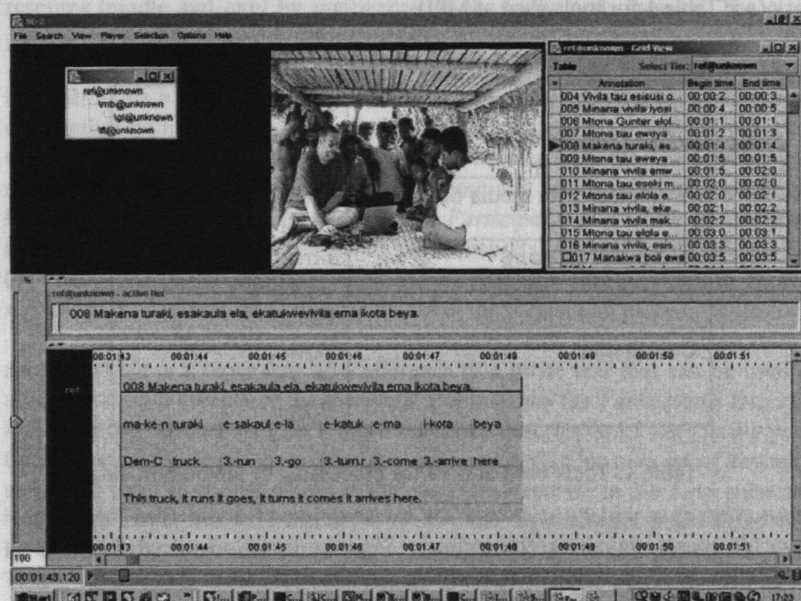


Figure 1: ELAN screen shot of example 008 (section 3 above)

In the case of our Kilivila corpus annotation data was originally collected and processed without using ELAN. Thus, as already indicated above, special processing was necessary to convert the data to an archive format which allowed the exploitation of ELAN's visualization and search capabilities.

First WAC (Word Annotation Converter) was used to convert interlinear texts in Microsoft Word (see the example 008 in Section 3 above) to an intermediate (XML) format. WAC is a Visual Basic script, created at the MPI, that uses a small formal description of the conventions used by the document's author to represent interlinear text to do this conversion. The intermediate format can be directly imported into ELAN and saved to an archive format (see section 6 for a discussion of archive formats). In this process the interlinear structure that is implicit in the vertical alignment of words and other tokens, is made explicit. ELAN itself was then used to attach each Kilivila event report to the correct recorded video scene (see Figure 1).

Depending on data that is already available and the researcher's tool preference many alternative workflows can be used for data processing. Efficiency of these workflows varies widely. Each time the end product is the same: an interlinearized and time-linked annotation document in a proper archive format, in our case EAF (Eudico Annotation Format). For each workflow a number of

basic tasks has to be done, the order and the tool that is used for the task can vary (see Table 1 for tools used at MPI).

Task	Tools used
Orthographic transcription	Microsoft Word
Phonetic transcription	Transcriber ELAN
Alignment with media time	Transcriber ELAN
Interlinear glossing	Microsoft Word Shoebox ELAN
Conversion	WAC ELAN Econv

Table 1: Tools used at MPI for processing of annotation data

*Transcriber* is a widely used tool for transcription of audio files (see: <http://www ldc upenn edu/mirror/Transcriber>). *Shoebox* is an advanced tool that is used a lot in field linguistics to create interlineared texts (see: (<http://www sil org/computing/shoebox/index.html>)). *WAC* and *econv* are conversion tools developed at the MPI (see: <http://www mpi nl/tools>). *WAC* helps converting Word documents that contain interlinear text to EAF format, *econv* does conversions between *Transcriber* and *Shoebox* formats, and facilitates conversion to EAF.

## 5 Metadata for linguistic corpora

Due to technological advancements both in storage technology and in recording and digitizing devices the number and complexity of available language and linguistic resources has increased enormously in the last decenium. The managing of this great mass of data can be facilitated by 'Metadata' descriptions, that is, by descriptions that provide data about data. Our linguistic corpora encompass not only digitised media files and annotation files, but also field notes, photographs, and the like. 'Metadata' describe these resources at different levels and thus provides, for example, information on location and time, information with respect to the language(s) used in the various subsets of a corpus, biographic information for the consultants, information with respect to the genres produced by the consultants, and so on.

First analyses of data especially within the field linguistic domain revealed that the resources to be found there are usually grouped in bundles. Each of these bundles – which we call a "Session" – pertains to one linguistic event or

action. Therefore, in our framework metadata are primarily connected to such a resource bundle and only by implication to the individual resource. The complete set of metadata used is shown in Table 2.

Most of the entries for the metadata speak for itself. The metadata set has as basic structure a division in: (1) General metadata: *Name* and *Title* of the session together with a specification where it was recorded (*Location*). Information on the *Project* and the *Collector* etc. is stored here as well. (2) The *Content* part stores information on what the session is about and offers a fine-grained linguistic categorisation system for this. (3) The *Participants* part stores information about the consultants whose linguistic performance is the subject of the session. (4) Finally the resources (annotations, media files) are administrated in the *Resources* part where we find information about the format and place of these files.

At several places in the metadata structure the user may give a *Description* subpart which is a prose text or a reference to a prose text that can be used to elucidate certain aspects of the metadata. For instance the *Participants.Description* field can be used to give a more complete description of the family relations between the participants. Also the user may define his own set of keyname and value pairs at several places in the metadata structure in the *Keys* substructure. If it is important, for instance, to specify that a participant wears a piercing, he can add the Keyname/Value pair: 'Piercing = true' to the participant part. This is a powerful means for customising the IMDI (ISLE MetaData Initiative) metadata set for specific projects and subdomains. This type of information is (usually) much smaller in size than the resources themselves, and the descriptions are easy to handle and to manipulate.

The technical group at the MPI has been developing an infrastructure framework for managing various linguistic resources. This framework is called 'ISLE MetaData Initiative' (IMDI). IMDI uses these metadata to create browsable hierarchies of corpora and subcorpora. On the basis of these metadata users can explore the various corpora. To do this, a special tool - the IMDI-BCBrowser - was developed. This tool makes it possible to navigate the browsable hierarchies and allows the user to access the various resources. The browsable corpus hierarchy of the MPIP corpora displayed in the IMDI-BCBrowser is shown in Figure 2 with the part relevant for the Kilivila language folded out.

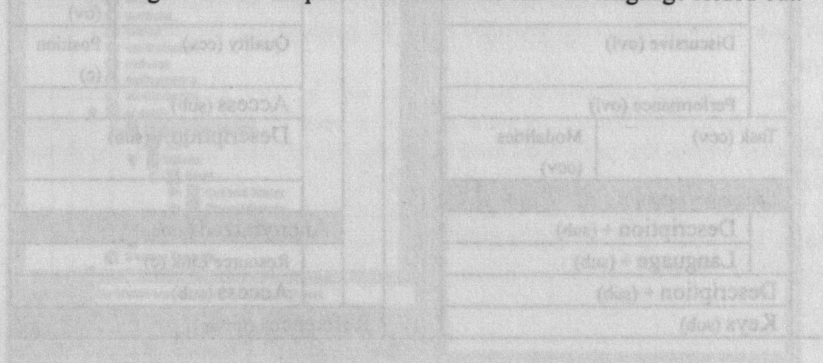


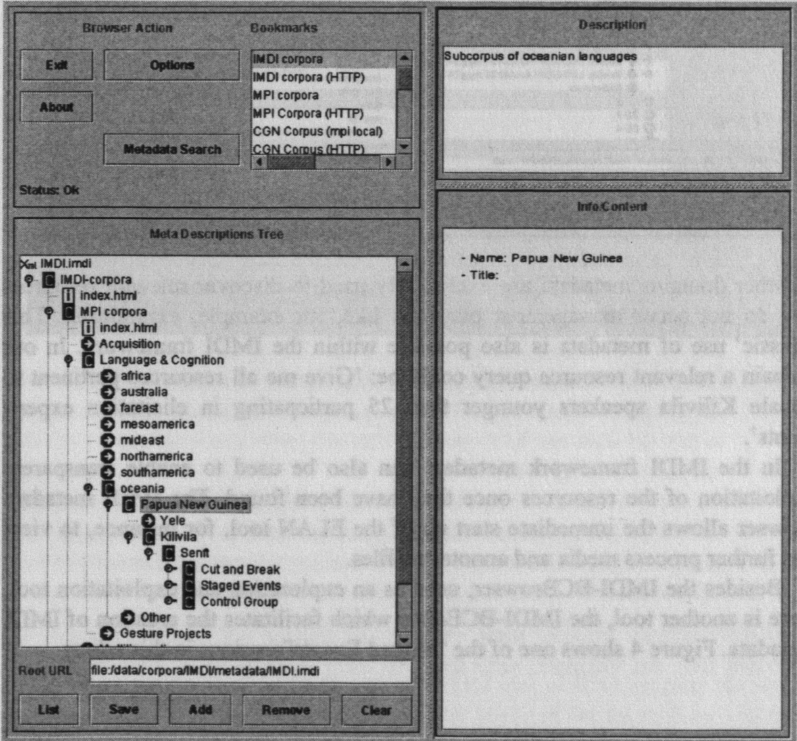
Table 2: The complete set of metadata for describing a 'session'

Session				
Name (string) *		Resources (group)		
Title (string)		Media File + (group)		
Date (c)		Resource Link (c)		
Location (group)		Type (ccv)	Format (ov)	
Continent (ccv)	Country (ccv)	Size (string)	Quality (ccv)	
Region + (string)	Address (string)	Recording Cond. (string)		
Description + (sub)		Position (c)		
Keys (sub)		Access (sub)		
Project (group)		Description + (sub)		
Name (string)	Id (string)	Keys (sub)		
Title (string)		Annotation Unit + (group)		
Contact (group)		Resource Link (c)		
Description + (group)		Media Resource Link (c)		
Keys (sub)		Annotator (string)	Date (c)	
Collector (group)		Type (ov)	Format (ov)	
Name (string)	Contact (sub)	Content Encoding (string)		
Description + (sub)		Character Encoding (c)		
Content (group)		Access (sub)		
Communication Context (group)		Language Id (ccv)		
Interactivity (ccv)	Planning Type (ccv)	Anonymized (ccv)		
Planning Type (ccv)		Description + (sub)		
Involvement (ccv)		Keys (sub)		
Genre (group)		Source +		
Interactional (ovl)	Discursive (ovl)	Id (string)	Format (ov)	
Discursive (ovl)		Quality (ccv)	Position (c)	
Performance (ovl)		Access (sub)		
Task (ocv)	Modalities (ocv)		Description + (sub)	
Languages (group)		Anonymized (group)		
Description + (sub)	Language + (sub)	Resource Link (c)		
Language + (sub)		Access (sub)		
Description + (sub)		References (group)		
Keys (sub)				



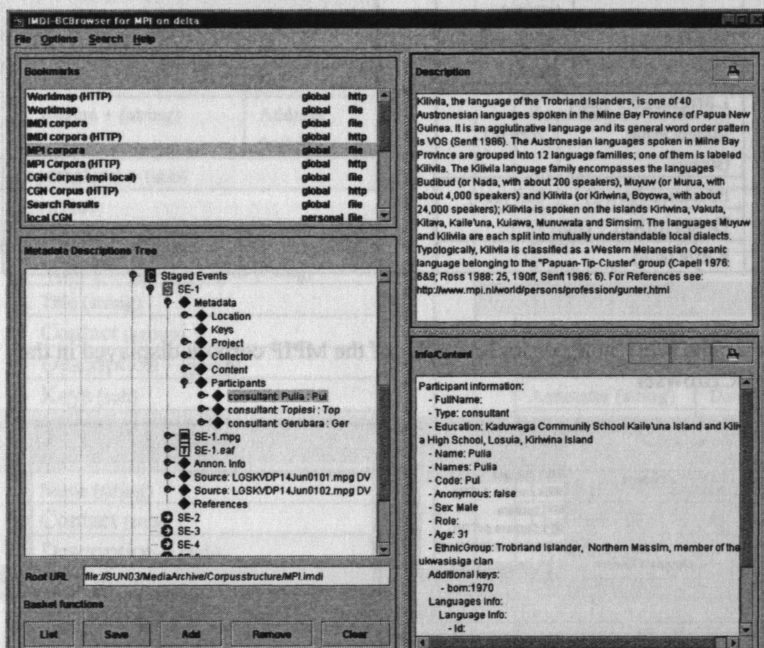
Participants (group)		Description + (sub)	
Description + (sub)			
Participant (group)			
Type (ov)	Role (ov)		
Name+(string)	Fullname (string)		
Code (string)	Anonymized (ccv)		
Language + (sub)			
EthnicGroup (string)	Age (c)		
Education (string)	Sex (ccv)		
Description + (sub)			
Keys (sub)			

Figure 2: The browsable corpus hierarchy of the MPIP corpora displayed in the IMDI-BCBrowser



The Browser can of course also be used to inspect the values of the metadata associated with resources. As an example the metadata content of a Kilivila session is shown displayed in Figure 3.

Figure 3: The metadata content of a Kilivila session



In other domains metadata are exclusively used to discover relevant resources and do not serve management purposes like, for example, exploration. This 'classic' use of metadata is also possible within the IMDI framework. In our domain a relevant resource query could be: 'Give me all resources pertinent to female Kilivila speakers younger than 25 participating in elicitation experiments'.

In the IMDI framework metadata can also be used to enable transparent exploitation of the resources once they have been found. The IMDI metadata browser allows the immediate start up of the ELAN tool, for instance, to view and further process media and annotation files.

Besides the IMDI-BCBrowser, used as an exploration and exploitation tool, there is another tool, the IMDI-BCEditor which facilitates the creation of IMDI metadata. Figure 4 shows one of the "Staged Events" sessions in this editor.

Figure 4:

File View Window Options Help



General Project Collector Content Participants Resources References



### Content

**Summary Content Type**



Task: Staged Events      Modalities: speech  
 Interactivity: interactive      Planning Type: planned  
 Interactional: description      Discursive:      Involvement: elicited  
 Performance:



**Content Type** Description Languages Keys



Task:   

Modalities:   



**CommunicationContext**



Interactivity:   



Planning Type:   




Involvement:   

**Genre**

Interactional:   

Discursive:   

Performance:   

 Clear Content     Open Content     Save Content

## 6 Data archiving and data access

Over the last few years we have been identifying and adopting a number of useful principles that hold with respect to archiving language data or accessing data in such archives (see e.g. Bird & Simons 2003). One of the main tasks of an archive is obviously to ensure that archived data are safely stored and can be adequately accessed for a long period of time. In what follows some of the guiding principles for archiving data and for making them accessible are listed.

### 6.1 Principles for data archiving

The following principles contribute to the long term securing of data

- Make sure that all data are physically duplicated and stored at different places. This principle requests the regular and periodical production of backup copies; however, it may also lead to the setting up of distributed archives.
- Copy data to different media at regular intervals, because media deteriorate over time and media types and formats become outdated in no time these days.
- Host archive data at organizations that can be expected to serve this function for as long as possible.

In general, data stored and documented in specific data formats have a much longer life time than the software tools used to create them. Therefore, knowledge about how to interpret the data should never be placed in tools, nor in programs or scripts that require a special tool to run them. The following principles can be formulated for archiving annotations and metadata:

1. Data should be archived in formats that are well-documented and can easily be validated. At present, this often means that XML (eXtensible Markup Language) is the best choice, and that formats are described by a DTD or schema.
2. If suitable standard formats exist, they should be used.
3. Any archive format should be interpretable by a human reader to ensure that the data can be re-used or that new software tools can be developed for re-using them.
4. As many of the data as possible should be represented explicitly. If information is represented implicitly, for example in the hierarchical structure of an XML file, it should be very obvious what exactly is represented.

Note that formats meeting these criteria are usually not optimal for specific purposes like, for example, searching. In such cases there is no objection to



creating derived special purpose formats, as long as they are not used for archiving or exchanging data.

## 6.2 Principles for accessing data

The following principles, most of which are already implemented in the IMDI, guarantee useful and save data accessing:

- Metadata must be freely accessible to the world (via the internet). This is essential to inform potential users about the existence of a specific resource.
- Metadata should be rich enough to allow potential users of a resource to judge whether or not it is relevant for their purposes.
- Access to resources themselves may be restricted to protect the interests of language communities and researchers. Thus, metadata should reflect under what conditions access can be granted and what procedures need to be followed. Without the researcher's and/or the data producers' explicit permission to use and publish (parts of) the resources, they are inaccessible by default.
- It can be necessary to anonymize real names of consultants, locations etc. in the metadata. To solve this problem a system is implemented in the IMDI that allows a user to define codes for *Full Names*. In the metadata description file these codes are stored and by default every user sees this code instead of the real *Full Name*. Only those users that have the code /*Full Name* mapping defined in their metadata tool environment will see the real *Full Name* automatically translated from the code.
- Any archive should strive for open and easy access. It should allow easy extraction of data for a variety of purposes like, for example, educational matters, research, and language revitalization.
- Ideally, archives should also store primary recordings and make them available to allow the validation of derived data.
- Archives should be open for storing additional data over a long period of time.

## 7 An outlook on future developments

Until now, our efforts concentrated mainly on the area of linguistic metadata and annotations with all the tool and archive construction aspects involved. With respect to lexical tools and data we are still in the analysis and design stage (if we neglect some small pilot projects and our past experience with CELEX). In the near future we would like to also give tool and archive support for lexical data in such a way that we can deal with all kinds of lexica without assuming

any predefined structure or content. It is clear that language resources of different types (like annotation documents, lexica, metadata descriptions etc.) can be dependent in numerous ways. For example, fields in some lexicon can be described by similar types as annotation layers in an annotation document, metadata descriptions can state that information of such a type is available in some resource. Specific lexical entries can refer to instances in some annotated corpus or vice versa. Metadata can describe some participant, some tier in an annotation document can then be associated with this participant. We intend to work on this kind of interlinking between resources of different types. It is also evident that there are huge structural and semantic differences between resources of the same type. Still we would like to compare these resources in an as adequate as possible way. To do this we have to solve the problem of how to map terminologies used and we have to establish adequate repositories for linguistic type information and controlled vocabularies.

Recently, language documentation and language archiving have been developing into domains attracting much interest and many activities by representatives of many different disciplines. We try to monitor all the developments within these domains and hope to continue playing an active participant role in these areas. It goes without saying that standards play an eminent role for documenting and archiving any kind of data. Standards relevant for our purposes are among others standards with respect to

- media formats (e.g. MPEG),
- metadata (DC, OLAC, IMDI), and
- language resources (EAGLES/ISLE, ISO, etc.).

We have been trying to contribute to developing these standards, especially standards that affect the lexicon, metadata activities, and (ways of linguistic and metalinguistic) annotation of the data. Our activities here keep in account user requirements for our tools and for the archives we have been establishing over the years (like, e.g., Spoken Dutch Corpus, European projects like ISLE, MUMIS, Intera, and ECHO, as well as DoBeS). Especially in the field of archiving endangered languages we have been closely collaborating with many national and international partners on a range of issues like tool and service development, archive services, best practices, legal and copyright issues, training courses, etc.

**Abbreviations:**

BCBrowser	Browsable Corpus Browser
CELEX	Center for Lexical Information
DC	Dublin Corpus
DOBES	Dokumentation Bedrohter Sprachen (documentation of Endangered Languages) sponsored by the VolkswagenStiftung
DTD	Document Type Definition
EAGLES	Expert Advisory Group on Language Engineering Standards
EAF	EUDICO Annotation Format
ECHO	European Cultural Heritage Online
Econv	Eudico Converter
ELAN	EUDICO Linguistic Annotator
EUDICO	European distributed Corpora
IMDI	ISLE MetaData Initiative
Intera	Integrated Language Resource Area
ISLE	International Standards for Language Engineering
ISO	International Standards Organisation
MUMIS	Multimodal Indexing and Searching
OLAC	Open Language Archives
WAC	Word Annotation Converter
XML	Extensible Markup Language

**Bibliography**

Bird, S. & G. Simons (2003): "Seven Dimensions of Portability for Language Documentation and Description". *Language* 79, 557–582.

**URLs:**

<http://www ldc.upenn.edu/mirror/Transcriber>  
<http://www.mpi.nl/world/data/fieldmanuals>  
<http://www.mpi.nl/IMDI/tools>  
<http://www.mpi.nl/tools>  
<http://www.sil.org/computing/shoebox/index.html>

Nijmegen Daan Broeder, Hennie Brugman, Gunter Senft

MPI for Psycholinguistics, PB 310, NL-6500 AH Nijmegen, The Netherlands,  
 Daan.Broeder@mpi.nl, Hennie.Brugman@mpi.nl, Gunter.Senft@mpi.nl