CHAPTER ELEVEN

BEST PRACTICES IN THE CREATION, ARCHIVING AND DISSEMINATION OF SPEECH CORPORA AT THE LANGUAGE ARCHIVE[1]

SEBASTIAN DRUDE, PAUL TRILSBEEK,
HAN SLOETJES AND DAAN BROEDER

1. Introduction: Digital Data and the Language Archive

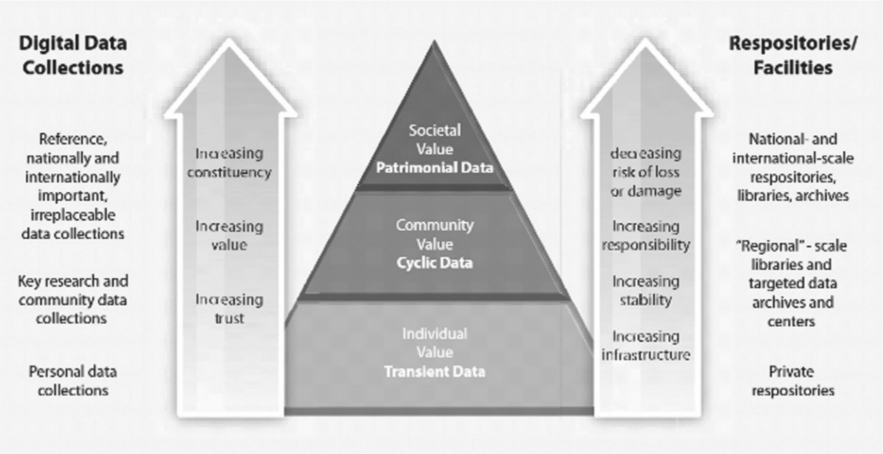
The amount of (digital) data created and used every day worldwide is increasing exponentially. Most of it is transient and of value only for very few people; such data, trillions of individual files, are mostly located on hard drives in personal computers or in places in networks with very restricted access (often local networks, although storing and sharing files "in the cloud" is becoming more and more common as storage costs decrease). Some data, in particular scientific data, are of a more general significance for a larger community, for instance for researchers that share data on a specific topic; such data are often organised in collections and stored for a longer period of time in *(sub)community repositories* (libraries, archives or similar data centres that serve a specific group of individual and institutional users).

But there are also data of national or even global significance, data that are irreproducible,[2] irreplaceable and of a high scientific or cultural value, forming humanity's incipient "digital heritage". For reasons detailed below, such data ought to be well-organized and kept safe in national or international-scale repositories. Although this makes up only a tiny

[1] Major parts of this paper are based on (or are even partly revised versions of) Drude et al. (forthcoming), and partly on earlier work.

[2] This holds mostly for "observational" data as opposed to "experimental" data. The latter can in principle be reproduced, but are often preserved for reasons of accountability.

fraction of all existing and future digital data, it amounts to a huge and also exponentially increasing set of material, posing serious challenges for repositories, which have the challenging responsibility to preserve it from damage or loss and to keep it accessible for future use.[3]


Digital Data Collections
Reference, nationally and internationally important, irreplaceable data collections
Key research and community data collections
Personal data collections
Increasing constituency
Increasing value
Increasing trust
Societal Value Patrimonial Data
Community Value Cyclic Data
Individual Value Transient Data
decreasing risk of loss or damage
Increasing responsibility
Increasing stability
Increasing infrastructure
Respositories/ Facilities
National- and international-scale respositories, libraries, archives
"Regional" - scale libraries and targeted data archives and centers
Private respositories

Source: Adapted from Francine Berman, UC San Diego, in Communications of the ACM.

Figure 11-1: The Data Pyramid—a hierarchy of rising value and persistency

One particular type of data of this third kind are data about (endangered) cultural and linguistic diversity, such as is typically obtained in fieldwork among minority groups around the globe. As a result of the activities in the new field of language documentation (see below), a wealth of data on the use of languages has been and is being produced. These data are not only relevant for further research on the languages in question and their speakers (even for—admittedly, limited—linguistic research when the language is not spoken any more), but also for language revitalization and stabilization initiatives, and for future generations of the speakers themselves. There is reason to believe that the amount of such data will continue to grow, as the value of the cultural and intellectual heritage "encoded" in the many languages is increasingly being recognised, and language loss is more and more often a concern on a national and regional level.

[3] For the "data pyramid" outlined here, see Berman (2008); see also High Level Expert Group on Scientific Data (2010).

Since the late 1990s, the Technical Group of the Max Planck Institute for Psycholinguistics (MPI-PL) has been investigating and building ways to make valuable scientific digital data available for future generations of researchers, speakers and the general public. These activities were strengthened by the participation of this group in the DOBES (DOkumentation BEdrohter Sprachen, "Documentation of Endangered Languages") programme by the German Volkswagen Foundation, where the MPI-PL is the technical centre and data archive for many individual projects that create lasting multi-purpose records of the use of endangered languages around the globe.

The background for language documentation is that the linguists' very research object, languages, are vanishing before their eyes, as they generally realised in the early 1990s: up to 80 per cent of the currently spoken languages could disappear in the coming 150 years or so.[4] As one response, the field of *language documentation* in a new sense (aiming at creating lasting multi-purpose corpora of annotated multi-media samples of small and understudied languages) was established from around the year 2000,[5] notably with the DOBES program (in the following years similar initiatives were implemented, such as the Hans Rausing Endangered Languages Project (HRELP), at the School of Oriental and African Studies (SOAS) in London). TLA's contribution included hosting the emerging DOBES archive, and giving technical support, including software development, for linguistic fieldwork. Since 2000, DOBES has funded around 70 individual major projects on more than 85 target languages. The DOBES funding programme will end around 2016.

The DOBES part of the TLA archive is an important part due to its (generally) high quality and broadness; however, there are data on many more languages in the archive: Metadata mention some 200 languages being used in data samples in the archive. One source consists of field work projects conducted at the MPI-PL outside the DOBES context (in particular in Stephen Levinson's *Language and Cognition* department). But donated corpora of older fieldwork and legacy data are of an increasing importance. TLA offers collaboration in digitizing and archiving important language resources.

[4] See Hale et al. (1992) and in particular Krauss, "The world's languages in crisis". The estimates have changed since then, depending on the criteria which have been refined; see Tsunoda (2004), Wohlgemuth & Dirksmeyer (2005).

[5] This novel use of "documentation" contrasts with the traditional meaning, which refers to the core of linguistics until the middle of the 20th century and the practices of studying and analysing the structure of languages, which we now call "description" (Himmelmann, 1998).

DOBES was a major factor for the on-going development of "Language Archiving Technology" (LAT) at the MPI-PL, which will be explained and detailed below, especially in section 5. Based on this work, TLA has begun to participate in larger infrastructure projects such as CLARIN, aiming at integrating different repositories and (language) technology and infrastructure in Europe and beyond.

This contribution presents the TLA approach for good practices especially in the archiving and dissemination of oral (multimedia) data from fieldwork research. The next section 2 provides information about the different kinds of corpora TLA works with, and about the curation of the same; by contrast, the various types of data files to be archived and of annotations, and a description of the tools used to work on the latter (ELAN and ANNEX) are discussed in section 3; section 4 focuses on the importance of metadata and on how to maintain their long-lasting accessibility; then, in section 5, we present the LAT suite of programs and web services, whereas in section 6 we discuss open access and legal and ethical issues. A short conclusion, in section 7, summarises the most important points.

2. Corpus Design and Curation

The Language Archive itself is not involved in collecting primary data and designing new corpora—this is done by individual researchers or projects with which TLA collaborates as technical support centre and archive. Still, some general comments can be made about corpora at TLA.

Some of the corpora are static (finished, closed), others are dynamic (still being worked on, extended). Among the static corpora that TLA hosts are the Spoken Dutch Corpus (Corpus Gesproken Nederlands), accessed with the specialized CORpus EXploitation software (COREX),[6] and the Dutch Bilingualism Database[7] (DBD). On the other hand the DOBES corpora, for instance, are ideally dynamic corpora where the researchers and more and more often also the speakers themselves make new, additional contributions in the form of more primary data (recordings) and/or secondary data (annotation, either adding more transcriptions and translations to hitherto not annotated material, or by adding or correcting other levels of annotation; see next section). It is also

[6] For more information about the CGN and/or COREX, please refer to http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI86949%23.

[7] For more information about the DBD, please refer to the IMDI Browser website (https://corpus1.mpi.nl/ds/imdi_browser/), and to the corpus node "DBD".

true that in many cases working on a corpus comes to dormant phase, (e.g., after funding runs out or a PhD dissertation has been written). In principle, however, the work on these corpora can always be taken up again.

Because each corpus in TLA is worked on by different researchers, often with different concrete intermediary goals, methods and theoretical backgrounds, different standards are being applied across (and sometimes even within) the corpora. This, of course, poses challenges for achieving (semantic) interoperability, which is needed to search for certain phenomena across different corpora (see next section). This is part of the curation process of the corpora, where TLA is involved in collaboration with the respective researchers.

The DOBES corpora, although created in the framework of one larger research and funding program, are, by design, multi-purpose: this is indeed one of the characteristics and goals of language documentation corpora (Himmelmann, 1998). That is because these corpora will often be the major or the only source of data about a specific language or people, not only for future linguistic research, but also for anthropology, ethno-musicology, -botany, -history, etc.

In linguistics, the corpora serve as the basis for extensive analysis and descriptions of the language system (grammars, dictionaries), and for typological and comparative studies. In elaborating these, the analysis may be refined or otherwise improved over time, which is one of the possible reasons for changes of the annotation and hence, the corpora.

3. Data Types for speech corpora

Three kinds of data objects (files) to be archived can be roughly distinguished: primary data, secondary data, and metadata.

Primary data are mainly audio and/or video recordings[8] of many different kinds of events—from merely observed cultural practices via explanations-while-doing, to dedicated sessions of storytelling or elicitation of linguistic forms. In his seminal paper, Himmelmann (1998) argued that data from the broadest possible spectrum of events should be collected for the documentation of a language to serve multiple uses, some possibly not yet identified.

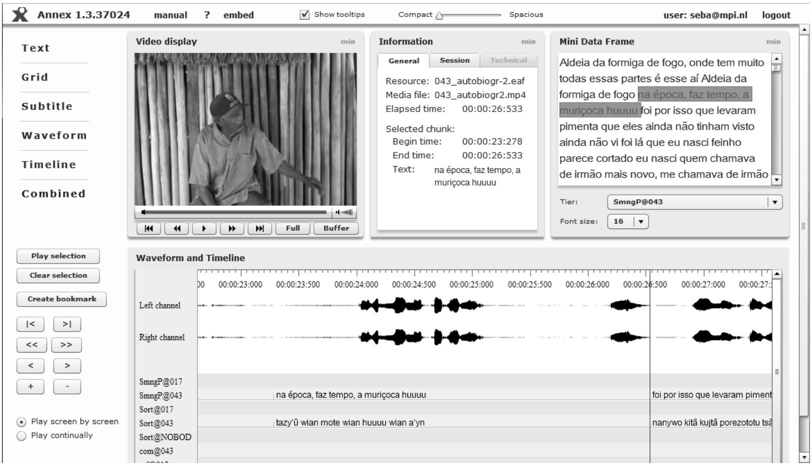
[8] Photographs and drawings or written texts produced by native speakers are other possible types of primary data. We focus here on the typical case of *temporal* primary data (i.e., audio or video recordings).

Secondary data consist of (temporal) annotation of the recordings, that is, they represent explicitly (usually in written or symbolic form) relevant properties of (specific temporal parts of) the recorded event.[9] Different types of properties require different layers of annotation—the basic layers of linguistic annotation are a transcription and a translation into a major language; other well-known types are part-of-speech labels for individual words, or glosses for individual morphs. Many other types of annotation can be added for various purposes, representing different phonetic, morphological, syntactic, semantic and pragmatic properties. Even for a single ontological layer, different types of annotation may (co)exist. For instance, there may be an orthographic, phonemic or phonetic transcription; it may either represent all pauses, false starts and corrections, or it can be "cleaned" to focus on morpho-syntactic structure, and so on.

In the context of language documentation, it is important to keep in mind that a communicative event is much broader than just speech alone—other "paralinguistic" channels, in particular gestures, play an important role and interact closely with the pragmatic and even the linguistic structure. The proper analysis of these phenomena requires again annotation on various levels. Technically, there might be tiers (roughly, lines) of annotation that code the hand shape and its position, the different phases of a gesture, etc., possibly separately for different body parts. In this context it is useful to be able to bundle tiers in a hierarchical structure, which is supported by the ELAN multimedia annotation tool developed at The Language Archive and widely used in language documentation and multimodal and sign language research. The ANNEX tool can be used to display and play specific parts of a recording, for instance one sentence of a text, together with its annotation; see Figure 11-2.

A challenge for the long-term usability and interoperability of such corpora is the huge heterogeneity of annotation conventions. Most projects produce as a first step what may be called "basic annotation"—a transcription, a translation into a major language (if it is not a major language itself), and possibly notes or comments. But beyond that, not only do linguists and other researchers use different sets of tiers to annotate their resources according to their research goals, but they also do not usually agree on the labelling of tiers and on the definition what kind of information each tier is to contain. More often than not, these definitions are not made explicit.

[9] Similarly, for non-temporal primary data, annotation relates to specific parts of the primary data, for instance to words or sentences of a written text which, in turn, relate to specific locations of an image of that text.



Resource: 043_autobiogr-2.eaf
Media file: 043_autobiogr2.mp4
Elapsed time: 00:00:26:533

Selected chunk:
Begin time: 00:00:23:278
End time: 00:00:26:533
Text: na época, faz tempo, a muriçoca huuuu

Aldeia da formiga de fogo, onde tem muito todas essas partes é esse aí Aldeia da formiga de fogo na época, faz tempo, a muriçoca huuuu foi por isso que levaram pimenta que eles ainda não tinham visto ainda não vi foi lá que eu nasci feinho parece cortado eu nasci quem chamava de irmão mais novo, me chamava de irmão

Tier: SmngP@043
Font size: 16

SmngP@017		
SmngP@043	na época, faz tempo, a muriçoca huuuu	foi por isso que levaram piment
Sort@017		
Sort@043	tazy'ũ wian mote wian huuuu wian a'yn	nanywo kitã kujtã porezototu tsã
Sort@NOBOD		
com@043		

Figure 11-2: An ELAN annotation file displayed in ANNEX.

There are only a few unofficial "*de-facto* standards" such as the Leipzig glossing rules[10], which specify conventions for two tiers in addition to the basic annotation tiers. One tier renders the original text (in the case of oral corpora, the transcription) split up into morphs. The other "glossing" tier contains some semantic gloss or functional label for each morph (or sometimes word form). We propose to call annotation consisting of basic annotation combined with such unit-related functional and semantic annotation "basic glossing". In some cases, basic glossing does not rely on morphs but rather on words as their minimal unit of functional and semantic annotation, whereas in other cases an additional line with some part of speech or analogous morphological label is also provided. The latter holds in particular for files where the basic glossing is produced with the help of the Toolbox program,[11] still quite popular among field linguists although for years it has not been officially supported by its producer, SIL International.

Annotation that goes beyond basic glossing can include any number of other (additional) tiers for different purposes, linguistic, paralinguistic (such as gestures, see above), cultural, etc. If glossing is involved at all, such annotation can be called "advanced glossing", or else just "advanced

[10] For further information on Leipzig glossing rules, please refer to http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

[11]For further information about Toolbox, please refer to http://www-01.sil.org/computing/toolbox/information.htm.

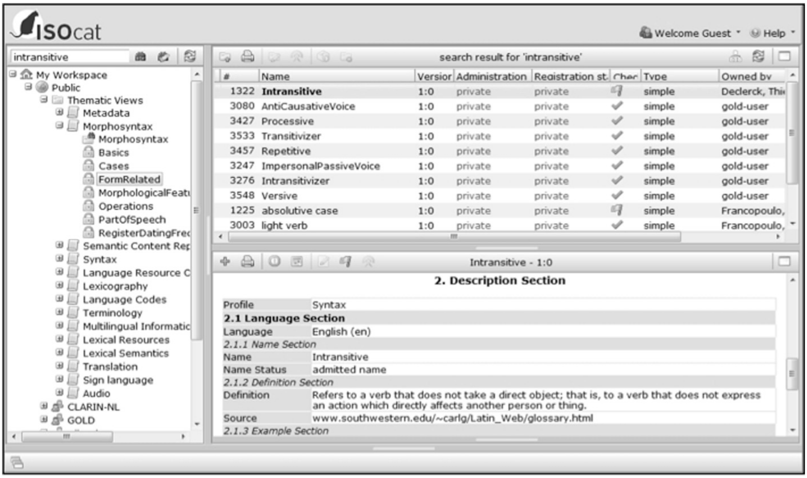
annotation". "Advanced Glossing" is also the name of a concrete proposal for a possibly maximal reference model for annotation on all structural-linguistic levels proposed by Lieb and Drude (2000) in the DOBES programme.

Within the tiers, even if they are analogous and intended to contain information of the same kind, the conventions applied unfortunately again vary a lot, in particular when it comes to terminology for linguistic concepts and their abbreviations. The purpose of the more recent ISOcat data category registry—by itself not part of LAT although also developed by, and hosted at, TLA—is to be a central location where definitions of terms for all areas of linguistics and language technology can be provided so that documents and other resources can refer to them. ELAN allows users to link tiers to ISOcat entries such that the tier's semantics are specified independently from their names. It is also recommended to link abbreviations to ISOcat entries so that their meaning is made explicit in an interoperable way. As such, ISOcat can be a central reference even for different user communities with their specific terminologies by creating "collections" of terms. The GOLD[12] (Farber & Langedoen, 2003) terms have been included in ISOcat by the RELISH[13] project and are available as one such selection.

With a relation registry, in the near future one will be able to define relations between different entries (e.g., the "substantive" in one framework can be very close to equivalent to the "noun" in another framework but still different enough to require two different ISOcat entries); language resources are being prepared for the semantic web (W3C, 2011; Good et al., 2010).

[12] GOLD: General Ontology for Linguistic Description. GOLD Community.

[13] TLA. (n.d.). RELISH. Retrieved on October 14th, 2013, from: http://tla.mpi.nl/relish/. RELISH: Rendering Endangered Languages Lexicons Interoperable Through Standards Harmonisation. Funded by the German Research Foundation (DFG) and the National Endowment for the Humanities (NEH). Goethe-Universität Frankfurt; Max Planck Institute for Psycholinguistics; Institute for Language Information and Technology (ILIT).



search result for 'intransitive'

#	Name	Versior	Administration	Reaistration st.	Chec	Tvpe	Owned bv
1322	**Intransitive**	1:0	private	private	👎	simple	Declerck, Thi
3080	AntiCausativeVoice	1:0	private	private	✓	simple	gold-user
3427	Processive	1:0	private	private	✓	simple	gold-user
3533	Transitivizer	1:0	private	private	✓	simple	gold-user
3457	Repetitive	1:0	private	private	✓	simple	gold-user
3247	ImpersonalPassiveVoice	1:0	private	private	✓	simple	gold-user
3276	Intransitivizer	1:0	private	private	✓	simple	gold-user
3548	Versive	1:0	private	private	✓	simple	gold-user
1225	absolutive case	1:0	private	private	👎	simple	Francopoulo,
3003	light verb	1:0	private	private	✓	simple	Francopoulo,

Intransitive - 1:0

2. Description Section

Profile	Syntax
2.1 Language Section	
Language	English (en)
2.1.1 Name Section	
Name	Intransitive
Name Status	admitted name
2.1.2 Definition Section	
Definition	Refers to a verb that does not take a direct object; that is, to a verb that does not express an action which directly affects another person or thing.
Source	www.southwestern.edu/~carlg/Latin_Web/glossary.html
2.1.3 Example Section	

Figure 11-3: The ISOcat data category registry.

4. Archiving and metadata: guaranteeing lasting availability

Although annotation is usually stored in separate files (at TLA, most often ELAN-XML files), it is important not to forget its relational character: annotation is annotation *of* a recording, and the relation between the primary and the secondary data has to be maintained when these objects are archived. This is one important reason for the existence of *metadata*, the third major type of data object. Another reason is to provide general global (not time-related) information about the recorded event (what, where, when, who, etc.) and properties of the primary and secondary data objects, such as their format, encoding, structure, provenance information, and content, and their digital identification and location in the archive. Metadata play a key role for archiving since it is via metadata that relevant pieces of data can be discovered and accessed later.

In the context of Language Archiving Technology (see next section), a group of interrelated files that are all based on the same event (or on a set of closely related events) is called a *session*. A session is described by a metadata file which itself is also part of the session. Figure 11-4 summarises the components of a typical session and the relations that hold among them.

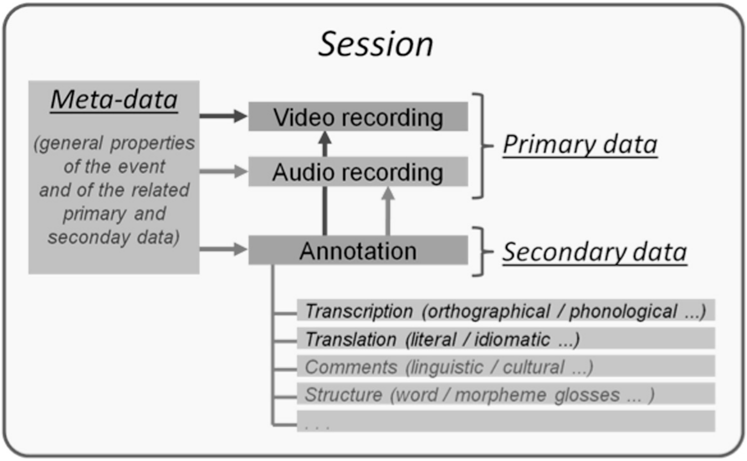
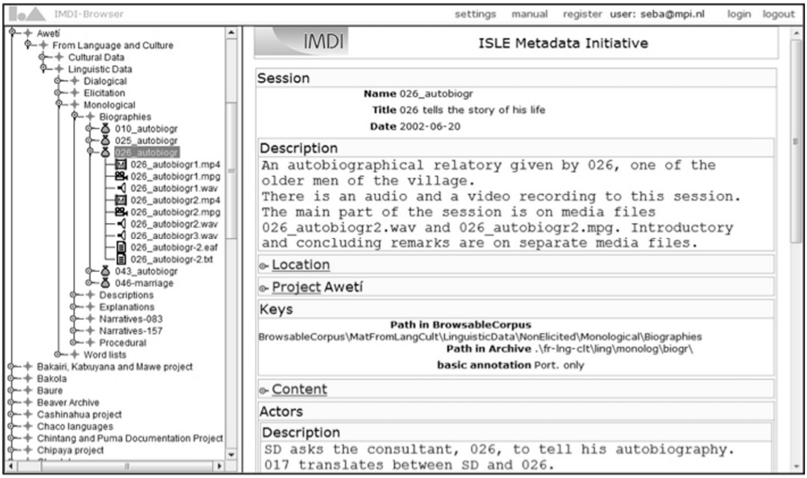

Session
Meta-data
(general properties of the event and of the related primary and seconday data)
Video recording
Audio recording
Annotation
Primary data
Secondary data
Transcription (orthographical / phonological ...)
Translation (literal / idiomatic ...)
Comments (linguistic / cultural ...)
Structure (word / morpheme glosses ...)
. . .

Figure 11-4: Components of a typical session.

Besides such sessions with observational data which account for the lion's share of the archive at MPI-PL, in particular the DOBES archive, there may be other sessions (in the sense of a group of files to be archived, described by one file with metadata), in particular, sessions with *derived* data, such as lexical databases or files containing field notes or even scholarly work analysing aspects of the language structure or culture. But in principle the "session" metadata concept was developed to represent collections of data about a linguistic event, where unity of place, time and action establish the desired granularity level for the sessions. In the context of CLARIN, a more flexible component metadata infrastructure (CMDI) (Broeder et al., 2011) has been developed which can be applied to arbitrary (humanities research) data types, where the notion of "session" has to be replaced by a more general notion such as "data bundle".

The organisation of data objects into sessions and their description by metadata is a first important step for guaranteeing that in the future the data can be located and identified as relevant, which is one aspect of providing lasting access to the data—the primary goal of a repository. Additional organisational measures such as file-naming conventions or grouping sessions into collections, which can be arranged in a hierarchical structure, can be realized by the depositors (usually, the researchers who generated the primary and secondary data). Such structures are also an important means to administer parts of the archive. For instance, for guaranteeing or restricting access to certain kinds of data according to practical or ethical criteria, these settings can be applied to an entire branch of related sessions.

Figure 11-5 shows a section of the metadata of a session in the IMDI-browser.



ISLE Metadata Initiative

Session

Name 026_autobiogr
Title 026 tells the story of his life
Date 2002-06-20

Description

```
An autobiographical relatory given by 026, one of the
older men of the village.
There is an audio and a video recording to this session.
The main part of the session is on media files
026_autobiogr2.wav and 026_autobiogr2.mpg. Introductory
and concluding remarks are on separate media files.
```

Project Awetí

Keys

Path in BrowsableCorpus BrowsableCorpus\MatFromLangCult\LinguisticData\NonElicited\Monological\Biographies
Path in Archive .\fr-lng-clt\ling\monolog\biogr\
basic annotation Port. only

Actors

Description

```
SD asks the consultant, 026, to tell his autobiography.
017 translates between SD and 026.
```

Figure 11-5: IMDI-browser view on an IMDI session in a LAT based language archive.

But there are two other challenges for guaranteeing lasting access to the data which need to be addressed. First, the carriers for digital data (tapes, magnetic hard discs, optical discs, flash media) are all very fragile and even under best circumstances have a limited lifetime measured in years or maximally decades. Second, the same holds true for the standards of formats and encoding, which tend to change quickly in the modern digital world. Both aspects make digital data highly vulnerable—if nothing is done, after just a decade digital data tend to become inaccessible, either physically or logically/structurally.

Therefore, all data have to be copied from one carrier to another before the first one becomes unusable—the more copies of the data exist, the better. For DOBES data, currently five copies of the bit-stream are created automatically at three locations. Also, selected data collections are being returned to the regions where they were recorded (see Broeder et al., 2011). At the end of 2013, the archive at MPI-PL contained some 90 terabytes of data, and some additional 220 terabytes are stored on the archive's servers, but are not described by metadata and their formats have not been checked. These data usually belong to one of two kinds: transient experimental data or still unexplored observational data. The amount of the latter is alarming (and we suspect that this still holds world-wide for large parts of valuable observational fieldwork data on minority languages and cultures). Currently, the archive grows by approx. 15–18 terabytes per year, but due to new data formats (such as lossless compressed mJPEG2000 video data and High Definition video recordings) this amount will increase enormously in the near future.[14]

Depending on archiving capacities and policies, an evaluation process of assessing the value and quality of archival material might be necessary, based on the metadata information and other criteria.

It is also important that all files in the repository adhere to current standards, in particular open and well-documented ones. If new formats emerge and the current ones become obsolete, all concerned files in the repository need to be "migrated"—(i.e., transformed so as to adhere to the new standard). In some cases (in particular, for primary data in lossy compressed formats) this implies a loss or distortion of information (i.e., modification without any actual factual basis, for example by introduction of "artefacts"—pseudo-data that was created by technical processes, such as the increasing granularity created by photocopying from a photocopy). Therefore, for audio-visual material, uncompressed or lossless compressed formats and for textual material, Unicode character encoding and XML-based formats should be used as archival formats whenever possible, and generally, closed, proprietary formats should be avoided. XML is particularly suitable as archival format because it is machine and human readable (W3C, 2008).

In sum, digital data need to be continuously migrated, both at the carrier level as well as at the structure/encoding level. This has to be automated as much as possible in order to minimise the costs and to avoid errors which threaten the integrity and authenticity of the data. As to physical integrity, it is best to use automatic copying to distinct locations according to safe protocols, making use of different software systems. Standards and an organised and planned procedure to ensure the integrity of large data collections are of utmost importance in this endeavour. All these tasks can best (or even only) be performed by large repositories in secure and trusted institutions.

The situation of digital on-line repositories is diametrically opposed to that of traditional archives of paper and other physical objects. The latter try to minimize the access and use of the archived materials, as each

[14] For these aspects, see also Wittenburg et al. (2012).

access entails (the risk of) damage to the archived objects. Digital data, on the contrary, should be “touched” and used as often as possible; not only do they not degrade in quality, but the frequent use encourages that the digital formats are kept up to date, as data in obsolete formats are discovered through usage, and can be migrated to more current formats. On the other hand, handling several versions of the “same” digital object poses new serious challenges to the repositories (for instance, persistent identifiers usually point to the original version of a digital object, but under certain conditions links can be redirected to the latest version, and usually all versions need to be preserved).

The Language Archive at MPI-PL has been built with sustainability and long-term-preservation in mind. It is one of the very few repositories which have an institutional commitment for the bit-stream preservation of the data (by the German Max-Planck-Gesellschaft, for at least 50 years). It uses persistent identifiers (PIDs[15] from CLARIN[16], using the EPIC handle system[17]) to ensure that objects can be cited and recovered even if the infrastructure and location of resources changes. Several local and regional archives worldwide are adopting the LAT infrastructure. Even if the technology is bound to change, new technology will be downward compatible (able to read, convert and/or integrate older formats) and many other independent developments will at least be interoperable with LAT and its successors.

5. The Language Archiving Technology suite and the larger infrastructure for digital data

Basic concepts as “browsable corpus” and the IMDI[18] metadata for multi-modal / multi-media resources and the “linguistic session” had already been developed by the years 2000–2003, but had until then been influenced by large language acquisition corpora such as the European Science Foundation Second Language (ESF SL) corpus and the Child Language Data Exchange System (CHILDES) corpus, or psycholinguistic

[15] CLARIN. (n.d.). Persistent Identifiers (PIDs). Frequently Asked Questions. Retrieved on September 22nd, 2013, from: http://www.clarin.eu/faq-page/268. See also Schroeder (2009).

[16] CLARIN. (n.d.). CLARIN ERIC. Common Language Resources and Technology Infrastructure. Retrieved on September 22nd, 2013, from: www.clarin.eu/.

[17] For further information on EPIC, please refer to http://www.pidconsortium.eu/.

[18] IMDI. (n.d.). IMDI Metadata. Language Archiving Technology. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/imdi-metadata/

experimental corpora. From 2001 on Language Documentation busted the development of additional technology.

The LAT[19] suite now comprises a well-known stand-alone tool for annotating audio and video language use data, ELAN,[20] an online service for creating and accessing lexical resources (see LEXUS[21]), and tools for metadata-based access to resources using the IMDI (in future: CMDI) metadata standard. Metadata can be created with a dedicated editor and now with the ARBIL[22] tool. The archive can be browsed and accessed with the IMDI-browser. The LAMUS[23] tool allows uploading resources to the archive while performing consistency checks as to file formats etc. User and access administration is done with the AMS[24] tool (see also the next section on legal and ethical issues). The resources can be explored online with tools such as ANNEX/TROVA[25] (for multimedia with annotation created with ELAN), LEXUS (for lexical data) and IMEX[26] (for images). Last but not least, and although not officially part of the LAT suite there is the central ISOcat data category registry,[27] which allows defining concepts all resources can refer to. In this way, different

[19] TLA. (n.d.). TLA tools. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/. Max Planck Institute for Psycholinguistics.

[20] TLA. (n.d.). ELAN. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/elan/. Eudico Language Annotator. Language Archiving Technology.

[21] TLA. (n.d.). LEXUS. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/lexus/. Lexus (Online Multimedia Lexical Database Tool): Language Archiving Technology.

[22] TLA. (n.d). ARBIL. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/arbil/.Metadata Editor, Browser & Organizer Tool. Language Archiving Technology.

[23] TLA. (n.d.). LAMUS. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/lamus/. Language Archive Management and Upload System: Language Archiving Technology.

[24] TLA. (n.d.). AMS. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/ams/. Access Management System: Language Archiving Technology.

[25] TLA. (n.d.). ANNEX. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/annex/, TLA. “TROVA” Accessed September 22nd, 2013, http://tla.mpi.nl/tools/tla-tools/trova/. Annotation Exploration tool: Language Archiving Technology.

[26] TLA. (n.d.). IMEX. Retrieved on September 22nd, 2013, from: http://tla.mpi.nl/tools/tla-tools/imdi_browser/. Image viewer tool. (By now part of the IMDI-Browser): Language Archiving Technology.

[27] ISOcat. (n.d.). Data Category Registry. Retrieved on September 22nd, 2013, from: http://www.isocat.org/. Data Category Registry.

terminologies can be made interoperable. In the following paragraphs we give more details on tools which have not been mentioned above.

As outlined in the last sections, the archive at MPI-PL (and in general any LAT-based archive) consists of sessions described by metadata in the XML-based IMDI format. They mostly contain equally XML-based ELAN annotation (EAF) files (secondary data) together with the multimedia (audio, video) recordings they annotate (primary data). The latter has to conform to selected widely used formats (uncompressed PCM for audio, currently for video MPEG compressed data has to be accepted, as the size of uncompressed or even lossless compressed video is still too big, let alone for high definition video).

A session and a single archived media or ELAN file can be referenced to by its PID. The online TROVA service allows searching the content of annotation, even across different tiers (layers of annotation).

Creating and exploring lexical databases (LDs) is the purpose of the LEXUS tool. Currently many LDs in LEXUS have been imported from other tools such as Toolbox[28], and interchange with other lexical database tools will continue to play an important role.[29] Still, differently from Toolbox, LexiquePro[30], FLEX[31] and other lexical tools, LEXUS is based on the ISO standard LMF[32] (Ringersma et al., 2010) for LDs and is designed to provide full multimedia support.[33] Although work on a stand-

[28] SIL International. (n.d.). Field Linguist's Toolbox. Retrieved on September 22nd, 2013, from: http://www-01.sil.org/computing/toolbox/. Toolbox Version 1.5.8: SIL International.

[29] The recently concluded RELISH project improved the interoperability of different lexical resources, in particular LEXUS (LMF) databases and LIFT-compatible databases.

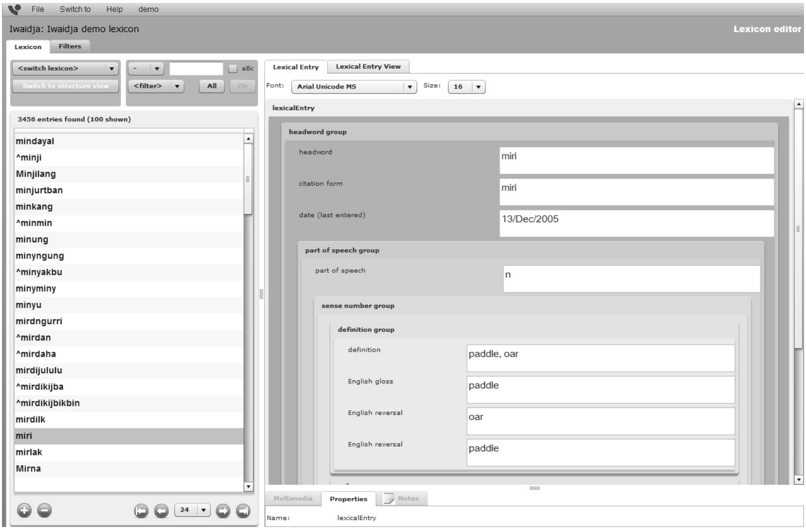
[30] SIL International. (n.d.). SIL Software Catalog. Lexique Pro. Retrieved on September 22nd, 2013, from: http://www-01.sil.org/computing/catalog/show_software.asp?id=92. Tool for Electronic and Online-Dictionaries: SIL International.

[31] SIL FieldWorks. (n.d.). Language Explorer (FLEx). Retrieved on September 22nd, 2013, from: http://fieldworks.sil.org/flex/. Language explorer as part of FieldWorks. Version 3.0: SIL International.

[32] Lexicalmarkupframework. (n.d.). Lexical Markup Framework (LMF). Retrieved on September 22nd, 2013, from: http://www.lexicalmarkupframework.org/ . Version 16, 2008. ISO-24613.

[33] Very recently, the backend of the LEXUS tool has undergone a complete re-implementation, and more major improvements regarding user interface and functionalities are foreseen for the next future. For instance, it is planned to integrate LEXUS with ELAN so that semi-automatic glossing of sentences and texts based on lexical data (minimally as it is known from Toolbox or FLEX)

alone version is making progress, LEXUS is fundamentally web-based so that integration with other tools and resources is easy to achieve.



Iwaidja: Iwaidja demo lexicon

3456 entries found (100 shown)

mindayal
^minji
Minjilang
minjurtban
minkang
^minmin
minung
minyngung
^minyakbu
minyminy
minyu
mirdngurri
^mirdan
^mirdaha
mirdijululu
^mirdikijba
^mirdikijbikbin
mirdilk
miri
mirlak
Mirna

Font: Arial Unicode MS Size: 16

lexicalEntry

headword group

headword	miri
citation form	miri
date (last entered)	13/Dec/2005

part of speech group

part of speech	n

sense number group

definition group

definition	paddle, oar
English gloss	paddle
English reversal	oar
English reversal	paddle

Name: lexicalEntry

Figure 11-6: View on a LEXUS lexicon

TLA has helped a growing number of "regional" archives at several locations worldwide by offering the LAT-suite to be installed at other institutions. These archives form a natural network of endangered language data archives. But TLA also cooperates with other similar initiatives, for instance in the context of DELAMAN (DELAMAN, 2006) and OLAC[34] (TLA participates via the "IMDI to OAI bridge").

TLA is also playing a leading role in different EU projects working on developing e-science infrastructure for the humanities, such as CLARIN (Common Language and Technology Research Infrastructure) and DASISH (Data Service Infrastructure for the Social Sciences and Humanities). CLARIN wants to create a single domain of Language Technology (LT) and Language Resources (LR) for the research community. To accomplish this CLARIN is creating a transparent infrastructure with four major pillars:

becomes possible. Still, this has limited its usefulness in a field work context, and generally the uptake of LEXUS is rather below expectations, so that its further development is not guaranteed at this point.

[34] OLAC. (n.d.). OLAC Misson. Retrieved on September 22nd, 2013, from: http://www.language-archives.org/. Open Language Archives Community, 2011.

(1) a single domain of metadata for LT and LRs using the CMDI technology; (2) using PIDs to refer to resources and services; (3) organizing access to resources and tools using federated identity management;[35] and (4) a set of recommended technology and (de-facto) format standards for LR that will further enhance interoperability between LT services and tools. Currently the CLARIN infrastructure is in its construction phase and the LAT tools will be adapted to be fully interoperable within it.

DASISH is a recently commenced EU project aimed at bringing together the European Strategy Forum on Research Infrastructures (ESFRI) with infrastructure projects, among these CLARIN, DARIAH[36] (Digital Research Infrastructure for the Arts and Humanities, from the wider humanities) and CESSDA[37] (Council of European Social Science Data Archives, from the social sciences). It will likely apply architectures worked out under CLARIN to other domains. For LAT this could be the opportunity for generalising new types of resources, a process already started when LAT needed to be adapted for the resources of the Neurobiology group at the MPI-PL.

Beyond DASISH, TLA is involved in the larger European project EUDAT which aims at creating common infrastructure elements for data from all scientific disciplines, and is even one of the corner pieces in the now emerging Research Data Alliance (RDA). These initiatives try to identify and implement concrete attainable tasks, standards, recommendations or infrastructure elements that help to integrate scientific data from all disciplines around the globe.

6. Open access and legal and ethical issues

Currently open access to research results is highly valued, not just access to the scientific publications but also to the data that form the basis of these publications. The Berlin declaration on Open Access to Scientific Knowledge was first published and signed in 2003 by representatives of most of the German research organisations, but has meanwhile been signed by more than 300 scientific organisations and universities

[35] This means that every researcher will have a single identity with which they can authenticate with all services.

[36] DARIAH-EU. (n.d.). DARIAH-EU. Digital Research Infrastructure for the Arts and Humanities. Retrieved on September 22nd, 2013, http://www.dariah.eu/, Digital Research Infrastructure for the Arts and Humanities.

[37] CESSDA. (n.d.). Council of European Social Science Data Archives. Retrieved on September 22nd, 2013, from: http://www.cessda.org/.

worldwide (Max-Planck-Gesellschaft, 2003). Certainly, there are many good arguments for making the outcome of research that has been funded with public money unrestrictedly available to the public and to other researchers. Giving access to the raw data on which publications are based would in principle allow anyone to verify the claims that were made and would allow the data to be reused for other analyses, saving resources that would go into organizing new recording sessions for example, which in the case of observational data on minority and endangered languages often may well be impossible.

In the case of linguistic observational data, however, the privacy of the human subjects who are recorded needs to be taken into account. Both informed consent and anonymization, often used in other fields, can be somewhat problematic in the field of documentary linguistics. Informed consent about making the data public on the World Wide Web would entail that the subject has a good understanding of what this implies, which cannot be guaranteed in all situations. As to anonymization, one could mask names of persons in file names and metadata and when mentioned in texts and perhaps even in audio recordings, but modifying audio and video recordings up to a point where the participating individuals can no longer be recognised would render them useless for many linguistic purposes. However, recordings in small communities sometimes require the researcher to protect the speakers in order to avoid conflicts within the communities. In the end, it is up to the individual researchers to discuss these issues with the speakers and to make careful decisions, taking both the open access principles and the privacy of the speakers into consideration and trying to establish informed consent to the greatest extent possible.[38] Of course, the access status of individual resources may change over time due to cultural and other necessities or possibilities.

In the DOBES programme, legal and ethical considerations were an important point of discussion from the very beginning. The legal situation in an international context is, however, very complex and no clear and reliable directives can be given even by experts.

Later intensive and serious discussions led to a number of conclusions:

- The DOBES program needs a proper basis to guide the interaction among all persons involved: speakers, collectors, archivists and

[38] There is a rich literature on ethical behaviour in fieldwork, see for instance Dwyer (2006).

users. The result was a Code of Conduct (Max Planck Institute for Psycholinguistics, n.d.), which was amended over the years.

- The roles of all actors in the complex system were defined and the expectations with respect to each actor were formulated. For the archivists it is the principal researcher who is responsible for specifying the access permissions (see below).
- The archive does not claim copyright in the stored material. However, it needs to have the right to archive in order to perform its task in a responsible way. With respect to users the archive will claim copyright on behalf of the data producers.
- No visible logos (e.g., watermarks) will be used in the video since they might obstruct the content.
- The responsible researchers and other designated persons always have access permissions to all material, and they (ideally, in the case of small ethnic groups, including representatives of these communities) can set access permissions for other persons. In particular, members of the speech community should be granted the rights and abilities to access the content.

Handling legal and ethical issues at a responsible level is a serious challenge. For instance, for culture-specific or other reasons, members of the speech communities may withdraw access permissions to certain material even though it was granted at a previous time. On the other hand, after years, necessary restrictions can be withdrawn by the depositor or by representatives of the speaker community. Opening the data as far as legally and ethically possible is generally a requirement, especially when the research was financed with public money. However, scientists and funding agencies or different community members may have different positions. To cope with all kinds of unexpected events a Linguistic Advisory Board consisting of highly respected field researchers was established that can be called upon by the archive to help solve potentially difficult questions.

Over the years, four levels of access privileges were agreed upon. These can be set with the AMS[39] tool, using the standard hierarchical organisation of the sessions—for instance, below a certain node in the "tree", free access can be granted to all audio, but not to the video material, or access to annotation can be limited to a certain user group while primary data are freely accessible. The four levels are:

[39] See footnote 24.

Level 1: Material under this level is directly accessible via the Internet;

Level 2: Material at this level requires that users register and accept the Code of Conduct;

Level 3: At this level, access is only granted to users who apply to the responsible researcher (or persons specified by them) and who make their usage intentions explicit;

Level 4: Material at this level will be completely closed, except for the researcher and (some or all) members of the speech communities.

Access level specifications for archived resources may change over time for various reasons, (e.g.. resources could be opened up a certain number of years after a speaker has passed away, or access restrictions might be loosened after a PhD candidate in a documentation project has finished their thesis).

The number of external people who requested access to “level 3” resources over recent years was not that high; but if requested, access has in almost all cases been granted. We need to see in the future whether the regulations that are currently in place can and should be maintained as explained. Access regulations remain a highly sensitive area, where the technical possibilities opened up by using web-based technologies need to be carefully balanced against the ethical and legal responsibilities which archivists and depositors have towards the speech communities. Despite almost 10 years of ongoing discussions and debate, no simple solution to this problem has yet been found.

Motivated by the Open Access initiative mentioned above, projects and infrastructures such as CLARIN are striving to maximize the number of resources available without access restrictions. Although they accept the need for a class of resources with limited access, they push to require the archives (or data depositors) to make the reasons for this limited availability transparent.

Generally, the key issue for a successful repository is trust. Trust needs to be established not only between the researchers and members of the speech communities, but also between these parties and the repository (the archivists) and ultimately also the users. Even if no easy answers to the intricate legal and ethical challenges can be given, taking these issues seriously and trying to address them is a first important step, (e.g.. by means of user and access management or an appropriate advisory board as explained above).

The usage of the archive is still in a rather initial but developing state. More than 2000 users have an account, and many others have accessed the archive anonymously in 2013. Unfortunately, it has not yet become widespread praxis to cite the data and its hosting institution, so we have as of yet only limited understanding of the usage people do with archive material. We believe that for the time being most users interested in a certain corpus have somehow participated in its compilation—cross-corpus or data-mining studies are only beginning to emerge. This is one of the biggest challenges for the archive, as despite TLA's promotion of standards (successful with regard to data and meta-data formats), annotation practices vary greatly, and over all the corpora at TLA are rather heterogeneous, usually requiring some pre-processing or further annotation before they can be employed in studies by other scientists. Still, showing current usage has now become increasingly necessary in the efforts to raise funds in times when the number of "clicks and downloads" often counts more than cultural heritage value.

7. Conclusion

The LAT suite, although still under constant development and far from complete, is already supporting most stages in the lifecycle of speech corpora, in particular language documentation data, and parts of it are also applied to other tasks. It allows for annotating multimedia recordings, compiling lexical data enriched with multimedia material, grouping bundles of related files into "sessions" and describing these in a single metadata file, and archiving these sessions in structured multi-purpose corpora.[40] Theoretical and technical concepts can be linked to a central concept registry which will allow for interoperability between resources applying different terminologies, as does the development and use of open and well-documented standards.

Such an archive has much better chances of being able to preserve the data over decades by automatically copying the bit-stream and by curating the data and converting it to new standards, guaranteeing its interpretability. In the work of scientists, such archives can serve as a basis for further studies and as a point of reference to illustrate and prove claims about the language structure, the culture or other relevant aspects of derived scientific analyses, making the latter accountable and empirically

[40] Improving the ease of making corrections of and additions to existing corpora so that this becomes possible even for not technology-savvy native speakers and others is one of the most important areas for needed future development.

well-founded. The rights of speakers and scientists have been taken into account as much as possible, and thus it is possible to control and restrict access to certain resources or resources of a certain type.

The work presented in this paper is being carried out at The Language Archive, a new unit at the Max Planck Institute for Psycholinguistics, a promising long-term institutional basis for maintaining and expanding the archive and further developing the associated tools. Other similar and related initiatives exist, and for some aspects other solutions may be applicable instead, or in combination. In the area of endangered language materials, TLA has arguably given an example with high standards, as is evinced by its strong role in DELAMAN, in which other archives participate. Only very few archives at the same time engage in software development (e.g., at PARADISEC, with Nick Thieberger's important contributions, as ExSite9 or EOPAS, cf. Schroeter & Thieberger, 2006).

Although competition between developments may prove more useful, in the long run, than just one solution (which would have to serve too many purposes, and may fail in some aspects, as may happen with any software development), it is important that the different projects and technologies are designed so as to allow for interoperability and cooperation, on the institutional and on the technical levels, and with respect to the content of digital archives. In this sense TLA tries to follow and establish good practice, for instance by applying and promoting standards such as for metadata (IMDI and CMDI, but also delivering to OLAC), practicing state-of-the-art software development, or systematic software testing.[41] Still, only this way the full potential of digital data on the world's linguistic and cultural diversity can be exploited in future.

References

Berman, F. (2013). Got data? A guide to data preservation in the information age. *Communications of the ACM, 51*(12)*,* 50–56. doi: 10.1145/1409360.1409376

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P. & Zinn, C. (2010). A data category registry- and component-based metadata framework. *Proceedings of the LRT standards workshop at the Seventh International Conference*

[41] Each software project has a life cycle itself, and is at first risky—some developments will not be taken up or soon be superseded, or do not arrive at a level of matureness and ease of use to be successful.

on Language Resources and Evaluation held in Valetta, Malta, 19-21 May, 2010. Retrieved on September 22nd, 2013, from http://www.mpi.nl/departments/other-research/research-projects/the-language-archive/tla-presentations

Broeder, D., Sloetjes, H., Trilsbeek, P., Van Uytvanck, D., Windhower, M. & Wittenburg, P. (2011). Evolving challenges in archiving and data infrastructures. In G. Haig, N. Nau, S. Schnell, & C. Wegener, (Eds.), *Documenting endangered languages: Achievements and perspectives*, (pp. 33-54). Berlin: Mouton de Gruyter.

DELAMAN. (2013). DELAMAN. *Digital Endangered Languages and Musics Archive Network*. Retrieved from http://www.delaman.org/

Drude, S., Broeder, D. & Trilsbeek, P. (2011). *The "Language Archiving Technology" solutions for sustainable data from digital fieldwork research. Research Report 2011-2012*. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics. Retrieved from http://www.mpi.nl/institute/annual-reports/archive-annual-reports/BiAnRep_2011_12_MPI_f_PSYL.pdf.

Drude, S., Broeder, D. & Trislbeek, P. (forthcoming). The Language Archive and its solutions for sustainable endangered languages corpora. *Book 2.0*.

Dwyer, A. (2006). Ethics and practicalities of cooperative fieldwork and analysis. In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of Language Documentation*, (pp. 31-66), New York: Mouton de Gruyter.

Evans, N. (2010). *Dying words. Endangered languages and what they have to tell us*. Malden, MA: Wiley-Blackwell.

Farber, S. & Langedoen, T. (2003). A linguistic ontology for the SemanticWeb. *GLOT International, 7 (3)*, 97–100. Retrieved from http://faculty.washington.edu/farrar/documents/article/FarrarLangendoen 2003.pdf.

Gippert, J., Himmelmann, N.P. & Mosel, U. (2006). *Essentials of Language Documentation*. New York: Mouton de Gruyter.

Good, J., Myers, T. & Nakhimovski, A. (2010). *Interoperability for Language Documentation. The Role of Semantic Web Tools*. Unpublished manuscript, University at Buffalo. Retrieved from http://www.acsu.buffalo.edu/~jcgood/GoodMyersNakhimovsky-Interoperability.pdf.

Hale, K., Krauss, M., Watahomigie, L.J., Yamamoto, A.Y., Craige, C., La Verne, M. & England, N.C. (1992). Endangered languages: On endangered languages and the safeguarding of diversity. *Language, 68*, 1–42.

High level expert group on scientific data. (2010). *Riding the wave: how Europe can gain from the rising tide of scientific data*. Retrieved from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics, 36*(1), 161–195.

Kraus, M. (1992). The world's languages in crisis. *Language, 68,* 4–10.

Lieb, H. H. & Drude, S. (2001). *Advanced glossing: A language documentation format*. Unpublished manuscript. Nijmegen: The Language Archive,.

Max-Planck-Gesellschaft. (2003). Open Access to Knowledge in the Sciences and Humanities. Berlin Declaration 2003. Retrieved on September 22nd, 2013, from: http://openaccess.mpg.de/

Max Planck Institute for Evolutionary Anthropology Department of Linguistics. (n.d.). The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses. Retrieved from http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Max-Planck Institute for Psycholinguistics. (n.d.). DOBES Code of Conduct. Retrieved from http://www.mpi.nl/DOBES/ethical_legal_aspects/DOBES-coc-v2.pdf.

Mosley, C. (2010). *Atlas of the World's Languages Danger*. Paris: UNESCO Publishing.

Ringersma, J., Drude, S. & Kemps-Snijders, M. (2010). Lexicon standards: From de facto standard Toolbox MDF to ISO standard LMF. *Proceedings of the LRT standards workshop at the Seventh International Conference on Language Resources and Evaluation held in Valetta, Malta, 19-21 May, 2010*. Retrieved from http://www.mpi.nl/publications/escidoc-446072/@@popup.

Schroeder, K. (2009). Persistent Identifier (PI) - ein Überblick. In H. Neuroth, A. Oßwald, S. Scheffel, S. Strathmann, & K. Huth (Eds.), *Nestor-Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.0* (9.4). Boizenburg: Verlag Werner Hülsbusch, 2009. Retrieved from http://www.nestor.sub.uni-goettingen .de/handbuch/index.php.

Schroeter, R. & Thieberger, N. (2006). EOPAS, the EthnoER online representation of interlinear text. In L. Barwick & N. Thieberger (Eds.), *Sustainable data from digital fieldwork* (pp. 99–124). Sydney: Sydney University Press, 2006. Retrieved from http://hdl.handle.net/2123/1297.

Tsunoda, T. (2004). *Language Endangerment and Language Revitalization*. Berlin: Mouton de Gruyter.

W3C. (n.d.). Extensible Markup Language (XML) 1.0 (Fifth Edition). Retrieved from http://www.w3.org/TR/REC-xml

—. (n.d.). Semantic Web. Retrieved from http://www.w3.org/standards/semanticweb/.

Wittenburg, P., Drude, S. & Broeder, D. (2012). Daten in der Psycholinguistik. In H. Neuroth, S. Strathmann, A. Oßwald, R. Scheffel, J. Klump & J. Ludwig (Eds.), *Nestor-Handbuch. Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme* (pp. 166-184). Boizenburg: Verlag Werner Hülsbusch, 2012. Retrieved from http://www.nestor.sub.uni-goettingen.de/handbuch/index.php.

Wohlgemut, J. & Dirksmeyer, T. (2005). *Bedrohte Vielfalt: Aspekte des Sprach(en)tods*. Berlin: Weißensee.