CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2010; **22**:2354–2368
Published online 28 August 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.1605

Archiving and accessing language resources

Peter Wittenburg*, †

MPI for Psycholinguistics, Wundtlaan 1, Nijmegen, The Netherlands

SUMMARY

Languages are among the most complex systems that evolution has created. With an unforeseen speed many of these unique results of evolution are currently disappearing: every two weeks one of the 6500 still spoken languages is dying and many are subject to extreme changes due to globalization. Experts understood the need to document the languages and preserve the cultural and linguistic treasures embedded in them for future generations. Also linguistic theory will need to consider the variation of the linguistic systems encoded in languages to improve our understanding of how human minds process language material, thus accessibility to all types of resources is increasingly crucial. Deeper insights into human language processing and a higher degree of integration and interoperability between resources will also improve our language processing technology. The DOBES programme is focussing on the documentation and preservation of language material. The Max Planck Institute developed the Language Archiving Technology to help researchers when creating, archiving and accessing language resources. The recently started CLARIN research infrastructure has as main goals to achieve a broad visibility and an easy accessibility of language resources. Copyright © 2010 John Wiley & Sons, Ltd.

Received 13 May 2009; Accepted 30 March 2010

KEY WORDS: archiving; accessing; language resources

1. INTRODUCTION

The Max-Planck-Institute for Psycholinguistics (MPI)[‡] was founded in 1976 to study how the human brain processes language and acquires language skills. Since its initiation it used two main methods to gain deeper insights into our language capacity: (1) on the one hand observations were made by looking into the structure of spoken interaction and into structural phenomena of different languages. (2) On the other hand experimental paradigms were invented to find out details of the mental processing of language. While the first was mainly based on sound and video recordings, the latter included all kinds of signals that could be detected ranging from reaction times on complex stimuli up to the recording of time series, such as eye tracking, motion tracking and more recently brain-imaging.

[*] Correspondence to: Peter Wittenburg, MPI for Psycholinguistics, Wundtlaan 1, Nijmegen, The Netherlands.
[†] E-mail: peter.wittenburg@mpi.nl
[‡] http://www.mpi.nl/

Establishing appropriate theories about the mind's language capacity requires evaluation of processing strategies for different languages and making comparisons. Owing to the distinctive properties, minority languages spoken mostly in non-industrialized areas are of particular interest for linguists. Therefore, MPI researchers have already started working for many years at different places of the world when the awareness grew that many of the 6500 languages still spoken will become extinct in the coming century [1]. In the 1990s this growing awareness led to the DOBES programme on documenting endangered languages starting from 2000[§]. This and other comparable initiatives[¶] led to deep discussions about how to facilitate the linguist's work to create and enrich language resources, how to manage their increasing amount and complexity of data and how to preserve this unique material for future generations. Two of the results of the current cultural migration streams will be a loss of structural language properties and a mixture of cultural semantics within a historically short period of time. However, we can expect that future generations will want to not only discover their roots, but will want to look back to the languages as they are spoken today. The situation with languages and cultures is similar to the situation with plants, the seeds of which are now stored in seed banks when we clearly anticipate the effects of gene manipulation techniques[‖].

The emergence of the Internet and the web brought researchers to the vision of a harmonized and integrated domain of language resources and technology. Of course it would be very desirable for a researcher working on a specific language to allow for integration of the resources about that language stored in various archives into so-called virtual collections—i.e. to create a virtual domain of resources without integration and interoperability boundaries. This vision brought together experts active in various sub-domains in linguistics (field linguists, speech researchers, natural language processing experts, etc.) to start the CLARIN (Common Language Resource and Technology Infrastructure) research infrastructure project[**]. CLARIN is meant to carry out a joint effort to overcome the fragmented situation and to create a persistent infrastructure to enable advanced eScience type of applications. It is understood that the problems are difficult to solve and that it will, therefore, take time to establish a seamlessly operating environment.

We will describe in Section 2 the requirements of the language documentation work, in Section 3 we describe in more detail the components of the Language Archiving Technology (LAT) software suite and in Section 4 the requirements for resource preservation. In Section 5 we will describe the achievements to date and finally address in Section 6 the future directions of the CLARIN research infrastructure initiative.

2. LANGUAGE DOCUMENTATION

Currently approximately 6500 languages are spoken worldwide with an uneven distribution across continents (Africa 2092, Americas 1002, Asia 2269, Europe 239 and Pacific 912)[††], most of them

[§] http://www.mpi.nl/dobes

[¶] http://www.delaman.org

[‖] http://de.wikipedia.org/wiki/Svalbard_Global_Seed_Vault

[**] http://www.clarin.eu

[††] http://www.ethnologue.com/

DOI: 10.1002/cpe

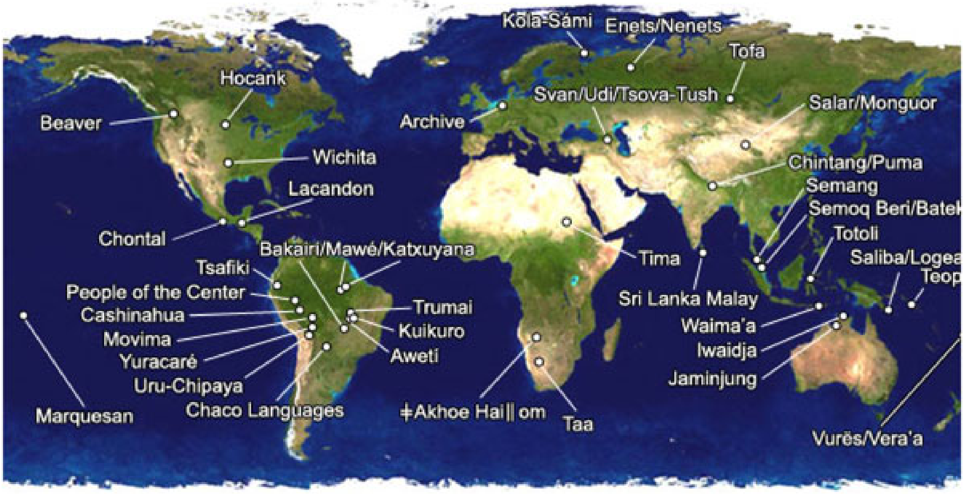


Figure 1. Places where the DOBES teams are operating and which languages they are documenting.

are spoken in areas that are not highly industrialized. Of these languages 96% are spoken by 3% of the people, which may be an indication of the eminent endangerment of most of the languages. Language change is considered normal. However, globalization puts even more pressure on many of these languages with the result that on an average every two weeks one of them becomes extinct.

Linguists are not the only ones to realize that with the death of each of these languages, thousands of years of evolution linguistic, cultural and environmental knowledge is also lost for ever. Therefore, a number of initiatives were founded in the last decade to revitalize and to document languages, the latter often being a premise for the first. The DOBES language documentation programme covers 45 documentation teams working all over the world (Figure 1) and covering mainly linguists, ethnologists, musicologists and ethno-biologists. The MPI is involved in two functions of DOBES: (1) some of its linguists are participating in the documentation work and (2) its technology team is setting up a central digital archive for all material that has been created.

The focus of the linguistic research at the MPI is not on documentation, but on human language processing and the relation between language and thought; much of the linguistic observation work of its researchers is devoted to also study 'special' languages: languages spoken by minorities living in different environmental situations, languages having different linguistic properties and languages as spoken by language learners. Therefore, the observations of the researchers of the MPI often include material about languages that are changing rapidly or will become extinct in a few years.

Almost all these observations, in the DOBES programme as well as at the MPI, are based on multimedia recordings as the primary sources. The reason for this orientation is the knowledge that language is not limited to verbal utterances. In particular, this is true for these 'special' languages and those languages that are spoken in limited environmental and cultural situations. Only combined audio/video recordings will give a comprehensive picture of how languages are used in human interaction and in concrete circumstances.

DOI: 10.1002/cpe

With respect to the linguistic analysis work we can identify some differences. The documentation work within DOBES is focusing on recording language in a balanced way, i.e. including different genres, different types of events, etc., and on describing essential aspects of the language. Therefore, DOBES established some guidelines for all participating teams[‡‡]. (1) For each recording an orthographic transcription and a translation to at least one of the major languages should be provided. (2) Morpho-syntactic glossing should be carried out for some of the recordings so that a reconstruction of the linguistic properties of the language in focus is possible. (3) A sketch grammar should be provided for the same reason as well as a lexicon focusing on the words as anchors for discovering meaning. Some teams also added other annotation tiers, for example music performances, aspects of gestures, semantic phenomena, etc. At the MPI created annotations are devoted to answer specific research questions and therefore they are much less systematic with respect to describing the linguistic properties of a language.

Owing to its technological capacities on the one hand and its research interests on the other hand, the MPI opened its digital online archive for other deposits. An increasing number of researchers are making use of the archive and are integrating their resources. For instance, the archive is now digitizing and integrating the recordings and annotations from Eibl-Eibesfeldt[§§], one of the most famous ethnologists who made comprehensive and unique studies of a number of cultures worldwide.

In total the language archive now covers approximately 47 TB, including more than 60 million of annotations and a large number of lexica emerging from the study of approximately 150 languages. This makes MPI currently one of the largest archives covering unique and non-reproducible cultural and language material.

3. LANGUAGE ARCHIVING TECHNOLOGY

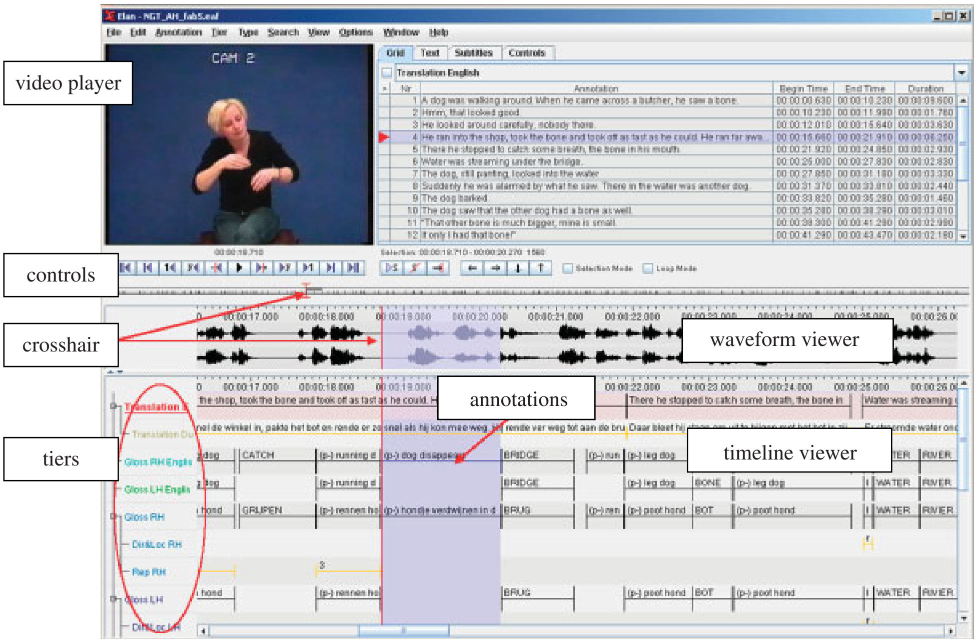
At the MPI we started rather early to create tools that support researchers in annotating multimedia recordings. In 1995 we offered our first multimedia annotation tool called MediaTagger [2], which was mainly used by researchers who studied the synchronization between gestures and verbal utterances and cultural differences in gesturing. The experiences with this first MAC-based tool led us to design and implement the ELAN annotation tool which started in the late 90s and which was continuously improved during that time (see Figure 2) [3].

The emergence of better computer tools and the reduced costs for hardware[¶¶] led to a revolution of linguistic practices. After maturization of MediaTagger we realized that other groups of researchers became interested in digital analysis methods and in using their personal computers as the central place of linguistic analysis. The result was an increasing amount of resources and complex relationships between them. Therefore, a strategic decision was made around 1998 to focus almost exclusively on digital methods and to work out strategies allowing users to better create and manage data. MPI's technology team was focusing on two major questions: (1) which tools do we need to build to improve linguists' efficiency in their daily activities when working with language

[‡‡] http://www.mpi.nl/DOBES/dobesprogramme/

[§§] http://erl.orn.mpg.de/∼fshuman/de/hpeibl.html

[¶¶] The costs per megabyte in 1977 are comparable with the costs per terabyte today.



Translation English

Nr	Annotation	Begin Time	End Time	Duration
1	A dog was walking around. When he came across a butcher, he saw a bone.	00:00:00.630	00:00:10.230	00:00:09.600
2	Hmm, that looked good.	00:00:10.230	00:00:11.990	00:00:01.760
3	He looked around carefully, nobody there.	00:00:12.010	00:00:15.640	00:00:03.630
4	He ran into the shop, took the bone and took off as fast as he could. He ran far awa...	00:00:15.660	00:00:21.910	00:00:06.250
5	There he stopped to catch some breath, the bone in his mouth.	00:00:21.920	00:00:24.850	00:00:02.930
6	Water was streaming under the bridge.	00:00:25.000	00:00:27.830	00:00:02.830
7	The dog, still panting, looked into the water	00:00:27.850	00:00:31.180	00:00:03.330
8	Suddenly he was alarmed by what he saw. There in the water was another dog.	00:00:31.370	00:00:33.810	00:00:02.440
9	The dog barked.	00:00:33.820	00:00:35.280	00:00:01.460
10	The dog saw that the other dog had a bone as well.	00:00:35.280	00:00:38.290	00:00:03.010
11	"That other bone is much bigger, mine is small.	00:00:38.300	00:00:41.290	00:00:02.990
12	If only I had that bone!"	00:00:41.290	00:00:43.470	00:00:02.180

Selection: 00:00:18.710 - 00:00:20.270 1560

Figure 2. Typical screenshot of the ELAN annotation tool with its different types of synchronized information panes.

resources and (2) which infrastructures do we need to create to allow linguists to simply focus on their research work.

Researchers quickly recognized the benefits of working in the digital domain, (1) such as having immediate access to the primary recordings, (2) not being dependent on someone's annotations which often include certain theoretical biases and (3) being capable of improving and enriching their resources at any time based on new insights or new research purposes. While in 1998 the total required storage capacity was on the order of 2 TB this has changed dramatically. Our current storage capacity increases by 12 TB per year. We mention this change in capacity only as a rough indicator, since the complexity of the stored resources (internal structure as well as the relations between the resources) is also increasing rapidly.

This has resulted in the LAT[‖‖], which is meant to support the whole life cycle of language resources as indicated in Figure 3. Three major tools will be supported to allow researchers to create, manipulate, analyze and visualize multimedia annotations (ELAN), lexica with multimedia annotations (LEXUS) [4] and syntactic annotations (SYNPATHY)[***] on their notebooks. These tools support generic models, such as EAF, the XML-schema based ELAN Annotation Format, LMF [5], the Lexical Markup Framework which is a new ISO standard and the TIGER format[†††] for

[‖‖] http://www.mpi.nl/lat
[***] http://www.lat-mpi.eu/tools/synpathy
[†††] http://www.program-transformation.org/Tiger/TigerSyntax

DOI: 10.1002/cpe

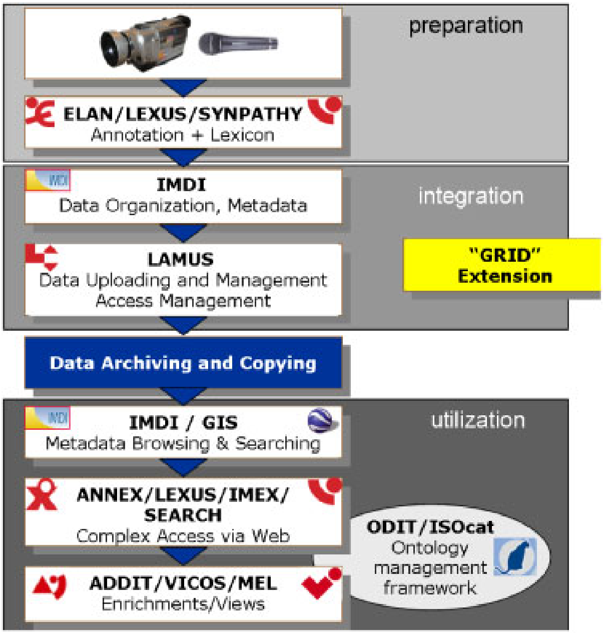


Figure 3. The LAT that was developed to support efficient language resource creation, description, archiving, accessing and linking. LAT is supporting the basic types of work processes, such as multimedia annotation and lexicon creation. While the preparation tools operate locally on personal computers, the other tools are designed as web applications with increasingly often web service APIs. All components are open source and freely available for academic users.

syntax tree descriptions. Knowing that there are many more excellent tools around that are used by linguistic sub-communities, converters were developed to import resources from Shoebox[‡‡‡], CHAT[§§§], Transcriber[¶¶¶] and some XML variants. Since the expressive power of EAF and LMF is greater compared to the legacy formats, the possibilities for backwards conversion are restricted.

With the help of the IMDI Editor[‖‖] researchers can create, manipulate and organize metadata descriptions that are used to manage and access the resources once uploaded into the archive. Here the widely used and openly available IMDI schema [6] and element definition documents are the formal basis. Other IMDI-based tools, such as a native XML browser, allows users to browse in local repositories as well. Recently, a new tool (LINORG) was provided that is meant to help users to graphically organize their material on their personal computer and to simplify the time-consuming metadata creation process. With the help of the LAMUS[****] gate keeper tool [7], resources including metadata organizations and descriptions can be uploaded into the

[‡‡‡] http://www.sil.org/computing/shoebox/
[§§§] http://childes.psy.cmu.edu/manuals/chat.pdf
[¶¶¶] http://trans.sourceforge.net/en/presentation.php
[‖‖] http://www.mpi.nl/IMDI/tools/
[****] http://www.lat-mpi.eu/tools/lamus/

archive. LAMUS takes care of aspects, such as format checks, consistency checks and proper access management, by the user. Since LAMUS is a web-based tool, authorized users can carry out these operations from remote locations, thus supporting decentralized operation.

To support true cyberinfrastructure work, an increasing number of web-based tools have been developed to discover, access, analyze, visualize and relate resources. Web-based versions of IMDI tools can be used to search and browse for resources, also GIS solutions [8], such as Google Earth overlays, can be used to access resources. ANNEX is the web-based cousin of ELAN, which will be extended stepwise to include similar functionality as ELAN[††††]. LEXUS, the lexicon tool, can also be used in the web-environment. IMEX[‡‡‡‡] is a tool that allows users to browse and visualize large photo collections in the archive. The TROVA[§§§§] search engine allows users to carry out fast context searches on structured annotations across the whole archive. TROVA is based on an index that is updated whenever new resources are uploaded into the archive via LAMUS. By combining metadata and content searches, researchers can carry out powerful operations in the parts of the archive to which they have permissions. It is evident that from the search hits, which can be presented in various forms including concordance layouts, one can immediately start accessing the resources.

Recently we started to support a set of new cyberinfrastructure tools that allows researchers to create annotations to any resource fragment and to draw relations between them. While ADDIT[¶¶¶¶] was designed to work with arbitrary fragments, VICOS [9] currently is limited to relate lexicalized concepts contained in LEXUS-based lexica and to bring them into conceptual spaces. These conceptual spaces offer completely new semantically defined views on collections, since VICOS allows users to graphically navigate in these spaces and to directly jump to resource fragments that explain or visualize the concepts and their relations. From collaborative projects including researches, technologists and language community people, we quickly learned that these communities are particularly interested in creating such conceptual spaces to bring concepts into the focus that are central for the culture. As such they can navigate in semantic spaces. Thus, the documentation is not purely linguistics anymore. It also has a cultural dimension. A redesign will allow us to integrate ADDIT functionality into VICOS so that users only need to cope with one tool and user interface to do these kinds of enrichments.

With all the components that are interacting in various ways, LAT is speeding up the time-consuming documentation and scientific analysis work which relies on annotations at various levels to extract and describe linguistic phenomena.

4. LANGUAGE RESOURCE PRESERVATION

One of the most urgent tasks currently is to manage the preservation of all material that is recorded and created by the researchers. The material can be thought of as snapshots of cultures and

[††††] The reason to maintain two separate multimedia annotation tool has to do with the fact that currently only local tools will allow users to create frame accurate annotations.

[‡‡‡‡] Not yet documented.

[§§§§] Part of ANNEX; http://www.lat-mpi.eu/tools/annex

[¶¶¶¶] http://www.lat-mpi.eu/tools/addit

DOI: 10.1002/cpe

languages which are continuously changing, covering even those that will soon become extinct. Therefore, this material is not recoverable and will be a treasure for future generations. Cultures and languages are not the only things that need preservation. It is also obvious that a large part of the created material (recordings, annotations, etc.) are endangered. According to a UNESCO study [10] approximately 80% of the material on cultures and languages created by linguists and ethnologists is stored in a way that it will be lost within the coming decades due to a deterioration of the magnetic substrates of the storage media. In the digital era this does not change, since all our opto-magnetic storage media have a limited lifetime when compared with the Sumerian cuneiforms, for example. Most of the researchers' knowledge about cultures and languages is gathered and stored on personal computers and it is obvious that this material is even more endangered than the primary recordings.

Therefore, the policy within the DOBES programme that all researchers need to give copies of their data to a digital archive is important. It is the task of such an archive to take care of long-term bit-stream preservation. The following picture indicates how we tried to address the problem. Four full copies are stored at geographically distributed computer centres, i.e. together with the own two copies there are in total six copies of the data. In addition, the MPI started establishing 'regional data centres' in a number of countries in close collaboration with local researchers. All data gathered in the respective regions are stored in these centres and they can be seen as nuclei where additional language resources from that area will be collected. Each of these centres is being operated under local control, but in general agreements with these centres guarantee that software maintenance can be carried out by the MPI team and that data are being copied to the Nijmegen centre to make it a part of the preservation machinery. Besides the function as nucleus for data preservation, these regional data centres also have the task of sharing copies of the resources with their places of the origin so that they can be maintained and enriched by the local experts and communities.

It is not only the survival of the bit-streams on some digital carrier that needs our attention, it is also the continuous change in formats that will hamper interpretation after a few decades. Some areas, such as video, are subject to extreme dynamics due to technological innovation. Within two decades we could observe a fast migration between formats ranging from Cinepak[‖‖‖‖‖] to now MPEG4[*****] with the H.264 codec with increasingly complex encoding schemes. While theoretically digital copying is lossless, transcoding between compressed formats can easily amount to the so-called concatenation phenomena. Therefore, archivists were looking for a 'master' format that will be stable for many decades and from which different presentation formats can be created. Recently, the movie industry agreed choosing lossless MJPEG2000 as a persistent format [11]. In addition, some language archives will now follow these trends, since the costs for buying appropriate technology are decreasing.

At the level of 'linguistic formats' much has happened during the last years. A number of generic formats, such as UNICOCE, XML, Lexical Markup Framework and Linguistic Annotation Framework [12], have been worked out in collaboration with W3C[†††††] and ISO TC37/SC4, i.e. they can be seen as reasonably stable. In addition, the data category registry model ISO 12620[‡‡‡‡‡] has been

[‖‖‖‖‖] http://de.wikipedia.org/wiki/Cinepak

[*****] http://de.wikipedia.org/wiki/MPEG-4

[†††††] http://www.w3.org/

[‡‡‡‡‡] http://www.isocat.org/index_bestanden/12620.html

DOI: 10.1002/cpe

adopted by ISO to register linguistic concepts and their definitions. All these developments will help to stabilize linguistic encoding and to facilitate linguistic interpretation hopefully for many decades.

All these measures in the area of bit-stream preservation and format and encoding explicitness will increase the chance of data preservation and interpretability, but they do not offer 100% guarantee. Since archiving is expensive, it is also obvious that coherence, consistency and ease of use of an archive will also influence data survival. Coherence and consistence of an archive with respect to formats, encoding principles and organization will make future transformations much cheaper, and ease of use will motivate communities and politicians to more easily provide necessary funding. The MPI language resource archive exhibits a high degree of coherence and consistency, since we always try to immediately transform existing data sets into the agreed standard formats, despite comparatively high costs. For example, transforming lexica created with Microsoft Word® in an unconstrained way sometimes requires a great deal of interaction between the involved researchers and technologists, the development of several smart scripts and several manual corrections. With respect to the ease of use of the archive, there is still much work to be done. For example, creating costly virtual exhibitions, technologists and archivists need to collaborate with design specialists, researchers and/or the language communities.

Without calculating the costs for the curation and integration of new resources, we can specify the costs for maintaining a complex digital archive as described above. Summing up the costs for local storage system migration (80 k€), for copies at two large computing centres (∼10 k€), for system and archive management (140 k€) and for maintaining the repository software (60 k€), resulting in annual costs of approximately 290 000 €. Of course economy of scale factors is valid. An even higher amount of language resources can be managed by one existing archivists if the archive is in a proper state.

5. ACHIEVEMENTS

In particular, the DOBES programme has pushed researchers, technologists and archivists to cross some barriers. Researchers had to accept that it is no longer appropriate to wait for 'final results' of their annotation or lexicon creation work before handing them over to an archive. Owing to various factors the products of linguistic work are continuously changing. Errors are being corrected, extensions are being made, new theoretical insights may require the revision of parts, etc. Researchers now understand that it is better to think in versions and to hand over the first 'stable' versions to an archive at an early stage so that others can use it and contribute to it. They have also started looking at the archive as a well-organized place compared with their personal computers which often contain a 'creative chaos' of resources.

It is now widely accepted that the resources need to be stored in a neutral and atomic form in an archive to guarantee separate accessibility and unbiased interpretation. It was a learning process for all participants to understand that the primary goal of an archive cannot be to create nice 'exhibitions' that combine resources into guided tours and to store them in presentation formats, since these exhibitions are biased in many dimensions. Although researchers like to think in terms of tools instead of formats, since the tools determine user friendliness and efficiency of work, there is now an increasing awareness for the need to adhere to certain standards, i.e. to prefer tools that support these standards. Here we can still see a large gap between our insights and what major software companies are producing.

We also indicated a change with respect to the willingness to share data and give other researchers access to often incomplete resources. Of course the language communities should be able to access 'their' material, but also cross-linguistic studies and interdisciplinary studies should be facilitated. The former view that the created data are the property of the responsible researcher changed stepwise. Yet much data are not openly accessible, for the purpose of ensuring privacy of the recorded persons, protecting access to recordings of ritual ceremonies, etc.

High-quality metadata creation remained a serious problem until recently. Four main reasons are responsible: (1) the creation of metadata descriptions is costly and those costs increase the longer people wait to create them; (2) metadata descriptions are primarily meant for others to find useful data and for organizing archives; (3) available tools were certainly not optimal and (4) the classification systems underlying certain metadata elements are not widely agreed.

We ran a statistics on 27 000 metadata descriptions [13] and found that language name, which is not standardized, was always filled in (100% compliance). In contrast the language ID, which is standardized and where researchers can do a lookup in the Ethnologue table offered as pick-up list in the tools, was only filled in 40% of the time. It cost the archive managers a lot of curation effort to unify the language names and to associate correct language IDs with the resources. Elements, such as 'genre', where there is no broad agreement were only filled in for 30% because the proposed vocabulary is not widely accepted.

Recently some new developments have helped change the negative image of metadata. In the DOBES programme more and more teams want to provide community-specific access portals to the archive. The team responsible for the documentation of Beaver[§§§§§] created a design for such a community portal in close collaboration with the people of the tribe. They wanted to have an interface that easily allows community members to select resources based on culturally important concepts. All participants realized that in a dynamic archiving situation with continuous enrichments, as it is true for digital archives, fixed links would not lead to stable and up-to-date presentations. Proper metadata descriptions and the execution of real-time queries will deliver better results. The success of such an approach as indicated in Figure 4 is convincing; an increasing number of researchers who properly filled in metadata will be benefitted for their own goals at the end.

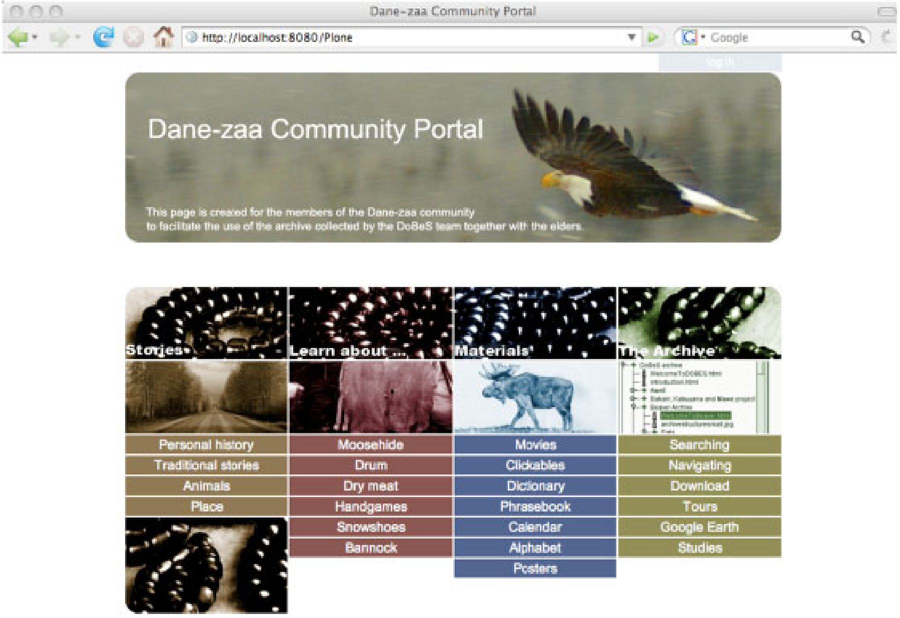
6. CLARIN INFRASTRUCTURE

Still a number of aspects have not yet been tackled within the DOBES/MPI framework, since they require collaboration between various archives. A few examples of activities that are currently not possible may illustrate the gaps in service provisioning:

- efficient searches in joint catalogues and joint archives
- building of virtual collections crossing institutional boundaries
- integration of lexical information seamlessly when working on texts
- easy alignment of transcriptions with sound files.

[§§§§§] http://www.mpi.nl/DOBES/projects/beaver

DOI: 10.1002/cpe



Dane-zaa Community Portal

http://localhost:8080/Plone

Dane-zaa Community Portal

This page is created for the members of the Dane-zaa community
to facilitate the use of the archive collected by the DoBeS team together with the elders.

Figure 4. The Beaver community portal as designed by the researchers in collaboration with the community members. It offers a more attractive view on the material stored in the archive since access is organized by important cultural concepts. When one of the items is selected a metadata query is carried out in real time to generate the proper and actual hit list.

The CLARIN research infrastructure initiative funded by the European Commission and a number of European member states want to tackle these aspects in collaboration with initiatives, such as ISO TC37/SC4[¶¶¶¶¶], TEI[‖‖‖‖‖] and others. Currently 148 institutions from 32 European countries support the CLARIN goals.

Its mission is to create an integrated and interoperable domain of language resources and technologies that can offer its services to all interested researchers in a persistent way. This is compliant with John Taylor who pointed out that 'eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it'. Thus, advanced research methodologies as indicated by the above-mentioned challenges can only be carried out when we manage to build new infrastructures that overcome the huge fragmentation of our field. This will not be easy and will only result in a stepwise process of improvements.

The CLARIN work has a number of dimensions that need to be tackled in parallel: (1) we need to get a deep and comprehensive understanding of the characteristics of language resources and technologies. (2) We need to understand how researchers from different disciplines, in particular from humanities and social sciences, work with textual material, since CLARIN's services should

[¶¶¶¶¶] http://www.tc37sc4.org/

[‖‖‖‖‖] http://www.tei-c.org/index.xml

be useful to all those interested. (3) We need to understand the technological aspects of creating an integrated and interoperable domain. (4) We need to simplify and harmonize the aspects of licensing, IPR and ethical issues to make it feasible for researchers to operate in virtually integrated domains. (5) We need to train and educate researchers to create appropriate resources and tools and to make use of new compliant resources—one of the most important aspects to prevent failure. (6) We need to define an appropriate organizational and funding model for a persistent service provider federation, which will be CLARIN at the end. CLARIN has taken up all of these challenges and has until the end of 2010 showed feasibility. Even so, some countries have already indicated funding support up to 2014.

With respect to the technological aspects of a research infrastructure, such as CLARIN, a few dimensions of work have been identified and first requirements specification documents have been worked out[******]:

- The basis of all persistent integration and interoperability work is a number of interconnected and collaborating centres that receive long-term funding support. For various reasons, in particular, since we can assume that national governments will be willing to support the languages spoken within national borders, we will have several such centres and not just one cloud computing centre. Economy of scale factors will mainly be considered when it concerns centres that will give services that are relevant for operating the European infrastructure itself. It is obvious that all these centres need to establish new business models that facilitate deposits from researchers and that do not establish new administrative hurdles for the depositors and his collaborators.
- These centres will form a service provider federation that agrees on a limited number of license models and defines the requirements for making trust agreements with the emerging national identity federations. Similar to the way publishers act, the Service Provider Federation will determine the attributes that need to be exchanged to create a domain of trust. In doing so a single identity and single sign-on domain is being created. This will provide ability for allowing users to build virtual collections crossing institutional boundaries.
- Another important point is to offer a service for registering and resolving persistent identifiers so that links to all resources and resource fragments remain reasonably stable. Collection building, semantic linking between resources and many other activities of researchers are based on the assumption that the defined typed relations are persistent. A service has been set up which is based on the Handle System[††††††], currently the most performant and robust system, which in contrast to DOI[‡‡‡‡‡‡], is not associated with an unsuitable business model for the millions of needed identifiers.
- More interesting for the researchers is the creation of a joint metadata domain—a 'virtual language resource and technology observatory'. The basis for these activities is making all components visible and creating proper descriptions supporting searching and browsing. The linguistic discipline has experienced now for more than a decade with metadata sets and

[******] http://www.clarin.eu/specification-documents
[††††††] http://www.handle.net/
[‡‡‡‡‡‡] http://www.doi.org/

DOI: 10.1002/cpe

schemas, such as DublinCore[§§§§§§], OLAC[¶¶¶¶¶¶], IMDI and TEI header elements. Based on the knowledge that strict schemas and restricted terminology will not be accepted by the community and to prevent the coverage which we want to achieve in CLARIN, a component-based model was designed and is currently being implemented. This model allows users to create their own components and profiles with the help of elements taken from accepted concept registries. The number of registered schemas could theoretically become infinite. A basis for interoperability is the requirement of re-using elements from concept registries, such as DublinCore and ISOcat[‖‖‖‖‖‖] , as it has been created by ISO TC37/SC4. Now tool builders need to be convinced to support these registries.

- The most difficult area is the design and development of what is called simply a scalable 'Service Oriented Architecture'. We still live in an era, where most people first download resources and tools and convert and adapt them so that they can carry out all operations locally. For some operations this may be the most suitable way of acting for the time being. However, it prohibits the users from taking benefits of real cyberinfrastructure applications and all its capabilities. We want to enable users to easily build chains of operations with the help of workflow tools that allow selecting resources and services from the catalogues and operate on/with them. Building an infrastructure where the different components will interact will not be easy. Machine-readable registries as described above need to be available and supported by the various tools, wrappers need to be provided to encapsulate services in such a way that they amend the registries contents for each annotation incrementally added, and the huge problems caused by the non-harmonized import/export formats and tag sets need to be overcome.

CLARIN follows a two-sided approach: (1) since the interoperability problems for workflow chains are not very well understood and have not been analysed in a more comprehensive manner, we study a number of well-known processing chains—short and long ones—analytically. In doing so we want to get a deeper understanding of how to do harmonization and define and standardize pivot formats about the type of converters needed and variation in the tag sets used. (2) In parallel, we currently implement a number of such workflows in a bottom-up-driven way in order to understand the practical issues that need to be solved.

Standards will play an enormous role in defining a successful infrastructure. In particular, in collaboration with ISO TC37/SC4 we are working mainly on two lines: (1) we are working on generic formats for important linguistic data types. For lexica the Lexical Markup Framework has been worked out and has the potential to act as pivot format. At the MPI some converters for known formats to LMF have already been implemented. With respect to annotations two suggestions are being discussed: Linguistic Annotation Format and Graphical Annotation Format [14], both having very much in common. (2) The standardization work on the model for a 'data category registry', which is actually a flat list of widely agreed linguistic concepts has been nearly finished, i.e. the model ISO 12620 can be seen as stable now. The ISOcat service is currently being implemented to support this model and a long list of categories has been already entered to act as suggestions for being integrated as reference categories. A process has been defined to formally refine and adopt

[§§§§§§] http://dublincore.org/

[¶¶¶¶¶¶] http://www.language-archives.org/

[‖‖‖‖‖‖] http://www.isocat.org/

definitions. The community does not expect that all linguists will accept the concepts as defined in the registry, but it is expected that an increasing number of linguists will refer from their schemas to ISOcat categories and in doing so cater for semantic interoperability. It is obvious that standards need to be supported by good and simple tools to be successful. Some tools, such as LEXUS, supporting LMF already allow users to re-use concepts from the ISOcat registry by supporting the web APIs.

7. SUMMARY

Most of the linguistically oriented support work at the Max Planck Institute emerged from the documentation of endangered languages and the study of psycholinguistic phenomena in ‘special languages’. Providing more efficient tools similar to those offered by high-tech companies, by the LAT and by other smart software developers not only motivated researchers to turn over completely to the all digital world, but led to an enormous speeding up in language resource creation and interlinking. As a consequence, the MPI’s technology team needed to invest time to develop professional resource management and archiving strategies and tools to free researchers from this task and to allow archive managers to work efficiently. Although long-term preservation of bit-streams cannot be guaranteed, we can improve the chances of data survival by distributing the resources to a number of data centres each of which applying migrations to state-of-the-art storage technology at regular time intervals. In addition, the MPI team has set up a number of regional repositories to increase the chance that non-digitized material will be uploaded in order to facilitate community enrichments.

To facilitate a true cyberinfrastructure type of work, the LAT software includes various components to operate on the archived data via web applications, including tools to create and visualize conceptual spaces. These components have the potential to allow users to build virtual collections where the resources are physically stored at different archives. At this time too many integration and interoperability hurdles need to be overcome to make this a feasible option. The domain of language resources and technology is highly fragmented along many dimensions. The MPI participates actively in the CLARIN research infrastructure initiative to help overcome the fragmentation. It also participates in various standardization initiatives, such as ISO TC37/SC4. The final goal must be to create more optimal research environments for the researchers working on the big research challenges in a stepwise manner. To date we do not understand all requirements and the optimal ways to achieve for example semantic interoperability, but this does not prevent us from taking concrete steps now to find solutions for tomorrow.

REFERENCES

1. David C. *Language Death*. CUP: Cambridge, 2000.
2. http://www.mpi.nl/world/tg/lapp/lapp.html [2009/2010].
3. Crasborn O, Sloetjes H, Auer E, Wittenburg P. Combining video and numeric data in the analysis of sign languages within the ELAN annotation software. *Proceedings of the Second Workshop on the Representation and Processing of Sign Languages*: *Lexicographic Matters and Didactic Scenarios*, Vettori C (ed.). ELRA: Paris, 2006; 82–87.
4. Ringersma J, Kemps-Snijders M. Creating multimedia dictionaries of endangered languages using LEXUS. *Proceedings of INTERSPEECH 2007*, Antwerp, 2007; 1529–1532.

5. http://www.lexicalmarkupframework.org/ [2009/2010].
6. Broeder D, Wittenburg P. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata*, *Semantics and Ontologies* 2006; **1**(2):119–132.
7. Broeder D, Claus A, Offenga F, Skiba R, Trilsbeek P, Wittenburg P. LAMUS: The language archive management and upload system. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC 2006*), Genoa, 2006; 2291–2294.
8. van Uytvanck D, Dukers A, Ringersma J, Trilsbeek P. Language-sites: Accessing and presenting language resources via geographic information systems. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC 2008*), Marrakech, 2008. CDROM.
9. Zinn C, Ringersma J, Cablitz G, Kemps-Snijders M, Wittenburg P. Constructing knowledge spaces from linguistic resources. *CIL 18 Workshop on Linguistic Studies of Ontology*, Seoul, 2008. CDROM.
10. Schuller D. Safeguarding the Documentary Heritage of Cultural and Linguistic Diversity. Available at: http://www.mpi.nl/LAN/issues/lan_03.pdf [2009/2010].
11. Trilsbeek P, Wittenburg P, Schafer R, Pavuza F, Schuller D. Video encoding and archiving in field linguistics. *IASA 2008 Conference*, Sydney. Available at: http://www.iasa2008.com/abstract/11.asp [2009/2010].
12. Ide N. *Linguistic Annotation Format*. Available at: http://www.cs.vassar.edu/∼ide/papers/ide-romary-clergerie.pdf [2009/2010].
13. Klassmann A, Offenga F, Broeder D, Skiba R, Wittenburg P. Comparison of resource discovery methods. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC 2006*), Genoa, 2006; 113–116.
14. Ide N, Suderman K. *GrAF*: *A Graph-based Format for Linguistic Annotations*. Available at: http://www.cs.vassar.edu/∼ide/papers/LAW.pdf [2009/2010].

DOI: 10.1002/cpe