Chapter 3
Evolving challenges in archiving and data infrastructures

Daan Broeder, Han Sloetjes, Paul Trilsbeek, Dieter van Uytvanck, Menzo Windhouwer and Peter Wittenburg

1. Introduction

Increasingly often research in the humanities is based on data. This change in attitude and research practice is driven to a large extent by the availability of small and cheap yet high-quality recording equipment (video cameras, audio recorders) as well as advances in information technology (faster networks, larger data storage, larger computation power, suitable software). In some institutes such as the Max Planck Institute for Psycholinguistics, already in the 90s a clear trend towards an all-digital domain could be identified, making use of state-of-the-art technology for research purposes. This change of habits was one of the reasons for the Volkswagen Foundation to establish the DoBeS program in 2000 with a clear focus on language documentation based on recordings as primary material.

The fact that more and more data is being collected poses some challenges for those who are dealing with this data in one way or another. The researcher who collects the material will need to maintain a coherent administration of all the relevant bits of contextual information surrounding the data. These "metadata" descriptions (see Section 4.2) are not just for the researchers own use but should also allow others to find the data once it has been stored in an archive and should allow others to assess whether the data suits their needs. Research data archives that are storing more and more large data collections will have to provide proper facilities and guidance for potential users of the data to find what they are looking for.

While technological advances have made it much easier to collect large amounts of audiovisual recordings, the automatic extraction of the relevant bits of information from these recordings is still very difficult and therefore needs to be done manually to a large extent. This causes a discrepancy be-

tween the amount of data that is being collected and the amount of data that ends up being analyzed and used to support research hypotheses.

Data archiving and sharing is currently on the agenda in all areas of science and the technical frameworks that are being developed are often based on the OAIS reference model (CCSDS 2002) that was originally designed for space data but can be applied more broadly. Different workflows and usage scenarios and differences in the nature of the archived data often require deviations from this abstract model though, in particular in the case of an online archive that gives users direct access to the archived material.

2. Issues and strategies in data handling

Since digital technology quickly offered ways to not only create large amounts of primary recordings but also several associated resources such as transcriptions, linguistic analyses, field notes, etc., it became obvious that new challenges appeared at the horizon: we needed ways to take care of proper lifecycle management of the archived data. In 2000 the MPI stored about one terabyte of digitized recordings, currently the data in the online archive and the data ready to be integrated take up about 74 terabyte. Due to technological innovation we are now able to process and store lossless compressed JPEG2000 video streams, which result in files that are a factor 20 larger than the MPEG2 files that were our highest quality archival copies until recently. This increase in file sizes results in an annual growth of the archive of about 18 terabytes currently, however with more and more researchers switching to high-definition video cameras we can expect another steep increase in annual growth in the near future.

In the humanities, sheer data volume specifications are not a good indicator for the data management challenges to solve. There are generally complex relations between the archived objects that need to be maintained in order to preserve all the knowledge about the objects. Each digitized recording is for example part of a hierarchy of semantically related objects. Often such objects are split into new objects for specific reasons such as presentations. Different layers of annotation of the linguistic content are created, perhaps even from different annotators at different times. Derived resources such as lexica are created that relate to a collection of archived objects (see Cablitz, this volume). Several versions and transformations of many objects might be created in the course of time. It is important to store the relationships between

all these objects, since in most cases context and provenance is essential for the interpretation of the objects' content. Handling the complexity in such collections is thus a true challenge.

From a UNESCO study we know that already for tape media the preservation of the stored information has turned out to become a huge and partly insoluble problem. About 80% of the existent recordings of languages and cultures created by ethnologists, linguists etc. are highly endangered because the physical carriers are deteriorating rapidly and the material is not in the hands of specialized archives (Schüller 2004). Digital technology moves on even faster, i.e. uncurated data is much more endangered than the traditional analog recordings. There is a great risk of losing parts of our cultural and scientific memory if we do not ensure that data formats and encodings are kept distinct from the software being used, if we do not use open standards such as XML (eXtensible Markup Language) for specifying structure and if we do not use widely agreed and thoroughly documented encoding schemes such as UNICODE, MPEG etc.

Digital data needs to be continuously migrated, both at the carrier level as well as at the structure/encoding level. How can we maintain integrity and authenticity - both essential pillars for the preservation of our contents - in such a dynamic world? Migration alone will not ensure data survival, since our media are very vulnerable and our software erroneous. Automatic copying to distinct locations according to safe protocols making use of different software systems is required as well to preserve our digital treasure. For DoBeS data, six copies are created automatically at three locations and in addition selected data is being returned to the locations where they were recorded.

For both aspects – migration and copying – there are no simple solutions that are safe enough and all procedures involving too many manual operations will not work in the end, since the costs would be much too high for the large volumes of data that we are creating and maintaining.

2.1. The influence of the DoBeS programme

One of the great outcomes of the DoBeS program in an early stage was that a few enthusiastic researchers and technologists sat together and contributed to the specification of a flexible metadata schema and infrastructure: ILSLE

metadata Initiative (IMDI[1]). It was quickly understood that metadata is the glue for maintaining the complex relationships that may exist between various objects in an archive. With the IMDI metadata infrastructure we were able to not only add descriptions to objects in order to make them retrievable, or to group them based on categories, but also to organize them into various collections. Each depositor constructs a hierarchical reference organization for their corpus, which forms the basis for all management and access permission operations, but alternative organizations are also possible. IMDI is still the basis for one of the largest online archives these days: the MPI language Archive of which the DoBeS archive is a well-organized part.

There is quite some discussion currently about proper data management and the challenges posed by what is now also called the "Data Tsunami". Just recently the European Commission founded a high-level expert group to bring out a report with the name "Riding the wave – How Europe can gain from the rising tide of scientific data" (High Level Expert Group on Scientific Data 2010) and to come up with actions to address the volume and complexity aspects. In the US, a final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (Blue Ribbon Task Force on Sustainable Digital Preservation and Access 2010) and the ASIS&T Summit on Research Data showed the relevance of the data curation and preservation challenges.

Looking back a decade we can state that the DoBeS program at an early stage made great contributions to address these questions. Excellent solutions were found given the early stages of the debates and contributions to the discussions about data management are still being made today. Principles of data archiving were worked out, the need of standards was articulated, a new organization framework based on metadata descriptions was invented, the issue of appropriate creation, management, access and enrichment tools was tackled and concrete actions were started along all these dimensions resulting in solutions that meet most of the requirements being discussed these days.

In the report of the EC high-level expert group, "trust" is indicated as one of the most fundamental principles for success. Obviously trust has many facets, but most essential is that (1) the depositors trust the archivists that they take care of proper preservation, curation and access and rights management; (2) that the archivists rely on the quality of the data provided by the depositors

1. http://www.mpi.nl/IMDI/

and (3) that the users can trust getting exactly those objects they are looking for in authentic quality. The last point has led to an important shift in the MPI archivists view on researcher involvement in data managing. Given the utterly dynamic era in which the DoBeS program and in particular the archiving part has been set up, we can state that this trust has eventually been established, even though it required some attitude changes both from the archivist's as well as from the researcher's side. The archivists for example had to become aware of the utmost importance that researchers attach to proper protection and presentation of their data, and the researchers had to get used to the idea of handing over their data to an online archive that has data sharing as one of its goals. It's in the nature of innovation that trust has to be continuously re-established.

3. Archive stakeholders and their needs

As indicated in Trilsbeek and Wittenburg (2006), a number of different parties typically interact with an archive, each from a different perspective and with different – sometimes conflicting – needs. Depositors require an easy way to deposit their material and to write metadata descriptions for their deposits, archivists need means to ensure consistent archive organization and data integrity, and various groups of users of an archive need easy means to navigate and access archival content. Particularly the latter group poses a challenge to the developers of access tools for an archive since it is a rather heterogeneous group ranging from interested members of the general public to journalists, to people from the speech communities whose recordings are in the archive, and to linguists who may or may not have specialized knowledge about the archived material.

It is almost impossible for an archive to cater for the access needs of every group of users, so it is important that it offers access to its resources in an atomic way and ideally also offers access to some basic web-services to explore the archived material. In this way, different web sites or portals with different looks and levels of complexity can be developed on top of the archiving infrastructure.

Offering access to archived material and services in such a way also becomes essential if an archive wants to become part of the various "e-research infrastructures" that are being developed at the moment in projects such as

CLARIN[2]. Agreements about standards and interchange formats for data and services are needed to ensure interoperability between various archives and tool providers.

4. Long-term preservation requirements

4.1. File formats and encodings

Long-term preservation of digital data involves both the physical preservation of the digital objects as well as keeping these objects interpretable in the long run. File formats and encodings change over the years up to a point where old formats cannot even be read any more on common hardware and software of a certain point in time. We have seen various examples of this in the past and we will no doubt see many more in the future. Keeping archived data interpretable therefore means that an archive needs to migrate its stored files to up-to-date formats before the old ones have become obsolete. Converting from one format to another, however, often involves loss of information or the introduction of artifacts. In audiovisual formats, transcoding between two lossy formats or even encoding a file again in the same lossy format will introduce artifacts and loss of information. To prevent loss of information or loss of quality, the archive should use formats according to the following principles:

– for audiovisual material, use uncompressed or lossless compressed formats whenever possible
– for textual material, use Unicode character encoding and XML-based formats whenever possible
– avoid closed, proprietary formats

For textual material and audio material it is quite straightforward today to follow these guidelines. Storing uncompressed or lossless compressed video, however, still requires a lot of storage capacity by today's standards, which is problematic for many language archives. One hour of standard definition MJPEG2000 lossless compressed video for example takes up about 70 GB of storage, for High Definition video this number would be even 4 times as high. The role of video in language documentation is growing since it provides a

2. http://www.clarin.eu

way to contextualize the spoken language and to analyze other communication channels such as gesticulation, however the quality requirements for video are much less straightforward than for audio. There are a lot of variables that play a role for the quality of the video signal but their importance may vary depending on the purpose of the recording. The recording equipment that is being used to acquire the video material also limits the quality that can be obtained to some extent; the resulting video quality will depend on the available budget and on the size and weight that is still practical in the recording situation. It's probably safe to assume that the price of digital storage will continue to drop at least at the same rate as it has during the past decade, so storing uncompressed or lossless compressed video will become feasible for more archives. The consumer camcorder market on the other hand is very hard to predict and is not driven by the needs of linguistic researchers.

Two XML-based formats for linguistic data that were developed with the help of the DoBeS program are the EAF format for linguistic annotations and the Lexical Markup Framework (LMF) format for lexica (ISO 24613:2008). EAF is the format that is used by the ELAN multimedia annotation tool for storing multi-layered annotations that are time-aligned to the audio or video files. The LMF format is a flexible format for creating structured lexica and is being used by the LEXUS lexicon tool. Both formats were designed as XML formats to allow for relatively easy conversions to other formats now and in the future.

4.2. Organization of data: metadata

When gathering and managing large amounts of data, be it in the form of analogue or digital resources, an additional layer of meta-information is indispensable. This might seem obvious for the classic case of a library full of books, but it is even more true for a digital archive where language resources are stored as digitized recordings and text files. Specific reasons for this are:

Digital resources are meaningless by themselves. On the lowest level they exist of bits (0 and 1). While digital storage systems themselves already provide for interpretation of the basic characteristics of the stored bit-streams, there are many other layers of interpretation necessary for keeping the data useful and manageable, each layer requiring explicit specific (metadata) information.

There are very many ways to organize digital language resources; one organization might be more suitable for a specific archiving or research purpose than another and fortunately the digital storage paradigm does not impose a single organization. Therefore we need the ability to impose different flexible organization models or views that match the interest of researchers or archivists. The richer the metadata available, the more possibilities there are for the end user to create these special views and explore the digital collection.

In the current landscape of digital repositories and archives, a number of specific metadata standards are prominent for the description of linguistic data. Such a standard usually specifies a set of metadata elements (sometimes called attributes) together with prescriptions for the values of these elements and also prescriptions on how the metadata elements and values should be put into a text format (schema).

The first one of these sets and probably the most widely used one, is Dublin Core,[3] which stems from the electronic library world. Dublin Core was later extended with some linguistic specializations into the OLAC standard[4] which has become popular for exchanging Language Resource metadata between archives. Around the same time the IMDI[5] standard was introduced and adopted by the DoBeS program. IMDI strives to allow detailed descriptions and several so-called specialized profiles were created for specific linguistic subdomains. A suite of tools to edit and use IMDI metadata was partly developed within the context of the DoBeS program.

At the time of writing (2011) a follow-up standard for IMDI, called CMDI[6] (Component metadata Infrastructure, cf. Broeder et al. 2010) is being worked out within the CLARIN framework. Rather than offering one single metadata schema it tries to offer the user a set of loose components that can be combined into a tailored metadata schema. This approach should allow for a detailed description while keeping the focus only on those metadata elements that are relevant. Apart from that, it also allows for partial re-use of existing metadata schemas and provides better mechanisms of semantic interoperability by requiring that the semantics of all used metadata elements are explicitly defined in an accepted concept registry. Using CMDI will hope-

3. http://dublincore.org
4. http://www.language-archives.org/OLAC/metadata.html
5. http://www.mpi.nl/IMDI/
6. http://www.clarin.eu/cmdi

fully increase metadata interoperability between linguistic research communities having different needs and traditions.

4.3. Other standards in language archiving

Both the LMF lexicon standard and CMDI metadata standard are prime examples of a trend to have standards which can be easily adapted to the needs of a specific resource type, as in use by a specific (research) community, or even a single resource. LMF provides a core meta model, with some extensions, which can be adorned with data categories taken from a data category registry to form the actual data model for a specific LMF lexicon. CMDI uses the same approach by storing pre-defined metadata components and profiles in a component registry. These components are also annotated with links into concept registries, from which the data category registry is one, to make semantic descriptions available and to share those. Registries are thus starting to play an increasingly prominent role in standards related to archiving. The MPI develops and hosts the following registries:

- ISOcat[7] (Kemps-Snijders et al. 2008) is the data category registry (ISO 12620:2009) for ISO TC 37, which is based on a grass-roots approach, allowing any linguist to participate in the specification and standardization of linguistic data categories.
- The CMDI component registry[8] for CLARIN-NL.
- RELcat is a registry to store (user-specific) relationships between data categories and possibly other concept registries.

A metadata example that is already in use illustrates the support for mapping from the IMDI to the Dublin Core metadata schemas by using these strategies. The metadata profile in ISOcat has been bootstrapped with the IMDI elements, which includes the */mimeType/*[9] data category. The specification of a data category can be very elaborate including translations in multiple languages, but at least an English name and definition should be available. The */mimeType/* data category is defined as the "specification of the mime-type of the resource which is a formalized specifier for the format included or a

7. http://www.isocat.org/
8. http://www.clarin.eu/cmdi/
9. http://www.isocat.org/datcat/DC-2571

mime-type that the tool/service accepts". In the CMDI component registry the cmdi-mimetype[10] component links the MimeType element to this data category. The format[11] element in the Dublin Core metadata schema actually plays the same role and is defined as "the file format, physical medium, or dimensions of the resource". In RELcat the equivalence relations between the ISOcat data category */mimeType/* and the Dublin Core format element can be specified using a simple RDF triple:

```
isocat:DC-2571 rel:sameAs dc:format
```

The metadata search that is currently under development in CLARIN can already exploit such semantic relationships to broaden the scope of a search. While this example zoomed in on the metadata domain, ISOcat is currently being populated with data categories for various other domains, e.g., morphosyntax and terminology, and it is expected that this will provide the same kind of flexibility to content search on resources created in these domains.

As usability, accessibility and interoperability are long-term goals of the archive, the persistency of the registries and the links to them is a major concern. Most of these registries do provide Persistent IDentifiers (PIDs) backed up by persistency strategies, which allow safe use of these identifiers in the metadata of resources or even the resources themselves. There are various PID frameworks available. To help an archive to choose among these frameworks, ISO 24619 "Persistent identification and sustainable access" (ISO 24619:2010) gives specific requirements these frameworks should meet to make them useful for archives of linguistic resources.

To promote the (re)use of the resources stored in the archive standards for harvesting metadata, e.g., OAI-PMH from the Open Archives Initiative, and standards on and agreements between archives about Authentication and Authorization Infrastructures are important. Large-scale infrastructure initiatives like CLARIN help to build up the federations of all involved organizations.

4.4. Versioning

When storing and archiving digital resources, an important policy decision concerns how to respond when a depositor offers a new "version" of a resource that is already present in the archive's holdings. There can be differ-

10. http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:c_1271859438106
11. http://purl.org/dc/elements/1.1/format

ent reasons for offering a new version: (a) The depositor has realized that the first version is simply broken or unusable, for instance in the case that the files were switched. (b) New insights make it necessary to change some annotations. (c) The format of a resource may need to be upgraded. For instance a codec used to encode a media-stream may become obsolete requiring resource replacement.

Depending on the archive organization and policies, it is possible to let the new version take the place of the old one in the existing network of relations with other resources and metadata. The old version may then be moved to background storage or, depending on archive policy, even deleted. Of course the relation between old and new versions needs to be stored and users should to be able to see that other versions exist.

We will not go into the question on what actually makes a resource a new version of another resource. This should best be left to the judgment of the depositor or caretaker of the original version.

It is however very important to realize that users may have created references to a resource in the archive, for instance as a link in a publication. Most users will expect that that reference will always link to the same version, while others may want to refer to the latest version. It is important that the archive is explicit about its versioning policy in this respect. The most flexible system is to always keep any reference to a specific resource version but to provide referencing to the latest version as a special service.

However it is known that some archives are unable to keep stable references to resources or resource collections due to legal or organizational obstacles. For instance, its legal owner might withdraw a resource from an archive's holding. In such cases the archive can only be as explicit as possible about such circumstances.

5. Open access vs. access restrictions

At the moment there is a large push towards open access to research results, not just the scientific publications but also the data that forms the basis of these publications. The Berlin declaration on Open Access to Scientific Knowledge[12] was first published and signed in 2003 by representatives of most of the German research organizations, but has meanwhile been signed

12. http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

by 294 scientific organizations and universities worldwide. In general there is a lot to be said in favor of making the outcome of research that has been funded with public money available to the public and to other researchers in an unrestricted manner. 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. In many fields of research however, data is collected by making use of human subjects, in which case the privacy of these subjects needs to be taken into account. Informed consent forms are often used to explicitly regulate the rights to publicize data of human subjects. Anonymization is another method to ensure that the privacy of the human subjects is respected, which in general means that all information that could be used to identify individuals is removed from the data. In social sciences for example it is common practice that the names and contact information of participants to a survey or an experiment are removed before the data set is published. Both informed consent and anonymization can be somewhat problematic in the field of documentary linguistics though. 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. Masking names in texts and in audio recordings is something that can be done but modifying audio and video recordings up to a point where the individuals can no longer be recognized would render them useless for many linguistic purposes. The fact that recordings are made within small communities sometimes requires the researcher to protect the speakers in order to avoid conflicts within the communities. It is up to the researchers working in these communities 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. Some of these issues, and possible solutions, are discussed in the following section.

6. Legal and ethical issues

As indicated above, when working with data collected in small language communities one has to carefully consider the rights and privacy of the interviewed contributors. In the DoBeS program, legal and ethical considerations were an important point of discussion from the very beginning. In its second year a workshop was organized with leading European law experts to determine a proper juridical basis for the DoBeS program and in particular the

online archiving ideas. The result however was disappointing from the practitioners' point of view, since it was concluded that the legal situation is much too complex to give clear juridical advice. The only advice the experts could come up with was to lock the material in a safe in a cellar, which was exactly the opposite of what was expected from the emerging archive – namely to be a place where authorized persons from all around the world could access and even enrich the stored data.

Intensive and serious discussions afterwards led to a number of conclusions:

- It was understood that the DoBeS program should have a proper basis to guide the behavior of all persons involved: collectors, archivists and users. The result was an elaborate Code of Conduct, 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 for example the access permissions etc. It is expected that the researcher responsible takes care of proper relationships with the communities and the interviewees and that all statements are based on informed consent. The archivist will adhere to the statements of the researcher responsible and provide access mechanisms that implement the requirements.
- The archivist declared that he does not claim copyright on the stored material. However, he needs the right to archive in order to perform his task in a responsible way. With respect to users the archivist will claim copyright on behalf of the data producers.
- It was decided to not use visible logos in the video since they might obstruct the content.
- The researcher responsible always has access permissions to all material and he 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 especially since communities may withdraw access permissions to certain material again although it was granted at a certain moment for culture specific reasons. Also other complicating issues may play a role requiring a high

degree of sensitivity of all actors involved. 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 solving potential difficult questions.

Over the years, when it became more obvious that more users may want to access material in the online archive, four levels of access granting were agreed upon:

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 researcher responsible (or persons specified by him or her) and specify their usage intentions;
Level 4: Finally, there will be material that 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 is done writing the thesis.

The number of external people who requested access to "level 3" resources over the last years was not that high. 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.

7. Data enrichment tools

Providing tools for tagging or annotating audio-visual media has been one of the focal points of software development at the MPI right from the start, from the Mac-only application MediaTagger, via a set of client-server based cor-

pus visualization programs to their convergence into the stand-alone, multi-platform annotation tool ELAN. This progression of tools was paralleled by the switch from data in a proprietary format to data stored in the evolving the evolving standard XML. After a thorough makeover (in 2003), marking the transition to the 2.x versions, ELAN further developed into an application supporting multiple videos in multiple formats, providing a growing number of import and export options, with increasing editing capabilities and available as both a Java Web Start and as a downloadable installer version.

ELAN allows enriching of audio and video recordings with multilayered, structured annotations stored in EAF (ELAN Annotation Format) files, a file format that can be uploaded into the archive as a constituent of a corpus. To make inspection and exploration of data in the archive more convenient than downloading a bundle of files and opening the software, the web application ANNEX has been created. It streams (chunks of) media recordings from the archive and visualizes associated annotations, not only those stored in EAF but other formats as well, in an interface resembling that of ELAN. ANNEX resembling that of ELAN. ANNEX is closely connected to TROVA, a search engine for structured search in annotation content. Queries can be executed in one or more corpora or parts thereof and from any search result or hit a jump to ANNEX can be made, showing that particular annotation in that particular file.

Processing multiple files simultaneously has recently become an important track of development of ELAN and it is expected that it will be in the years ahead. This type of operation improves productivity enormously and stimulates consistency within a (local) corpus. More generally, reducing the number of mouse clicks and keystrokes and steps that have to be performed manually will be a future goal. Semi-automatic annotation by pattern-recognition based software components is expected to become available for everyday language research soon.

Another data enrichment tool developed by the MPI is a flexible online lexicon tool called LEXUS[13] (Ringersma and Kemps-Snijders 2007). The LEXUS lexicon schema can be based on the meta-model of the LMF standard, but actually users have extensive freedom to construct a rich lexicon schema appropriate for the language to be described. Elements in this schema can be linked to the data category registry, ISOcat, and can thus have explicit, and shareable, semantics. Import tools allow loading existing lexica in

13. http://www.lat-mpi.eu/tools/lexus

various formats, e.g., MDF, into LEXUS. In principle lexica exist in a user-specific workspace. However, LEXUS allows sharing these lexica with other users thus enabling collaboration on the development and population of a lexicon. Cablitz (this volume) gives a detailed account of the implementation of LEXUS in an actual documentation project.

The LEXUS frontend has evolved over time using different browser-based technologies into the current FLEX version, which due to its use of the Adobe Flash plug-in provides a similar look&feel across a wide variety of browsers and platforms. The rendering of lexical entries has always been very flexible, allowing users to construct templates for both list and entry views. A new version of the LEXUS backend is currently close to completion and next to providing increased stability and performance will also allow to more easily add new output formats, e.g., a printable version of the lexicon.

Making different tools like LEXUS, ANNEX/TROVA and ELAN cooperate as seamlessly as possible is another important line of development. Separation of metadata and annotation content has its merits, but at some point they will have to come together e.g. in a combined data-metadata search in TROVA. Some annotation editing options, especially those that are executed on multiple files (like find-and-replace in many files), make perfect sense in the context of ANNEX. The combination of ELAN and LEXUS will on the one hand allow lookup and retrieval of information from a lexicon while annotating, and on the other hand will enable the user to start building a lexicon while annotating.

8. Accessing data

8.1. Meta data searching and browsing

Access to archived resources is generally offered by means of search and browse functions for the metadata catalogue. Search functions can be implemented in various ways, e.g. as free text Google-like search across the entire metadata catalogue, as an advanced search for searching in specific metadata fields, or as a "faceted search" that allows one to narrow down search results by selecting values of a number of pre-defined fields. Searching within a metadata catalogue that makes use of a single metadata scheme is fairly straightforward. The only problem here is that there is a certain degree of variation of metadata values that actually refer to the same kind of data, if

the metadata field does not require the use of a controlled vocabulary. Some kind of mapping would need to be performed in order to find all variants of the same value. The situation becomes more complex if one needs to search across different catalogues with different metadata schemes. It is hoped that the use of links to the ISOcat data category registry in metadata schemas and value sets will make cross-archive searches more manageable. As an example, archive A may use a metadata schema that contains the element "gender" for speakers for which the values can be "F" and "M", archive B may use the element "sex" for basically the same concept and uses the values "female" and "male". If both metadata schemas would refer to the proposed ISOcat term "HumanGender" and the values "feminine" and "masculine" (with the definitions that this relates to the gender of a person rather than grammatical gender), it would be possible to search across both archives using either terminology using an ISOcat-aware search tool.

8.2. Content searching

A more elaborate search for the actual content of the resources is required if one wants to find specific examples of language use that cannot be described in the metadata. At the moment this content search will be limited to textual resources (annotations to audio/video) but possibly in the future this could be extended to a limited set of features in the audio or video material itself. Searching for annotations can also be done in varying levels of complexity. The TROVA content search tool for example offers a simple search mode to search across the entire annotation file, it offers a "single layer" mode to search for sequences within a single annotation layer and it offers a "multiple layer" mode to search for sequences both within and between annotation layers. Content search tools can be used to find specific examples in a language corpus, but can also be used to perform statistical analyses on a corpus by finding all cases of a certain linguistic structure. Also in textual content search tools, the variation in terminology that occurs within and between archives can be an issue. Here also the ISOcat registry can play a role by allowing search tools (and users) to create mappings between different terms that actually have the same meaning.

8.3. Portals

While metadata and content search tools are generally suitable for specialists to find material that they are interested in, members of a speech community or members of the general public have other requirements when accessing archived content. If the search services and archive access framework are set up in a rather generic way and can be called via standard web-service interfaces, it is possible to create an additional "layer" on top of the archive that serves a specific user group. This layer or "Portal" can have an appealing graphical design and it can direct people to certain pre-defined searches that have been set up or interesting resources that have been selected. Within the European research infrastructure projects that are currently running such as CLARIN, more and more tools are being made available as web services. To what extent these web services will be of use for certain user-specific portals remains to be seen, but at least they open up a wide range of possibilities to combine resources and services together in a web interface.

9. New challenges

Life cycle management of data can be split into three major and related phases: creation, curation/preservation and access/utilization. With respect to all three phases we will see accelerated technological innovation which on the one hand has positive effects in so far that research can make use of newest inventions and products and on the other hand has negative implications with respect to the stability of the solutions found. The trick will be to define the islands of stability in a very dynamic environment and to participate stepwise in the innovation process. This holds for the archive as well as for all software being written. In all phases of the data life cycle, the challenging ethical and legal situation needs to be taken into account.

Creation Phase: The creation process will benefit from further sophistication in recording equipment, where in particular three developments will have their implications: (1) miniaturization of data storage leading to increased capacity; (2) resolution; (3) connectivity. Miniaturization will lead to continuously increasing storage capacities allowing researchers to make high-resolution recordings with portable devices. Miniaturization also will simplify field work in so far that direct annotation will be easier with help of smart and small devices demanding less power. The resolution of recording

devices will be increased so that soon high-definition video cameras can be expected in the low cost sector. Also connectivity will become better – even at remote places. This will mean that digitized (in fact digital recording is now the norm in the vast majority of fieldwork contexts) recordings can be transmitted earlier and faster. It also will mean that the possibility of downloading or accessing archive material will be improved.

Preservation/Curation: A big step for video preservation has already been done by introducing lossless MJPEG2000 recently. This indeed means that we are able to store a master file from which other formats, e.g. for presentation purposes, can be generated without risking serious transformation effects. Information technology (channel bandwidth, storage capacity, CPU power) will allow us to deal with the increased data amounts.

Long-term preservation is very much dependent on "safe" replication where every operation on a data object will automatically lead to check whether the copied instance is indeed the same as the original one. It is widely agreed now that the extensive use of externally registered persistent identifiers associated with checksum information is the only way to ensure data integrity and authenticity in distributed and thus more complex data management scenarios. The DoBeS archive is prepared to participate in such state-of-the-art archive federation scenarios, since for some years it is already based on persistent identifiers and automatically generated checksum information. Together with the computer center in Garching (RZG) it has been testing actively a switch to a rule-based safe replication strategy based on the iRODS software and it seems that the system can be put into operation in 2011. This will be a major step ahead also to support the open deposit service of the MPI offered to all researchers with language resources.

Also in 2011 the component based metadata tools will come into place, which offers much more flexibility for the researchers to design a metadata profile that is suitable for their resources. Interoperability will be guaranteed by making use of categories defined in the ISOcat registry. The ARBIL editor, which has now replaced the IMDI metadata editor, is already supporting this component structure and will hopefully motivate researchers to provide better metadata descriptions, since they will be the key for the application of advanced analysis tools and for generating portals designed for the special community in mind.

Utilization Phase: We expect many developments in the improved utilization possibilities of the stored data as long as access is being granted and

as long as the quality of the metadata and the data is high – quality will actually be the crucial point for many advanced operations. One big concern is that the amount of recorded media streams that is not being touched (annotated in some form to make it ready for analysis) is increasing continuously which means that much of the stored data will effectively not be of much use to anyone other than the person who collected it. A new attempt to use state-of-the-art speech and image processing technology is required that does not build on holistic stochastic models but on detectors that react to comparatively simple patterns in media streams and create annotations. There will be several of these detectors all with different characteristics that may also be specialized on specific quality types of recordings. The resulting lattice of annotations could be the base for linguistic evidence and theorization if there are smart tools allowing the researcher to look for specific patterns and to easily navigate in it.

We can indicate a few other areas where we expect new opportunities in the coming months and years:

- Semantically based weaving of content (creating relations and navigating in the resulting conceptual spaces) is very attractive for finding linguistic evidence. However, this work is hampered by the huge effort required to create meaningful relations. Better usage of existing ontologies for automatic support in creating the relations would make this work practically feasible.
- Archive federations are being set up, metadata has been standardized, resource formats are being much more harmonized and improved tools to foster semantic gateways will make it easier to carry out cross-archive and cross-corpus related work.
- More and more tools are being turned to web services or at least support web-based interactions. Since the programming interfaces are also currently being harmonized, there is great hope that in a few years researchers will be able to combine useful algorithms to chains of operations on texts (annotations, etc.), audio and video streams and even other type of data to carry out work that currently is only possible when large scale expert knowledge is directly available. For these advanced operations, the quality of metadata and data will be of crucial importance.

Much funding is currently invested in creating infrastructures that will increase the integration and interoperability of resources and tools. CLARIN is

the initiative that aims to achieve these goals in the linguistic domain. Such infrastructure work can only be achieved when we apply standardization and harmonization where possible without hampering the research progress. The DoBeS community was one of the driving forces to apply open standards and foster new standards. If this positive attitude is continued, the work on endangered languages will profit in many ways from new technological developments in the coming years.

References

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. San Diego: BRTF-SDPA. Online version: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Broeder, Daan, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A data category registry- and component-based metadata framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010), Valetta, Malta, May 19–21, 2010, 43–47.

Consultative Committee for Space Data Systems. 2002. *Reference Model for an Open Archival Information System (OAIS)*. Washington DC: CCSDS Secretariat, NASA.

High Level Expert Group on Scientific Data. 2010. *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*. Brussels: European Commission. Online version: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

ISO 12620:2009. 2009. Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources.

ISO 24613:2008. 2008. Language resource management – Lexical markup framework (LMF).

ISO 24619:2010. 2010. Language resource management – Persistent identification and sustainable access (PISA).

Kemps-Snijders, Marc, Menzo A. Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. ISOcat: Corralling data categories in the wild. In *Proceedings of the Sixth International Conference on Language Resources*

and Evaluation (LREC 2008), Marrakech, Morocco, May 28–30, 2008, ed. European Language Association (ELRA).

Ringersma, Jacquelijn, and Marc Kemps-Snijders. 2007. Creating multimedia dictionaries of endangered languages using LEXUS. In *Proceedings of Interspeech 2007*, eds. Hugo van Hamme and Rob van Son, 65–68. Baixas, France: International Speech Communication Association.

Schüller, Dietrich. 2004. Safeguarding the documentary heritage of cultural and linguistic diversity. *Language Archives Newsletter* 1(3):9.

Trilsbeek, Paul, and Peter Wittenburg. 2006. Archiving challenges. In *Essentials of Language Documentation*, eds. Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 311–335. Berlin, New York: Mouton de Gruyter.