The Language Archive at the MPI: Contents, Tools, and Technologies

Peter Wittenburg, Romuald Skiba, Paul Trilsbeek
MPI, Nijmegen

The language resource archive at the MPI is not a classical archive that is only concerned with long-term data preservation, nor is it a strictly digital library that only gives access to the data. It is a modern digital archive that combines safeguarding of the stored material such that future generations can access it and providing immediate access to those who are currently interested in the material. The basic technological requirements are driven by long-term considerations, i.e. representational aspects. Presentation aspects are seen as something that will be derived from the standards-based representation (Wittenburg et al. 2004).

History

The MPI Language Archive in its current version has existed since 2000; however, the original archive is older than that. The local-access intranet-based language data samples that have existed since the early nineties were hybrid: partly HTML-based structures and partly UNIX-based file structures. The first corpus accessible via the internet was the "ESF Second Language Acquisition Corpus", which was published in 1996 and was subsequently added to the IMDI metadata corpora. The 2000 version of the archive was entirely IMDI-metadata based (see Wittenburg & Broeder 2002). All data in the archive is described by IMDI-session metadata, i.e. the primary language data (recordings and/or annotations) is described as a bundle of metadata information and also as part of the archive tree structure (IMDI-corpus metadata). The metadata information is freely accessible. One can browse through the metadata tree (Browsable Corpus) and finally reach all the available information on the primary data. Access to the primary data can be restricted by the owner of the data (the researcher, group or institution).

Contents

Currently the MPI corpus contains more then 20,000 sessions (the entire IMDI-based corpus contains twice as much, Broeder 2004:12). There are data from the Max Planck Institute for Psycholinguistics, from the DoBeS program (Documentation of Endangered Languages, links: DoBeS), ECHO (European Cultural Heritage Online), and other projects. Child, sign, and second language acquisition data, are represented, as well as data for many endangered languages. The contents are described in information files at the entry point of the corpora (links: IMDI-corpora).

Data formats

Recording formats

The sources of the data are recordings on many different media, representing the long history of technological development of recording devices. For video, this ranges from celluloid film to analog and digital video formats such as U-matic, VHS, Hi-8 and DV. Audio material ranges from recordings made on reel-to-reel tapes to recordings made with the latest flash-memory recorders. The MPI stores these recordings in the archive as high-quality digital audio and video files, making conversions when necessary.

Digital data formats and codecs

In order to preserve digital data for future use, the file formats and codecs that are used should be explicitly documented and preferably conform to open standards (Wittenburg et al. 2004:3).

For the MPI, there are different rules with respect to the permitted file formats and codecs for the different corpora. For example, for the DoBeS archive, there are more restrictions than for some of the MPI-internal corpora, due to the fact that certain agreements have been made within the DoBeS program to maximize the probability of long-term data preservation. Currently, the archive contains data in the following data formats and codecs:

- Audio: WAV files of linear 16-bit PCM audio with a sampling frequency of 44.1 kHz (CD quality) or 48 kHz (DAT quality). Also lower quality recordings, e.g. made with a minidisk recorder, are converted to this format in order to have a more consistent archive and provide a greater chance of data survival (even though the conversion doesn't add any extra quality). As well as the WAV files, MP3 files are created as a web presentation format.

- Video: MPG files containing MPEG1- or MPEG2-encoded video are the main formats. Even though MPEG video is compressed and reduces the quality of the original recording, it is currently the only feasible encoding for the MPI archive. It will still take a number of years before digital storage technology will allow us to store uncompressed video. MPEG4 video is created as a web presentation format.
- Images: JPEG files containing JPEG compressed data are still the standard for digital photography at this point and are therefore accepted as an archive format. However, digital photography is moving towards uncompressed formats, which will be the preferred formats in the future.

Metadata formats

Metadata descriptions are in accordance with the IMDI standard, a scheme-based XML format (Wittenburg & Broeder 2002). While the IMDI standard contains a large set of fixed metadata description elements, it also allows for user extensions.

"Info" files can be seen as another kind of metadata, created to provide a more detailed description of the content of a specific point in the corpus. The structure of these files is open, but the file formats are restricted to HTML, plain text, and PDF.

Annotation formats

Annotations are typically documents that refer to the content of media files. These documents conform to various standards and formats. The primary format of the archive is the EAF (ELAN Annotation Format, Brugman & Russel 2004), which is XML-schema based. The ELAN annotation tool (see below) creates annotations in EAF format. Another format is Shoebox/ Toolbox (links: Shoebox:31). Shoebox data can be imported and manipulated using ELAN. The CHAT format (MacWhinney 1995) is also supported by ELAN. Some less-represented annotation data are in PDF and MediaTagger format. The recommended character encoding is Unicode (UTF-8). We are trying to convert all annotation formats to the EAF format, i.e. XML-based with Unicode UTF-8.

Lexicon formats

Lexicon data related to EAF annotations are in LMF format (supported by LEXUS, see Wittenburg & Kemps-Snijders 2005; see also section "Accessing Archival Content" below). Shoebox and CHAT-lexicon files are accepted, but conversion to LMF format is recommended. Character encoding for lexica is again Unicode (UTF-8).

The architecture of the MPI archive

The archive distinguishes four groups of people involved in digital archiving: (1) donators; (2) corpus managers; (3) system managers; and (4) users. Donators and users should not see the physical storage structure; they should operate in virtual domains defined by linguistic terminology and should have the flexibility to create their own personalized virtual archival domain. Systems managers must ensure that the data can reliably meet the criteria for long-term storage: accessibility and protection against unauthorized access. Corpus managers create and maintain a canonical organization of all material, map virtual and physical data organization, ensure that the archive is consistent, and mediate between the different groups involved.

To meet these requirements the archive uses the IMDI metadata framework to organize and maintain all material, and also to offer the user a browsable and searchable virtual archive. The virtual IMDI domain is the core of all donator, user, and management activities. Underlying this is the physical structure maintained by the system managers. This principal distinction allows system managers to migrate and copy data without affecting users. To meet the requirements of long-term storage the archive has several layers of redundancy. In the center is a Hierarchical Storage Management system (HSM) running on two parallel multi-processor SUN servers. These servers are connected to a two-layer RAID system - a faster one for small files and another, slower, for large media files - and a tape library. Via fast European network connections, complete copies of the data are dynamically created at two large computer centres at the Max Planck Society and the MPI for Evolutionary Anthropology. The Max Planck Society has committed to being responsible for a portion of the data (e.g. DoBeS data) for a period of 50 years. An additional group, the developers, ensure that there are methods for uploading resources into the archive, for maintaining the archive, for monitoring its consistency, and for accessing its content at different layers. These layers are described below.

Archive technologies and tools

Digitization, capturing, and transcoding tools

Capturing and transcoding of audio and video is performed using various hardware and software. Information about current procedures is available via the Web (links: A4 guides).

Metadata tools

The IMDI infrastructure is the core of the archive's architecture. For the user it is like a browsable and searchable catalogue for discovering resources; for the archive manager it is the basis of maintenance operations. The developers provide a professional editor that allows the creation of valid IMDI files for various resource types; a native IMDI browser operating in the domain of linked XML-based metadata descriptions; and a server program for browsing the IMDI domain as if they were HTML pages. All IMDI tools are freely accessible via the Web (links: IMDI-tools).

Annotation tools

The main annotation tool for creating the standard EAF archiving format is ELAN. ELAN allows the creation of media annotation files in an open format. ELAN can import and export other accepted formats, such as CHAT and Shoebox files. (As stated above, the ultimate goal of the archive is to unify all formats to XML-based formats. For the moment we also accept and archive annotations in their original format.)

Accessing archival content

There are various methods of accessing the archival content. The browsers allow the downloading of individual or groups of resources once they are identified by their metadata. Download allows for local operations on resources, such as playing audio or video, viewing annotations, or using tools such as ELAN or LEXUS on the resources found. Currently, a web-based exploration framework is being developed.

Uploading and management

Until now the corpus managers had to upload and check the data manually. Metadata descriptions were also checked and integrated, and media resources were digitized and linked with the metadata descriptions. With an increasing number of teams contributing to the archive, this process proved to be no longer efficient and so LAMUS (Language Archive Management and Upload System) was developed. It offers a web-based interface so that donators as well as managers can use a standardized interface to prepare and check the material and then upload it.

Access management

The archive is bound by a number of rules restricting access. The archivist has the right to store the data, but copyright is held by the donators and consultants. Therefore, the researcher is seen as the central figure in determining access to the material. The archivist assumes that the researcher has obtained the consent of the language community and/or consultants for making the resource available to others. Potential users are required to accept a Code of Conduct developed for the DoBeS program and to declare their intended use before they are granted access to the resources. To support this, the Access Management System has been developed to deal with access management issues electronically. In addition, it allows the delegation of the right to grant access to others, such as the responsible researchers. It is efficient insofar as it allows the granting of access rights to groups of people with a single command from a particular node in the linked IMDI domain.

Future aspects

The DoBeS archive, hosted at MPI, intends to intensify its cooperation with other archives within the DELAMAN initiative (links: DELAMAN). It will focus on integrating its collection with others to make it more useful to users and will seek to improve the chances of data survival by distributing copies of its data worldwide.

References

Broeder, D. 2004. 40,000 IMDI sessions. *Language Archive Newsletter* 4, p. 12.

Brugman, H. & Russel, A. 2004. Annotating multi-media/ multi-modal resources with ELAN. In: F. Ferreira, R. Costa, R. Silva, C. Pereira, F Carvalho, M. Lopes, M. Catarino, & S. Barros (Eds.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004, Lisbon), X., pp. 2065-2068. European Language Resources Association (ELRA), Paris. CD-ROM version available.

MacWhinney, B. 1995. *The CHILDES Project: Tools for Analyzing Talk.* 2nd Edition. Erlbaum, Hillsdale, NJ.

Wittenburg, P. & Kemps-Sn ders, M. 2005. LEXUS - a new lexicon tool. *Language Archive Newsletter* 5, p. 10.

Wittenburg, P. & Broeder, D. 2002. Management of language resources with metadata. In: L. Romary, C. Galinski, N. Ide, & K.-S. Choi (Eds.), *Proceedings of the Third international Conference on Language Resources and Evaluation* (LREC 2002). Workshop on international standards of terminology and language resources management, pp. 49-53. European Language Resources Association (ELRA), Paris.

Wittenburg, P., Skiba, R. & Trilsbeek, P. 2004. Technology and tools for language documentation. *Language Archive Newsletter* 4, p. 12.

Links

A4 guides: http://www.mpi.nl/corpus/a4guides/

DELAMAN: http://delaman.org

DoBeS: http://www.mpi.nl/DOBES/

IMDI-corpora: http://corpus1.mpi.nl/BC/IMDI-corpora/

IMDI-tools: http://www.mpi.nl/tools/

Shoebox: *The Linguist's Shoebox: Tutorial and User's Guide.* SIL International. http://www.sil.org/computing/ shoebox/ShTUG.pdf