The LACITO Archive : its purpose and implementation

Michel Jacobson, Boyd Michailovsky
LACITO, CNRS, Villejuif, France
jacobson@idf.ext.jussieu.fr, boydm@vjf.cnrs.fr

Introduction

The LACITO Archive project has as its goal the conservation and the diffusion of linguistic documents, mainly in little-known, often endangered languages, including both sound recordings and annotation such as transcription and translations.

The original aim of the project, when it was first conceived in about 1992[1], was to conserve recordings which were in danger of deterioration or dispersal. Linguists at the LACITO possessed hundreds of hours of field recordings, collected at some expense, some of which had been exploited for publication, but for which there existed neither a clear policy of conservation nor resources for implementing one. It is safe to say that this is still the case, not only at the LACITO but in many research groups or academic departments. The Archive project was partly inspired by the relatively new possibility of digital archiving — at the time on the newly accessible recordable CD medium — at a cost far lower than that of archiving analog magnetic recordings, and with the possibility of producing multimedia resources accessible to computer processing.

Before briefly reviewing the implementation of the Archive, and demonstrating it, I will expand slightly on some issues involved in defining the content of the archive, and in the project objectives of conservation and diffusion.

Content

The purpose of the project is to archive multimedia linguistic documents, with annotation synchronized with recorded sound. Document preparation requires considerable effort on the part of document providers. Not to discourage potential contributors, who have made recordings for a variety of research purposes and from a variety of theoretical perspectives, the project avoids imposing a particular annotation content. Standardization, for the Archive project, is a question of adopting information processing standards (see below) and not of imposing particular linguistic standards.

One model for the annotation of archived documents has been the traditional linguists' "interlinear glossed text", and the synchronization of such material, some of it published, with sound recordings has been a major project activity. ("Interlinear", however, is a notion that has to do with display and not with content, and we do not regard "interlinear text" as a data type.) But the project does not require that archived documents be glossed, or even translated: the minimum content envisaged is a recording and a transcription. *A fortiori*, it has not been a goal of the project to develop a uniform linguistic annotation. Some currently archived documents have morphological annotation, however, and the open document format makes enrichment of existing annotation possible.

The promise of synchronized, random access to recordings and annotation, along with accessibility to corpus linguistics tools, are major incentives to document-providers.

[1] By J. B. Lowe, Martine Mazaudon and B. Michailovsky.

Conservation

We envisage the following threats to conservation:

- Physical deterioration or obsolescence of media.
- Obsolescence of data format.
- Impermanence of the conserving institution.

Physical danger to the zeros and ones of digitized data is perhaps the least worrisome. Digitized data can be copied without loss form one medium to another. It is relatively easy to insure against a local disaster simply by keeping it in several places at once, and by transferring it to new supports as old ones become obsolete. This becomes an institutional rather than a physical or chemical problem.

Obsolescence of the data format is a separate problem. Protection against this danger lies in the adoption of a standard, open, and explicitly structured data format. Such a format has a high likelihood of remaining useful over time — and the certainty of becoming obsolete. Before this latter eventuality, explicitly structured data can be converted to another format. The more widespread the original format is, the more likely it is that conversion tools will be readily available.

For a CNRS research group, by nature impermanent, perennity of the conserving institution is particularly worrisome. Although a properly conceived digital archive may require relatively little maintenance, we are still prospecting for permanent institutions such as digital libraries or *phonothèques* who find our Archive sufficiently interesting to assume this responsability. We are simultaneously trying to make the Archive more tempting. We would like if possible to interest several such institutions in the Archive.

Diffusion

The obvious vehicle for diffusion of digitized material is the Internet. However, simply putting up a web page is not sufficient to insure diffusion. Two obstacles to diffusion that have concerned us are the following:

- The variety of equipment used by potential Archive users.
- The ability of potential clients to find the Archive by content-based searching on the Internet.

Concerning the first, the response (again) is to remain as standard as possible. Since our data structure conforms to World Wide Web standards (W3C recommendations), it is accessible to standard browsers, which are available for virtually any platform. The editors of these browsers take on the responsibility of adapting them as systems and platforms evolve.

The second problem is analogous to cataloging. Cataloging data, called "metadata" in the digital, multimedia world, must be standardized to be useful, so, as a small project, we have not attempted to design our own. Rather, we have adapted to the protocol of the Open Archive Initiative, and had ourselves listed by centralized service providers such as the OAI, the Open Linguistic Archive Community, Linguist List, etc. We have also participated in the prototype European ISLE Metadata Initiative (IMDI).

Implementation

The project has designed an XML data structure (Michailovsky 2001) in which annotation is time-aligned with digitized sound recordings, usually at the utterance or sentence level, making both accessible simultaneously. Apart from time-alignment data, annotation may be minimal — e.g. a simple transcription — or more complex, including

different kinds of transcriptions, translations and grammatical information at different levels (utterance, word, morpheme).

In accessing a document, the user chooses between a number of "views", defined by XSLT stylesheets (Jacobson and Michailovsky 2000), on the available data. He may choose to see only the transcription, or the transcription and the free translation, or the transcription and the aligned interlinear word-glosses; he may choose the language of the translation if more than one is available. Once a document is opened, clicking on a displayed text element causes it to be highlighted and the corresponding sound to be played. The user may choose either to listen to one segment at a time or to the whole of the remainder of the text. Basic tools of corpus linguistics, such as word searches and concordance, have also been implemented as stylesheets. Responses to queries remain linked to the sound resource: for example, clicking on a concordance line causes the appropriate sound to be played.

An experiment linking a corpus of archived texts to a lexicon is described in a separate paper (Jacobson and Michailovsky, this volume).

The Archive uses only standard data formats and requires no proprietary software either to implement or to access. The key to implementation of the server lies in a servlet which calls an XSLT processor to apply stylesheets to the XML documents, extracting the relevant sound from the sound resource, and directing the output to the client. HTML pages produced by the processor are displayed on the client machine and receive user input.

Archive documents are prepared in various ways, depending on the source material. Existing structured or implicitly structured annotation, such as Shoebox files, can be directly converted to XML and Unicode by script, ready for time-alignment and archiving (Michailovsky 2001). Other computerized material can also be converted after some preparation. Alternatively, annotation can be entered directly in XML using an XML editor. Time-alignment data is produced and marked up interactively, most conveniently with the aid of the SoundIndex tool developed by the project (Michel Jacobson).

Over 150 text/sound documents of spontaneous speech recorded for the purpose of linguistic research, mainly in little-known or endangered languages, have been archived to date (4/2002). Currently, 71 documents in 15 different languages, prepared by 8 linguists associated with the LACITO, are freely accessible, along with all of the project software and documentation, on the Archive project website. Access is over the Internet, using standard browsers and multimedia players on a variety of platforms (Windows, Mac, Linux).

References

Jacobson, Michel and Boyd Michailovsky. 2000. A Linguistic Archive on the Web. Workshop on Web-Based Language Documentation and Description. Philadelphia. December 2000. (http://morph.ldc.upenn.edu/exploration/expl2000/papers/michailovsky/index.htm)
[This is the most up-to-date description of the overall project architecture, and particularly of the use of XSL.]

Michailovsky, B. 2001. The LACITO Archive project markup. Linguist List Workshop: The Digitization of Language Data: The Need for Standards. Santa Barbara, June 2001. (http://linguist.emich.edu/~workshop/markup/lacito/Lmarkup.pdf)

Other descriptive material and all software plus the XML and sound documents (MP3) can be downloaded from the project website:
http://lacito.archivage.vjf.cnrs.fr.