HAL
open science

A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive

Boyd Michailovsky, Alexis Michaud, Séverine Guillaume

To cite this version:

Boyd Michailovsky, Alexis Michaud, Séverine Guillaume. A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive. 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA 2011), Oct 2011, Hsinchu, Taiwan. pp.14-23, 10.1109/ICSDA.2011.6085973 . halshs-00620893

HAL Id: halshs-00620893

https://shs.hal.science/halshs-00620893

Submitted on 12 Sep 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author manuscript, published in "2011 International Conference on Speech Database and Assessments (Oriental COCOSDA 2011), Taiwan, Province Of China (2011)"

halshs-00620893, version 1 - 9 Sep 2011

Keynote speech at Oriental-COCOSDA 2011*, Hsinchu, Taiwan, October 26th-28th 2011*

A SIMPLE ARCHITECTURE FOR THE FINE-GRAINED DOCUMENTATION OF ENDANGERED LANGUAGES: THE LACITO MULTIMEDIA ARCHIVE

*Boyd Michailovsky** *Alexis Michaud* °* *Séverine Guillaume**

*LACITO-CNRS, France °CEFC-CNRS, Taipei

ABSTRACT

The LACITO multimedia archive [1] **provides free access to documents of connected, spontaneous speech**, mostly in "rare" or endangered languages, recorded in their cultural context and transcribed in consultation with native speakers. Its goal is to contribute to the documentation and study of a precious human heritage: the world's languages. It has a special strength in languages of Asia and the Pacific.

The LACITO archive was built with little personnel and less funding. It has been devised, developed and maintained over two decades by two researchers assisted by one engineer. Its simple architecture is based on current standards: Unicode character coding and XML markup; and Dublin Core/Open Language Archives Community recommendations for metadata. **The data can be consulted online with any standard browser.** The technical simplicity of the tools developed at LACITO makes them **suitable for the creation of similar databases at other institutions**. (For instance, tools from this archive were successfully adapted in the creation of the Formosan Languages archive [2].)

提要：法國科學院口傳語言與文化研究所（CNRS-LACITO）的多媒體資料庫[1]提供自由網上訪問及下載服務，主要內容爲瀕臨消失語言的長篇語料文件：錄音或錄像以及詳細的記音、註釋、翻譯。這些資料都源於第一手的本地調查：在當地的自然語境中錄製并與發音人共同進行後期註解工作。目的在於世界瀕臨消失語言的檔案化及科學研究。該語言庫的亞洲及太平洋語言區域資料尤爲豐富。

LACITO 多媒體數碼化語言庫成立至今已有接近二十年的歷史，在人手極其有限(兩名研究員及一名工程師)的情況下堅持不懈，使得資料庫初具規模。它的管理工具模板、技術架構以 XML，Unicode，Dublin Core ，OLAC 等通用標準爲基礎，便於更新簡化。此次演講著重于 LACITO 資料庫工具模板的簡化技術，使得此資料庫工具模板可以應用在眾多不同的研究機構，例如"台灣南島語數位典藏庫"[2]就曾經成功地借鑒了 LACITO 資料庫工具模板爲己所用。

Index Terms: Multimedia corpora, language documentation, endangered languages, spontaneous speech, interlinear glossing, online databases, long-term preservation

1. INTRODUCTION: THE NEEDS ADDRESSED BY THE LACITO ARCHIVE

The goal of the LACITO Linguistic Archive project is to conserve and disseminate recorded and transcribed oral literature and other linguistic materials in (mainly) unwritten languages, **giving simultaneous access to sound recordings and text annotation**.

The necessity to document the world's languages is now well-known to linguists and the general public. Fewer people are aware of the dismal current state of linguistic documentation in many research institutions. "**Enormous amounts of data – often the only information we have on disappearing languages – remain inaccessible both to the language community itself, and to ongoing linguistic research"** [3], and eventually disappear.

> "The data that we create (…) should be reusable, both by ourselves and by others. First because any claims that we make based on that data must themselves be replicable and testable by others, and second, because the effort of creating a digital representation of the data should not be duplicated later by others, but used as a foundation that can be built on. The same issue is being faced by scholars in many disciplines (…). **This is all the more important when a linguist makes the only recordings for an endangered language–one that may no longer be spoken in the near future**." [3]

The present paper, which provides an update of earlier published descriptions of the LACITO archive (in particular [4]), presents the basic structure and client-server architecture of the database. We also point out some possibilities for further development, in the belief that the technical simplicity of the architecture makes it **suitable for the creation of similar databases at other institutions**: for instance, tools from the LACITO archive were successfully adapted in the creation of the Formosan Languages archive [2].

halshs-00620893, version 1 - 9 Sep 2011

With the increase in storage facilities, linguists now record considerable amounts of audio and video recordings. An individual researcher often has hours (or tens of hours) of recordings. These data are highly diverse, reflecting the variety of the linguist's interests: continuous speech (folk tales, life stories…), dialogues, elicitation sessions, etc. The LACITO (Langues et Civilisations à Tradition Orale) laboratory of the French CNRS (Centre National de la Recherche Scientifique) has undertaken the archiving of the data collected by laboratory members (also proposing this service to colleagues from other laboratories) in order to **facilitate research on these materials** as well as to ensure their **long-term conservation**. **Over the last ten years, the number of languages in the LACITO Archive increased from 20 to more than 60.** Currently, the archive contains over a thousand audio files in 67 languages, of which 350 have a transcription and annotation.

A major advantage of **digital storage** is that it **allows the recordings and their textual documentation to be kept together**. As the hardware evolves, it is essential to maintain the link between these two components of the linguistic documents. The annotation, and in particular the transcription of the recorded speech, is typically worked out in the field in consultation either with the original speakers or with other members of the same speech community. It is just as irreplaceable as the recordings themselves and requires far greater effort to prepare. As a rule of thumb, the production, verification and formatting of the transcription and glosses for a ten-minute recording requires ten to fifty hours.

The link between sound and text annotation is established through metadata, eliminating the risk of separation of the recorded speech and its annotation. Each resource has its own metadata that describe the content of the resource and list the linked resources, using a unique identifier and its associated URL.

The logical next step was **the synchronization of sound and text annotation**. This allows for simultaneous access to sound and text annotation, which greatly facilitates browsing the data and querying them using computer-aided research methods. It also facilitates the confrontation of the annotation with the recorded speech, and thus its correction and enrichment.

The LACITO Archive thus hosts two types of documents: (i) recordings and (ii) the corresponding textual annotations -- transcriptions, glosses and translations. In addition, metadata files contain descriptions of all resources available.

2. TECHNICAL CHOICES: ADHERING TO WEB STANDARDS

2.1. Why XML?

The LACITO Archive was started at a time when the information industry had moved toward new standards in the areas of structured text markup, multi-byte character sets, and access to multimedia, i.e. areas that correspond to the needs of our project. Unicode promised a solution to linguists' perennial font problems [5]. We were influenced by the Guidelines for Electronic Text Encoding and Interchange [6], which proposed SGML markup (Standard Generalized Markup Language [7]) for a wide variety of document types used in literary and linguistic studies. **We did not adopt the Text Encoding Initiative piecemeal because the predefined markup went far beyond our needs in most areas while remaining rudimentary in the domain of speech annotation.**

As the project was getting started, the slightly simplified XML (eXtensible Markup Language) dialect of SGML was being developed, for web applications in particular [8]. When the World Wide Web Consortium (W3C) adopted XML as a recommendation, it triggered **widespread development of free and public domain software**, a phenomenon that had previously occurred around the display language HTML (HyperText Markup Language) but never around the original SGML. **The Archive project** (with a number of other academic projects, including a renewed TEI consortium) **adopted XML markup in 1997 for both text annotation and metadata** (which are contained in two separate but cross-referenced files).

All the formats and standards chosen for the resources in the LACITO Archive are described below.

2.2. Why choose WAV/PCM for audio files?

Most of the audio files in the LACITO Archive are in the WAV/PCM format (*WAVEform audio file format / Pulse Code Modulation)*. This recommendation ensures easy format migration in future, as current formats become obsolete. For a good quality recording, sampling and quantization of an audio file are recommended at 44,100 Hz and 16-bit (or higher settings). We expect contributors who put huge amounts of time and effort into the transcription of analysis of speech data to pay attention to the quality of the recordings: avoiding formats that involve lossy compression (such as MP-3) and following "good practice" in recording, as taught in *Field Methods* classes worldwide. In our experience, **data that are collected with long-term preservation in view tend to be of higher quality than those whose archiving only came as an afterthought**. This has considerable advantages for research, since data sets with technically inappropriate recordings, incomplete information for the metadata file or unclear transcriptions are obviously much harder to use in research.

3. THE DOCUMENTS CONTAINING THE TRANSCRIPTION AND TEXT ANNOTATION

3. 1. Markup and rules

The XML markup used for the text annotation makes the structure of documents fully explicit. For example, the <S>

halshs-00620893, version 1 - 9 Sep 2011

level ("sentences") is identified explicitly by delimiting tags containing a label that we defined -- <S> at the beginning and </S> at the end -- and not implicitly by its position on the page, font, style, etc. **Text marked up** in this way **with labels identifying the functional, rather than the typographical, nature of each element is "logically structured text"**.

The structure of documents reflects traditional practice in interlinear glossing, which will be familiar to linguists. In linguistic publications, this structure is implicit in the typographical format, as in (1) below. When the data are transformed into an XML document, this structure is made fully explicit.

(1) (…) dy˥-ku˩lu˩ | mæ˧ʈʂ˥ ɖʂɯ˧-i˧ lɑ˧-ʂɯ˧ ||

[王母娘娘] 在地（天下）找了一个女婿。

[The Heavenly Mother] looked for a son-in-law down on earth.

dy˥	ku˥lu˥	mæ˧ʈʂ˥	ɖʂɯ˧	i˧	lɑ	ʂɯ˧
地	里面	女婿	一	量词	实施	寻找
earth	inside	son_in_law	one	CL	ACCOMP	look_for

The basic annotation is the transcription of the original language. Typically, the transcription of an entire **text** is divided into **sentences**, and sentences into **words**. Translations into any number of languages (here: Mandarin and English) may be provided, aligned with the transcription at different levels: glosses for words, and free translation at the sentence level and the text level. In the example above, only one level of glosses is provided, but two levels can be distinguished: the **word** and the **morpheme** (e.g. in an annotation of an English text, the word 'comes' can be analyzed into two morphemes: the verbal root, 'come', and an inflectional morpheme for the 3rd person singular). There exist widespread conventions for interlinear morpheme-by-morpheme glosses, in particular the Leipzig Glossing Rules [9]; for a thorough presentation of morphemic glossing, see [10]. Interlinear glosses aid in the linguistic study of the text by indicating the contribution of each word. Also, **multilinear annotation brings out formal differences between the forms of words in context and in lexicon entries**: for instance, in (1), the word-by-word glosses make it clear that /lɑ/ (glossed as ACCOMPLISHED marker) does not have a lexical tone of its own, unlike the other words in this excerpt. Here, it surfaces with a Mid tone (indicated by the IPA mark for Mid tone : ˧). Also, the postpositions /ku˥lu˥/ 'inside', which have a lexical High tone, surface with a changed tone (Low tone) in this context: /ku˩lu˩/.

The free, "sentence"-level translations serve as an intelligible, if not always smooth or elegant, translation. A more reader-friendly translation can be added for the entire text -- again, in any number of languages.

In the XML document, the text annotation data are organized into elements whose beginnings and ends are identified by structural tags. Between the start-tag and the end-tag, an element may contain other elements or text annotation data, or both. For instance, a translation as 'son' for a word is marked up as <TRANSL>son</TRANSL>. In addition to tags, the structural markup may include attribute-value pairs. For example, the language of the gloss 'son' can be indicated by an attribute as <TRANSL xml:lang="en">. Provision is made for reference to external resources, which may contain non-XML data such as sound or images. As mentioned, an element may contain other elements, but any contained element must be entirely nested in the containing element; there can be no overlap. The structure of an XML document is thus equivalent to a rooted tree, the unique root element containing all the others.

The structural properties of an XML document or of a class of such documents can be declared in a Document Type Declaration (DTD) or in an XML Schema Definition (XSD). The latter has become more common in recent years, and our project accordingly shifted from DTD format declaration to XSD format declaration.

XML document-processing software begins by parsing the document to verify its "well-formedness", that is, its conformity to the syntactic rules of XML. A "validating parser" goes a step farther and verifies the conformity of a document to a particular XSD. **The structure of the LACITO XML markup is now defined in an XSD which we continue to call the "LACITO DTD".** The entire XSD is available online [11]. Fine details in the XSD are of no consequence for our software tools, including authoring tools, as these have only minimal expectations about XML documents.

The text annotation data is contained in elements labeled TEXT, S, W and M, corresponding to the "text", "sentence", "word" and "morpheme" levels of structure. Each S element is uniquely identified (via the "id" attribute). The metadata concerning the whole text, such as its title, the date, the speaker, the language, etc. are stored separately. The main elements of the LACITO Schema Definition for text annotation are presented in (2).

(2) The only required element is the TEXT markup and it has two required attributes, *xml:lang* which declares the language spoken in the whole text and *id* which is the unique identifier of the resource (OAI identifier [12]). TEXT can contain 4 types of elements: FORM which contains the whole transcription, TRANSL for the translation of the text, NOTE to add some comments if necessary, and S.

The markup S divides the text into sentences. It has one required attribute which is the identifier *id* of the sentence (which allows playing the audio file sentence by sentence). S can contain differents elements: FORM to add a transcription, TRANSL for a translation and W to indicate the division of the sentence into words.

Similary, the element W can contain FORM, TRANSL and M, the latter indicating the division of the word into morphemes (each of which can in turn receive a

halshs-00620893, version 1 - 9 Sep 2011

gloss/translation: a morpheme contains the elements FORM and TRANSL).

Also, as mentioned above, the element TRANSL can have an attribute *xml:lang* which declares the language used for the translation or the glosses using the ISO 639-3 Standard for Language Codes. This system allows multiple translations for each sentence, word or morpheme, in a fully unambiguous way. Furthermore, it is possible to provide several transcriptions at any level (TEXT, S, W, M), for instance an orthographic transcription and a phonetic/phonological transcription.

The markup of a fragment of a text annotation is shown in (3), corresponding to the data shown in (1) above.

(3)
```
<S id="FemmeCelesteS3">
<FORM>dy˥-ku˩lu˩, mæ˧y˥ ɖɯ˧-i˧ lɑ˧-ʂɯ˧.</FORM>
<TRANSL xml:lang="cn">在地（天下）找了一个女婿。
</TRANSL>
<TRANSL xml:lang="en">[The Heavenly Mother] looked for a son-in-law down on earth. </TRANSL>
<W><FORM>dy˥</FORM>
<TRANSL xml:lang="cn">地</TRANSL>
<TRANSL xml:lang="en">earth</TRANSL>
</W>
<W><FORM>ku˥lu˥</FORM>
<TRANSL xml:lang="cn">里面</TRANSL>
<TRANSL xml:lang="en">inside</TRANSL>
</W>
<W><FORM>mæ˧y˥</FORM>
<TRANSL xml:lang="cn">女婿</TRANSL>
<TRANSL xml:lang="en">son_in_law</TRANSL>
</W>
<W><FORM>ɖɯ˧</FORM>
<TRANSL xml:lang="cn">一</TRANSL>
<TRANSL xml:lang="en">one</TRANSL>
</W>
<W><FORM>i˧</FORM>
<TRANSL xml:lang="cn">量词</TRANSL>
<TRANSL xml:lang="en">cl</TRANSL>
</W>
<W><FORM>lɑ</FORM>
<TRANSL xml:lang="cn">实施</TRANSL>
<TRANSL xml:lang="en">accomp</TRANSL>
</W>
<W><FORM>ʂɯ˧</FORM>
<TRANSL xml:lang="cn">寻找</TRANSL>
<TRANSL xml:lang="en">look_for</TRANSL>
</W>
</S>
```

The choice of a simple structure for the digital documents in the LACITO Archive relates to important characteristics of these documents. In the literature on speech annotation, it is usual to consider transcriptions as source data, reserving the term "annotation" for translations and analyses. The transcription of spontaneous speech in little-known languages, however, is built on a set of linguistic hypotheses many of which are devised by the transcriber: the phonemic analysis and the choice of symbols (sometimes including a practical orthography) may change over the years, as well as the morphological analysis.

In our experience, **a simple embedded structure (text, sentences, words, morphemes) provides a convenient backbone for the annotation**. Users may want to add more information, on the basis of theoretical commitments, personal preferences, cross-linguistic differences or research purposes, e.g. part-of-speech tagging for syntactic research, or phoneme-level alignment with the recording for phonetic research. Sentences may be grouped into broader units such as turns of speech or "oral paragraphs". **No claims are made that the division into sentences is "water-tight" in any sense.** We are fully aware that different decisions can be made about the division of the same text into sentences, and do not provide any explicit criteria on how the division is to be made. **Our <S> is just a label for an element which can just as well be an "utterance"**; many contributors divide their texts into relatively short <S> portions, on the order of three seconds each. This is left open to the contributor's decision.

The point here is that "LACITO DTD" is a solid, functioning model that can easily be expanded in the direction of additional structural levels or additional data categories (e.g. part-of-speech tagging). **Once the text annotation has been prepared according to the "LACITO DTD", the data can be archived, and displayed on any web browser; they can then be used for numerous specific research purposes, at which stage the annotations can be straightforwardly converted to other formats and further enriched.**

Revising the "LACITO DTD" itself is also quite possible. Since the LACITO markup has the same structure at each level, it would be possible, in a new version of the DTD, to have any amount of numbered levels. At each level, a heading element would specify what kind of item that level represents. Stylesheets and scripts of the type currently used could be adapted to operate over this new structure.

To recapitulate, the structure is hierarchical; it can be represented as a tree. At the same time, the annotation has linear structure in that the transcription units at each level are sequentially ordered. This linear structure is put into correspondence with the temporal structure of the recorded sound.

3.2. Time-alignment

Sound recordings can be segmented as finely or as coarsely as desired, certainly more finely than required to detect any linguistically significant element. The main decision to be made in aligning sound and text annotation is

halshs-00620893, version 1 - 9 Sep 2011

the *granularity*, that is, the length of the smallest text elements to be time-anchored, and hence the length of the smallest segments of the sound resource which can be accessed. Since the documents under discussion here consist of connected text, it was decided to anchor units above words or phonemes. On the other hand, because the languages of the texts are unfamiliar to most users, it was desirable to keep these units short. **The simplest choice was to anchor the S-units, although anchoring to the word is also permitted by the schema**.

The original recordings are kept whole, with the portions corresponding to each S identified by their starting and ending time-offsets as measured from the beginning of the sound resource. Accessing segments of a sound file has required some in-house markup and software development. To mark up the alignment data, an AUDIO element is incorporated into each time-aligned element, with start- and end-time offsets as attributes in the XML text annotations. The oblique at the end of the empty AUDIO element obviates the need for an end-tag. We will discuss below how these offsets are interpreted.

In the transcription, **the start-offset of each S may typically coincide with the end-offset of the preceding S, but this is not necessary, and overlap is possible**. Thus, in a conversation, speaker B might start before speaker A has finished. In this case, the XML structure does not reflect the overlap, and indeed XML would not allow partial overlap between the end of <S who="A"> and the beginning of <S who="B">. Nevertheless, the temporal overlap in the recorded sound can be indicated, because the values of these attributes are not XML structural elements and are not verified by the XML parser.

Time-anchoring of elements at hierarchically different levels, for example S and W, does not imply partial overlap, so the XML markup can mirror the temporal structure. Note that the parser will not verify that W time-offsets are nested between those of the containing S.

The markup for the time alignment is shown in (4), corresponding to a fragment of the data shown in (1) above.

(4) <S id="FemmeCelesteS3">
<AUDIO start="4.2" end="6.9"/>
<FORM>dy˥-ku˩˩˨˩, mæ˧˥ɣ˥ ɖɯ˧-i˧ lɑ˧-ʂɯ˧.</FORM>
<TRANSL xml:lang="cn">在地（天下）找了一个女婿。
</TRANSL>…

This time alignment markup indicates that the sentence is pronounced between 4.2 seconds and 6.9 seconds in the recording.

The metadata file will specify which recording is linked with which text annotations, allowing the user to listen to the recording while reading the transcription and annotations.

3.3 The metadata

Metadata provide a description of the data. They allow for the organization and cataloguing of all the resources in the archive, such as the time and place of recording and the participants involved in the recording as well as in text annotation. In addition, metadata allow for the creation of links between available resources.

The standard chosen by the LACITO Archive for cataloging its digitalized sound and text resources is the XML-based Dublin Core metadata set [13], as adapted by the Open Language Archives Community (OLAC) [14].

The Dublin Core provides a small set of fundamental pieces of information by means of which most resources can be described and catalogued. Using only 15 basic tags, a Dublin Core metadata record can describe physical resources such as books or digital materials (video, sound, image, text files…) and more. All the Dublin Core tags begin with *dc:* or *dcterms:*.

OLAC is a community where a consensus on best current practice for the digital archiving of language resources is worked out and developed. The OLAC standards are based on the complete set of Dublin Core metadata terms, but this format allows for the use of extensions to express community-specific qualifiers. For example, with Dublin Core one can define a list of contributors who have worked on a given resource, but one cannot specify their role, whereas OLAC gives the possibility to specify the role of each contributor, such as speaker, researcher, translator, or depositor, e.g. <dc:contributor olac:code="recorder" xsi:type="olac:role">Michailovsky, Boyd </dc:contributor>.

The metadata of all resources appear in an XML document where each resource is described in an *item* tag. To distinguish resources, each item is defined by its unique resource identifier and the date of its importation in the archive repository.

For the LACITO archive the minimum information required for the metadata for each resource is described below:

- *dc:title* (title)
- *dc:subject* (language being documented)
- *dc:language* (for all languages spoken in the recording or used for translation on text annotations)
- *dcterms:spatial* (the place of recording)
- *dcterms:accessRights* (Access right to specify whether a given resource is with limited access or freely available for non-commercial use)
- *dc:identifier* (OAI identifier for retrieving the resource in the archive repository)
- *dcterms:isRequiredBy* or *dcterms:Requires*, for creating links between resources when its necessary. The most obvious link is between the audio file and its text annotation. All links are based on unique identifiers (OAI unique identifier).

halshs-00620893, version 1 - 9 Sep 2011

It is possible to add much more information for a given resource, as explained on the Dublin Core website [13].

Here is an example of metadata set, for a text annotation corresponding to the data shown in (1):


```
<crdo:item crdo:datestamp="2011-05-13" crdo:id="crdo-NXQ_FEMMECELESTE">
<dc:title xml:lang="fr">Femme céleste</dc:title>
<dc:subject xsi:type="olac:language" olac:code="nxq">Lazé</dc:subject>
<dc:language xsi:type="olac:language" olac:code="nxq"/>
<dc:language xsi:type="olac:language" olac:code="fr">français</dc:language>
<dc:language xsi:type="olac:language" olac:code="cn">chinois mandarin</dc:language>
<dcterms:spatial>Muli, Liangshan, Sichuan</dcterms:spatial>
<dc:contributor xsi:type="olac:role" olac:code="depositor">Michaud, Alexis</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="researcher">Michaud, Alexis</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="speaker">Tian, Xiufang</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="interviewer">Michaud, Alexis</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="annotator">Michaud, Alexis</dc:contributor>
<dc:publisher>Lacito/CNRS</dc:publisher>
<dc:type xsi:type="olac:discourse-type" olac:code="narrative"/>
<dc:type xsi:type="olac:linguistic-type" olac:code="primary_text"/>
<dc:format xsi:type="dcterms:IMT">text/xml</dc:format>
<dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
<dcterms:conformsTo xsi:type="dcterms:URI">oai:crdo.vjf.cnrs.fr:crdo-dtd_archive</dcterms:conformsTo>
<dc:identifier xsi:type="dcterms:URI">http://crdo.risc.cnrs.fr/exist/crdo/michaud/nbf/FemmeCeleste.xml</dc:identifier>
<dcterms:isFormatOf xsi:type="dcterms:URI">http://crdo.risc.cnrs.fr/exist/crdo/michaud/nbf/FemmeCeleste.xhtml</dcterms:isFormatOf>
<dcterms:requires xsi:type="dcterms:URI">oai:crdo.vjf.cnrs.fr:crdo-NBF_FEMMECELESTE_SOUND</dcterms:requires>
<dcterms:accessRights>Freely available for non-commercial use</dcterms:accessRights>
<dc:rights>Copyright (c) Michaud, Alexis</dc:rights>
<dcterms:isPartOf xsi:type="dcterms:URI">oai:crdo.vjf.cnrs.fr:crdo-COLLECTION_LACITO</dcterms:isPartOf>
<crdo:collection>Lacito</crdo:collection>
</crdo:item>
```


In the LACITO Archive, **the metadata are stored in a repository which follows the Open Archives Initiative Protocol for Metadata Harvesting** (OAI-PMH). The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The essence of the open archives approach is to enable access to Web-accessible material through interoperable repositories for metadata sharing, publishing and archiving. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest the metadata. This OAI-PMH repository hosts not only the metadata of LACITO Archive but also those of various other institutions: it was created to facilitate access to a large number of language resources, on major languages as well as on less-documented languages. The LACITO is both a data provider and a service provider.

4. SOFTWARE TOOLS AND IMPLEMENTATION

4.1. Authoring tools

Supposing you are by now determined to try out formatting your data along the lines described above, the next question is how to go about doing so. There are several solutions for preparing a text transcription/annotation in XML format.

(i) Users who are familiar with writing scripts can type the text and interlinear glosses as plain text, and then run a script (e.g. a Perl script) that adds the XML markup. This is what the authors of the present paper currently do; a script is available online, with an example [15].

(ii) Users of Toolbox [16], ELAN [17] or other authoring software can convert their data automatically to the "LACITO DTD" by means of scripts.

(iii) Users with little previous experience of working with linguistic markup can use the authoring tool ITE (Interlinear Text Editor), developed by Michel Jacobson at LACITO [18].

Once the annotated text is ready, the time-offsets must be added, to align the transcriptions with the recordings. An authoring tool, SoundIndex [18], was developed (also by Michel Jacobson) to determine these time-offsets and add them into the markup. SoundIndex incorporates a sound editor and a text editor. It plays a sound file (a wide variety of formats is supported, including WAV and AIFF) and simultaneously displays the waveform and the transcription prepared by the linguist. It allows the user to mark on the waveform the start- and endpoints corresponding to each segment of the transcription; the corresponding time-offsets are then recorded. The current version works directly with XML documents. The user specifies the sentence to be aligned. When the corresponding part of the waveform is identified, the program generates an AUDIO tag with the appropriate time-offsets and incorporates it into the XML annotation document. SoundIndex uses a standard XML parser and can open any well-formed XML document on condition that it respects the LACITO DTD rules. It reads the text from a standard Document Object Model (DOM site) interface and uses a Unicode-aware text editor to display the content, highlighting the tags. The current version was developed in Tcl/Tk using the Snack sound extension and the Tk text widget, following the example of Transcriber [19].

Although the process of time-alignment might seem tedious, field linguists who use SoundIndex generally profit from it to improve their transcriptions. The quick and precise access which the program affords to time-aligned texts greatly facilitates verification.

halshs-00620893, version 1 - 9 Sep 2011

4.2. Browsing tools

To facilitate dissemination of the project documents, the browsing system is built around scripts, stylesheets and standard web browsers. The documents are accessible to any client equipped with a standard browser and a soundcard. In practice, some minor adjustments may be necessary, e.g. downloading software for playing the recordings (such as Quicktime) if no appropriate software was available on the computer.

After choosing a language the user has to make a choice of resource. Once a resource is selected a screen appears with an applet at the top for playing the audio file with text annotations below.

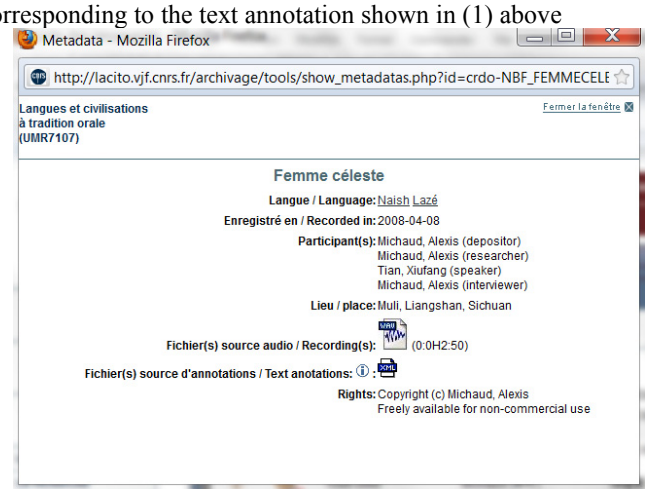
Through the use of stop and play button placed before each sentence, the user can choose to listen to one sentence at a time. The audio playing applet also allows the user to listen to the whole text from the beginning to the end.

Text annotations can include several transcriptions (phonetic, phonologic, orthographic), translations in many languages and glosses for words and morphemes. That's a lot of information on the screen so the option of hiding some parts of annotations information was added. Furthermore, if there is more than one translation, the user can choose to see all translations or only the one which interests them (see Fig. 3).

The system operates in a way which may sound paradoxical but which saves a lot of time and effort. The LACITO Archive's metadata are fed into the OAI repository as new documents are added to the database. Then we harvest our own metadata from the OAI repository: the LACITO Archive's harvester gets the LACITO Archive's metadata from the OAI repository and puts them automatically into an XML file. All the necessary information can be retrieved from this metadata file: where the recording is, which text annotation is linked with it, and which pieces of information are available about the resource. Thus we do not need to make any in-house database management.

When a user requests a resource, an XSL stylesheet is applied to the data. A generic XSLT processor on the server uses the chosen stylesheet to format the required elements of the XML document as HTML, which is transmitted to the client machine. The first stylesheet creates an HTML view of the metadata for a specific resource when the user tries to access it.

Fig. 1: A view of the metadata of the recording corresponding to the text annotation shown in (1) above



The second stylesheet creates an HTML view of both linked recording and text annotation. The HTML document contains references to the audio file and to a playing applet, which are called by the client browser. It also contains JavaScript functions which handle user input (e.g. the mouse-clicks which select text for listening) and communicate with the applet. The applet handles access to segments of the audio file, a function which the browser cannot handle directly because it is not provided for in HTML.

Fig. 2. Part of an XSL stylesheet to transform XML text annotation into HTML view

```
<xsl:template match="annot:TEXT">
<table width="100%" border="1" bordercolor="#993300" cellspacing="0"
cellpadding="0"><tr><td>
<table width="100%" border="0" cellpadding="5" cellspacing="0"
bordercolor="#993300" class="it">
<tbody>
<xsl:for-each select="annot:S">
<tr class="transcriptTable">
<td class="segmentInfo" width="25">S<xsl:value-of
select="position()"/>
</td>
<td class="segmentContent" id="{@id}">
<a href="javascript:boutonStop()"><img src="stop.gif" alt="stop"/></a>
<a href="javascript:playFrom('{@id}')"><img src="play.gif"
alt="écouter"/></a>
<xsl:if test="annot:FORM">
<div class="word_sentence">
<xsl:for-each select="annot:FORM"><xsl:choose><xsl:when
test="@kindOf"><xsl:if test="@kindOf='phono'">
<div class="transcription1"><xsl:value-of select="."/></div>
</xsl:if><xsl:if test="@kindOf='ortho'">
<div class="transcription2"><xsl:value-of select="."/></div>
</xsl:if><xsl:if test="@kindOf='phone'">
<div class="transcription3"><xsl:value-of select="."/></div>
</xsl:if><xsl:if test="@kindOf='transliter'">
<div class="transcription4"><xsl:value-of select="."/></div>
</xsl:if></xsl:when><xsl:otherwise>
<xsl:value-of select="."/></xsl:otherwise></xsl:choose>
<br/>
```

halshs-00620893, version 1 - 9 Sep 2011

```
</xsl:for-each></div></xsl:if><xsl:if
test="not(annot:FORM) and (annot:W/annot:FORM or
annot:W/annot:M/annot:FORM)"><xsl:for-each select="annot:W">

<div class="word_sentence"><xsl:choose>
<xsl:when test="annot:M/annot:FORM">
<xsl:for-each select="annot:M/annot:FORM">
<xsl:value-of select="."/><xsl:if test="position()!=last()">-</xsl:if>
</xsl:for-each></xsl:when><xsl:when test="annot:FORM">
<xsl:value-of select="annot:FORM"/>
</xsl:when></xsl:choose></div></xsl:for-each></xsl:if>
<br/>
<xsl:if test="annot:TRANSL">
<xsl:for-each select="annot:TRANSL[@xml:lang='en']">
<div class="translation1"><xsl:value-of select="."/></div>
</xsl:for-each>
<xsl:for-each select="annot:TRANSL[@xml:lang='fr']">
<div class="translation2"><xsl:value-of select="."/></div>
</xsl:for-each><xsl:for-each
select="annot:TRANSL[@xml:lang!='fr' and @xml:lang!='en']">
<div class="translation3"><xsl:value-of select="."/></div>
</xsl:for-each></xsl:if><br/>…
```

In the remainder of this section, the operation of the browsing system is described in more detail.

Fig. 2 is an excerpt from one of our XSL stylesheets, whose function is to produce HTML. The first "rule" (contained in the first xsl:template element) instructs the processor, when the TEXT element is encountered, to create an HTML table. For each S element detected in the text, the number of the sentence and a *play* and a *stop* button are displayed -- if the user clicks on these buttons, a JavaScript function (play(id) or stop(id)) is called and launches the recording of the given sentence exactly at the point where the sentence begins.

The next step is to detect if the sentence has a transcription. The markup FORM is first searched at the sentence level. If there are transcriptions, they are displayed. In case there is no transcription for the sentence, the program will search at the word level -- W. If W has a transcription, a concatenation of all the words of this sentence is made to create a transcription. If it does not, the program will continue searching and do the same at the morpheme level -- M.

So, on the first line of the table we can see the id of the sentence, the *play* and *stop* button, and the transcription. If there is more than one transcription, each transcription is displayed on a different line.

On the second line of the table, the program looks for a translation at the sentence level -- markup TRANSL. If it finds one or more translation, each translation appears on a separate line, just like for transcriptions.

On the third line of the table, the glosses of words or morphemes are displayed. If an M markup is detected and if M contains a transcription and a translation or glosses, the transcription of each morpheme is put in the first line of a table and the glosses in the second line. Each morpheme is in a different cell, so each transcription of a morpheme appears above its glosses. This reflects the structure of the sentence and brings out the correspondence between a morpheme and its glosses. When glosses are provided in several languages, each one appears on a separate line. If no glosses are detected on the morpheme level, the same process is applied for the word level.

Application of the stylesheet (Fig. 2) to the sample document produces the HTML document shown in Fig. 3.

Fig. 3: HTML resulting from application of the XSL stylesheet

The Archive Project browsing and interrogation systems are entirely distinct from the authoring tools. **Linguistic documents prepared by the project are browsed and queried using standard browsers through a standard web interface, facilitating their dissemination.** The browsing system is relatively stable, but the data management and querying software is in an early stage of development, in part because of a lack of defined standards and hence of standard tools. General searching, word-indexing, and concordancing modules have been developed in prototype form.

5. WEB HOSTING AND LONG-TERM DATA PRESERVATION

Finding solutions for perennial archiving (long-term data preservation) and web hosting is a central concern for the creators of digital open archives. The technical solution

halshs-00620893, version 1 - 9 Sep 2011

used by the LACITO Archive is in keeping with the same principles that guide the technical choices explained throughout this paper: building a solution of our own was out of the question, because (i) our laboratory does not have sufficient resources, and (ii) a research laboratory, unlike a library, is not in principle a perennial institution that can guarantee data preservation over very long periods of time. **The solutions for web hosting and long-term archiving used for the LACITO Archive are those jointly elaborated for digital data by institutions of national scope: the CINES** [20] **for archiving, and the computing centre of IN2P3** [21] **for web hosting.** The same solutions are used by other multimedia archives of linguistic data, such as the archive based at Aix-en-Provence [22].

This implies that once the data are archived, we need not worry about server problems. The team at IN2P3 takes charge of this 24/7 for a great number of research laboratories in all areas of science. Also, we do not need to be constantly watching out for new storage technologies: as the storage media evolve, data migration is taken care of by CINES.

This aspect of our project will not be discussed here in any detail, as the emphasis is on the structure of the documents and the tools for their creation. However, it may be useful to point out that **adhering to web standards in preparing linguistic documents greatly facilitates archiving and long-term preservation**. Archivists cannot reasonably accept deposits whose format is not sufficiently explicit, as they cannot guarantee that these data will be readable in future. The data formats that we use (XML for text data and WAV/PCM for audio, also allowing for the possibility of AIFF format) meet the requirements that archivists place on digital documents, ensuring that the process of archiving goes smoothly. By contrast, commonly used formats such as those of MS-Word documents (.doc, .docx, etc) cannot be archived across-the-board. They can be converted into PDF files and archived in PDF format, but this entails severe limitations in terms of data querying and browsing. Structured text is clearly the way forward.

6. CONCLUSIONS

6.1. Practical advantages of the LACITO Archive

The LACITO Archive Project has developed a method of digitizing the traditional linguists' annotation of oral texts in XML, including time-alignment with digitized recordings, and a system for browsing and querying the resulting documents over the Internet. 350 texts annotations have been made publicly available on the project web site. Texts will continue to be prepared and to be made publicly available (except in cases where issues of intellectual property constitute insuperable obstacles; this constitutes a topic of its own).

From the technical point of view, the key to the success of the Archive Project is the adherence to standards, XML in particular, and the use wherever possible of generic tools which are maintained by others to keep up with changing hardware, web standards, etc. Reliance on generic software allows us to concentrate our limited technical resources on the development of specifically linguistic tools. We are able to choose between several different generic XML parsers and XSL processors, changing them as improved versions are distributed. The fact that many generic tools are freely available is an obvious advantage; the fact that they are open-source has been helpful in that it has allowed us to add specific modules to them while retaining their generic functionality. We have taken advantage of this to incorporate regular expression matching into a query processor and are planning to incorporate complex sort order parametrization into XSLT.

The LACITO Archive Project is one of a number of projects which are converging on common principles: the use of logically structured markup, in particular XML, at least as an exchange and processing format, and the use, to the extent possible, of generic software so that processing is accomplished through high-level stylesheets and queries. An important consequence of the simplicity of the structure is that the software developed at the onset of the project can still be used today -- a life span that is above average, and that saves much-wanted engineering time.

As explained above, it is not the aim of the project to propose a single, standard annotation. The text corpora are in typologically diverse languages and have been prepared by linguists with differing research interests. Under these circumstances it is illusory to imagine that agreement could be reached on a single markup, however detailed. Although it may be possible to maintain a single DTD to cover the annotations, this, too, could become undesirable if it makes it more difficult to ensure the coherence of any one corpus annotation or to adapt tools to process it. **In our view, standardization is to be sought in adherence to the principle of logically structured text --which makes it relatively easy to transform one markup into another-- rather than in particular markup conventions.**

As more projects adhere to common standards (e.g. PARADISEC [23] and EOPAS [24]), we expect that distributed development and interoperability of software for field linguistics and other specialities will become a reality.

6.2. Usefulness for research

Linguists are invariably led to **improve their transcription when they have time-aligned access to the recordings** -- proof, if one were needed, of the scientific value of time-aligned documents.

Having the data prepared in this format is admittedly time-consuming, but it is highly rewarding for researchers, as it creates a solid empirical basis for many research questions. For instance, the author of the corpus query processor CQP [20, 21] pointed out to us that, while the

halshs-00620893, version 1 - 9 Sep 2011

data sets that we have for each language are relatively small, **the quality of their annotation is in fact higher than that of most of the huge databases that are available for national languages**. This makes annotated data sets for 'small languages' highly promising resources for automated queries. To explore this new prospect, the Limbu corpus (from the LACITO archive) was integrated in the CQP-Web database. The transfer (conversion to the format used in CQP-Web) could be entirely automated thanks to the fully explicit nature of the markup that we use. The small size of the corpus, which comprises 7 texts in all, amounting to slightly over 8,000 words, is not an obstacle to creating concordances and statistics on frequency of occurrence, employing the very best query tools. This is just one example showing that it is indeed possible to "do great things with small languages" [to borrow the magnificent motto of N. Thieberger and R. Nordlinger: 22], even with limited resources.

7. ACKNOWLEDGMENTS

Many thanks to Martine Mazaudon, John B. Lowe and Michel Jacobson for their continuing participation; to the various structures whose support makes this project possible (InSHS; LACITO; ADONIS, CINES and CC-IN2P3); to Bernard Bel (CNRS-LPL) for useful comments on a draft version; and to the contributors to the Archive.

8. REFERENCES

[1] "The Lacito Archive." [Online]. Available: http://lacito.vjf.cnrs.fr/archivage/presentation_en.htm.

[2] "中央研究院 - 台灣南島語數位典藏 Academia Sinica Formosan Language Archive." [Online]. Available: http://formosan.sinica.edu.tw/.

[3] "Australian Research Council grant DP0984419: Doing Great Things with Small Languages." [Online]. Available: http://linguistics.unimelb.edu.au/research/projects/greatthings.html.

[4] M. Jacobson, B. Michailovsky, and J. B. Lowe, "Linguistic documents synchronizing sound and text," *Speech Communication*, vol. 33, pp. 79-96, 2001.

[5] "Unicode Consortium." [Online]. Available: http://unicode.org/.

[6] L. Burnard and M. Sperberg-McQueen, "The design of the TEI Encoding Scheme," *Computers and the Humanities*, vol. 29, no. 1, pp. 17-39, 1995.

[7] "Overview of SGML Resources." [Online]. Available: http://www.w3.org/MarkUp/SGML/.

[8] "Extensible Markup Language (XML)." [Online]. Available: http://www.w3.org/XML/.

[9] B. Comrie, M. Haspelmath, and B. Bickel, "Leipzig Glossing Rules." [Online]. Available: http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

[10] C. Lehmann, "Interlinear morphemic glossing," in *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband*, G. Booij, C. Lehmann, J. Mugdan, and S. Skopeteas, Eds. Berlin: de Gruyter.

[11] "XML schema of the LACITO Archive, adopted as part of a national pilot project." [Online]. Available: http://crdo.risc.cnrs.fr/schemas/archive.xsd.

[12] "OAI-PMH Implementation Guidelines - Specification and XML Schema for the OAI Identifier Format." [Online]. Available: http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm.

[13] "DCMI Home: Dublin Core® Metadata Initiative (DCMI)." [Online]. Available: http://dublincore.org/.

[14] "Open Language Archives Community." [Online]. Available: http://www.language-archives.org/.

[15] "Resources for electroglottography and goodies for LACITO Archive users." [Online]. Available: http://ed268.univ-paris3.fr/lpp/pages/EQUIPE/michaud/en/ressources/.

[16] "Toolbox." [Online]. Available: http://www.sil.org/computing/toolbox/. [Accessed: 20-Aug-2011].

[17] "ELAN — a professional tool for the creation of complex annotations on video and audio resources." [Online]. Available: http://www.lat-mpi.eu/tools/elan/.

[18] "Tools of the Lacito Archive." [Online]. Available: http://lacito.vjf.cnrs.fr/archivage/outils_en.htm.

[19] "Transcriber: a tool for segmenting, labeling and transcribing speech." [Online]. Available: http://trans.sourceforge.net/en/presentation.php.

[20] "CINES: Centre Informatique National de l'Enseignement Supérieur. Overview -- Long-term data preservation -- Training Workshops -- Services -- HPC-Europa2." [Online]. Available: http://www.cines.fr/.

[21] "Computing Centre of IN2P3 laboratory, CNRS - Centre de Calcul de l'IN2P3." [Online]. Available: http://cc.in2p3.fr/?lang=en.

[22] "SpLanDR -Speech and Language Data Repository." [Online]. Available: http://sldr.org.

[23] N. Thieberger, ": PARADISEC: Pacific and Regional Archive for Digital Sources in Endangered Languages." [Online]. Available: http://paradisec.org.au/.

[24] "- EOPAS - Audio Annotations." [Online]. Available: http://www.eopas.org/.

[25] A. Hardie, "CQPweb Main Page." [Online]. Available: http://cqpweb.lancs.ac.uk/.

[26] A. Hardie, "CQPweb - combining power, flexibility and usability in a corpus analysis tool," forthcoming.