WHEN BEST PRACTICE ISN'T NECESSARILY THE BEST THING TO DO: DEALING WITH CAPACITY LIMITS IN A DEVELOPING COUNTRY

John Bowden
The Australian National University

John Hajek
The University of Melbourne

Introduction

Waima'a is an endangered Austronesian language with a few thousand speakers, spoken on the outskirts of Baucau, East Timor's second largest town. The Waima'a language documentation project began in 2002. It is but one of an increasing number of such projects that have been inspired by work by Himmelmann (1998), Woodbury (2004), and others, and helped into being by the availability of funding from sources such as the Hans Rausing Foundation and, in our case, by the Volkswagen Stiftung based in Germany. Language documentation is distinguished from more traditional language description by its aims of providing not just a grammar and dictionary (and perhaps some supporting texts) but also a richly annotated corpus of recordings of a variety of culturally significant events. Recent years have seen a growing awareness that the objects of linguistic study (languages) are dying out at an unprecedented rate across the world. The UNESCO Ad-hoc experts group on endangered languages (2003, p. 3) has estimated that about 90% of the world's roughly 6000 languages may be replaced by dominant languages before the end of the 21st century. These developments have impressed upon linguists and others the increasing urgency of providing a lasting data source for future reference. Digital recording, the advent of the internet, and other technological innovations have brought with them greater simplicity in reproducing and disseminating original recordings. However, they have also brought with them a plethora of different recording formats and other standards, many of them with very limited longevity. The international archiving community has adopted a number of standards to cover not just file formats, but also accessibility to the archives, through
agreed standards of metadata. Organisations such as the Open Language Archives Community (OLAC) and Electronic Metastructure for Endangered Languages Data (E-MELD) have helped to spawn a worldwide network of linguistic archives that are in basic agreement on an appropriate set of standards for the long-term preservation of linguistic data. An emerging consensus on the appropriate kinds of annotations that should accompany the archived recordings is also starting to emerge, with publications such as Gippert, Himmelmann and Mosel (2006) laying down the foundations of such a framework.

While the aims of most academics in preserving an appropriate record of the world’s linguistic heritage may be largely scientific, members of threatened linguistic communities may be expected to have more personal reasons for their interest in language preservation, and often come to the task with different aims from those of the scientifically inclined academic. Community members typically start out documenting their languages with very different levels of expertise and familiarity with the tools of language documentation than do academic practitioners. Limited exposure to computers, particularly in the developing world, means that following best practice in the way envisioned by academics is not always an easy thing to achieve. As a result, the simplest and most effective solutions regarding language documentation for these communities are often not the ones deemed best practice by OLAC and the like.

Although the academic documentarian may see a suitable electronic archive as the best way of preserving and disseminating material, for local third world community members, such an electronic product is rarely of direct value. When community members have no access to electricity, let alone technical equipment, computers and the internet, an electronic archive serves little or no practical use at all in the short to medium term. If local communities are to get any benefit at all from the documentation process, other more traditional ways of archiving, presenting and disseminating materials must be found, and in most cases, this means paper-based print format.

Even when academics are the main instigators of linguistic documentation, it goes without saying that their enterprise cannot take place without the active cooperation of community members. Sometimes the differing perspectives and backgrounds of academics and community members can lead to difficulties in translating good intentions into best practice. Figure 1 illustrates a possible best practice work-flow that would
encompass something like ideal practice from the documentarian's perspective.

1. make digital recording

2. transfer recording to computer in archivable format

3. Use computer tool for transcribing, segmenting, interlinearising and creating lexical (and other?) database in XML format

4. Create metadata and link to interlinearised media files

5. Store all components of documentation in archive

6. Use components of linked XML files to create hard-copy publications, distributable recordings, and so forth.

Figure 1: A representation of an ideal workflow for a language documentation project

In the following section, we look at how these issues have played out in the course of the Waima’a documentation project. In particular, we draw attention to the organisational complexities of a multinational and multilingual project required to meet the challenge of best practice for
Background to the Waima’a project

In 2002, the authors began work on a Volkswagen Stiftung-funded language documentation project with Nikolaus Himmelmann from Germany. The project focuses on Waima’a as spoken in Caisido, a small village to the west of Baucau. The electronic archive for the project (Belo et al., n.d.) can be seen at the DOBES website, housed by the Max Planck Institute for Psycholinguistics. When the first project fieldtrip took place in November 2002, East Timor had been independent for six months, but the country was still lacking in much of the most rudimentary infrastructure that one might expect to find elsewhere. Baucau, the main location of our work and the second largest town in East Timor, only had and continues to have an erratic town electricity supply. Fortunately, we were able to negotiate an arrangement with the Catholic Teachers Training College whereby we rented office space in the campus and took advantage of the College’s own generators to run our computers and charge the batteries of our recording equipment, and so forth.

Project workers were distributed across four sites in three different countries: East Timor, Australia, and Germany. Maurício Belo, a native-speaker of Waima’a, was based in Baucau, and in Germany Nikolaus Himmelmann and his research assistant, Jan Strunk, were based in Bochum. In Australia one of the authors was in Canberra while the other was in Melbourne. An East Timorese research assistant, Alex Tilman, also had a part-time position working in Melbourne. We also had occasional access to additional research assistance in Melbourne. The distribution of workers across all these different places meant a number of complications in terms of how materials were processed.

The complexity of the project was also considerably increased by the requirement of the East Timorese Instituto Nacional de Linguística (INL, National Institute of Linguistics) that annotations of all documented materials, whether for local or international consumption, should be available not only in Waima’a and English, but also in Tetum and Portuguese, the official languages of East Timor. Although the INL had no particular interest in seeing our documentation available in Malay, it was necessary for our own practical reasons to include Malay annotations. Our local team member, Maurício Belo, educated during Indonesian times,
was more competent at reading and writing Malay than any other language. In addition, foreign team members Bowden and Himmelmann were more proficient at Malay than any of the INL sanctioned languages. As a result, Malay was indispensable as a working language for the team.

Although each of the team members was proficient in a selection of the languages used in the documentation, no-one was proficient in all of them, so different parts of the glosses and other annotations had to be completed by different people, at different times, and in different places. The fact that individuals also had varying levels of technical competence in each of the tasks that needed to be completed, and the fact that at different stages of the project, the available software was not always appropriate for best-practice, added further complications. Since some project workers were using Macintosh computers, for example, no Unicode-compliant interlinearisation program was available and we used the SIL's Shoebox program rather than Unicode-compliant Toolbox.

Intensive local involvement was not necessary for all of this work, but it was imperative to get local input on one important detail at the initial stages of the project: to wit, the development of an orthography (see below for further details). It had to be acceptable to the local community as well as national East Timorese norms for vernacular writing systems.

The next section of the paper outlines in some more detail the workflow of the project, from initial recording to archiving, and production of printed materials for local consumption.

The workflow

During the initial project fieldtrip undertaken during November/December 2002, Himmelmann and Bowden were responsible for setting up the initial workflow arrangements with Belo, as well as working arrangements at the Catholic Teachers' College in Baucau. At the beginning of the project, as we have already noted, Belo had had no previous computer experience. One of the major tasks was also then to train Belo to the point where he could work successfully on his own when no foreign project members were present in East Timor.

The initial trip went exceedingly well. By the end of it, we had collected a dozen or so hours of recordings, and Belo was successfully transferring digitised audio recordings from the video tapes to his computer, transcribing them into Word files, then transferring the Word files into Shoebox and creating initial glosses and free translations in both
Indonesian and Tetum. Fortunately, learning to use Shoebox was not quite as hard as it might have been if we had been working on a different kind of language. Waima’a is largely isolating, so learning to deal with complex affixation in the Shoebox interlinearisation process was not a problem.

At this stage, Bowden and Himmelmann also created an initial Shoebox lexicon using materials that had been collected by Bowden and Hajek on earlier fieldtrips, and data from spreadsheets that had been provided by Geoffrey Hull of the INL. Much of this material had been compiled before Unicode fonts were available, and in any case, as already noted, we were not able, because of platform differences, to use Unicode-compliant Toolbox for interlinearisation.

Before Belo could be left to transcribe texts on his own, the team first had to settle on an orthography for Waima’a. The language’s highly complex segmental phonology meant that deciding on a spelling system was not quite as simple as settling into work on Shoebox, however. With four manners of articulation for stops, voiceless nasals as well as plain ones, and over thirty individual consonant phonemes, it was clear that a number of digraphs would have to be employed in any usable orthography. Waima’a community members who were exposed to the orthography, though, soon became adept at using it. After an orthography workshop held at the Caisido school in November 2002, attending members of the language community were using the proposed writing system successfully in a matter of half an hour, and expressed satisfaction with what we had proposed. Later, in 2005, when we distributed printed materials to members of the community, people began reading them successfully more or less straight away. Belo was already using the orthography for creating his transcriptions before the workshop and continued to use it after the event. Luckily also, in spite of the complexity of the phonological system, an orthography was devised that only used basic Latin characters, so Unicode was not necessary for this.

At the time we began work on the project, using ELAN for the initial transcription and segmentation of texts was not possible. As already noted, our work required glossing in four languages apart from Waima’a itself, and these glosses had to be done by different people in different places. Early versions of ELAN had problems with outputting Shoebox files consistently in such a way that they could be reimported into Shoebox, so it was essential that we completed all glossing and translation
in Shoebox before attempting to segment the underlying audio and video files.

Belo also received training on the initial creation of metadata. It was important that he be able to create some initial metadata if he was going to be able to record local events while others were away from the field. He began by entering information into tables in a Word document, which was later transferred by other project participants into a database using the IMDI software developed by the Max Planck Institute for Psycholinguistics at Nijmegen. Belo also learned to use the video camera and set up microphones and he was left with the responsibility of making video recordings of any events that he thought were of cultural significance while other project members were out of the country.

A regular task performed by the foreign researchers when they visited East Timor (or when Belo visited Australia) was a thorough check of the initial transcriptions, looking for inconsistencies, incorrect glossing of homonyms automatically generated by Shoebox and the like, as well as making sure that we had a proper understanding of the meanings of the texts so that we could do further glossing in English. It was also necessary that the meanings of texts be clear for our assistant Alex Tilman. He is an East Timorese permanently resident in Australia. A native speaker of Tetum, he is also able to speak all of the languages, involved in our corpus, including Portuguese but not Waima’a.

One of the most critical tasks to be performed on each visit was reintegrating separate Shoebox lexicon files into one master file that would serve as the basis for subsequent work. Since initial glosses into Malay and Tetum were made in East Timor, but English and Portuguese glosses were done in Australia and sometimes in Germany, separate lexicon files were being used in different places. The complications of version control and the reintegration process that needed to be done from time to time led us to the conclusion that it would be best if no more than two versions of the lexicon were being amended at any one time. Given that communication with Belo was difficult between visits, and since transcription and glossing were major work tasks, Belo always had access to a lexicon file, but access to the other lexicon file was limited to one project participant from outside East Timor at any one time. Prior to a team member’s departure for East Timor, work on that file would cease until a reintegrated file was returned from the field. This process worked smoothly enough until at one point it was decided to make major changes to the structure of the
lexicon file to facilitate production of a printed glossary. Reintegration
after the structure was changed took quite a few days of work in Baucau –
by this time there were almost 5000 lexical entries in the file. The ideal
solution to version control over the lexicon would have been to have just
one version of the file available via an internet connection and to update it
on-line. This would never have worked in East Timor, though, given that
an internet connection was not available in Baucau until the project had
almost ended, and even then was only available via a very slow and
unreliable dial-up connection.

Once more or less complete transcriptions and glosses had been made,
the Shoebox files were imported into ELAN and time alignment of the
texts to the media files was performed. In addition, complete metadata
was created using the IMDI tool. Most of this work was performed in
Germany by Himmelmann and Strunk. Finally, after having been sent
around the world several times in many cases, the materials were ready for
submission to the archive in Nijmegen.

The other major local initiative was the production of a number of
publications that were printed for local distribution. These were an
alphabet primer for school children (Belo & Bowden, 2005), a collection
of short animal stories for early primary school children (Belo, Belo, Belo,
Bowden & Himmelmann, 2005), a longer folk-tale for older children
(Belo, Bowden, Himmelmann & Cardoza da Silva, 2005) and a
multilingual Waima’a glossary (Belo, Bowden, Hajek, Himmelmann &
Tilman, 2006). These were all based on the materials we collected for the
electronic documentation project, produced as PDF documents and
printed in Baucau. The primary objective of these materials is to provide
locally available and easily accessible documentation that can be circulated
around the community. They could also be, if the community so wished,
used as literacy materials in Waima’a. How long our printed materials will
survive in a difficult, humid environment is not known, but short- and
medium-term documentation and access are assured. Electronic versions
of all documents are also archived for future access or reprinting. The
preparation of these materials required considerable additional effort,
including different kinds of software (despite some shared use of
Shoebox/Toolbox) and a range of other skills not necessarily associated
with digital documentation.

It is important to realise that as far as most members of the local
community were concerned, these publications were the only significant
product of the whole project. While future generations of Waima’a speakers may eventually be able to use the archived electronic resources, for the moment and for the interim future, there is and there will not be access to digital materials produced for the archive, and electronic archival best practice remains an irrelevance for them.

Conclusions

We need not spend too long going over what are seen by the researchers as their aims in compiling a suitable language archive: the problems of longevity and accessibility are now generally well understood by the international language documentation community. Although some of the data collected at the earliest stages of the project were not always compliant with best-practice standards—for example, the earliest phonetic transcriptions were not made with Unicode fonts—the international researchers have eventually been able to rectify these problems and the finalised electronic archive will conform to best-practice principles to the best of our ability.

These kinds of technical requirements for long-term digital archiving, however, are very esoteric as far as Waima’a community members are concerned. Much more useful for them are the printed materials which were produced as a by-product of the archiving project.

Our local team member, Maurício Belo, made truly astonishing progress in his initial computer training. He became a productive team member very quickly, but some aspects of computer work remained difficult throughout the project. Computer file management was never fully managed in situ, and tasks such as the final creation of metadata were best left to international participants, as were other aspects of the project such as font handling and selection of appropriate file formats.

With respect to the ideal workflow seen earlier in Figure 1, some stages are more complex than others. Most time, not surprisingly, has been spent on stage 3 activities (transcription, interlinearisation and related treatment of recorded materials). These are rendered more complicated by specific factors we have already identified, such as the dispersed nature of our project team, the unusual multilingual requirements of the project, and frequent lack of internet access to and from Baucau. Moreover, we have already, for instance, noted the inherent difficulties in managing the lexicon file. Nevertheless, from our perspective as academic
documentarians, the project has been very successful, with a large collection of worked up materials now archived.

Perhaps the biggest problem with 'best practice' in linguistic archiving is that its relevance for communities such as the Caisido one is extremely remote. This is not simply because the village has no electricity, computers, or internet connections. It is also because even if community members do live in towns like Baucau, where some of these things are available sporadically at least, no-one had been fully acculturated to modern technology in the way that the international participants in this particular project had been. There is a play-off between ease of use of the technology and maintaining best-practice, and when community members have no way of clearly seeing the utility of best practice, working with more difficult technological practices is not something that anyone can clearly see the benefits of. This is especially the case when there are more simple solutions to immediate problems.

It is likely that in times to come, when greater development has reached East Timor, and when members of the Caisido community have readier access to things like electricity and computers, Waima’a speakers will be more aware of the need for best-practice documentation and all of its ramifications. The problem for the community about this state of affairs is that the kind of economic development required to achieve it will probably also lead to a massive decline in the numbers of people who speak Waima’a. As a result, there may well be little left to document.

What is the way around this? We would suggest that one place to start would be by following the lead of international development practitioners in recent years (see, for example, Boven and Morohashi (2002) for an overview of 'participatory development' practices), and by consulting more widely with community members about what they would like to see in their archives, how they would like them to be compiled, and what sorts of products they would like to see. It would also be useful to see what sorts of tools local participants would like to be able to use in order to achieve their goals in a locally appropriate manner, and work towards the development of tools that make sense not just to technologically sophisticated academics in developed countries, but also to members of minority language communities in relatively impoverished and technologically underdeveloped regions of the third world.
References


