LINGUISTIC DISCOVERY
DARTMOUTH COLLEGE

Volume 5
Issue 1
2007

Ngbugu Digital Wordlist: A Test Case for Best Practices in Archiving and Presenting Language Documentation

Gary F. Simons
SIL International

Kenneth S. Olson
SIL International and University of North Dakota

Paul S. Frank
SIL International

doi: 10.1349/PS1.1537-0852.A.314

url: http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/1/xmlpage/1/article/314

Linguistic Discovery
Published by the Dartmouth College Library

ISSN 1537-0852
linguistic-discovery.dartmouth.edu

Ngbugu digital wordlist: A test case for best practices in archiving and presenting language documentation[1]

Gary F. Simons
SIL International

Kenneth S. Olson
SIL International and University of North Dakota

Paul S. Frank
SIL International

Language documentation faces challenges of data preservation and accessibility. Data can be lost due to physical deterioration (e.g. field notes or tape recordings) or outdated format (e.g. Microsoft Word 3.0). Archived data is typically difficult to access, and it is sometimes found that the archived information is inadequate for research purposes. Increased interest in language documentation has coincided with advancements in digital technologies, offering hope for meeting these challenges. This paper discusses the archiving of a 204-item wordlist of Ngbugu, an Ubangian language spoken in Central African Republic, employing best practice recommendations. Our solution includes: TIFF digital imaging of the original handwritten transcription, WAV digital recording of the wordlist, descriptive markup encoding of the wordlist in XML employing Unicode transcription, viewing and playback via an XSLT style sheet that renders the information in HTML, publishing metadata for resource discovery with the Open Language Archives Community (OLAC), and depositing the original materials and digital representations in an institutional archive committed to long-term preservation and access.

1. Introduction

Significant strides have been made recently in the documentation of the world's languages, but along with this have arisen certain challenges. Much of this documentation remains unpublished and is therefore inaccessible to others. Some of that documentation, e.g. audio tape recordings, will eventually be lost due to physical deterioration. Materials can be easily converted into digital form to increase accessibility, but as Bird and Simons (2003) observe, unless steps are taken to ensure its longevity, "much digital language documentation and description becomes inaccessible within a decade of its creation" (p. 557).

These challenges point to the need for principled approaches to making language documentation long lasting, accessible, and re-usable. In responding to this need, linguists have begun to concern themselves with the development of best practices for digital language documentation. The Open Language Archives Community (OLAC) [http://www.language-archives.org] exists for this purpose, and the concern is also a driving force behind the E-MELD project (Electronic Metastructure for Endangered Language Data) [http://www.emeld.org].

This paper discusses the digital archiving of a 204-item wordlist in Ngbugu, an Ubangian language spoken in Central African Republic by approximately 95,000 people (ISO 639–3 code: [lnl], Gordon 2005, [http://www.ethnologue.com/show_language.asp?code=lnl]). The process employed is a test case for various best practice recommendations concerning the archival documentation of language resources, especially those of Bird and Simons (2003), Plichta and Kornbluh (2002), and MATRIX (2001). Many of these best practice recommendations have been

[1] We wish to thank Fraser Bennett, Kent Spielmann, and two anonymous referees for helpful comments on the text, and Kazuaki Maeda, part of the TableTrans development team, for programming assistance. An earlier version of this paper was presented at the Fourth International Conference on African Languages, University of Maryland Eastern Shore, November 4–7, 2004.

elaborated at the E-MELD School of Best Practices in Digital Language Documentation [http://www.emeld.org/school]. The paper builds on the process laid out in Frank and Simons (2003). We report the results of our project to prepare these materials for long-term archiving (the "archival form") and to provide present-day access of the information via the Internet (the "presentation form") (Simons 2006).

It may be helpful to begin by showing the final results. Examples (1) and (2) provide links to the presentation form and the archival form, respectively, of the Ngbugu digital wordlist.

(1) Presentation form (Olson and Mbomate 2007)
Ngbugu wordlist as a web page, with integrated metadata, images, and recordings

(2) Archival form (Olson 2006)
OLAC metadata record (3 kilobyte XML file)
Wordlist with time-aligned transcriptions (48 kilobyte XML file)
Digitized original audio recording (82 megabyte WAV file)
Digital images of original wordlist forms (2 TIFF files of 8 megabytes each)

Definitions of XML (Extensible Markup Language), WAV, TIFF (Tagged Image File Format), and other pertinent terms employed here are available at [http://emeld.org/school/glossary.html].

First, click on the link in (1) to see the form that has been developed for publishing the Ngbugu wordlist as an interactive web page using today's presentation technologies. On that page, click on the links in *The Wordlist* section to see the resource description (i.e. metadata) and the images of the original transcriptions. Then click on a loud speaker icon to hear the pronunciation of the word as recorded. The icon is linked to a WAV file; thus your web browser will attempt to play it with the sound program that is set up as the default WAV player on your computer.

After exploring all the features of the presentation form, click on the two links in (2) to see the XML form of the metadata and the XML form of the wordlist from which the presentation form was generated. The other two elements in the archival form (the complete recording and the images of the field transcriptions) are too large to make available via the web medium at this time. The complete set of results is available on a CD-ROM from the SIL Language and Culture Archives, 7500 W. Camp Wisdom Rd., Dallas, TX 75236–5699, USA, archive_dallas@sil.org.

We give an overview of our approach in section 2. In section 3, we enumerate the best practice recommendations we followed. Section 4 presents an overview of the process we followed to convert the original materials into archival and presentation forms. Finally, in section 5, we offer concluding remarks.

2. Solution

The original Ngbugu wordlist materials collected in Central African Republic included two items: a two-page wordlist form and a 16-minute recording on audio cassette. The wordlist form presented the standardized wordlist of 204 items from Moñino (1988). For each item the form provided a prompt in French and a space for the transcription of the elicited form. In this case, the form was filled in with handwritten Ngbugu orthographic transcriptions by Jacques Vermond Mbomate, a Ngbugu speaker literate both in Ngbugu and French. Some items included a suggested alternative pronunciation in parentheses, or an indication of uncertainty concerning the data. The second author verified the list in consultation with Mr. Mbomate, and together they produced a transcription employing the International Phonetic Alphabet (IPA 1999).

The second author then created a revised list in Microsoft Word for Windows 2000 that included the French prompt, the orthographic rendering, and the IPA rendering. The accompanying audio cassette contained a recording of Mr. Mbomate repeating the revised list. He produced the French prompt first, then the Ngbugu equivalent. The recording was made with a Marantz PMD 420 monaural cassette recorder and an Audio-Technica ATM 33a microphone. The recording session took place on March 6, 2004, at the ACATBA center (l'Association Centrafricaine pour la Traduction de la Bible et d'Alphabétisation) in Bangui, Central African Republic.

As Simons (2006) observes, a linguist must do two things to ensure that language documentation will persist far into the future. First, in order to ensure that the materials will still be readable from the software point of view, they must be put into a file format that software of the future will still be able to interpret. Second, in order to ensure that the materials will still be readable from the hardware point of view, they must be deposited with an institutional archive that will ensure that they are migrated as needed to fresh media lest they perish on media that become obsolete (e.g. 5 ¼" diskettes) or unreadable (e.g. the limited shelf-life of CD-Rs, cf. Byers 2003).

In order to meet the first objective of putting the material into a format appropriate for long-term archiving, we set out to do the following:

- Make archive-quality digital images of the original handwritten wordlist forms,
- Make an archive-quality digital version of the audio recording,
- Create an archive-quality digital encoding of the transcribed wordlist as an XML file that encodes French and English glosses of each word, the transcription in Ngbugu orthography and IPA (using Unicode for the encoding), any additional notes, and the start and stop times in the digital audio files for each Ngbugu utterance, and
- Create a metadata description of this set of archival materials.

In order to meet the second objective of ensuring the on-going availability of the materials far in the future beyond the life of the current hardware and media, our plan was to:

- Place all of these materials in the SIL Language and Culture Archives for long-term preservation.

But long-term access is not our only objective. We also want the materials to be available today in an easy-to-access form on the Internet. In order to meet this objective, we also set out to:

- Generate versions of the digital audio and digital images that are suitable for browsing on the web,
- Develop an XSLT (Extensible Stylesheet Language) script that generates an HTML (HyperText Markup Language) presentation form for viewing the digitally encoded wordlist that provides links to display the digital images of the original transcriptions and to play back the recording of each utterance,
- Publish the presentation form of these materials on the Internet to enable linguists to inspect them, and

- Make these materials known through OLAC using their metadata standards.

3. Best practice guidelines

In this project, we were guided by best practice recommendations for digitizing the audio recordings (Plichta and Kornbluh 2002), for digitizing the images of the transcription (MATRIX 2001), and for creating digital language documentation and description in general (Bird and Simons 2003). The following tables summarize relevant aspects of these recommendations and indicate the degree to which this project was able to adhere to these recommendations.

Table 1 addresses guidelines for digitizing audio recordings. Plichta and Kornbluh (2002) recommend a sample rate of 96,000 Hz and a bit depth of 24 bits as a standard for archive-quality digital audio, but also note that 44,100 Hz, 16-bit is adequate for technical purposes. Nearly all acoustic information pertinent to language is below 11,000 Hz, and the upper limit of hearing for most people is 22,000 Hz (Ladefoged 2003:18, 26). Since the highest frequency that can be reconstructed from a digital recording is half the sampling rate (aka "the Nyquist frequency", cf. Nyquist 1928, Shannon 1949), the 44,100 Hz rate is sufficient for speech and hearing. Since our digitizing equipment was not adequate at the time for the higher sampling rate and bit depth, we opted for the lower resolution. (See the EMELD glossary for an explanation of the terms "sample rate" and "bit depth" [http://emeld.org/school/glossary.html].)[2]

We avoided the use of minidisc recorders and MP3 files, because these involve compression techniques that result in loss of sound quality. This is of particular concern for those wishing to perform acoustic analysis of the recordings, who would want the data to be as unprocessed as possible.

Recommended Best Practice	*Ngbugu Wordlist Project*
Recommended for archival purposes -sample rate: 96,000 Hz -bit depth: 24-bit	Lack of appropriate hardware prevented us from following these recommendations.
Sufficient for technical purposes -sample rate: 44,100 Hz - bit depth: 16-bit	This is the standard that we followed.
Oversampling delta-sigma A/D converter with dither added prior to sampling	Lack of appropriate hardware prevented us from following this recommendation.
WAV file format	This is the format that we used.

Table 1: Recommendations for digital audio (Plichta and Kornbluh 2002)

Table 2 addresses guidelines for digitizing images of textual materials. MATRIX (2001) proposes separate recommendations for master images and for access images. We followed the

[2] Since the writing of this paper, several additional documents have been published that also address the question of best practices for digitizing audio recordings (Casey and Gordon 2007, IASA-TC03 2005, IASA-TC04 2004, Pohlmann 2006). The two reports from the technical committee of the International Association of Sound and Audiovisual Archives (IASA-TC03 2005, IASA-TC04 2004) recommend a minimum digital resolution of 48,000 Hz, 24-bit for analog originals and note that 96,000 Hz, 24-bit has become widely adopted in heritage/memory institutions. They recommend use of the Broadcast Wave Format (BWF) file type, which is an extension of the WAV format.

former recommendations for generating the archival form and the latter for the presentation form.

	Recommended Best Practice	*Ngbugu Wordlist Project*
Master images		
Bit depth	8-bit grayscale or 24-bit color	8-bit grayscale
Scanning resolution	300 dpi for original documents if smaller than 11" × 17", 200 dpi if larger than 11" × 17"	300 dpi
Image size	Size of original document at scan resolution	Original image size of 8.5" × 11" is preserved.
Format	Uncompressed TIFF	Uncompressed TIFF
Access images		
Bit-depth	8-bit grayscale or 24-bit color	8-bit grayscale
Scanning resolution	72–90 dpi depending on character height	72 dpi
Image size	Original size, at 72–90 dpi	Original image size of 8.5" × 11" is preserved.
Format	For documents smaller than 8.5" × 14": 4-bit interlaced GIF for 8-bit grayscale images or 8-bit interlaced GIF for 24-bit color images For documents larger than 8.5" × 14": 8-bit greyscale JPEG for grayscale images or 24-bit color JPEG, RGB mode for color images.	8-bit interlaced GIF

Table 2: Recommandations for digital imaging (MATRIX 2001)

Table 3 addresses the problem of the portability of digital language resources. This includes the synchronic portability of resources across a multiplicity of present-day computing platforms as well as the diachronic portability of today's resources to the computing platforms of the future. Bird and Simons (2003) identify seven dimensions of portability and propose best practice guidelines designed to maximize the ability of digital language resources to move across different computing platforms and to remain usable far into the future. Two of the seven dimensions—*citation* and *preservation*—are mostly relevant to the institutions that publish and archive resources. The other five, however—*content*, *format*, *discovery*, *access*, and *rights*—are particularly relevant to the linguists who create language resources. The following table repeats some of Bird and Simons' key recommendations in these five areas and describes how the current project has responded.

Recommended Best Practice	*Ngbugu Wordlist Project*
Content	
When texts are transcribed, provide the primary recording (without segmenting it into clips)	The full elicitation is provided in a 16-minute WAV file.
Transcriptions should be time-aligned to the underlying recording in order to facilitate verification.	Each response is time-aligned.
Format	
Use open formats supported by multiple software vendors.	The formats used for archival forms are open: XML (for transcription), WAV (for audio), TIFF (for images).
Use Unicode for character encoding.	IPA transcriptions are encoded in Unicode.
Use a descriptive markup system (preferably XML) for textual information.	The whole wordlist (including glosses, transcriptions, and time alignments) is represented in an XML file with descriptive markup tags.
Provide a human-readable version of the same information in a suitable presentation format.	The XML wordlist is transformed to HTML for presentation in a web browser.
Discovery	
Describe the resource using the metadata standard of OLAC.	An OLAC-conformant resource description is supplied.
Make the resource known to the world at large by publishing the metadata description with an OLAC data provider.	The metadata are published via the OLAC data provider for SIL's Language and Culture Archive.
Access	
Make resources accessible to all interested users.	Presentation form is published on the web.
Publish in such a way that users can access the original materials to manipulate them in novel ways.	Archival form of full resources is available by ordering a CD-ROM.
Rights	
Make a clear statement of terms of use so that users know what they may do with the material.	The resource description states the materials are copyrighted and available to all under standard terms of Fair Use.
Identify and protect any sensitivities inherent in the material.	There are no known sensitivities and this is stated in the resource description.

Table 3: Recommendations regarding the portability of language documentation and description (Bird and Simons 2003)

4. Results and process

Following Simons (2006), we distinguish three forms of data in this project:

- Working form: the form in which information is stored as it is created and edited,

- Archival form: the form in which information is stored for access long into the future, and
- Presentation form: the form in which information is presented to the public.

Part of the design of this project is to distinguish archival and presentation formats for the information. In this section we describe these two forms of the data in detail. In addition, we describe certain aspects of the processes we used to create them.

4.1 Archival form

Following the various best practice recommendations, the archival form for the digital files are in open formats—XML for the textual data (using the UTF-8 encoding for the Unicode character set), WAV for the audio data, and uncompressed TIFF for the graphic images of the original written documents. Open formats have greater longevity and can be transformed into presentation formats that are more "reader friendly."

The XML files are derived from the XML output of the TableTrans software (Bird, et al. 2002) that was used to organize the French and English glosses, Ngbugu orthographic and IPA transcriptions, and WAV file time-alignment data. Figure 1 shows entries 14–16 in the XML file, including the item number, French and English glosses, start and stop times for the Ngbugu utterance in the master sound file, and the two transcriptions of the utterance. Where there are two alternate Ngbugu words or pronunciations for a given prompt, these are recorded in separate "response" blocks of XML data within a single XML <item> element, as seen in item 15.

```
<item n=" 14">
        <gloss xml:lang="fr"> bouche</gloss>
        <gloss xml:lang="en"> mouth</gloss>
        <response>
                <audio start="51.775000" end="52.325000"/>
                <orth> ma</orth>
                <form> mà</form>
        </response>
</item>
<item n=" 15">
        <gloss xml:lang="fr"> bras/main</gloss>
        <gloss xml:lang="en"> arm/hand</gloss>
        <response>
                <audio start="54.900000" end="55.575000"/>
                <orth> könô</orth>
                <form> kōnó</form>
                <note>arm</note>
        </response>
        <response>
                <audio start="57.075000" end="57.675000"/>
                <orth> tchâneû</orth>
                <form> tʃánə́</form>
                <note>hand</note>
        </response>
```

```
</item>
<item n=" 16">
        <gloss xml:lang="fr"> brouillard</gloss>
        <gloss xml:lang="en"> mist; fog</gloss>
        <response>
                <audio start="59.575000" end="60.225000"/>
                <orth> ndrö</orth>
                <form> ndʁō</form>
        </response>
</item>
```

Figure 1: Entries 14–16 in the XML archive file for the Ngbugu data

The XML file contains the information from the wordlist as a structured plain text file. It uses the UTF-8 encoding of the Unicode character set. Software capable of displaying data in UTF-8 should be able to render the phonetic data faithfully, given a Unicode font that includes the IPA block of characters. We used the Doulos SIL font in this project, as it conforms to our specifications [http://scripts.sil.org/DoulosSILfont].

One problem that arises given the present state of computing concerns the proper rendering of diacritics. In order for diacritics to be properly placed when using Unicode characters, both the font and the software need to have “smart font” capabilities, e.g. using OpenType or Graphite technology [http://scripts.sil.org/RenderingGraphite]. In our data, this concerns the placement of acute, macron, and grave diacritics (used to mark tone in the IPA) above a character with a dot (such as “i”), and the stacking of a tone mark above a tilde (used to indicate nasalization). The diacritics may not always be displayed properly; for example, they may be superimposed rather than stacked. This is a problem that should go away as future versions of software will likely incorporate these capabilities. Despite this short-term problem, we still consider it best to opt for Unicode encoding of the IPA characters rather than using a custom font. This is because the underlying data is preserved for long-term storage following a standard that is true to the original transcription, even if the rendering on some systems may be less than ideal at present.

4.2 Presentation form

Clearly, a presentation form of this data set is also needed, and we have chosen to prepare a form of these data for web presentation. For this purpose, we need three things: an XSLT style sheet to render the XML file in HTML, individual WAV files for each of the utterances that are linked to the data for playback, and GIF versions of the scanned images of the original transcription. The rationale for creating individual WAV files and employing GIF files for the presentation form is that they are much smaller in size than the archival WAV and TIFF files. This is necessitated by present-day limitations of storage for computers and especially bandwidth for Internet access. In the future, these limitations will likely cease to exist, in which case the archival standards could be employed for the presentation form. A sample of the HTML presentation form of the first ten words in the wordlist is given in Figure 2.

1.	abeille	bee	wräto	[wʁātò]	
2	acide (vb)	tart	kpï, kï	[kpī]	
3.	aile	wing	mbrö	[mbʁō]	(adjective)
4.	aller	go	e	[ʔè]	
5.	amer (vb)	be bitter	chü	[ʃū]	(adjective)
6.	animal	animal	gia	[già]	
7.	année	year	ngû	[ngú]	
8.	appeler	call	e tchô	[ʔè tʃó]	
9.	arbre	tree	yö	[jō]	
10.	attacher; lier	attach	i (reu)	[ʔì]	

Figure 2: Presentation form of the first ten words in the Ngbugu wordlist

4.3 The process

A major aspect of the Ngbugu project was linking the transcription to the audio material. In the past, this would have been a cumbersome process requiring finding each utterance and noting its start and stop times in the master audio file using conventional audio software. However, several programs are now available that make this process easier. These include Transcriber [http://trans.sourceforge.net] and ELAN [http://www.mpi.nl/tools/elan.html], among others. We chose to use the TableTrans program (Bird, et al. 2002) from the Linguistic Data Consortium [http://www.ldc.upenn.edu]. It allows the linguist to follow a “play cursor” through a graphic representation of the waveform as the recording is played back. The linguist may stop the playback at any point, use the mouse to select a region of the wave, and then enter annotations about that particular bit of the recording into a table with user-defined fields.

In our application, we defined fields for the following six annotations: the item number, the French prompt, the English translation, the orthographic transcription of the Ngbugu utterance, the IPA transcription of the Ngbugu utterance, and any additional notes that were entered on the original wordlist forms. TableTrans is able to import the annotation data from a standard comma-delimited file. We were able to exploit this feature since the annotations were already available to us in a Microsoft Excel for Windows 2000 spreadsheet; we saved that in comma-delimited format and then used TableTrans to add the time alignment to the original recording. The fully annotated and time-aligned data set can then be exported from the program in either comma-delimited (CSV) or XML format. We then used an XSLT script to transform the XML “annotation graph” output of TableTrans into the descriptive wordlist format shown in Figure 1. Additionally, TableTrans automatically created the individual sound files that correspond to each of the segments that have been identified in the transcription process. (This functionality was ultimately the reason we chose to use TableTrans, as it was not available in the other programs.) These are used in the web-based presentation form of the data so that the playback of a single utterance involves only the download of a small WAV file.

In this project, the sample rate of the recordings in both the archival form and the presentation form was the same. However, it is likely that researchers archiving data at a higher sampling rate (e.g. 96,000 Hz) will want to convert the recordings to a lower sample rate (e.g. 44,100 Hz) for the presentation form. When doing so, it is important to apply a low-pass pre-filter to the data to remove all frequencies above the Nyquist frequency in order to prevent aliasing (Ladefoged 1996:139–140). Some audio processing programs (e.g. Cool Edit 2000) contain filters to do this.

4.4 Archive depositing and resource discovery

Included with the archival form was a metadata record of that resource. We followed OLAC's standard for the format of this metadata [http://www.language-archives.org/OLAC/metadata.html], simply using a standard text editor to create the record. Alternatively, the metadata record could be created using the OLAC Repository Editor [http://www.emeld.org/tools/ore.cfm]. The metadata was then verified using OLAC's metadata validation service [http://www.language-archives.org/tools/metadata/freestanding.html].

We deposited a CD-ROM of the archival form of the Ngbugu digital wordlist (including the metadata record) with the SIL Language and Culture Archives, located in Dallas, Texas. This institutional archive was then responsible for entering the resource into its own metadata catalog and publishing a record of the resource on the Internet [http://www.ethnologue.com/show_work.asp?id=47050]. Since the archive is a data provider to OLAC, the resource should appear on pertinent searches done with OLAC search engines: [http://linguistlist.org/olac/] and [http://www.ldc.upenn.edu/olac/search.php]

5. Discussion

One success of this project was in preparing the same data in both archival and presentation formats. The presentation format of the textual data can be automatically generated from the archival form, thus avoiding the need to maintain two distinct data sets. The archival data is primary: without care in the preparation of the archival form of the data there is a high likelihood that the information would be unusable within just a few years because of changes in technology. With the current approach, it is possible to have multiple scripts for generating multiple presentation formats. While the archival format stays constant over time, future generations can generate new presentation formats to take advantage of advances in presentation technology.

Transcriptions and recordings exist for many other languages. The TableTrans files prepared in this project could serve as a template for preparing similar electronic data sets for these other languages. The new transcriptions can be entered for a given language, the appropriate sound file associated with it, and the time-alignment done. Other tools developed for this current project could be easily adapted to facilitate the preparation of presentation forms of wordlists for these other languages.

In addition, the general principles underlying the development of the electronic version of the Ngbugu wordlist could be applied to other types of language documentation, especially the distinction between archival and presentation formats, the use of XML and Unicode for the textual data, and the time-alignment of the audio information and textual information.

References

Bird, Steven, and Gary F. Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79/3.557–582. Preprint available at [http://www.ldc.upenn.edu/sb/home/papers/0204020/0204020-revised.pdf].

Bird, Steven; Kazuaki Maeda; Xiaoyi Ma; Haejoong Lee; Beth Randall; and Salim Zayat. 2002. TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse tools built on the Annotation Graph Toolkit. Proceedings of the Third International Conference on Language Resources and Evaluation. Paris: European Language Resources Association. Online. URL: [http://arxiv.org/abs/cs/0204006]. TableTrans and the other related tools can be downloaded by selecting "AGTK Windows" from [http://agtk.sourceforge.net/].

Byers, Fred R. 2003. Care and handling of CDs and DVDs: A guide for librarians and archivists. (National Institute of Standards and Technology Special Publication No. 500–252). Gaithersburg, MD: National Institute of Standards and Technology/Washington, D.C.: Council on Library and Information Resources. Online. URL: [http://www.itl.nist.gov/iad/894.05/docs/CDandDVDCareandHandlingGuide.pdf].

Casey, Mike & Bruce Gordon. 2007. Sound directions: Best practices for audio preservation. Bloomington, IN & Cambridge, MA: Indiana University & Harvard University. Online. URL: [http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/sd_bp_07.pdf]. (Accessed 17 December 2007.)

Frank, Paul S., and Gary F. Simons. 2003. Sáliba wordlists: A test case for best practices in archival documentation of an endangered language. Paper presented at the Society for the Study of the Indigenous Languages of the Americas annual meeting, Atlanta, Georgia, January 2–5, 2003.

Gordon, Raymond G. (ed.) 2005. Ethnologue: Languages of the world. 15th edition. Dallas: SIL. [http://www.ethnologue.com].

IASA-TC03. 2005. The safeguarding of the audio heritage: Ethics, principles and preservation strategy, version 3. Online. URL: [http://www.iasa-web.org/IASA_TC03/TC03_English.pdf.] (Accessed 19 November 2007.)

IASA-TC04. 2004. Guidelines on the production and preservation of digital audio objects. Aarhus, Denmark: International Association of Sound and Audiovisual Archives.

International Phonetic Association. 1999. Handbook of the International Phonetic Association. Cambridge: Cambridge University Press.

Ladefoged, Peter. 1996. Elements of acoustic phonetics. Second edition. Chicago: University of Chicago Press.

Ladefoged, Peter. 2003. Phonetic data analysis: An introduction to fieldwork and instrumental techniques. Oxford: Blackwell.

MATRIX: The Center for Humane Arts, Letters and Social Sciences Online at Michigan State University. 2001. Digital imaging for archival preservation and online presentation: Best practices. ms. Online. URL: [http://www.historicalvoices.org/papers/image_digitization2.pdf].

Moñino, Yves (ed.) 1988. Lexique comparatif des langues oubanguienne. Paris: Geuthner.

Nyquist, Harry. 1928. Certain topics in telegraph transmission theory. Transactions of the American Institute of Electrical Engineers 47.617–644.

Olson, Kenneth S. 2006. Ngbugu digital wordlist: Archival form. SIL-LCA-47050. SIL Language and Culture Archives, Dallas, Texas.

Olson, Kenneth S., and Jacques Vermond Mbomate. 2007. Ngbugu digital wordlist: Presentation form. Linguistic Discovery 5/1:40-47 Online. URL: [http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/2/xmlpage/1/article/317].

Plichta, Bartek, and Mark Kornbluh. 2002. Digitizing speech recordings for archival purposes. ms. Online. URL: [http://www.historicalvoices.org/papers/audio_digitization.pdf].

Pohlmann, Ken C. 2006. Measurement and evaluation of analog-to-digital converters used in the long term preservation of audio recordings. Paper presentat at the "Issues in Digital Audio Preservation Planning and Management" roundtable discussion, Washington, DC, March 10-11, 2006. Online. URL: [http://www.clir.org/activities/details/AD-Converters-Pohlmann.pdf.] (Accessed 17 December 2007.)

Shannon, Claude E. 1949. Communication in the presence of noise. Proceedings of the Institute of Radio Engineers 37/1.10–21.

Simons, Gary F. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. SIL Electronic Working Papers 2006–003. Dallas, TX: SIL International. Online. URL: [http://www.sil.org/silewp/2006/003/SILEWP2006-003.htm]. 5 May 2006.

Authors' contact information:
Gary F. Simons
SIL
7500 W Camp Wisdom Rd
Dallas, TX 75236
E-mail: http://www.sil.org/~simonsg

Kenneth S. Olson
SIL
7500 W Camp Wisdom Rd
Dallas, TX 75236
http://www.sil.org/~olsonk

Paul S. Frank
SIL
7500 W Camp Wisdom Rd
Dallas, TX 75236
E-mail: Paul_Frank@sil.org