Sharing data in small and endangered languages

Cataloging and metadata, formats, and encodings

Nicholas Thieberger and Michel Jacobson

Speakers of small or 'under-resourced' languages often first contact the world of Information Technology via the effort of field linguists. Good practices in linguistic data management include the separation of structure and content and of data and metadata formats. Primary outputs of field research (lexicon, transcripts and interlinear glossed text collections, and their associated media) need to be coded and preserved. Long-term access to these data is addressed by the establishment of archives that also act as the locus for training and advocacy for well-formed data. In this paper we discuss two such archives, one in Australia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), and the other in France, the "Archiving Project" from the LACITO/CNRS.

For speakers of most small or "under-resourced" languages, their first contact with the world of information technology is via the effort of a field linguist. In recent years, increased emphasis by linguists on language documentation has led to a greater focus on good practices in computerization. These include linguistic data management; separation of structure and content; separation of data and metadata formats and codings for the primary outputs of field research, which are the lexicon, transcripts and interlinear glossed text (IGT) collections, as well as the media on which these are recorded. Long-term access to these data is being addressed by the establishment of archives that act not only as data repositories but as the locus for training and advocacy for well-formed data. In this paper we discuss two such archives: one in Australia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC); and the other in France, the "Archiving Project" from the Laboratoire de Langues et Civilisations à Tradition Orale (LACITO) of the French Centre National de la Recherche Scientifique (CNRS). These archives

are preparing data on small and endangered languages and participate in efforts to support linguists to produce well-formed archival data.

1. Introduction

The field records linguists produce are meant to endure and to be available to the people we record and their communities, as well as to fellow researchers well into the future. Archiving is no longer something we do at the end of our fieldwork. It is apparent now that it should be integrated into everyday language documentation work and that it is a crucial aspect of documentary linguistics. We have learned to separate form and content in the representation of linguistic data in order to establish archival forms (e.g., a Toolbox lexical database) with derived representations (e.g., a printed dictionary or a set of HTML-encoded files). Recent technological advances have pointed to the importance of planning data management and workflow for ethnographic recording. This, in turn, has facilitated an expansion in documentary linguistics and archiving. Recordings should always be of high quality, but it is in the context of small and endangered cultures and languages that the quality of recording takes on new significance (quality here refers both to the content and the form of the recording). If we are the only recorders of the last remaining speakers or performers, then we must, right from the moment of recording, be concerned with making good documents and placing them into a suitable archive for storage and discovery. Thus, we can distinguish archival practice, a process resulting in well-formed archival data, from archival storage in a repository.

An example of this is the making of the initial recordings and their digital representation, citable by means of a persistent identifier, which allows further work to be located with reference to that primary data. Typically this further work involves annotation of the data and the construction of dictionaries, text collections, and grammatical descriptions. In all primary material, the content is plain text structured in a standardized format with an explicit and unambiguous coding to allow it to endure into the future. Description of the data with standard metadata terms allows its discovery in the long term. All of these procedures facilitate repatriation of the data to the communities from which it originates, as they are able to locate the data once they have been archived.

Archives have an image of being repositories of old stuff, and usually old stuff that comes from old people. A colleague, when asked if he was considering depositing materials with our archive, said, "Did I look as if I was going to die any minute when you last saw me?" For him, as for many people, archiving is something done at the end of one's career, when there is time to go back to fill in gaps

and make the entire collection of data more presentable. This view of archiving imagines that boxes of stuff can be delivered to the archive to be held in perpetuity sometime after the linguist has finished with them. The recent focus of linguistic archives, informed by the discussion of language documentation, is that the stuff deposited must be of sufficient quality and sufficiently well described that it can be useful into the future.

Current archives train and provide advice in response to the need for such a service in the community of documentary linguists. These archives are primarily long-term repositories that take well-structured data and provide the infrastructure for securely holding and locating it over time. An archive is also the point of reference for a network of practitioners who want advice on how to proceed. It is the archive's role to agree on standards that seem most appropriate and to assist in their adoption by the broader community of linguists. Given that none of the current archives has the resources to edit items in their collections, they rely on depositors to produce material that is well formed from an archival point of view. Such data have an explicit structure, encoded, for example by labels (as in a Toolbox lexical file), or tags (as in XML), or written in stand-off markup (as in time-aligned transcripts), or in the form of a relational database. The data are also archived in a nonproprietary form that can be read on any platform, now and in the future, and can be converted globally when new ways of working with it appear in the future (either new software or new media).

The fact that the most common current working tools for transcription and time alignment[1] are coming out of this same effort indicates that archives are central to the promotion of new technologies as a means for ensuring that normal linguistic fieldwork will result in the best possible archival form. The two projects described here have evolved to deal with separate and complementary approaches to archiving linguistic data. PARADISEC's primary goal was to make old recordings safe for eventual access, while LACITO has focused on methods for accessing media via its transcripts using a practical XML-based system.

2. The perspective of PARADISEC

PARADISEC is a digital archive based in virtual space between Sydney, Melbourne, and Canberra in Australia. It was established in 2003 by a group of

1. Specifically, ELAN (http://www.lat-mpi.eu/tools/elan/) from the Max Planck Institute for Psycholinguistics in Nijmegen, or Transcriber (http://trans.sourceforge.net/en/presentation.php) from La délégation générale pour l'armement [General Delegation for Ordnance] (DGA), with support from OLAC via the Linguistic Data Consortium.

linguists and musicologists concerned with the lack of a repository for material recorded outside of Australia by Australian researchers. For those working with indigenous Australian languages, there is a national archive (the Australian Institute for Aboriginal and Torres Strait Islander Studies, or AIATSIS), which has been operating since the 1960s. National Australian cultural institutions, such as the National Library and the National Film and Sound Archive, do not have a mandate to keep field recordings from outside Australia. In particular, PARADISEC was concerned about audiotapes recorded since the 1950s that were not being stored in any suitable repository and were physically deteriorating. Thus, the initial focus was on the preservation of existing so-called "legacy" material, and as of late 2009 we had digitized some 2,500 hours, or 4.5 terabytes, of data. However, once we started processing these tapes, it became clear that there was a huge demand from current researchers who wanted to work with their data in a digital form and wanted high-quality archival representation of their media before they conducted most of their analysis.

At PARADISEC, we encourage practitioners (whom we take to include mainly linguists, musicologists, and indigenous language workers) to deposit media material by ensuring that they will have a high-quality digital version of their data in the short term. If an archival form of the file is created first and is then used as the basis for the subsequent effort of transcription and time aligning, the resulting work has a citable source that should persist into the future. We have been encouraging postgraduate students to lodge their tapes with PARADISEC as soon as they return from fieldwork (and we have had DVDs lodged directly from a fieldwork location to provide backups of the primary data). We digitize or capture their data and provide both an archival (usually at 96 kHz/24-bit BWF [Broadcast Wave Format]) and a representational (linear MP3) copy with its persistent identifier in our collection. This gives them a digital file to work with, but, more importantly, it gives them a citable form of archival data with persistent identification. Their intellectual effort of annotating this primary data can then build on a firm foundation for both their own immediate goal (typically a dissertation) and the long-term needs of having richly annotated primary data safely archived.

We also spend considerable time with many old tapes, preparing them for data transfer by cleaning and, in some cases, baking or placing them under vacuum. We also run training workshops of half a day to several days' duration on the use of software tools and on data management. We use these as a means of advocating a workflow for language documentation that builds archiving into the normal everyday work of the field linguist. Otherwise, it can be an onerous addition, or a task left until the weight of the cumulative research effort becomes unbearable at the end of a researcher's career. For example, Thieberger developed a tool

called *Audiamus* for building a media corpus that he used in presenting data with his documentary grammar of South Efate, a language of Central Vanuatu (Thieberger 2004). Once a time-aligned media and text corpus is developed, it is a straightforward task to prepare audio CDs for return to those recorded, or to place all media into a media server such as the widely used iTunes software, from which speakers can make their own selection of "tracks" for their own CDs.

Archives rely on the relationships they have established with their communities, including both the depositors and users. In general, the benefits of depositing are clear, in particular because we digitize analog tapes and hold copies at no cost for members of our consortium. The ability to be "trusted," as a repository should be, arises from a number of factors, but the key for us has been the ability to provide advice and training needed to ensure the quality, both technical and in content, of recordings and associated derived materials (transcripts, glosses, dictionaries, and so on). The rationale is that, if we want high-quality recordings and well-structured archival data, then we have to provide training in their creation. We run workshops in using Toolbox, which is still the only tool that creates structured lexical files linked to IGT. As tools like Transcriber and Elan are produced by our colleagues, we introduce them to a community of users in our region at occasional workshops, both in our universities and in community-based language centers.

The ability to enforce standards on depositors extends to the description of the data, or the metadata, that allows the data to be discovered. PARADISEC mainly works with legacy data, so the quality of its metadata can be quite variable, often no more than a few lines on a tape box, together with contextual information about the collection from which the item will be identified. At PARADISEC, we use a cataloging system that provides a description of both the item and the process it undergoes from accession. All of this metadata can be output in various forms, one of which is the OLAC metadata set. Exporting to OLAC metadata has increased the visibility and, therefore, the discoverability of the material in our collection. Moreover, its ease of use meant that we were able to move our metadata system to an Open Archives Initiative (OAI) conformant metadata repository after a few months of operation.

We encourage users to develop a persistent naming convention using fairly standard ASCII characters and to avoid unnecessarily long names. If we can then take the users' names for their own files and incorporate them into our persistent identification, it makes it much easier to keep track of the relationships between the notes and the media files. Our persistent file names follow the directory structure of the mass storage system on which the files will reside; they are composed of a collection identifier, followed by an item identifier and then a specific local identifier (like "A" or "B" for the side of a tape). These are then followed by a three-letter extension indicating the file type.

Working with legacy material, we sometimes see what small additional steps researchers could have taken to make their recordings more useful. Obviously, collections vary greatly in the accompanying documentation. In some cases there is no specific information about the tapes we have located in a box or filing cabinet, and, while there may be accompanying field notes, we do not have the time or the personnel to work through field notes and to establish their relationships to field recordings. Instead, we take digital images of notes and put them into an

Table 1. Comparison of Earlier Methods of Handling Data with those Advocated Here

	Previous	Current
Data	Analog	Digital
Copyright in material clarified	Rarely	Consent forms signed by interlocutors (because deposit in an archive is envisaged as part of the process)
File names	Arbitrary	Persistent identifiers
Data structure	No explicit structure (implicitly marked by fonts and styles)	Explicit structure is used as the basis for derived forms (e.g., as in lexical files in Toolbox)
Archival accession of primary data	After use of the material by the researcher (typically after retirement or death of the researcher)	Incremental accession, ideally before use of the material by the researcher
Annotation of primary media	Little done, usually by hand	More comprehensive annotation, using time alignment and interlinearizing
Archival accession of annotations	Typically after retirement or death of the researcher	Work in progress archivable and overwritten by subsequent versions (safe backup)
Persistent identification to support citation forms of data	Maybe in fieldworker's notes, hampered by lack of discoverability	Assigned by archive and persistent identifier resolved to an item in the archive
Metadata standard	Library/MARC (large existing infrastructure)	DC/OLAC (support for small, collector-based archives)
Metadata discovery	Library catalogs (not always interoperable)	Open Archives Initiative, subject specialized searches
Persistence of data	Analog tape in one location	Digital simulacra/copies (Lots Of Copies Keeps Stuff Safe [LOCKSS])
Relation between items	Ignored or treated in catalog	Treated in metadata and instantiated where possible (e.g., tape/transcript)
Repatriation of copies	Copies of tapes provided from a single location	Digital copies of tape/transcript in linked form; available for download from the Web or provided on CD

online delivery system that permits researchers to propose metadata descriptions. We plan to allow online annotation of media that will enrich the existing collection. Simple descriptive metadata then allows us and potential researchers to locate the relevant material and reintegrate it with the field notes.

Table 1 summarizes an ideal approach to data taken in the current initiatives, compared to an idealized earlier approach.

3. The LACITO archiving project

LACITO (Laboratoire de Langues et Civilisations à Tradition Orale) is a research group of the French Centre National de la Recherche Scientifique (CNRS) where researchers (linguists, anthropologists, and ethnomusicologists) have been working for some 30 years to describe languages, many of which were previously unwritten. As a result of their fieldwork, they have collected recordings, usually audio but some video, as well as transcriptions, translations, and other associated material, made in association with their local collaborators, speakers of the language. These recordings are the basis of their further research when they return from fieldwork. The analysis based on these recordings is principally phonetic, using IPA (International Phonetic Association) symbols. They are further translated, using interlinear glossing. In addition to the glossed texts, the data typically contains elicitation sessions, word lists or dictionaries, songs, and so on. All of this material represents some hundreds of hours of recording.

Only a very small part of the recorded material ends up being used in a publication, such as a monograph description of a language. The rest of the material is typically unpublished, not referenced, and left in unmanaged collections in the hands of the researchers. These recordings – in particular, analog tapes – degrade over time and are at risk of becoming unreadable. As time passes, linguists discover that they cannot access their own data, either due to the deterioration of the tapes themselves, or because of the increasing lack of tape recorders like Uher, Revox, or Nagra, or the inability to maintain them.

3.1 The archiving project

Our laboratory has undertaken a large-scale project with two principal aims: safeguarding the data and its annotations, and enabling its appropriate diffusion. These aims are clearly internally linked, and to achieve them we have to undertake a process of standardizing the encoding and format of the data. We do not discuss organizational or legal issues here except in passing, despite their

importance in achieving our aims. The success of our project has led us to broaden our scope to data from other organizations.

An appropriate response to the deterioration of analog magnetic tapes is digitization. This involves conversion of the analog signal (which was the dominant mode of recording until recently), using equipment that is now well known and readily available. We have chosen to use the CD-audio standard sampling rate and sample size (44.1 kHz, 16 bits). As the digital file is identical to all of its copies, there is no longer a true original or copy. Preservation of the data is thus only possible by proper management of the mass storage systems, which are constantly audited for data errors and migrated as required to new media.

Annotations of this data, once typically handwritten in notebooks, are now usually created in one of a number of digital forms. Conservation of this data involves describing its contents and standardizing its encoding and format. Today there are many useful encodings and formats for linguists. For example, the IPA symbols have been incorporated into Unicode (ISO-10646), XML is generally accepted as an exchange format, PCM (pulse-code modulation) is the standard for audio data, and so on.

We have chosen the markup language XML as the formal representation for all the annotations of the documents in our archive. This choice was based on a number of factors: the encoding characters are Unicode; XML is easily integrated into Web architecture; there are many tools for working with XML; and there is widespread agreement on its adoption.

Our laboratory used various formats in the past, and so we had to develop a number of conversion tools to take these files to well-formed XML. For manuscripts, this means digitizing the paper versions. While it is possible to scan them and then connect them to text using Optical Character Recognition (OCR) software, it is actually better to reenter them, especially given the low volume of material that typically results from linguistic fieldwork.

Having elected to use XML, we then have to describe the logical structure of our annotations. This structure can be expressed by a Document Type Definition (DTD) or an XML schema, both of which constrain the structure of a document, including the name and the type of permitted elements and their attributes. Other constraints control the content of elements, including the order of their appearance, controlled vocabularies, optional or obligatory status, and so on. This formal syntax should reflect the type of analysis the data needs to undergo, but normalization of the data is made more difficult by various theoretical approaches to the data.

Further, there is little consensus in the community of linguists regarding what constitutes the objects in the data. Many ontologies have been proposed and attempts at creating encoding standards – notably the Text Encoding Initiative (TEI) or the Corpus Encoding Standard (CES) – have been made. At the moment, at the ISO, the TC37/SC4 working group is attempting to solve this question. As there was no intellectually satisfying solution, we have chosen to create a specific DTD that is very simple, but based on the TEI, to facilitate interoperability with it.

In LACITO's DTD, there are five hierarchical levels, which are defined by the element names – ARCHIVE, TEXT, S, W, and M, corresponding to corpus, text, sentence or phrase, word, and morpheme. Each level can contain one or more items from the level immediately below it. Thus, a phrase is composed of words, which are composed of morphemes, but it is impossible to have a morpheme in a phrase that is not part of a word. Each of these levels can include transcriptions (FORM), translations (TRANSL), and a time code (AUDIO). The translations have to specify the target language. At the level of words and morphemes, the translation corresponds to what is usually called a gloss. Transcriptions have to be specified by type (phonetic, phonological, orthographic, transliterated); the name of the transcriber and the date of the transcription can be added in case of multiple transcriptions for the same object. Phrases can contain contextual information, such as the name of the speaker, in the case of dialogues. If needed, words and morphemes can carry typological information – like part of speech, class, and so on – but these have to be free text rather than controlled vocabularies, since each linguist tends to use his or her own preferred system. General notes can be included anywhere in this system.

Temporal links are made by inserting an AUDIO element at the level required, with the attributes "start" and "end," which indicate milliseconds from the start of the file. These links can express temporal events that are:

1. chained: phrases follow one after the other;
2. embedded: words of one phrase are embedded in this phrase;
3. overlapping: more than one person speaking at the same time.

The annotation of time codes that we propose here is based on the recommendations of the Text Encoding Initiative. It relies on the hierarchical nature of XML elements to represent their inclusion in a temporal frame, and the order of elements to represent their successive order over time. All elements, at any point in the hierarchy, can be linked by time codes, but they do not have to be. Those which have no time code are therefore considered to occur within the time codes of the next highest element, a recursive process that applies until the highest level is reached. For example, a word is always located between the beginning and end of the phrase in which it occurs. A non–time-coded element will be located after the end of the preceding element of the same level and before the next, in the order in which they occur in the text. In contrast to the hierarchy of levels, elements at the same level can break their linear position by means of their time codes. In

a narrative, for example, phrases follow on from each other, the end of one generally corresponding to the beginning of the next.

Field linguists are generally interested in morpho-phonology, and so we decided to limit annotations to this level. This can be considered a "gross" or base-level analysis. Further analysis of parts of these base-level documents can be achieved by use of XLink[2] pointers in other documents. This solution allows a simple DTD to be focused on the field materials, while leaving it open to other analyses, even potentially contradictory ones.

The LACITO archives contain linguistic resources: recordings and their annotations. All of these resources are catalogued using as fine-grained metadata as possible, with the help of descriptors established by OLAC. This metadata is also encoded in XML, and its dissemination conforms to the standards of the Open Archives Initiative (OAI), which provides a relatively simple exchange protocol.

Once the data are stored safely and in a standard form, it becomes possible to share these resources with the larger community. We chose the Web as the means for publication of the data because it is virtually cost free. As the data are stored on a server and not with the user, it can regularly be corrected if necessary. We can also supply tools for dynamic analysis of the data and, like the data, maintain the tools more easily if they are server based. Furthermore, the Web reaches the greatest number of people and is the most relevant multimedia platform for archiving, especially because one does not have to write the tools from scratch. Access to the data in the archive is via the OAI conformant catalogue. The data itself is downloadable or viewable via a Web interface. This interface transforms the XML-encoded data on the fly from the server and uses multimedia based on time codes with the help of some client plug-ins and JavaScript applications.

4. Sharing data: How and why?

As linguists we want to be able to use our data ourselves, meaning that the linguist who collected the data wants to have access to it and to use analytical tools in order to continue his or her postfieldwork research. Similarly, other participants in the research need to access the data, especially those recorded who may need to veto or edit what has been recorded.

Further, the data can be shared with other academics, usually linguists working in the same region or on the same language family. At a broader level, all linguists, including Natural Language Processing (NLP) practitioners, may be interested in seeing the primary data. In addition, the data should be available to

2. XLink is XML linking language, used for creating hyperlinks in XML documents.

those who want to use it for pedagogical purposes. Given how few resources are available for small languages, it would be counterproductive not to make them available to a broader community of users (though they would always be subject to normal access agreements).

The intention to share data is not itself enough, and it takes some effort to establish a mechanism for sharing data. The first step in sharing is listing what resources exist in a generally available catalogue that is legible not only by humans but by machines. The Open Archives Initiative (OAI) specifies an architecture with two or three levels (data providers, aggregators, and service providers). Metadata harvesting is done by the aggregators and the service providers. They centralize all the metadata from the selected data providers and offer services like search engines on the all-metadata databases. Both projects discussed here subscribe to this architecture as data providers. The description of resources must, as much as possible, be standardized in order to facilitate research and exchange as well as federated searches over all conformant catalogues. The current aim is to conform to Dublin Core (which is the minimum for OAI) or OLAC metadata systems.

The next aspect of sharing is achieved by standardizing formats and normalizing data in order to build a homogeneous corpus and associated tools for research and editing of the corpus. The basic level of sharing requires standardized encodings, such as IPA for transcriptions as suggested above. A higher level of interoperability of data can be achieved by using encoding systems like that recommended by the Text Encoding Initiative, which is not very detailed when it comes to oral transcriptions, or the standards of the working group on ISO TC37/SC4.

The third aspect of sharing is at the organizational or institutional level. We will not be able to share data with future generations unless we can protect it not only from normal deterioration, but also from political and technological changes. There is still room for progress in corpus construction, and especially in recognizing its value as an activity in itself, equivalent to other forms of publication.

PARADISEC and LACITO are engaged with large sets of legacy linguistic data and with currently created digital data, which is becoming increasingly important. Both projects use the recommended standards. We also advocate that our user community provide its data in the best possible form to enable it to undergo the kinds of processes typically required to make linguistic data usable to the broader community and to the researchers themselves. Secure long-term storage of well-described linguistic records is crucial to language documentation and also has the potential to provide corpus data for NLP efforts.

Linguists are active in their support for language renewal and revival, but the ultimate determining factors in the ongoing use of these languages are typically political and economic. The role of the linguist is primarily as documenter of language use in as many domains as possible (see Himmelmann 1998, Woodbury

2003). The results of this documentation need to be safely housed and made available for ongoing research or repatriation to the speakers of the language or their descendants, especially for use in heritage language programs. From a linguistic perspective, these records may reveal points of typological interest and will also provide invaluable information for comparative and historical studies as they capture a view of a language in use at a particular point in time. They also record an important aspect of global diversity and ensure that many languages from a range of the world's language families are represented in repositories and research libraries for the future.