8

Anxious Respect for Linguistic Data: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD)

Nick Thieberger

8.1 Introduction

It is clear that the wind of change blowing through the discipline of linguistics in the recent past is resulting in a heightened focus on fieldwork methodology. Leaving aside that part of our discipline that has never concerned itself with field recording, there is a change in the way that linguists can now conceive of linguistic data, in particular, but not solely, as it relates to endangered languages. Three main determinants of this change are: the realization of the urgency to record as much as possible before speakers stop speaking small languages; developments in technology¹ which allow us to work with data with an immediacy not previously possible; and a growing involvement in linguistic projects of speakers or their descendants who want access

This chapter was written while I was supported by the following Australian Research Council grants: SR0566965—Sharing access and analytical tools for ethnographic digital media using high speed networks, DP0450342—New methodologies for representing and accessing resources on endangered languages: a case study from South Efate, LEO453247—Digital archiving equipment for PARADISEC research archive of Asia-Pacific region audio recordings, LEO560711—PARADISEC, the Pacific and Regional Archive for Digital Sources in Endangered Cultures: Accessibility and Decentralization.

¹ 'New technologies' is used here to refer to digital recording, software, tools and methods that have become available in the past decade.

to reusable linguistic data. These three factors point to the importance of making a high quality record of a language where quality refers both to the content and the form of the recorded data. This chapter arises from my current research on developing workflows for linguistic fieldwork, based in my own recording of South Efate (a southern Oceanic language of the Malayo-Polynesian branch of Austronesian) in central Vanuatu, and also on the development and operation of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). This is a linguistic archive primarily providing for researchers working in the Pacific, Papua New Guinea and Indonesia and so contains a significant collection of Austronesian material. The associated training and advice network, the Resource Network for Linguistic Diversity (RNLD), was set up to ensure that good techniques for working with linguistic data were propagated both within and outside of academia, resulting in better documentation suited for long term archiving.

8.2 Grammars and Reusable Linguistic Records

This chapter discusses the notion of preservation of linguistic records and asks to what extent that is achieved by writing a grammatical description of a language. I show that, like most grammars, Austronesian language grammars are not currently based on a corpus, do not cite primary recordings nor embed their analysis within the data and do not prepare recordings for future reference. I will discuss methods for citing data (including creation of a repository and permanent identification of files), and finally discuss the need for training and advice to propagate these methods among linguists and language workers. While in this chapter I focus on the form in which data is best created and stored, it should be noted that the content should be linguistically as comprehensive as practicable, including a range of communicative types (monologue, conversation, meeting, procedural text, etc) with a range of speakers of different ages (as argued, for example, by Himmelmann 1998). The data needs to be in accessible formats, which means that data files are not held in proprietary or compressed formats. Typically this means plain text which may be structured using some kind of markup (such as Toolbox's 'backslash' codes or XML). Further, good file-naming conventions and data structures will enable us and others to locate files we refer to in our analysis.

A major reason for considering these issues is that the data should be legible and reusable by others in future, that is, it should be portable (cf. Bird and Simons 2003), and not just by linguists, as it is 'less and less viable for linguists to think of stakeholders in language documentation to be constituted only of vaguely-conceived scientific posterity' (Woodbury 2003: 38–9). There has been a notion that grammars are themselves a sufficient record of the language, but they are only a partial record, especially if they are written in a narrow and currently fashionable theoretical framework. While it is not in dispute that a grammatical description is an essential part of the documentation, it is not the only part, and it can now, with the popularization of digital tools, be embedded within the data in a way that was not possible in the past. While it is now

possible to adopt a means of writing a grammar that includes the development of broader documentary materials in its process, it is still not common practice and has not been rewarded sufficiently by the linguistic community. This is especially the case for PhD dissertations for which students are not currently given any recognition if they provide data in the form of texts, a dictionary or media.

In an attempt to understand how grammar writers to date have conceived of the relationship between their grammar and the data on which it is based I searched through a range of grammars written over the past generation for references to the source data. I found that typically grammars do not provide data beyond example sentences. Occasionally there are a few texts, but they do not cite their source and are not given as recordings but only as transcriptions. Rather than listing all the grammars consulted, it is preferable (and quicker) to acknowledge those that do cite data so that their readers can confirm hypotheses and work with the data in ways not necessarily envisaged by the author. An early example is Jeffrey Heath's Nunggubuyu grammar (Heath 1984), texts (Heath 1980) and dictionary (Heath 1982) in which each example in the grammar is sourced from the (paper) texts. More recently Morey's (2004) Tai grammar and my grammar of South Efate (Thieberger 2006) have provided media with transcriptions, allowing the reader to hear audio or see video examples used by speakers and to verify that examples actually occurred and in what context. While it may be argued that a small corpus (mine contains around 125,000 tokens linked to media) is not a suitable or representative sample of a language, we are operating within constraints of time and budget that will not permit creation of large corpora as are constructed for metropolitan languages (which can be in the order of millions of words—the Reuters corpus is 80 million words). The converse, that is, the lack of any corpus or citable data accompanying grammars, should no longer be acceptable for linguistic research in that none of the observations made by the linguist can be confirmed by the reader and none of the data worked on intensively by the linguist can be made available for future access, especially by the people recorded.

Two recent overviews of issues in grammar writing are instructive for the way in which they ignore the role of the data in the grammatical description. They are both fine collections of papers on the craft of writing a grammar, but, at the beginning of the twenty-first century, neither of them seriously engages with the use of new technology. In the first, Ameka, Dench, and Evans (2006) briefly discuss the issue of new technological means for developing a corpus, and note that, '[w]here the published text collection once served as the grounding evidence for a linguistic analysis, the digital archive will come increasingly to fill that role' (Evans and Dench 2006: 25). However, they go on to caution that the increased speed of producing digitized recordings is unlikely to be accompanied by detailed analysis of the concomitantly larger collection (ibid). This emphasis on the problems involved in the use of new technologies is unfortunate and misses the basic issue that we can build new techniques into our work to create much better outcomes. Further, Evans and Dench (2006: 29) discuss 'new forms of grammars' as hyperlinked documents, again requiring more than the

'normal' effort of writing a grammar. There is vast gap between a rich hypertextual encyclopedic grammar/dictionary/text collection and the current model in which data is not presented. Creating well-formed data ensures there can be a number of output formats, so we should be adopting standards where they exist (e.g. the Multi Dictionary Formatter standard for use with Toolbox) and, as a community with shared interests, developing standards where they do not yet exist.

The second recent work on these issues is the special issue of *Studies in language* edited by Payne and Weber (2006), titled 'Perspectives on grammar writing'. In the introduction to this collection, Payne notes that in recent years 'the emphasis has shifted back to the centrality of data' (Payne 2006: 237), but this observation does not resonate with the collection of papers in the volume which cover traditional grammar writing and barely mention the relationship of the grammar to the data, or how that relationship could be made explicit. The contributions that seems to deal with the possibilities offered to language documentation by new technologies are two by David Weber (2006a, 2006b). For example, he observes that, '[i]deally language data would be managed—and grammars written—in a computational framework that integrates grammar and corpus, with examples existing in the corpus but accessed from the grammar. Examples would not be 'taken from' a text but displayed therein' (Weber 2006a: 449). Further, Weber (2006b:427) briefly discusses the organization of data into a corpus and the need for the grammar to be encoded in an archival interchange format, such as XML. But there is no discussion of citation of examples or texts to a source, even under the subheading 'some future possibilities', nor is there a discussion of data repositories for curation of the data recorded by field linguists. In concluding, Weber (2006a:459) acknowledges that, 'Grammar writers need hospitable authoring environments, with tools that are powerful and flexible, yet reasonably easy to learn and use. Until these are available we labor under the limitations of ink-on-paper.' This is an unfortunate conclusion that reflects the fact that we have not yet put enough effort into the creation and dissemination of tools that are relatively easy to use and that address the functions required by field linguists. It also reflects the lack of incentive for field linguists to be using existing methods and tools.

It is significant that neither of these collections on grammatology conveys the central issue facing endangered language data, that is the longevity of the data on which the grammar is based, and the means for interacting with that data. Critics of the use of new technologies often fail to understand the basic notion that digital data will survive better than analog data and that we may as well be creating good data from the outset both for that reason and for the ease with which we can access the data in our analysis. Or, as Grenoble and Whaley (2002:2) put it: 'The advent of digital technology and growing global access to the world wide web has the potential of radically changing the way that we store and access information on languages, and thus of how we linguists and anthropologists think about the data we collect.'

In 1992, some fourteen years prior to these volumes being published, J. Randolph Valentine published software called *Rook* which provided a grammar authoring

environment, linking texts to the grammar via a lookup table. So, for example, by clicking on a word in a text, the lookup would determine what part of speech it was and present that section of the grammar which was hierarchically organized. The text was imported from an interlinearization program (*IT*), and was presented together with audio. This system was written in HyperCard, and has now been superseded, but it still stands a fine example of the kinds of links we want to be able to make in a grammar writing and presentation tool. The Rook template was used to prepare a brief grammatical sketch of the Australian language Warnman linked to the audio of a short story (Thieberger 1994). *Rook* is an important early model but work has been proceeding to determine ways of encoding data types found in grammatical descriptions—typically descriptive prose (Burnard and Sperberg McQueen 1995/2002), interlinear texts (Bow, Hughes, and Bird 2003), paradigms (Penton and Bird 2004) and lexicon and the structure of each of these data types. This makes sense both from the perspective of making new kinds of grammatical descriptions in which the naturally interlinked elements of the language can be explored and explained in context, and from the need to make explicitly structured digital documents that will endure into the future.

8.3 Preservation

Linguists often claim that the grammatical descriptions they write will preserve endangered languages but unless they take steps to secure their field recordings then the only thing that will be preserved is the grammatical analysis. This is better than no record, but is not providing the data on which its claims are based. Further it is not providing narratives, songs, conversations, or examples of language in use that speakers of the language could reasonably expect to be results of fieldwork on their languages. Crowley expresses his dismay at ‘overinflated’ linguists who ‘claim to “save” some particular dying language, when in fact all that person is going to do is document the language. It is more than unhelpful when such linguists claim to be doing precisely what we are not in a position to do; it is, in fact, unethical’ (Crowley 2007: 187–8). We can provide good records as a result of our fieldwork but as outsiders we cannot preserve a language. To provide good records implies that we have strategies for *data management* together with the *infrastructure* to hold the data, and the means to *locate* it and *cite* it. It means creating *well-formed data*, that is, data that will itself not deteriorate over time and will be legible in future. Knowing what we now do about the transient nature of digital data, we need to build effective means for *storing* and *reusing* data, and ensuring it is *interoperable* as much as possible. Clearly, to do all of this we need to be *training* both new and old researchers in appropriate methods. I will elaborate on each of these themes below.

8.4 Data Management

Keeping track of all the recordings, notes, photos and videos that we make in the field requires disciplined use of a good database which will allow the internal relationships

within the data to be mapped. For example, it is most likely that you have recorded a small number of speakers on a number of different occasions. A table of the database will be the speaker names, another table will be the recordings, another will be the list of derived stories for production in Toolbox and then publication, another will be a list of pictures taken and so on. A field in the database can also be a checkbox for the stage of processing that a file is at, so that an untranscribed recording can be readily identified. A database will also help enforce standard ways of referring to the items, for example, allowing a controlled list of speaker names to be used to avoid variant spellings or versions of names.

Naming conventions need to be thought out in advance so that you are able to name the files created in ways that allow them to be retrieved later, and so that they can be transferred directly into an archive without renaming. If you use these names from the moment of creation of the file then you will not have to keep track of renamed versions later on. At PARADISEC 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 user's 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, and are composed of a collection identifier (e.g., NT1), followed by an item identifier (200601) 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 (so NT1-200601-A.wav).

You will also need to record good metadata about the file at the time it is created or as soon after as possible—who is on it, when and where it was recorded, what the language is and so on. Again, using standards (e.g., the ISO standards for language names (ISO 639-3, based on Ethnologue) or for dates (yyyy-mo-da)) can save a great deal of effort later on. The Open Language Archives Community (OLAC) metadata set is a simple standard that provides for the basics that linguists want to include in descriptions of field recordings—roles of participants; language names; dates; linguistic type of the data among others. The IMDI metadata set is far richer and offers the ability to operate at a more granular level with the data, but requires more work on the part of the creator than does the OLAC metadata set.

Data needs to be backed up, preferably being sent to an archive which will take care of backing it up in perpetuity. In one case at PARADISEC we had a fieldworker in the Philippines sending DVD copies of data to us directly from the field (as seen in Plate 14). To do this, she developed a naming convention for the files that then carried through all of her work. We have been encouraging PhD students to deposit material in the archive as soon as they return from fieldwork as this not only secures their data, but gives them a citable form of their data for their analysis (as discussed below).

It is important to understand that the data resulting from fieldwork has to flow through various tools and processes as the linguist proceeds with analysis. This

Figure 8.1 Workflow for file transformations (Nash 2005)

workflow is described in Thieberger (2004, 2005) and elaborated on by Nash (2005) as seen in Figure 8.1. It tracks typical operations carried out in the course of analysing linguistic data given the current available tools, beginning with a digital recording (either audio or video), adding metadata and archiving the file before commencing analysis, transcribing with time-alignment and various software for annotation and exploration of the data (including *Audiamus*, a tool I wrote to work with a corpus of media via a textual representation). Finally there are outputs resulting from this workflow which include a collection of annotated texts linked to media, a dictionary and so on.

This is the workflow that suits me and it may not be for everyone. However, it makes clear to those who are starting out what the various tools do and how they fit together, and how the data should flow through them, with value being added at each step, but allowing the data to be kept separate from the applications that operate on it.

8.5 Infrastructure

In order to preserve linguistic records in the long term and to undertake the various tasks listed in this chapter there is a need for shared infrastructure (to avoid duplication of effort) to create a data repository. PARADISEC is an example of this kind of infrastructure and includes digitization suites in Sydney, Melbourne, and Canberra which convert analog tapes recorded as early as the 1950s. An online catalog tracks these items and ensures they have basic metadata before they are sent to a server in Sydney for inspection, processing as appropriate to the type of data (e.g. sealing of audio files with an XML textual header and production of MP3 copies for delivery)

and automatically backed up to tape. A further automatic process copies files to the mass data storage system² in Canberra. Weekly, monthly, quarterly and annual reports are generated from the Canberra storage system³ that allow us to ensure the integrity of the data. We have assisted a number of linguists to enter their own metadata and then to deposit their recordings and we have been refining the software over time. The ease of use of this process is critical, as we know very well that linguists have enough to do and will not take on extra work unless it is fairly painless and they can see a benefit to themselves.

8.6 Field Helper

While it is tedious to enter metadata, there are ways of automating metadata entry and extraction from files, as for example with the tool called Field Helper⁴ in which files can be dragged onto icons associating them with certain metadata items. Thus the user may define a set of icons to represent linguistic types (lexicon, text, primary description and so on) and then be able to drag items (such as transcripts or texts) from folders onto those icons, thereby assigning those characteristics to the metadata for those items. The underlying XML file need never be seen by the researcher, but, due to its well-formed structure, can be copied to repositories which can then incorporate it into their catalogs.

8.7 Locatable and Citable Data

Searching for linguistic data requires that it has been described adequately and in a form that is predictable so that search engines can retrieve it. The catalog or metadata information minimally has to provide for standard language names to be used, and for a means for expressing non-standard names for those languages that are currently outside of the ISO system. Geographic data can also be associated with the entry, in PARADISEC's case using a geographic bounding box on a map (min/max latitude/longitude), but possibly also using the Getty Thesaurus of Geographic Names (TGN).

Once an item has been accessioned into the archive it will be provided with persistent identification which means that its name and location will not change over time, allowing it to then be cited in other work. With new techniques for transcribing with time-alignment it is possible to cite to the level of a word or utterance within the collection, which is the practice I developed in my PhD thesis (published as

² A hierarchical file system (HFS) with robotic tape server at the Australian Partnership for Advanced Computing (APAC).

³ This system, called 'Babble' was written by Stuart Hungerford at APAC.

⁴ Field Helper (or FIDAS) (available at <http://acl.arts.usyd.edu.au/FIDAS/> and described at http://acl.arts.usyd.edu.au/index.php?option=com_content&task=view&id=178&Itemid=211) is being developed collaboratively between PARADISEC and the Archaeology Computing Lab at the University of Sydney. At the time of writing it is in alpha release.

Thieberger 2006). Citation of archival data will become the norm as younger linguists adopt the kind of methodology advocated in this chapter and as more repositories are established.

8.8 Well-Formed Data

Archival forms of data are those that we predict will be usable over time. We know that digital data can be locked up in proprietary formats, so to avoid losing data when companies change their software, PARADISEC archives open and uncompressed formats following standards established by far larger interest groups of librarians and archivists than our own. We will not accept multimedia presentations or files in proprietary formats as we do not have the staff to extract data from them, but we can provide advice on how to create good files in the first place, or how to convert files into a suitable format. The archival version of a file is the primary output of recording (as seen in the workflow in Figure 8.1) and derived forms of the data can be converted for delivery, as for example MP3 files are smaller and easier to move around than are their corresponding WAV files. Similarly, a lexical database may contain much more information than the derived dictionary, which may be in a number of formats (paper, computer-based CD, or web page), and a set of interlinear texts may have more information than their online representation does.

8.9 Reusable and Interoperable

If you have ever had to rework a dictionary created by someone else you will appreciate the value of having such work in a digital text form rather than just on paper. As speaker communities become more literate and want, for example, to provide resources to support use of the language in schools, they will revisit the work done by a linguist in that community five, ten, twenty or more years ago. If the work is held in a suitable repository it can be reproduced in a new form. If, however, the linguist used proprietary software like Microsoft Word or lost the only copy of the file on their computer then the work of revising the dictionary is so much harder. Once reusable data is made locatable via general search engines it can be put to unexpected uses (as long as they conform to deposit conditions). For example, linguist Tom Dutton recorded a radio play by a Papua New Guinean author in 1979. This tape became part of his collection and, after accessioning at PARADISEC it became locatable via an internet search engine. It was found by a drama researcher in New York who knew the playwright's work and could find no other copy of this play, either as a written script or as a performance. She signed our access agreement, downloaded the files and transcribed them, subsequently depositing her transcripts with us, resulting in the script becoming part of the material available in the collection.

Interoperability of data is more difficult to achieve than simple reusability. For data to interoperate with other data (and this means automatic interoperation by computational processes) it has to be constructed in a strictly standardized format. Once we have standard formats then we can benefit from tools based on those formats, and further, we can visualize the data in new ways once it starts to work together with other data or we can locate new sources of data that we were previously unaware of. An example of interoperable data is the set of metadata created by an archive that is part of the Open Language Archives Community (OLAC). The PARADISEC catalog exports a selection of fields to an XML encoded document which conforms to the OLAC standard. The strength of XML is that the document declares in its header what kind of standard it adheres to (just as an HTML document begins by declaring itself to be HTML) but it also provides a means of validating the structure of the document against a published schema (which is a kind of phrase structure grammar). This allows it to then be harvested by a number of search engines which would not be able to access it as an unstructured document, except in the way that Google accesses data, that is, as a set of largely undifferentiated text.

A further benefit of reusability is the return of material to the host or source community. It is particularly the case when a language is no longer widely spoken that future generations may want to relearn it, but it is also the case that these small communities have little in the way of written or recorded history and greatly value the recordings made by linguists. As Grenoble and Whaley (2002) point out, and my experience at PARADISEC confirms, dissemination of data in digital form and via the internet may be the best means of making research data available to the source communities. During a fieldtrip to Erakor village in Vanuatu in 2006 I was asked for copies of recordings I had made with an old man who had since died. Because I had prepared my data in the way advocated in this chapter I was able in a matter of hours to produce ten copies of an audio CD with a set of 15 'greatest hits'—stories and songs from the village, together with cover notes drawn from the existing metadata catalog. This would simply not have been feasible had the media been in analog form.

8.10 Online Presentation of Annotated Text and Media, the Ethnographic Online Presentation and Annotation System (EOPAS)

In 2005 and 2006 I led a team developing a method for representing interlinear glossed text (IGT) linked to media. IGT is a convenient means for presenting bilingual annotations together with a free translation. The issue here was how to develop an archival form for interlinear text and then how to stream media (in open formats) and present interlinear text with the media it transcribes, locally and on the internet. Some linguists use Toolbox to produce interlinear annotations of texts; however, there is still

no published standard for structuring these interlinear files despite a good typology of interlinear texts developed by Lehmann (1983), Bow, Hughes and Bird (2003), and Hughes, Bird and Bow (2003). Toolbox uses the 'backslash' markup in text files, so the files are quite reusable, but because there is no constraint on the fields that can be used, the order they occur in or on their names, the form of the data cannot be predicted for automatic processing. Exports from Toolbox in XML require a suitable internal structure within Toolbox to provide an XML document with a predictable structure.

The multidisciplinary and multi-institutional project called *Sharing access and analytical tools for ethnographic digital media using high speed networks*, or simply EthnoER, worked on a model that would ultimately allow diverse data types to interoperate, with grammar, dictionaries, texts, and media all interlinked and navigable. In this first step we focused on a means for viewing, concordancing and searching simple transcriptions or interlinear text associated with media, combined with the ability to access the media itself. The types of linguistic outputs addressed by EthnoER are time-aligned transcripts (from ELAN and Transcriber) and IGT (from Toolbox). Media is streamed using Annodex which can call time segments within the media which itself is automatically converted to an open format (Ogg).

We provided a template for Toolbox users so that data exported from Toolbox is well-formed XML and we developed a schema (called EOPAS) against which the XML could be validated as it is uploaded. Validation means that the structure of the document is known and predictable and can then be used in the browser we created for IGT, in the same way that a web browser presents HTML documents. This project has made explicit the structure of simple interlinear text, a common data type in linguistic methodology, and also provides an archival form for interlinear text in which the internal relationship between a morpheme and its gloss, or an utterance and its translation, are made explicit by XML tags. The resulting form of the data is both reusable and interoperable.

8.11 Storing Data: PARADISEC Case Study

PARADISEC is a collaborative research project established in 2003 between the University of Sydney, the University of Melbourne, and the Australian National University (the University of New England joined in 2004). The main motivation was the need to digitize audio tapes recorded since the 1950s by researchers who had nowhere to deposit them and who were storing them in places neither designed nor appropriate for long-term housing of analog tapes. As noted above, PARADISEC stores data in several locations and hopes to extend the location of copies via an international collaboration known as DELAMAN, which includes a number of language archives (among them are the Archive of the Indigenous Languages of Latin America (AILLA) in Texas, Endangered Languages Archive (ELAR) in London, Dokumentation

Bedrohter Sprachen (DoBeS) in the Netherlands, and the Alaskan Native Languages Centre (ANLC) in Alaska). At April 2008 we store 2,400 hours of media (over 4.1Tb of data). There are 5,200 records in the PARADISEC catalog with data from 600 languages and 50 countries. We spend considerable time with many old tapes, preparing them for data transfer by cleaning and, in some cases, baking or placing them under vacuum to readhere the media to the backing tape. We return all tapes to the depositors as we have no conservation funds and no physical storage space. However, not all items in our collection represent legacy material. Some, and an increasingly large part, is 'born digital'—that is, it results from current fieldwork using digital recorders and requires less work on the part of a digital archive, especially if depositors take care to use archival formats of media and enter their own metadata online.

Working with legacy material we can observe what small additional steps a researcher could have taken to make their recordings more useful to us now. In some cases there may be no specific information about the tapes we have located in a box or filing cabinet, and, while there are 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. We are taking digital images of notes and putting them into an online delivery system that permits researchers to note the relationship that obtains between the notes and the tapes. We plan to allow online annotation of media that will enrich the existing collection, using the expertise of local experts. Simple descriptive metadata then allows us and potential researchers to locate the relevant media files and to reintegrate them with field notes.

The ability to enforce standards on depositors extends to the description of the data, or the metadata that allows the data to be discovered. We use a cataloging system that provides a description of the item as well as of the process it undergoes from accession. The catalog is critical to the organization of the collection and includes a number of controlled vocabularies (including roles of people involved in producing the item, a standard language list and country list, for example) to ensure consistency of data entry accessed via pop-up menus. Geographic data can be associated with each entry via a box drawn onto a map (as seen in Map 8.1), or by reference to the country or language name, depending on the level of detail required.

All of this metadata can be output in various forms, one of which is the Open Language Archives Community metadata set. We provide a dynamic repository (using a PHP query of our SQL database) of our metadata for OLAC searches in order to increase the visibility and so the discoverability of the material in our collection. This has been a fairly easy process that has meant that we were able to move our metadata system to an Open Archives Initiative conformant metadata repository in a few months.

We encourage practitioners (whom we take to mainly include linguists, musicologists, and indigenous language workers) to deposit recorded media by providing them with a high quality digital version of their data in the short term. If an archival form

Map 8.1 Density of items in the PARADISEC collection, an image from the search tool

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, so we have been encouraging researchers to lodge their tapes with PARADISEC as soon as they return from fieldwork. We digitize or capture their data and create an archival (usually at 96khz/24bit Broadcast Wave Format or BWF), a working copy (data or audio CD), 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 descriptive analysis or a dissertation) and the long-term needs of having richly annotated primary data safely archived.

8.12 Austronesian Languages and PARADISEC

Initially PARADISEC was conceived as being a repository for language material from the region around Australia. This was, in part, due to the recognition that Australian researchers working in the region had nowhere to deposit their research material, unlike those who worked with Australian languages for whom the Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS) was the location for depositing their data.

While we have accepted data from languages much further afield than this, the bulk of the collection is from Papua New Guinea, Indonesia and the western Pacific. Of 6,040 items in the collection (as of January 2009), 3,600 are from Papua New Guinea, Indonesia, the Philippines, and the Pacific, as can be seen in Map 8.1 drawn from the geographic search tool in our catalog (density of shading indicates the number of items

Table 8.1 A sample of Austronesian data in the PARADISEC collection

Alexander Adelaar	24 tapes. Salako, Embaloh	Indonesia
Amanda Brotchie	11 items. Tiraq	Vanuatu
Terry Crowley	8 items. Paama, Sye, Ura, Malakulan languages	Vanuatu
Mark Durie	314 tapes. Acehnese.	Indonesia
Tom Dutton	295 tapes, in a range of PNG languages, including some Austronesian languages	PNG
Domenyk Eades	12 tapes. Gayo	Indonesia
David Goldsworthy	89 tapes. Music.	Indonesia and Malaysia
David Goldsworthy	16 tapes. Music	Fiji
Ralph Lawton	7 items. Kiriwina	PNG
Claire Moyse-Faurie	11 tapes. Xârâgurè, Xârâcùù	New Caledonia
Laura Robinson	58 items. Dupanangan Agta, Eastern Cagayan Agta.	Philippines
Leonard Drile Sam	11 tapes. Dehu	New Caledonia
Cynthia Schneider	62 items. Abma	Vanuatu
Al Schütz	6 tapes. 1966 Nguna	Vanuatu
Jeff Siegel	31 tapes. Fijian	Fiji
Jeff Siegel	3 tapes. Nauruan pidgin	Nauru
Wolfgang Sperlich	Niue and Namakir dictionaries	Vanuatu
Tote Tepano	21 items. Rapa Nui	Easter Island
Nick Thieberger	364 items. 1990s South Efate	Vanuatu
Stephen Wurm	161 items in various languages, mainly from Reefs/Santa Cruz	Solomon Islands

in the collection). Of these roughly 1,300 represent Austronesian languages as can be seen in the list given in Table 8.1.

While each collection is different, some examples of items in the collection will illustrate the range of material held. Stephen Wurm's estate includes several hundred tapes, including 120 Solomon Islands tapes and transcripts or field notes from the 1970s. Arthur Capell's estate includes 114 tapes and 30 archive boxes of field notes from the Pacific, Indonesia, and Papua New Guinea. As a trial of how to manage the papers in these collections we took images of pages of notes, with a naming convention for files that allowed them to be assembled into an online presentation format. Some 14,000 images of Capell's work are now online and have already been used by researchers who would otherwise have had to travel to Sydney to see them. A copy has also been provided to the Vanuatu Cultural Centre for local use. Similar work has been begun with Stephen Wurm's papers and with the missionary Calvin Roesler's papers in Asmat. While there is much more of this kind of material that could be made more accessible, due to a lack of funds we have not proceeded further with this work.

8.13 What PARADISEC is Not

PARADISEC is a project operating in virtual space between a number of sponsoring institutions. It was only ever funded at the bare minimum required to establish a digitization suite, server and the human infrastructure to support those activities. Much as we would like to, we have never been in a position to 'value-add' to our collection, and instead have to rely on the form in which depositors create and supply their data. We do provide guidelines and assist in reformatting if necessary, but we simply do not have the staff to do this kind of work. As a result, we have good infrastructure for what has been called 'thin' metadata—what we consider the minimum to make the data discoverable. We have not provided for in depth querying of particular datasets. While it may be useful, for example, to locate all subordinate clauses used by a male under 50 in a particular region, this is not the kind of query supported by our metadata.

8.14 PARADISEC: Future Plans

PARADISEC began as a project between academic institutions, but has now become an incorporated entity (PARADISEC Inc) with deductible gift recipient status in Australia which allows us to seek sponsorship to cover the cost of digitizing collections. With a more secure funding base we could actively locate endangered collections of field recordings and preserve them by visiting small cultural centres which have analog tape collections and providing digital copies to them while safely storing archival versions. In addition to the large backlog of academic research outputs, we regularly receive requests for assistance with small collections, for example those made by patrol officers, missionaries, or travellers, often from the children of the recorders who recognize the value of the collection, but have nowhere for it to be housed.

We participate in a number of initiatives to make linguistic data easier to create in forms that will be interoperable and archivable (such as the EOPAS and Field Helper projects mentioned above) and consider these kinds of initiatives to be integral to the future of language documentation.

8.15 Training

Archives rely on the relationships they have established with their communities, both depositors and users. In general, the benefits of depositing are clear, in particular as we are digitizing analog tapes and holding 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 a key for us has been the ability to provide advice and training to ensure the quality, both technical and in content, of recordings and associated derived material (transcripts, glosses, dictionaries, etc). The rationale is that if we want high quality

recordings and well-structured archival data, then we have to provide training in its creation. We run workshops in using Toolbox, which is still the only tool that creates structured lexical files linked to interlinear text, and, 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 centres.

Training graduate students, current practitioners and speakers of endangered languages to use these technologies is critical to their uptake and to the successful creation of linguistic resources that will endure into the future. New methods and tools keep appearing and collective wisdom is required to keep track of the best of them. A further issue is that there is and will be much more language recording than we in academia know about. It is therefore important for us to reach out to anyone doing this kind of work and to provide advice and training so that their efforts are as productive as possible. To this end Margaret Florey and I established the Resource Network for Linguistic Diversity (RNLD) which promotes sharing of information and has sponsored conference sessions and training workshops at a number of locations. The mailing list, launched in April 2004, has had around 730 messages since then, with 230 subscribers. The strong links between our archiving project (PARADISEC) and RNLD help to promote good archival storage of results of research and repatriation of older collections to their source communities.

RNLD has run many training courses in the use of linguistic software, including some with an Austronesian focus at the University of Hawai'i at Manoa and the University of New Caledonia.

8.16 Conclusion

Field researchers can now have instant access to field recordings both to support analysis and to provide citable archival forms of data. This digital data is ideal for repatriation to the host community as it is easier to produce (CD copies can readily be duplicated) and it has been enriched with time-aligned transcripts and is more useful than tapes alone used to be.

Linguistic practice benefits from exposure of primary data as linguists become more accountable for their analyses and at the same time provide richer data for more fine-grained analysis. Archival data is critical for long-term preservation of linguistic information and there is a need for more linguistic archives collaborating in the international network. These archival projects need to be collaborative at a national or international level because we have so little in the way of resources and we have similar needs. This, in turn, implies that we use common methods based in accepted standards among the community of linguists. To produce suitable forms of data from fieldwork we need to develop and propagate methods and tools that are easy to install and use,

and don't add too much to the existing workload together with archival infrastructure to support practitioners in language documentation efforts.

In 1938 Edward Sapir, discussing the nature of linguistic data created by fieldworkers, said:

I'm not particularly interested in 'smoothed-over' versions of native culture. I like the stuff in the raw, as felt and dictated by the natives. [...] The genuine, difficult, confusing, primary sources. These must be presented, whatever else is done. [...] There are too many glib monographs, most of which time will show to be highly subjective performances. We need to develop in cultural anthropology that anxious respect for documentary evidence that is so familiar to the historian, the classical scholar, the Orientalist. (Letter from Sapir to Fay Cooper-Cole, 25 April 1938, quoted in Darnell 1990:139)

Perhaps our anxiety today is that so much is possible in the way of recording and producing documentation of small languages and expectations of our outputs are also increasing. Today's fieldworker must be anxious both about doing sufficient work on the analysis of the linguistic data, as well as producing well-formed documentation.

My experience at PARADISEC is that researchers will engage with new techniques for recording and annotating field materials when they understand that it will provide better outcomes from their work. It is critical for successful curation of digital linguistic data that we train practitioners in new methods so that they can create archival data as part of their normal practice, depositing their material before analysing it, and so turning the old understanding of archiving on its head.

In addition to the importance of preparing data for repatriation and for re-use in future, the methods described above allow us to create citable forms of our research data in order to do better research now, providing data together with an analysis and allowing our claims to be tested. Finally, our responsibility to the people we record is a motivation for creating locatable and accessible data repositories. There is a confluence of new technological tools with an awareness among many linguists of the need to deal responsibly with data they collect. Linguists can take advantage of new tools and infrastructure to develop our own work while leaving sufficient records so that we can earn the respect, anxious or otherwise, of posterity.

Appendix: A list of resources discussed in this chapter

Annodex	http://www.annodex.net
Capell papers online	http://paradisec.org.au/fieldnotes/AC2.htm
Digital Endangered Languages and Musics Archives Network (DELAMAN)	http://www.delaman.org

EOPAS	http://wiki.arts.unimelb.edu.au/ethnoer/Main_Page . The EOPAS schema is here: http://ethnoer.unimelb.edu.au/eopas.xsd
EthnoER	http://ethnoer.unimelb.edu.au This project was funded by the Australian Research Council (SR0566965).
Eudico Linguistic Annotator (ELAN)	http://www.mpi.nl/tools/elan.html
Field Helper	http://acl.arts.usyd.edu.au/FIDAS/
IMDI editor	http://www.mpi.nl/IMDI/tools/
IT	http://www.sil.org/computing/catalog/it_dos.html , http://www.sil.org/computing/catalog/it_mac.html
OLAC Open Language Archives Community	http://www.language-archives.org/OLAC/metadata.html
Resource Network for Linguistic Diversity (RNLD)	http://rnld.org
Roesler papers online	http://paradisec.org.au/fieldnotes/ROES/web/roes.htm
Rook	http://www.sil.org/ftp/software/mac/rook101_sea.hqx
TGN—Getty Thesaurus of Geographic Names	http://www.getty.edu/research/conducting_research/vocabularies/tgn/
Toolbox	http://www.sil.org/computing/toolbox/index.htm
Transcriber	http://trans.sourceforge.net/en/presentation.php
Wurm papers online	http://paradisec.org.au/fieldnotes/SAW2/SAW2.htm

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Thieberger, N

Title:

Anxious Respect for Linguistic Data: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD)

Date:

2010-05-01

Citation:

Thieberger, N. (2010). Anxious Respect for Linguistic Data: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD). Florey, M (Ed.). Endangered Languages of Austronesia, pp.141-158. Oxford University Press.

Persistent Link:

<http://hdl.handle.net/11343/31158>