Language Documentation and Description

ISSN 1740-6234

Audio responsibilities in endangered languages documentation and archiving

DAVID NATHAN

Audio responsibilities in endangered languages documentation and archiving

David Nathan

1. Introduction[1]

Today the world is facing the impending loss of at least half of its languages. Many linguists are addressing this challenge through the emerging discipline of documentary linguistics (Himmelmann 1998, 2006, Woodbury 2003). Documentary linguistics (also called 'language documentation') focuses on data, and how data is acquired, represented, presented, and preserved, in contrast to the analytical and theoretical concerns of much of linguistics (Austin 2006, Austin and Grenoble 2007). And since many endangered languages are not written, the majority of this documentary data is audio. In turn, this raises new and interesting questions, such as: what audio data needs to be collected to count as a record of a language that is likely to disappear? Are standard corpus concepts of coverage and balance applicable to endangered language documentations? How can quality be measured? For what purposes and by whom will the data be used?

For those of us concerned with the evolution of documentary linguistics, there are four key audio-related issues. The first is *audio quality*; typically, linguists need considerable training in order to make good audio recordings. To help address this, the Hans Rausing Endangered Languages Project at SOAS has developed and run audio training courses in several locations, including London, Lyon (France), Tokyo (Japan) and Winneba (Ghana). The second issue is the role and nature of the *symbolic data* that accompanies audio. While linguists are increasingly using standardised software tools for annotation, transcription, and metadata creation, there are still debates about methodologies and wildly varying practices. There is also no clear agreement about the roles that symbolic data play in archiving, processing and presenting endangered languages data. The third issue is what we call *mobilisation*: the practical development of resources and products that make use of collected data to serve purposes such as language revitalisation (Nathan 2006). While examples such as pedagogical multimedia can be effective, in general methods for creative and effective presentation and navigation of audio remain limited, being drawn from other areas such as games. The fourth issue, *protocol*, arises from the fact that audio directly captures and represents

[1] An version of this paper is to appear in the *Taiwan Journal of Linguistics*, 6(2).

David Nathan 2009. Audio responsibilities in endangered languages documentation and archiving. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 6*, 101-116 London: SOAS.

individuals in a way that written data does not. For endangered languages communities, which are often under a range of social pressures, we have to enhance the ways we deal with sensitivities and how we implement protocol in audio access and distribution.

2. Endangered languages and documentation

Documentary linguistics is a subfield of linguistics that emerged a decade ago as a response to predictions that thousands of human languages will disappear within a century (e.g. Krauss 1992). It aims to develop “methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language” (Gippert, Himmelmann and Mosel 2006:v). Language documentation weaves its focus on endangered languages together with ‘traditional’ descriptive linguistics and an emphasis on the appropriate use of media and information technologies. It also adds the ethical dimension of involving language speakers and considering their rights and needs (Grinevald 2003). Austin and Grenoble (2007) identify the core features of documentary linguistics as the following, after Himmelmann (2006:15):

- *focus on primary data* – documentation consists of collecting and analysing an array of primary language data which is also made available for a wide range of users
- *accountability* – access to primary data and representations of it makes for more transparent evaluation of linguistic analyses
- *long-term preservation* – a focus on archiving to ensure that documentary materials are available to a range of potential users into the distant future
- *interdisciplinary teams* – documentation requires input and expertise from a range of disciplines and is not restricted to linguists alone
- *involvement of the speech community* – collaboration with community members not only as consultants but also as co-researchers

The outcomes of documentation are sometimes described in terms of lists of interaction types, genres and styles. For Wittenburg et al (2002), for example, “the corpus should consist of a variety of text types and genres” as in the following list of (from Johnson and Dwyer 2002):

- *interaction* – conversation, verbal contest, interview, meeting/gathering, riddling, consultation, greeting/leave-taking, humour, insult/praise, letter

- *explanation* – procedure, recipe, description, instruction, commentary, essay, report/news
- *performance* – narrative, oratory, ceremony, poetry, song, drama, prayer, lament, joke
- *teaching* – textbook, primer, workbook, reader, exam, guide, problems
- *analysis* – dictionary, word-list, grammar, sketch, field notes
- *register* – informal/conversational, formal, honorific, jargon, baby/caretaker talk, joking, foreigner talk
- *style* – ordinary speech, code-switching, play language, metrical organization, parallelism, rhyming, nonsense/unintelligible speech

In addition, audio (or video) recordings are generally at the centre of a documentation, and "should be associated with an orthographic or phonemic transcription, a translation in one of the major languages of the world, and/or glossings in a local lingua franca and English" (Wittenburg et al. 2002).

Nevertheless, due to a lack of settled conventions in the field, or perhaps in defiance of the recommendations of Himmelmann and others, documenters often characterise their documentation corpus in terms of number of hours of audio/video recordings made and the percentage of it that they have transcribed or annotated (all too frequently only 10 or 20 percent). Funding bodies can also impose quantitative specifications or expectations on the documentary work they are willing to support, such as number of hours recorded or transcribed (Dobrin, Austin and Nathan 2008).

However, a survey of the goals and practices of documentary projects that the Endangered Languages Documentation Programme (ELDP)[2] has sponsored indicates that in fact many projects have a specialised focus on particular linguistic or cultural phenomena or practices or genres[3]. This should be regarded as welcome: it is not realistic to expect documenters to do 'everything'; and even if they did, their results are likely to be consequentially thin. As this survey suggests, the content of documentary recordings depends on many factors, including the particular situations, personalities and preferences of the researchers and language consultants (and their families and communities). Recordings and representations of specific phenomena will be of more interest to the researcher, their consultants, and the language

[2] ELDP is a component of the Hans Rausing Endangered Languages Project at SOAS and is currently one of the world's largest funders of endangered language research.

[3] For ELDP-funded examples, see www.hrelp.org/grants/projects.

community.[4] A more realistic view of documentation outcomes is that they are unique, situated, negotiated collections that depend on the specific people and processes that gave rise to them (see also Dobrin, Austin and Nathan 2009 section 5).

3. Data and archiving

The activities of documentary linguistics as described above suggest some degree of shared interest with corpus linguistics research. But the specific context of language endangerment limits such similarities. Although a corpus of a million words or more is recommended for analytical purposes in corpus linguistics, this cannot be attained for most endangered languages – in other words, for the majority of the world's languages. There are simply too many undocumented languages, and too few documenters. Languages situations inhibit the amount of data that can be collected, whether due to small numbers of speakers, a moribund state of the languages, or the conduct of documentation activities being limited by community sensitivities or their physical remoteness. Endangered languages are typically not written[5] so that there are few extant texts to collect and limited literacy traditions to draw on. Thus the content of documentations is likely to be local, particular, opportunistic, and uneven; quite the opposite of the large well-designed, balanced samples and hypothesis-driven nature of many corpus linguistics collections.

Archives are increasingly playing a role in documentary linguistics, providing not only preservation but several other services (Nathan 2008). Many language archives disseminate materials, functioning as specialist electronic libraries that are equipped to deal with the new genres of documentation. They also provide knowledge about changing technologies for recording, data management, and multimedia publishing. Ultimately, given the scale of language endangerment, language archives are likely to become the repositories of much of the world's linguistic and cultural heritage, and their holdings will provide the only possible basis for reviving many languages.

Current endangered languages archives have a range of emphases. Some are for local community use only, such as the archive of the Squamish Nation in Canada, some have regional coverage (e.g. Archive of the Indigenous Languages of Latin America, Paradisec) and others are international (DoBeS,

[4] Pedagogical effectiveness, however, is rarely taken into account – see Nathan and Fang 2009.

[5] See Csató and Nathan 2007 for a counterexample.

ELAR). Some are associated with a research institute (eg. the Aboriginal Studies Electronic Data Archive of the Australian Institute of Aboriginal and Torres Strait Islander Studies), while some are associated with documentation funding bodies (DoBeS, ELAR). Some archive only digital resources (e.g. DoBeS, ELAR), while others also hold paper and other 'legacy' materials (Alaska Native Language Centre). For further information about these and other archives, see the Appendix.

For most of these archives, limited funding means decisions have to be made about which materials to curate and preserve. For ELAR, which is mainly a repository for ELDP grantees, quality control is mainly achieved through a competitive applicatin process that leads to the successful award of funding. However this process has its own dynamic and may not be sustainable in the longer term; for example, among ELDP applicants there has been an escalation of the number of hours of audio and video recordings that many say they plan to make, presumably because they think it will better their chances of receiving a grant. However, many of the plans are totally unrealistic given the realities of the speakers, communities, and field situations. In the case of video, not only are documenters planning to overburden themselves (and their consultants), but it is now clear that some documenters are shooting poor quality video (poor both aesthetically and technically), and that the resulting large volumes of low-value data threaten to overwhelm our data storage resources in the medium term.

Fundamentally, archiving consists of managing relationships among providers, users, and the archive itself. For an endangered languages archive in particular, the relationship between the depositor and archive should not stop at the point of depositing, but should be ongoing, because such languages and the information about them are rapidly changing; for example, we encourage depositors to supplement or update deposited materials.

4. Audio and archiving

Fifteen years ago, while working in language education support in South Australia, I began to create interactive multimedia language learning materials. Looking for resources amongst fieldwork recordings, I was shocked by the typically poor quality of linguists' audio recordings. Eventually I realised that these fieldworkers were approaching recording from a different perspective from me. Recording was, to many of them, a 'side effect' of what they saw as the real task of transcribing and analysing languages. It was often approached with little skill and little thought about the nature of the recording being made. Their principal results were those written in their field notes and noted in their minds; only occasionally later would the audio tapes or cassettes be used to jog their memories, or to serve as 'proof' that they had actually done the field elicitations. Recording methodologies were unknown: many used cheap devices and their cheap built-in microphones, as often as not placed in random positions on tables, and frequently right next to the papers that linguists shuffle while doing their elicitation.

In one collaborative multimedia project, my colleague and I decided to include the text and audio for a narrative from an Australian Aboriginal language that had previously appeared as a text in her published grammar. She lent me the original reel-to-reel audiotapes that she had recorded in the 1970s so I could digitise the relevant segment to provide the audio component. However, no matter how hard I listened, I could not locate the stretch of audio that contained the story. Instead, I had to reconstruct it by editing together various fragments, repetitions, and rephrasings, which was, of course, just what she had previously done to create the published story text. In other words, her recordings were *evidence* of a story rather than a *performance* of it.

Cases like this show that audio played little part in the epistemology of linguistics (except for phonetics) before the arrival of documentary linguistics. The materials of linguistics – its data – were written materials, such as dictionaries, grammars, and texts. Audio was (where it played any part) mainly an inconvenience on the route to analysis. This view caused a tragic loss of much linguistic information that would be highly valued today; in Australia, some linguists were even instructed by their funders to reuse tapes (i.e. record over previous recordings), and to not ‘waste’ tapes by recording narratives and conversations![6]

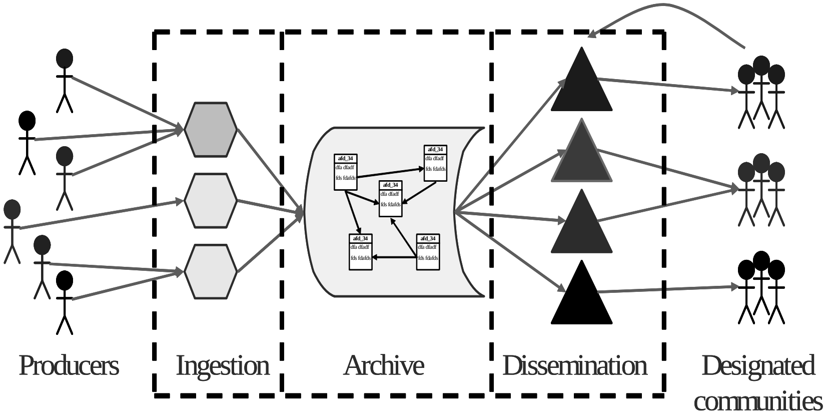
Subsequent developments have improved this situation. Documentary linguistics brought new activities and reprioritised existing ones; in particular, it emphasised the collection and curation of primary data, most often audio (but also including video). Since the events that are recorded are often unique, it became clear that they should be captured in as much detail and quality as possible, and that in turn the recordings must be properly archived for the long term. Newly established archives have increased field linguists’ access to technical expertise in recording and data management (many of these skills arguably should already have been part of the field linguists’ skill set, but at least the new developments have provided a means of addressing the deficits).

The influence of the broader digital archives environment has been positive, for example by emphasising the role of metadata and encouraging depositors to collect and manage it. Some archives, such as ELAR, are based on an architecture developed by the Open Archives Initiative (OAIS 2002), which provides a model extending beyond preservation to dissemination (see

[6] Personal communication, Luise Hercus.

Figure 1). It therefore defines audiences to be served ("designated communities"), and the kinds of formats and materials that each audience might need. By providing such centres for the discovery and dissemination of materials, today's archives are helping to fulfil Bird and Simons' accountability objective (Bird and Simons (2003:563), Thieberger 2004).

Figure 1: OAI model (Munro and Nathan 2005), after OAIS (2000)



In addition, ELAR at SOAS has emphasised mobilisation – the development of deposited materials into practical resources that can be used by language communities trying to combat the decline of their languages. The rationale for archive involvement in mobilisation is that preserving materials should not mean reducing the ability of communities to use them when they need or wish to do so. In addition, archives often have the relevant technical expertise to adapt electronic materials to diverse needs. We currently do this type of activity through training and collaborative multimedia development work, and plan to increase our contribution in this area in the future.

Documentary linguistics has also benefited from changes in media and information technology. The technology that has seen the greatest improvement in terms of increased quality at lower prices in recent years is audio recording equipment. Only five years ago, language documenters were using minidisc, DAT, cassettes or direct-to-CD; only a very few early adopters were using solid state devices. The situation has changed so thoroughly that in a documentation workshop held at the Tokyo University of Foreign Studies early in 2008, most participants arrived equipped with their own Edirol R-09 solid state recorder! The opportunities provided by new high quality, compact digital recording equipment, powerful but cheap computers and software, new sources of advice and training, and the popularisation of audio processing, mean that it is now reasonable to expect fieldworkers to create high quality recordings. However, the field has been slow to respond by gaining the appropriate skills for making recordings at the quality levels that are now possible (and appropriate for documentation goals). The field currently experiences a state that I call ‘Edison’s complaint’, which could be formulated like this: in 1878, the American inventor Thomas Edison gave the world his invention of the recording phonograph, and predicted that it could be used for “the preservation of languages”. Imagine his frustration, if he were alive today, to find that despite huge advances in audio apparatus available to linguists (as well as the added benefits of reduced size, weight and price), recording quality remains patchy and there have been no notable developments in genres for presentation and usage of audio.

5. Archivism

Documentary linguistics relies extensively on electronic technologies. Audio and video recording, data management, and many other activities including transcription, annotation and lexicography, are all performed using electronic devices and computers. Recordings and data must be digitally archived.

A technology focus has had important benefits, such as raising awareness about data management, especially ‘portability’ (Bird and Simons 2003) and its various components such as consistency, explicitness, use of standards, and care for primary data. The degree to which documenters can undertake data management methods that achieve portability will be a determining factor for the sustainability of digital language archives; most language archives have limited human resources for the conversion of incoming materials to archival formats. It is thus true that the outcomes of documentation and archiving depend on the ways in which documenters deploy technologies.

However, many documenters, rather than taking a holistic, artisan-like approach to the skills involved in meeting their linguistic and humanitarian goals, have come to believe that their methodologies are largely governed by a selection of technical desiderata such as audio resolutions and file formats. I use the term ‘archivism’ to describe such formulations of documentary linguistic practices that focus on such particular technological or quantitative criteria.

The substitution of awareness of technical parameters for deeper understandings of the art and science of audio recording is easily found in documentation literature and amongst accounts from documenters that I meet at training workshops, conferences and other events. For example, many have a basic awareness of audio file parameters and an abhorrence of compressed audio, but little or no knowledge of effective recording methods (especially about microphone types and handling, which are the greatest determiners of audio recording quality), acoustics, or managing noisy recording environments. One of our trainees even expressed the opinion that the cheapest two-dollar microphone was sufficient because he worked in a very noisy environment! A general result of these technically-focused formulations is that a narrow range of properties such as recording hours, data volume and file parameters have come to be seen as reference points for the 'quality' of documentations, or for meeting 'best practice' (Austin, Dobrin and Nathan 2009). It is not surprising that Dietrich Schüller, Director of the Vienna Phonogrammarchiv[7] has described linguists' audio recording methodology as some of the least scientific practice of all disciplines.[8]

6. Symbolic information

Audio materials are generally accompanied by some associated symbolic information. In music publishing, this symbolic material consists of song title, artists' names, publisher, and perhaps lyrics and other information. In documentary linguistics, it typically consists of metadata together with content-related or time-related material such as a time-aligned transcription or translations and annotations. While metadata, as generally understood, can be distinguished from transcriptions due to its primary use in cataloguing, all symbolic information associated with language recordings can be considered to be metadata (Nathan and Austin 2005). In practice, metadata means different things to different people. To linguists, the term 'metadata' is rather like a reminder to collect and manage contextual information about an event such as details about speakers, settings, equipment, rights, and permissions. Given documentation's emphasis on primary data for a range of communicative events, recording metadata might be thought to have priority over transcription, which can potentially be made later once the researcher's knowledge of the language increases, and which can continue to be worked on and refined[9]. However, in practice, making transcriptions is part of the documenter's language-learning process in the field, and, in addition, documenters increasingly transcribe in collaboration with speakers (and/or

[7] See http://www.pha.oeaw.ac.at/home_e.htm.

[8] At an ELAR Workshop on 'Audio Recording, Digitisation and Archiving' held at SOAS, 13th February 2006.

[9] The same is true of annotation and translation – see Woodbury 2007 concerning the "ongoing, contingent, interpretive, hermeneutical quality of the documentation of meaning" and his description of the cycle of refinement in the translation of one Cup'ik text he and others worked on.

train community members to transcribe). As a result, the anticipated order is reversed: transcribing tends to take place in the field setting and metadata creation is (unfortunately) often left until later.

For the archivist, symbolic information is crucial for the operation of the archive. Without symbolic data, custodians and users of digital media are plunged back into some kind of dark age, equivalent to the time before books were invented, when the only way to access information was to experience events in real time and hope to hear something useful! If the documenter never creates or provides sufficient metadata or transcription, the resource is left in the dark, barely findable and unusable, forever (or until someone else provides the symbolic information). Ideally, the richness of symbolic information should be proportionate to the potential value of the materials to users and to the high costs of digital storage. See section 8 for further information about metadata.

A disciplinary area that has a particular interest in symbolic endangered languages data is linguistic typology, where the focus is on large datasets from a variety of languages. The value of such data for typologists is greatest where they are classified using standard codes (e.g. for language names or morphological glossing) to make statistical comparisons easier. Typologists have strongly urged documenters to develop and apply standard ontologies for coding linguistic phenomena (such as the GOLD ontology developed by the EMELD project). Although standards can provide a foundation for good practice, while thousands of languages remain undescribed it is premature to propose or prescribe standard ontologies. Human languages and the people who venture to describe them are so diverse and eccentric that flexibility, creativity and uncertainty need to be features of the documenter’s representational apparatus.

7. Representation and protocol

At the Endangered Languages Archive (ELAR), we use the term ‘protocol’ as a shorthand for the concepts and processes that apply to the respect and implementation of language speakers’ rights and sensitivities. Protocol has long been part of corpus linguistics methodology; for example, recorded subjects are asked whether their identity can be revealed and measures such as anonymisation are undertaken. For endangered languages, protocol issues are heightened. Endangered language communities are typically under social pressures, and vulnerable, so we have to enhance the ways we deal with sensitivities and implement protocol in audio access and distribution. Protocol involves more than seeking permissions and applying anonymisation. In small communities it is almost impossible to be anonymous; many within the community know each other very well, so even the briefest remark can reveal someone's identity. This is exacerbated by the priorities of documentation: the most valuable recordings are those of casual conversation, which are most likely to be peppered with personal comments. Even though such materials might be effectively anonymised to outsiders, if they are used within the community to support local language goals, they can have unintended consequences.

People whose voices have been recorded may express sensitivities and restrictions of various kinds – political, religious, personal – or pertaining to ownership by themselves or some wider group. Therefore it is important that fieldworkers elicit and record protocol information and convey it along with the documentation, including to the archive.

The coding of protocol information needs to be flexible and detailed enough to capture what is important to speakers, but at the same time formalised enough to be able to be effectively implemented by the archive. At ELAR we researched and developed a protocol grid which has worked well so far.[10] Soon we will support the implementation of restrictions not only at the deposit level (i.e. to all items in a deposit) but to individual files and even parts of files. This is important because it would be against the spirit of our work if depositors have to, for example, deny access to a one hour audio recording because within it there are one or two minutes of sensitive material. We have yet to implement the full range of protocol processes I have described here, but plan to do so over the next few months.

Protocol information is not immutable: it changes over time. Language endangerment is inevitably connected with communities under stress, and sensitivities and permissions change from time to time, depending on social and cultural factors. For example, name taboos following death apply in many Australian Aboriginal communities, so that names should be suppressed for an appropriate period following a death, and then restored after sufficient time has passed. ELAR is thus building a web-based system for depositors to manage their protocol and other metadata.

It is worth noting that on the positive side, there are real advantages to the fact that audio (and video) can, unlike written data, directly represent individuals in an unmediated way. The ability to present direct voices and identities of speakers to end-users is a valuable aspect of multimedia language learning resources (Nathan 2006).

[10] See www.hrelp.org/archive/depositors/depositform.

8. ELAR

The Endangered Languages Archive (ELAR) at SOAS provides digital archiving and associated services for ELDP grantees and others working with endangered languages. We are focused on digital preservation and providing local facilities, but dissemination of materials is also a priority. Currently, we are working on an innovative online dissemination system which will be operational by the end of 2009. In addition, we also participate in various mobilisation projects to help create usable language materials for communities.

We are increasingly involved in delivering documentation training to various groups, inclduing ELDP grantees, students from the Endangered Languages Academic Programme (ELAP), and at international documentation training workshops held in the UK, France, Ghana, and Japan. ELAR partners with ELAP staff and students in many activities, and also participates in various international collaborations including in the DELAMAN network, an umbrella body for archives engaged with research on endangered languages and cultures worldwide (see Appendix).

ELAR currently holds about 50 deposits with a total volume of approximately 4 TB. The average deposit is about 80 GB. However, sizes vary widely, with a small number of very large deposits; the median size is around 10GB. We expect the total volume to nearly double over the next year as more funded projects are completed. Table 1 illustrates some data types of interest for a small but representative sample of holdings:[11]

[11] Since this analysis was done, new deposits have included a very large video corpus of Australian Sign Language (AUSLAN) documentation that has increased the proportion of video materials.

Table 1: Data types by number of files and volume (representative sample, about 60% of the ELAR collection as at February 2008)

Data type	Files	Volume (MB)
audio	6,312	360,411
image	2,221	28,592
video	895	208,995
text	781	32
msword	404	223
trs	246	5
eaf	176	33
pdf	134	196
lex	29	9
imdi	26	1
xls	19	1

For its metadata, ELAR has taken a 'middle path' approach. We have provisionally defined the archive's metadata as a set of about 40 elements, which are more comprehensive than the OLAC set (which slightly extends Dublin Core's 15 elements)[12] but less numerous than the approximately 70-element IMDI set created for language documentation by the Max Planck Institute for Psycholinguistics, Nijmegen.[13]

On the other hand, we also hold depositors' metadata in a variety of formats. In the early days of ELAR's development, it was decided that because language documentation is an emerging rather than a mature field, it would be fruitful to observe what happens when documenters are encouraged to produce metadata that caters to their own research environments and needs.

[12] See http://www.language-archives.org/OLAC/metadata.htm.

[13] See http://www.mpi.nl/IMDI/. Details of the ELAR set will become available on our website http://www.hrelp.org/archive.

As a result, from a survey of approximately 40 early data deposits, we can now state that:

- *each documentation project can have its own unique 'recipe' for metadata, depending on factors ranging from the language's typology to preferences of researchers and consultants, to community values*
- *each language documenter has their own skills and priorities that influence what metadata they wish to encode and how they can best encode it*
- *since our goal is to maximise the quality and quantity of metadata for each deposit in its own terms, it is wise to support diversity.*[14]

9. Conclusion

As documentary linguistics has developed over the last ten years, it has benefited from the knowledge and experience of other disciplines. Perhaps documentation has now gathered enough experience to be able to offer useful advice to others. This survey of audio and archiving issues in documentation has attempted to identify issues which most spoken corpora will face, especially those concerned with endangered languages materials. Whatever might be around the corner, we may discover together.

[14] Of course this also imposes costs. To attain robust and portable formats for preservation (Bird and Simons 2003), we will need to convert and migrate various document formats. For example, some documenters find that Excel spreadsheets provide the right balance between their skills and their representational needs; these documents will need to be converted to marked up plain text for preservation.

References

Austin, Peter K. 2006. Data and language documentation. In Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.) *Essentials of Language Documentation*, 87-112. Berlin: Mouton de Gruyter.

Austin, Peter K. and Lenore A. Grenoble. 2007. Current trends in language documentation. In Peter K. Austin (ed.) *Language Description and Documentation*, Vol. 4, 12-25. London: SOAS.

Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79: 557-582.

Csató. Éva Á. and David Nathan. 2007. Multiliteracy, past and present, in the Karaim communities. In Peter K. Austin (ed.) *Language Documentation and Description*, Vol. 4, 207-230. London: SOAS.

Dobrin, Lise, Peter K. Austin and David Nathan. 2009. Dying to be counted: commodification of endangered languages in documentary linguistics. This volume.

Gippert, Jost; Himmelmann, Nikolaus and Mosel, Ulrike (eds.) 2006. *Essentials of language documentation* (*Trends in Linguistics. Studies and Monographs*, 178). Berlin: Mouton de Gruyter

Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In Austin, Peter (ed.) *Language Documentation and Description*, Vol. 1, 52-71. London: SOAS.

Himmelmann, Nikolaus. 1998. Documentary and descriptive Linguistics. *Linguistics* 36: 161-95.

Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.), *Essentials of language documentation*, 1-30. Berlin: Mouton de Gruyter.

Johnson, Heidi and Arienne Dwyer. 2002. Customizing the IMDI metadata schema for endangered languages. Proceedings of LREC 2002.

Krauss, Michael. 1992. The world’s languages in crisis. *Language*. 68:4-10.

Munro, Robert and David Nathan. 2005. Introducing the ELAR information system architecture. The Third meeting of the Digital Endangered Languages and Music Archive Network (DELAMAN III), Austin. Online at http://www.robertmunro.com/research/munro05elar.pdf (accessed 19 April 2008).

Nathan, David. 2006. Thick interfaces: Mobilizing language documentation with multimedia. Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.) *Essentials of Language Documentation*, 363-379. Berlin: Mouton de Gruyter.

Nathan, David. 2008. Digital archives: essential elements in the workflow for endangered languages documentation. In Peter K. Austin (ed.) *Language documentation and description,* Vol. 5, 103-119. London: SOAS.

Nathan, David and Fang Meili. 2009. Language documentation and pedagogy: a mutual revitalisation. This volume.

Nathan, David and Peter K. Austin. 2005. Reconceiving metadata: language documentation though thick and thin. In Peter K. Austin (ed.) *Language Description and Documentation*, Vol. 2, 179-187. London: SOAS.

OAIS 2002. Consultative Committee for Space Data Systems (CCSDS). CCSDS 650.0-B-1. Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. Published on line at: http://public.ccsds.org/publications/archive/650x0b1.pdf. (accessed 19 April 2008)

Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Peter K. Austin (ed.). *Language documentation and description,* Vol. 2, 169-178. London: SOAS.

Wittenburg, Peter, Ulrike Mosel and Arienne Dwyer. 2002. Methods of language documentation in the DOBES project. Proceedings of the international LREC workshop on resources and tools in field linguistics, Las Palmas 26/27 May 2002.

Appendix: Listing of some endangered languages archives

Aboriginal Studies Electronic Data Archive, Australian Institute of Aboriginal and Torres Strait Islander Studies. http://www1.aiatsis.gov.au/ASEDA/

Alaskan Native Language Center Archives (ANLC) University of Alaska. http://www.alaska.edu/uaf/anlc/

Archive of the Indigenous Languages of Latin America (AILLA), University of Texas. http://www.ailla.utexas.org/site/welcome.html

Digital Endangered Languages and Musics Archives Network (DELAMAN). http://www.delaman.org/

Dokumentation Bedrohter Sprachen Archive (DoBeS), Max Planck Institute Nijmegen. http://www.mpi.nl/DOBES

Endangered Languages Archive (ELAR), School of Oriental and African Studies. http://www.hrelp.org

Langues et Civilisation et Traditions Orale (LACITO), Centre National de la Recherche Scientifique. http://lacito.vjf.cnrs.fr/archivage/index.htm

Leipzig Endangered Languages Archive (LELA), Max Planck Institute Leipzig. http://www.eva.mpg.de/lingua/resources/lela.php

Pacific and Regional Archive for Digital Sources in Endangered Cultures (Paradisec), University of Melbourne/University of Sydney. http://paradisec.org.au/

Rosetta Project, Long Now Foundation. http://www.rosettaproject.org/