Language documentation and archiving: a work in progress[1]
Workshop introduction

DAVID NATHAN
School of Oriental and African Studies

1. A FICKLE RELATIONSHIP

As Woodbury (2011:163) points out, language documentation has been practised in a recognisable (and still very influential) form for well over a century. However, while documents, field notes and recordings from Boas, Sapir and other early documenters have been preserved (Johnson 2004:140), it is only recently that archiving has become a distinguishing mark of documentation.

For those involved with endangered languages today, whether of a theoretical or applied orientation, the terms 'language documentation' and 'archiving' slip off the tongue together as if they have always been connected. But they have been systematically linked only since the late 1990s, when Nikolaus Himmelmann, in his seminal paper for documentary linguistics (Himmelmann 1998:168), stated that:

> Language Documentation … is concerned with compiling, commenting on, and *archiving* language documents (emphasis added DN).

and foresaw many of the issues that continue to occupy us (Himmelmann 1998:191):

> technical problems … such as the choice of an appropriate recording and presentation technology (sound recording, video, multi-media applications, etc.), the problem of archiving and maintaining documentations, and the problem of providing and controlling access to documentations

In their influential paper Bird and Simons (2003) described the same issues in terms of 'portability', the sustainability of digital documentation across different computing environments and over time.

The pairing of documentation and archiving also appears in several other contexts, including ethics, access and training. United via ethics we find, for example, Dwyer (2006:40) emphasising that '… properly archiving collected data is far more respectful to a speaker community than piling it in the back of a closet'. Dwyer (2006:35) also identifies archiving as a 'phase' of documentation that carries forward and fulfils language speakers' preferences:

> '[d]uring the archiving phase, the researcher must carry though the wishes of the consultants in terms of anonymity and recognition … [and] on user access to the materials (community, scientific researchers, general public)'

In a recent chapter on 'Archiving and language documentation', Conathan (2011) interweaves documentation and archiving through considering access and intellectual property issues, where documentation and archiving intrude on and affect each other's practices. Nathan (to appear), drawing on an analogy with libraries, describes archivists and depositors as 'joint librarians' for endangered languages materials, where depositors play the major role because they are the ones who

[1] I am grateful to all the authors represented in these Proceedings for their stimulating ideas, and to Peter K. Austin for additional comments and encouragement.

understand the materials' context. Archiving and documentation regularly appear as a duo in the curriculum of training courses such as those run by DoBeS, HRELP and InField.

This *prima facie* relationship between documentation and archiving has not, however, received anything like the same scrutiny as that between documentation and *description*. The latter pair have been theorised by Himmelmann and discussed and debated by many since in a large number of conferences and publications (cf. Austin and Grenoble 2007).

It is surprising, therefore, that this present Workshop appears to be the first fully fledged occasion that is symmetrically targeted at *both* language documentation and archiving. We hope that this workshop challenges the directions of our disciplines. The papers assembled here offer a glimpse into the future, not only of endangered languages archiving and language documentation, but perhaps even the survival of particular languages.

2. THE PAPERS

The papers in this volume are grouped into three thematic sections: *New methods for creating and structuring archive content*, *Enhancing archive usage and effectiveness*, and *New models for archiving*.

New methods for creating and structuring archive content includes three contributions which focus on more formal and computational-related aspects of our field. Ulrike Gut's paper highlights the relationship between documentation and corpus linguistics. While many documentary linguists have used the terms 'corpus' and 'documentation' more or less interchangeably[2], Gut locates documentation within corpus creation through the methods and capabilities of the 'Pacx' set of software tools. While corpus approaches tend to be more formalist and aimed at academic practitioners (Cox forthcoming), Gut shows that Pacx software also meets some of the more community-oriented criteria highlighted by later authors. For example, while Pacx remains within the 'language resource' framework that prioritises structured morphological annotation, it also allows data to be decentralised and ongoing, 'allowing continuous additions and changes'.

Jeremy Nordmoe's paper continues the software theme, but focuses specifically on metadata, tackling the problem that metadata is often not provided together with documentary materials, or is formulated out of the documentary context and therefore provides a lesser record. Nordmoe takes a pragmatic approach to deal with the problem he calls 'so much metadata, so little time', and describes SIL's Language and Culture Archives' RAMP software which aims to make metadata entry easier. Strategies include removing 'roadblocks' identified though user surveys, separating the composition of metadata from its upload to the central SIL repository, and streamlining the set of metadata categories and the process for entering them.

Sebastian Nordhoff and Harald Hammarström are interested less in software than in the underlying logical structures of information, in particular the information found in grammars. Their chief interest is in reinterpreting grammars firstly as

[2] Although note the distinctive DoBeS/IMDI 'schema of metadata elements … specifically directed towards describing multi-modal … and written language corpora'; IMDI 2003.

‘grammatical descriptions’ and then as ‘granular annotations’, in order to widen the applications and usages of grammatical information, especially via the semantic web, where information can be searched and processed in terms of its logical and ontological structures. After surveying the structure of typical printed grammars, and dispensing with their printed-page-only properties, they propose a 21st century approach to digital grammar-writing as ‘nonlinear database[s] of micropublications’.

The second set of papers *Enhancing archive usage and effectiveness* begins with Paul Trilsbeek and Alexander König’s concern that our digital endangered languages archives are under-used. They identify the main audience categories as the academic and speaker communities, and turn to the latter as new, so far untapped, providers of language documentation. While many communities are keen to participate in the new media landscape characterised by YouTube, the authors examine some of the problems that archives would face in handling community-sourced uploads to public portals, whether to sites such as YouTube or to extensions of archives such as DoBeS. In particular, they doubt that other sectors of the audience would be confident of the provenance, veracity and ethical conduct associated with such resources ‘contributed by unknown depositors’. The authors hint at an interesting reversal in the properties of community-resourced versus researcher-sourced materials: while a YouTube-fired zest for public exposure could give rise to masses of freely available but less accountable material from community members, many academic researchers tend to over-apply access restrictions on their (presumably) more accountable materials. The authors therefore call for a greater willingness among researchers to share access to their materials, while respecting source community wishes.

Joshua Wilbur presents a unique, semi-biographical account of his interactions with various archives in Saami country where he conducted his documentary fieldwork. With the explicit goal of making his documentary materials more accessible to Saami people, he negotiated with three different archives, and presents here his experiences and observations. He finds that smaller, local archives, although having the greatest potential to reach community members, have very particularised (in some cases, limited) capacities, resources, skills, policies and preferences. These mean that in order to reach local communities through such archives, the documentary linguist may need to invest considerable time and effort not only in negotiating with archive management and technical staff, but also in acting as a technical consultant to them. Wilbur also makes an important distinction between ‘discovery’ and ‘promotion’. Discoverability is the ability of potential users to identify a relevant resource, typically through metadata-based search; it is the oft-stated rationale for certain types of (standardised) metadata schemes and data aggregating portals (Bird and Simons 2001). However, Wilbur finds that ‘usage of materials … is not guaranteed by their mere presence in an archive’, regardless of metadata, and that archives need to ‘actively promote the language materials they have been … trusted with’.

The final section *New models for archiving* includes contributions by Mary Linn, Edward Garrett, and the plenary paper by Tony Woodbury.

With the goal of proposing CBLA (Community Based Language Archiving), Mary Linn draws interesting parallels between models for participatory linguistic fieldwork and new models for archiving. She raises a number of innovative, indeed challenging, ideas such as ‘decentralised curation … [where] there is no need for an archivist at all’. This might resonate with some archives who delegate much of the

curation process to depositors through issuance of guidelines and software that governs structures and formats. However, Linn takes curation to go far beyond checking formal properties of data. She considers 'radical user orientation', where the archivists' primary curatorial task, namely contextualisation, is centred on the context of *the users themselves*, because it is they (especially as community members welcomed into the archive ecology) who ultimately determine the success of archives in meeting their goals. Such departures from classical archiving approaches form the basis of Linn's proposal for CBLA, in which the language community is involved in every step, from documentation planning to curating to dissemination. Linn provides a case study showing how such approaches have had a positive social impact on communities and revitalisation through increased archive usage and resultant language activities.

Edward Garrett's software[3] will demonstrate innovative ways of including language community members in the documentation and archiving process. He proposes decentralised, web-based functions that allow language speakers to interact with archives' existing resources. They can add further materials, comments, or contextualisation. They can identify themselves or their relatives in order to claim their moral rights[4] in recordings and other materials. And finally they can make corrections to erroneous data, interpretations, and attributions. With this suite of functions, which he characterises as 'collect and correct', Garrett offers a view of how some of Linn's proposals might work in practice, and addresses some of Trilsbeek and König's questions about how crowdsourced materials might be included in archives' collections. Garrett sees crowdsourcing as a way for language speakers to establish real links with resources, rather than being merely 'participant metadata'. Garrett's concrete proposals expose the weakness of calls for including 'communities'. Because services are supplied to individual persons, Garrett takes care to recast the problem as providing software and infrastructure for 'language speakers'. Garrett will also demonstrate his 'speech bubble annotator', a new way of presenting transcribed video resources. It not only presents documentation more accessibly, but also continues his emphasis on the representation and acknowledgement of language speakers as individuals.

The plenary paper by Tony Woodbury, like those by Linn, and Trilsbeek and König, pays considerable attention to archives' audiences. Woodbury also shares Trilsbeek and König's concern about the apparent under-usage of archives by their potential audiences. However, Woodbury takes a slightly different turn, being more interested in thinking about the *nature* of 'audience' as providing desiderata for what ought to count as 'good' documentations, or, as his title puts it, documentations that people 'understand and admire'. Paradoxically, he starts out considering paper documents and paper archives, to remind us, perhaps, of the dictum that in using new technologies we should not forget what previous technologies did very well. The body of his plenary consists of a timely and much-needed set of proposals for rethinking the genres, content, and arrangement of language documentation. Bringing the discussion full-circle back to audiences, he suggests that audiences can most fruitfully be 'critics' of documentation, for example as reviewers, thus reinforcing calls for journal reviews of documentation to be added to the ecology of language documentation. This could be taken as another reminder not to readily abandon familiar and effective genres. Peer reviews, as part of an evaluation and feedback loop, are an indispensable

[3] Garrett did not submit a paper to these Proceedings but will demonstrate software at the workshop.
[4] See http://www.ipo.gov.uk/types/copy/c-otherprotect/c-moralrights.htm [accessed 2011-10-27].

component of the scientific process and the evolution of ideas (see, e.g., Allen et al. 2009).

Woodbury has also quietly exposed the issue of *archivists'* contributions to the presentation of materials, a question that most endangered languages archives have barely grappled with. Archivists' contributions relate not only to contextualisation of materials but also to the software and design issues involved in creating screen interfaces that appropriately delineate whose 'voice' the audience is reading/hearing. Although the production and presentation of contextualising matter and finding-aids and the preparation of exhibition materials are standard fare for archivists in 'traditional' archives, in our field they have been slowly re-delegated to documenters (or left un-done). This may perhaps be due, as Linn describes, to most present-day endangered languages archivists having their backgrounds in linguistics rather than archiving (or museum studies, or related fields).

3. DISCUSSION

A number of issues recur in this set of papers and I have selected some of the more innovative and challenging ones for further discussion.

Community curation

Several authors write about what could be called, following Christen (2011), 'community curation'. The new sharing and participatory practices and environments proposed by Linn, Garrett, and Woodbury (and appearing to some degree in other papers too) represent paradigm-changing challenges. Linn countenances archivists (in the conventional sense) being dispensed with altogether, while Garrett and Woodbury see language speaker audiences as correctors and critics. We are presented with a radical inversion: the archive concept of 'context' is no longer that of the materials, or their (supposed) provenance, but of the *users*. And these users are principally the language speakers, who are also, in a virtuous circle, central participants in the documentation and archiving processes.

Not forgetting research communities, Woodbury and Nordhoff and Hammarström make proposals about models and structures of information and documents, opening up existing documentation to the possibility of contributions such as annotations being attached by researchers other than the original creator(s), and bringing together disparate sources of related information through the semantic web.

Synthesising these types of participation will pose a challenge. Our tendency so far has been to polarise them. Trilsbeek and König highlight the contrast made by existing archives between their depositors, who are accountable and 'internal', and community-sourced and crowd-sourced materials that are fraught with unknowns. Wilbur injects a rarely considered class of participants: small, local and unique archives, who may be the best option for reaching communities, after all.

Promotion

Wilbur's paper introduces the idea that archives need to do more than acquire, curate, preserve and disseminate materials. To reach target audiences for language revitalisation, archives also need to actively *promote* the materials they hold. This suggests that archives need to develop relationships with their audiences that are not based purely on access to language materials, for the success of archive outreach may

depend on first developing contact, relationships and trust in order to encourage usage or other participation with the materials. And quite independently of their dissemination function, the promotion of language materials in the public sphere by archives helps 'valorize' and thus sustain the very languages represented in collections. Woodbury also advocates promotion, for example through exhibitions and reviews, wishing that AILLA's compelling materials 'could interest and intrigue many more people'.

Contextualisation

Contextualisation of materials is at the heart of archiving, but in many of our contemporary archives, the art of contextualisation has given way to the science of software development. However, communities may wish to play a role in framing the interpretation of *their* materials to *others* (cf. Christen 2011:197). Similarly, community access to materials is not reducible to file transfer, but in reality entails access to *meaning* (Christen 2011:194; Nathan, to appear). Linn's CBLA proposal, therefore, is not about negotiation of metadata, or even about promotion. It is baldly about sharing or handing over management of the archives themselves, since 'when communities and families know what's in archives and how they work, the collections get used more.'

The form of documentation

Despite extensive theorisation of documentation in previous work, there has been little discussion of the *form* of documentation: its granularity, structure, organisation, links, and how it is to be navigated. Several papers here do address the issue. Nordhoff and Hammarström give a detailed alternative view of the shape of grammars for a networked age. Woodbury's paper describes 'a Noah's Archive, a one-time sampling of the uses of a language for a grammar, dictionary, or thumbnail linguistic ethnography', with detailed proposals for content. He also describes materials by Knut Bergsland from 1959 which, although printed, are essentially models in hypertext and text retrieval techniques that today's digital environment can easily deliver but language documentation has not yet conceived for itself.

Publishing

A corollary of Woodbury's many proposals, including increased archivist contributions, attention to genre, exhibitions, promotion, and reviews of documentations, is that what archives do is expressed better as 'publishing' rather than 'dissemination'. The idea that 'archiving is a form of publishing' may have appeared first in Johnson 2005 (see also Nathan 2011).

General observations

More generally, there are interesting tensions and divergences which hint at future 'forking points'. For example, some authors advocate a corpus-based framework with its emphasis on a 'representative sample' for 'scientific study', while Linn's CBLA model and Garrett's proposals evoke a more participatory-organic-evolutionary framework (for a close examination of the resonances and dissonances between corpus and documentary approaches, see Cox, forthcoming). Similarly, some believe that further codification and implementation of standards is a key to greater sharing, while others are wary of standards (see also Christie 2005). Finally, there is the question: are language documentation and archiving a cross-disciplinary affair, or two

parts of the same discipline? The papers here seem to range along this one-or-two disciplines axis, so the question remains open, at least for now.

4. CONCLUSION

All the papers here share the goal of exploring how the linguistic and archive communities can provide effective and ethical responses to language endangerment. We all wish to raise the quality of documentary materials, and the effectiveness of our technologies. The key seems to be recognising that neither of these is meaningfully measurable without considering audiences and usage. Engagement with language speakers will be necessary for progress to be made. Here is an example: current web trends are towards mash-up pages, mobile 'apps', and aggregating portals.[5] These gather resources based on a particular user's preferences, and display them according to topic, geographic location, or language. But what happens when the user wants to view information connected to a specific person, say language speaker X? Unless speaker X is truly 'part of the system', as a member, owner, or curator, rather than a mere meta-data-point, then such a speaker-centred page will be an incomplete and insipid representation, with distorted access because nobody except speaker X can properly decide who he/she wants to share with. We are fortunate, therefore, to see the maturation[6] and continued rise of online social networking and innovative 'apps' which personalise individuals' interactions with a multitude of resource providers and provide further exemplars for implementing the participatory models suggested by several authors here. Whether dedicated 'language-resource' platforms are going to be effective is yet to be confirmed; when it comes to matters of rights, communications and sociality it is likely that well-designed systems that work for everyone will be the best ones for language speakers too.

One thing seems clear: for too long we have proffered (and accepted) glib statements about the advantages of the internet, that it solves our problems by reaching everyone. They might have struck a chord in 1996 but today they are digital prehistory. Thomas Friedman, celebrated journalist and author, recently observed that less than 7 years ago 'Facebook didn't exist, Twitter was a sound, and Skype … was a typo.'[7] All of these are, of course, platforms for social interaction. Our future successes will be about communities, not the internet.

As Woodbury says, there is 'much, much to be done'. Ours is still a work in progress, perhaps barely begun.

---

[5] These will be even more useful for endangered languages community members with the increasing use of mobile devices.

[6] It is only in 2011 that dynamics for online privacy and sharing have become deeply and widely debated, and as a result are evolving to become more nuanced and conventionalised. The release of Google+ (Google's platform for social networking), as the first real competitor to Facebook, has precipitated the first true 'battle for ideas' on the territory of ethical practices for sharing and privacy.

[7] Thomas Friedman, 'What went wrong with America?' Highlights of an address given at the Melbourne Town Hall on 29 July 2011. Big Ideas, ABC Radio National, 8 September 2011.

REFERENCES

Allen L, C. Jones, K. Dolby, D. Lynn & M. Walport. 2009. Looking for Landmarks: The Role of Expert Review and Bibliometric Analysis in Evaluating Scientific Publication Outputs. PLoS ONE 4(6): e5910. doi:10.1371/journal.pone.0005910

Austin, Peter K. and Sallabank, Julia (eds). 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

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

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

Bird, Steven and Gary Simons. 2001. The OLAC Metadata Set and Controlled Vocabularies. *Proceedings of the ACL Workshop on Sharing Tools and Resources for Research and Education.* Toulouse, France, 7-18.

Christen, Kimberly. 2011. Opening archives: Respectful Repatriation. In *American Archivist*, Vol 74(1):185-210.

Christie, Michael. 2005. Aboriginal Knowledge Traditions in Digital Environments. Unpublished ms. Charles Darwin University. Available at http://www.cdu.edu.au/centres/ik/pdf/CHRISTIE_AJIEpaper.pdf.

Conathan, Lisa. 2011. Archiving and language documentation. In Peter K. Austin & Julia Sallabank (eds), *The Cambridge Handbook of Endangered Languages*, 235-254. Cambridge: Cambridge University Press.

Cox, Christopher. forthcoming. Corpus linguistics and language documentation: challenges for collaboration. In John Newman, Harald Baayen & Sally Rice (eds) *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Rodopi: Amsterdam/New York. Available at www.ualberta.ca/~aacl2009/PDFs/Cox2009AACL.pdf.

Dwyer, Arienne. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Jost, Nikolaus Himmelmann and Ulrike Mosel (eds) *Essentials of language documentation*, 31-66. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus. 1998. Documentary and Descriptive Linguistics. *Linguistics* 36, 161-195.

IMDI (ISLE Metadata Initiative). 2003. Metadata Elements for Session Descriptions. Version 3.0.4. Available at http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed) *Language Documentation and Description, Volume* 2, 140-153. London: SOAS.

Johnson, Heidi. 2005. Corpus Management 101: Creating archive-ready language documentation. Talk given at Linguistic Society of America annual meeting. Available at www.ailla.utexas.org/site/docs/corpus_mgmt.ppt.

Nathan, David. to appear. Access and accessibility at ELAR, a social networking archive for endangered languages documentation. In *Oral Tradition,* Special Edition on Archiving orality and connecting with communities. http://journal.oraltradition.org/.

Nathan, David. 2011. Archives as publishers of language documentation: experiences from ELAR. Presentation at Second International Conference on Language Documentation and Conservation, University of Hawaii, February 12, 2011. Available at http://hdl.handle.net/10125/5223.

Nathan, David 2006. Thick interfaces: mobilising language documentation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds) *Essentials of Language* Documentation, 363-379. Berlin: Mouton de Gruyter.
Woodbury 2011. Language Documentation. In Peter K. Austin and Julia Sallabank (eds) *The Cambridge Handbook of Endangered Languages*, 159-211. Cambridge: Cambridge University Press