ABSTRACTS

Research, records and responsibility (RRR): Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

(Listed by first author surname)

Alexandre Arkhipov *XML databases and XSL transforms in language documentation workflows*

More and more programs use XML formats for internal data storage, not only for interchange. This includes both general-purpose tools like MS Office and OpenOffice/LibreOffice and specialized linguistic software such as ELAN, EXMARaLDA, FLEx, Speech Analyzer, Arbil, WeSay, SayMore and so on. Thus more and more linguistic data are being created in XML, not just convertible to XML.

Although not ideal (verbosity, high processing time), XML formats have a number of benefits to boost workflow efficiency. Importantly, XML documents can be processed with XSL transforms to get new data, remaining still in the realm of XML (the XSL transforms themselves are also XML and can be transformed by other XSL...), displayed as HTML or published into PDF. Finally, there are now mature free native-XML databases like eXist-db and BaseX which offer the full cycle of operations in one application with browser-based interface: store existing documents, browse and query data, create and edit data online, apply XSLT to publish.

I will illustrate this with examples of transformations we used in language documentation workflow to convert interlinear texts in Archi (East Caucasian) between various formats including OpenOffice and FLEx. A connected issue which will be addressed is the need for an interchange standard format for interlinear texts.

Linda Barwick and Amanda Harris *PARADISEC*

Founded in 2003 by a team of linguists and musicologists at the University of Sydney, Australian National University and the University of Melbourne, the Pacific and Regional Archive for Digital Sources in Endangered Cultures is now celebrating its ten-year anniversary. In the first decade of operation, PARADISEC created an archive of more than 3000 hours of audio recordings of language and music recordings from the Pacific, Asia and beyond. The digital collection now contains over 6 TB of data with a queue of audio, video and image files still to be ingested into the archive, and PARADISEC was recently included in the UNESCO Memory of the World Register. In this

presentation, we will trace the development of PARADISEC from its origins as a storage facility for precious research data. We will discuss the ways that uses of the archive have grown and changed over the years, and reflect on prospects for PARADISEC's future in a fragile funding environment.

Andrea Berez *Integrating archiving into the language documentation postgraduate curriculum at the University of Hawai'i at Manoa*

The language documentation and conservation (LD&C) track in the Master of Arts program in the Department of Linguistics at the University of Hawai'i at Manoa (UHM) is unique in the United States and one of only a small handful of graduate programs in the world offering advanced degrees in the documentation of endangered languages. A key component of the program is the integration of the Kaipuleohone University of Hawai'i Digital Ethnographic Archive into the curriculum.1 This paper discusses the development of Kaipuleohone and its increasing role in the professional development of students in the LD&C program at UHM. Kaipuleohone was started in 2008 by Nick Thieberger (Albarillo & Thieberger 2009). The original mission of the archive was to provide a permanent secure home for digitized language recordings from scholars affiliated with UHM over the five decades since the Department of Linguistics was created. During the first phase of Kaipuleohone, hundreds of recordings from eminent field linguists like Derek Bickerton and Robert Blust, as well as the collection of the Charlene Sato Center for Pidgin, Creole and Dialect Studies, were digitized and ingested. Now in its second phase, Kaipuleohone has increasingly become an archive for materials actively being collected, especially by students in the LD&C program. The core curriculum stresses the importance of archiving in the language documentation workflow, and Kaipuleohone provides an opportunity for students to develop good habits of consistent metadata collection and regular deposit, even from the field. Students are required to consider issues surrounding data longevity, access, and multipurpose value early in their careers, better preparing them for achieving the best practices of contemporary language documentation as professionals. In addition, good archiving practices among our students allows us to require the proper citation of documentary source materials in doctoral theses via permanent handles, furthering the scientific goal that linguistic claims be verifiable by data, and thus increasing the quality of scholarship in the Department. Kaipuleohone conforms to international archiving standards for digital archives. Audio files are stored at high resolution and the metadata conforms to the Open Language Archives Community, Open Archives Initiative and Dublin Core. All digital files are curated by the Library system at the University of Hawa'i's D-Space repository, ScholarSpace.

Reference

Albarillo, Emily A. & Nick Thieberger. 2009. Kaipuleohone, the University of Hawai'i's Digital Ethnographic Archive. *Language Documentation &*

Conservation 3(1): 1-14.

1 Kaipuleohone is Hawaiian for ‘gourd of sweet words’. We are grateful to Laiana Wong for suggesting this name and for allowing us to use it as the name of this archive.

Bruce Birch *The Ma! Project: Crowdsourcing Software for Language Documentation*

The Ma! Project’s (themaproject.org/) first app + database package is a crowdsourcing lexicon development system consisting of a smartphone/tablet app which allows users to sync audio, video, text and image data to an online database for the purposes of building dictionaries. Synced data is curated via an online moderator control panel. Approved user contributions are used as the basis for new entries, or for modification of existing entries, which are then published to the app. The next time any user syncs, the new or modified entries will be added to the dictionary on their device. The project aims to engage younger speakers of endangered languages in the documentation process. The app has been made available for Android and iOS. So far we have piloted in three endangered language contexts: Iwaidja (Northern Australia); Mokpe (Cameroon); and Bena Bena (PNG). We are currently developing versions for Gamilaraay (NSW), Somali, and the 17 languages used by the Darwin-based Aboriginal Interpreter Service, all of which are expected to be operational by the end of 2013. The presentation will consist of a live demonstration of the app and moderator control panel. For those wishing to preview the app, search for 'Ma Iwaidja' on Google Play or the App Store. .

Steven Bird *Aikuma*

Aikuma is a free Android App designed for recording and translating oral literature. Listen to stories, dialogues and songs, all sorted by language and location. Make your own recordings and instantly share them with other Aikuma users over the Web. A special feature of Aikuma is its voice-driven translation mode. Hold the phone to your ear and listen, and interrupt to give a commentary or translation. The phone records what you say and lines it up with the original. Now the meaning is also preserved. The app is downloadable from GooglePlay. See the website: http://lp20.org/aikuma/

Cathy Bow, Michael Christie and Brian Devlin *Mobilising the Living Archive of Aboriginal Languages*

Originally started through an impetus to preserve thousands of books produced in Aboriginal languages during the years of bilingual education in the Northern Territory, the Living Archive of Aboriginal Languages (www.cdu.edu.au/laal) has the basic structure of a traditional archive. However those involved in its development – academics, linguists, educators, language owners and literacy workers – are concerned at each step to develop structures and strategies

whereby local knowledge and language authorities can supervise the development and use of their own collections. This goal is in part to provide resources for ongoing language and culture work at the local level (in schools and the wider community), but also to connect interested researchers worldwide with the knowledge authorities who can speak for and enrich the collection through collaborative research. This presentation will consider the work involved in negotiating the structural arrangements within the emerging archive, as well as some of the technical, social and political aspects involved in bringing the archive to life in the 20 communities of origin in conjunction with the 25+ language groups whose literature is represented in the archive.

Shubha Chaudhuri *The community and the archive*

The Community and the archive – preservation, ownership and dissemination. Archives had been thought of remote ivory tower spaces with dim vaults and dusty shelves. However archives have been changing as what is archives has changed from state documents to include audio visual documents and cultural expressions. With this change , there is also the shift of who such an archive is for and who uses it and for what. The relationship of archives to the communities that it interacts with is one that has been undergoing change in the past decade, as the concept of the community comes to the centre of the discourse in many areas. The UNESCO Convention for the Safeguarding of Intangible Cultural Heritage places the community and its rights at the centre in many of its directives, WIPO works on community rights for archives, libraries and museums, and so forth. These are not initiatives of large international bodies. The concept of heritage has gone from that of 'high art' to cultural expressions, Masterpieces have been replaced by Representative Lists, the voice of the subaltern, the concept of “bottom up” approaches are at the centre of discussion in many areas, and community archives is not a term that is uncommon any more. I plan to discuss some of these issues as they relate to archives, and trace the path taken by an ethnomusicology archive through its development and its changing aims and profile.

Andrea Emberly *Repatriating childhood: Issues in the ethical return of Venda children’s musical materials from the archival collection of John Blacking*

In academic disciplines that engage in ethnographic field research there have been dramatic shifts in the ways in which scholars approach work with children and young people, guided by changes to institutional Human Ethics Protocols. These shifts have impacted how ethnomusicologists approach the study of children’s musical cultures and must also be addressed when framing issues of repatriation that centre on materials collected from children and young people. Whilst the field of ethnomusicology is increasingly concerned with issues surrounding repatriation and the extensive ethical and community considerations involved in returning materials to cultural heritage communities,

there has only been peripheral consideration of how repatriation might impact the lives of children and youth represented in archival collections. In the late 1950s ethnomusicologist John Blacking collected significant materials documenting the musical lives of young children in remote communities in Limpopo province, South Africa. Blacking's research became the foundation for his seminal work on musicality and Venda children's musical cultures in particular. Blacking's field recordings (including video, audio and photographic materials) and extensive documentation and analysis are currently housed in two archival collections at Queen's University in Belfast and the Callaway Centre Archive at the University of Western Australia (UWA). The Callaway Centre at UWA is currently examining ways to repatriate selections of Venda materials to communities in Limpopo. This paper will examine ethical and methodological considerations for repatriating materials that document childhood, including issues such as informed consent, changing methods for documentation of children's lives, and the return of historical childhood materials to research subjects who are now adults. An exploration of the materials in the Blacking Collection at UWA will be used as a case study to examine issues that may arise in ethnomusicological research that involves the return of archival records of childhood to communities of origin.

Gary Holton *Thanks for not throwing that away: How archival data unexpectedly inform the linguistic and ethnographic record*

Witnessing the explosion in the amount of digital data over the past decade many authors have concluded that not everything can be preserved, that we must instead develop strategies for prioritizing objects for digital preservation (Ooghe and Moreels 2009). Digital language archives have been at least partly immune to these arguments, owing both to the nature of the data they preserve and to their status as early adopters. From the outset language archives have worked closely with the documentary linguistics community to develop standards for data portability which greatly simplify preservation and access (Bird and Simons 2003). The products of modern language documentation are by design much easier to archive than, say, eBooks or video games. Moreover, digital language archives have generally had privileged access to large computing infrastructures, often through particular arrangements with cyber-infrastructure built for hard science data storage and analyses. As digital archiving comes of age and digital language archives are brought within the fold of larger digital preservation efforts, the pressure to prioritize preservation goals will increase. Before we decide to discard materials as superfluous, it is useful to consider some of the ways language archives are being used. In this paper I review some current uses of materials housed at the Alaska Native Language Archive (ANLA). Though designed exclusively as a repository of linguistic knowledge, ANLA is now increasingly recognized by its user community as a rich source of ethnographic information. Language

documentation is for the most part a holistic effort, and though language documenters may not be specialists in topics such as botany, kinship, or geography, they are often the only ones to record this knowledge. Hence the value of language archives as repositories of traditional knowledge. Of course, ANLA is also a rich source of more traditional linguistic documentation. This is not surprising in cases where little or no published documentation exists. However, increasingly we are discovering important information which was excluded from published reference works, ostensibly because it was not thought to be important at the time. Archival documents have revealed errors and oversights in the published records for even the most well-documented Alaskan languages. While anecdotal, these experiences demonstrate the value of preserving all linguistic data, even in cases where good published documentation exists. Digital language archives must resist pressure from the wider library and archives community to prioritize preservation efforts and triage collection. Fortunately, digital language archives are already ahead of the curve, having developed inter-institutional frameworks which stress regional focus and avoid duplication of preservation efforts (Barwick 2004, AIMS Working Group 2012). On this tenth anniversary of PARADISEC it is encouraging to note the great progress which has been made in the development of digital ethnographic archives; however, we must also be prepared for a new era in which digital archiving is a quotidian effort and we face increasing pressure to discard materials.

References

AIMS Working Group. 2012. AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.

Barwick, Linda. 2004. Turning It All Upside Down . . . Imagining a distributed digital audiovisual archive. *Literary and Linguistic Computing* 19.253-63.

Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3).557-82.

Ooghe, Bart, Heritage Cell Waasland and Dries Moreels. 2009. Analysing selection for digitisation. D-Lib Magazine 15(9/10).1082-9873.

Cat Hope, Lisa MacKinney, Lelia Green and Tos Mahoney *The Western Australian New Music Archive Project: Performing as remembering*

In 2013, Edith Cowan University partnered with The State Library of Western Australia, The National Library of Australia, Western Australian music advocacy body Tura New Music, The National Library of Australia and ABC Classic FM on the Western Australian New Music Archive (WANMA), an online archive of Western Australian music from 1970 to the present day. This paper discusses the initial phases of the project and the challenges inherent in creating a digital archive that includes materials which reflect contemporary recognition of improvisation and sound art as composition, recordings as an alternative score, and video as an important documentation device for sound art

and installations. Complex copyright and intellectual property issues are also being addressed, as is the future of the archive beyond its period of project funding.

Daniela Kaleva *Are Collections Scholarly Outputs? Re-defining Research Quality and Impact of Cultural Performative Findings*

The paper explores user-centred cultural heritage collections which are rooted in community life and scholarship, and searches for new terminology and rationale to define them as research outputs first by examining notions of content and quality according to current ARC definitions and ERA categories of research outputs. Considering the performative paradigm as an opportunity to re-define knowledge that is not numerically or textually encoded but contains performative signs incorporated into captured cultural records through audio and/or audiovisual mediums, the paper proposes that the performative paradigm allows scholars to view performative signs not as mere data or by-products of their research but as a wealth of performative evidence that complements quantitative and/or qualitative methodologies and therefore is an integral part of their scholarly outputs. This implies the need for consolidation of the close collaborative processes between scholars, community members and information professionals, and further development of databases such as PARADISEC and AusLit to enhance discovery, description and citation of audio-visual materials while at the same time hosting hybrid online journals that publish text-based exegeses with the capacity to link to specific portions of cultural heritage content. Lastly, the paper discusses the assessment and measurement of research impact which will be the new focus of government assessment exercises, highlighting the advantage of cultural heritage knowledge banks due to their digital accessibility and potential for a variety of citation statistics that enable assessment of both quality and impact.

Renée Lambert-Brétière and Lynn Drapeau *Making the most out of the past: Retrieving and archiving old records of the Innu language*

In a situation of language endangerment, communities express a strong need for the documentation and preservation of their language which is increasingly threatened as the older generations of monolinguals pass away. This paper presents the documentation project of Innu, an endangered Algonquian language spoken by roughly 13,000 people in 11 communities spread out over Northeastern Quebec and Labrador in Canada. It will specifically address the issues related to retrieving and converting a large body of existing linguistic materials into digital format, and making old language records available for the benefit of the people in their efforts of revitalization. Since the beginning of the 20th century, hydroelectric, mining and forestry projects are undertaken on the ancestral lands of the Innus. Many of these projects were carried out without their consent. However, since the late seventies, strategies for successful

negotiations with government and private developers include discussions with the Innus to ensure a constructive dialogue with the province's economic development partners. Innu Elders play a critical role in these exercises by providing traditional ecological knowledge, e.g. about waterways, and documenting where people hunt, fish, trap, gather and camp. The outcome of these surveys constitutes precious information about the Innu culture, traditions, as well as language, since many of the Elders were monolinguals. As an illustration, in the early eighties, the Attikamek-Innu Council led a major investigation on territorial use and occupancy. The collected material comprises the testimony of more than 400 Innus and over one thousand hours of recordings. This documentation and all other existing analog records are precious for the collective memory and history of the Innu communities, and need to be located, retrieved, digitalised, and archived for the long-term sustainability of the language. We undertook this enterprise in partnership with the Innu Cultural Institute. With the constitution of an archive comprising old records and original materials, the Innu language documentation project aims to make an important contribution to the ongoing developments in language documentation research and a major step in building a valuable tool for language maintenance and revitalization.

Paul Monaghan *Towards a Nhangga hermeneutic: breathing life into written archival materials*

This paper is concerned with issues arising from the translation of a dreaming story from English/Aboriginal English back into an extremely endangered Aboriginal language. The story of Bilarl (Sooty Bell Magpie) was recorded by Daisy Bates in the early years of the twentieth century during her time at Eucla (WA) or Fowlers Bay (SA). The story was uncovered in the Daisy Bates Collection as the Barr Smith Library and was translated by a group of Wirangu people (Nhanggas) and the author at Ceduna (SA) in 2011. The translation process, the first attempted for the Wirangu language, revealed a number of surprises relating to what I call a Nhangga hermeneutic. That is, the way that Wirangu people perform the task of translating into their heritage language reflects local ways of making meaning that differ substantially from methods informed by normative literacy practices. In outlining this hermeneutic, this paper explores the processes of liberating oral narratives from written documents.

Stephen Morey, Mark W. Post and Victor Friedman *The language codes of ISO 639: A premature and possibly unobtainable standardization*

ISO 639 is an ambitious attempt to standardize and organize various types of references to the languages of the world. It is designed to be fully comprehensive and permanent; as such, it promises to greatly enhance the precision and reliability with which language materials can be archived,

catalogued, and referenced in the literature, as well as the ease and precision with which such materials and references can be processed by machines and effectively located via search queries. There are, however, a number of serious problems with several components of ISO 639 as they are currently conceived. At a minimum, these are:
(1) use of both The Ethnologue as the basis for ISO 639-3's "three-letter codes" and of SIL as its registration authority is problematic for a number of reasons
(2) in-principle "arbitrary" (but in fact not arbitrary) "mnemonic" labels of ISO 639-3 have the potential to enshrine offensive designations for language communities, and in fact currently do so
(3) decision-making processes in ISO 639-3 are currently excessively centralized and privilege the views of a minority of the linguistics community
(4) the in-principle "permanency" of language codes such as those of ISO 639-3 is fundamentally incompatible with the nature of human languages, which are demonstrably impermanent
(5) the structure of ISO 639-3 has a serious potential to be misunderstood, misused, and in fact abused by decision-making bodies (such as arms of government in various political contexts)
(6) ISO 639-5, which attempts to catalogue the genetic affiliations of the world's languages, is highly premature, since there is nothing approaching agreement among specialists in a great number of cases
(7) ISO 639-6, which attempts to catalogue language variation, is in principle impossible, unless it aims to extend to an analysis of the language use of every human being on Earth, living or dead

On the basis of these observations, which we will illustrate by means of three detailed "case studies" from the Eastern Himalaya, the Burmese/Indian border region, and the Balkan region, we will argue that ISO 639 must be substantially re-conceived and re-organized before it can be supported by linguists.

Simon Musgrave and John Hajek *Linguistic scholarship in the data-driven 21st century*

Linguistic scholarship in the data-driven 21st century Two important forces have been acting on the discipline of linguistics since late in the twentieth century: technological changes which allow the capture and dissemination of high quality multimedia data efficiently and at a reasonable cost, and an emphasis on the collection of primary data as a response to deep concerns about the reduction of linguistic diversity across the world. The convergence and interaction of these two forces is driving changes to scholarly practice in the discipline. With an organisation such as PARADISEC celebrating its tenth anniversary, we can identify some aspects of those changes which are stabilising and it is therefore possible to speculate in an informed fashion about what linguistic scholarship will look like in the coming decades and to consider

the implications. In this paper, we will argue that there are implications for technical infrastructure which are being addressed at least to some extent, but that the implications for the social infrastructure of our discipline, particularly channels of dissemination for scholarly work, are much more profound and are not yet being adequately addressed. We suggest that 21st century linguistics will be increasingly based on access to primary data. By this we mean access at all stages of the process of scholarship: access to shared data in well-organised repositories as well as the possibility of directly citing data in our publications. The infrastructure for archiving exists (witness PARADISEC); additional elements such as servers which allow clients to address specified segments of media files on demand are being discussed and developed (RNLD List April 2013). On the other hand, the less tangible infrastructure to support these changes is not yet so prominent. Direct citation of primary data means moving fully to electronic publishing; by this we mean not merely making work available online in a format such as pdf, but reconceptualising our forms of scholarly communication as essentially freed from text on paper. This in turn implies a reworking of the systems of gate-keeping and prestige which are associated with the current publication models, and such changes must also include the recognition of the deposit of properly curated data as an accepted part of scholarship. Moves in these directions have begun: last year saw the appearance of a volume devoted to the topic of electronic grammaticography (Nordhoff 2012), and the Australia Linguistic Society is engaged in a dialogue with the Australia Research Council about the recognition of data deposits as research outputs (Thieberger, Margetts, Morey, Musgrave and Schembri 2012). We suggest that the benefits of a linguistics which is closely linked to primary data are evident and moves in this direction are therefore inevitable. But the concomitant changes to the institutional structures of scholarship will be profound and complex.

References:

Nordhoff, Sebastian (ed.). 2012. Electronic Grammaticography. Hawai'i: University of Hawai'i Press.
Thieberger, Nick, Anna Margetts, Stephen Morey, Simon Musgrave and Adam Schembri. 2012. Assessing curated corpora as research output. Paper presented to the 2012 Conference of the Australian Linguistic Society, University of Western Australia, November 2012.

Simon Musgrave, Linda Barwick, Michael Walsh and Andrew Treloar *Language identifying codes: remaining issues, future prospects*

The work of organisations such as PARADISEC is crucially dependent on accurate and reliable identification of the languages which are represented in resources. For efficient discovery of resources to be possible, an identifying system which is accurate and stable in itself is necessary, as is wide agreement to use the system across the relevant communities (archivists and researchers

from various disciplines). ISO 639-3 is such a system and acceptance of it is now widespread; this should not, however, be taken as meaning that no problems remain and in this paper we draw attention to some of the remaining issues and the potential role of Australian researchers in working towards their solution. ISO 639-3 reflects the reality of language differentiation more or less accurately depending on the region in question. A process for requesting revisions to the codes exists and is being used quite extensively by scholars working on Australian languages. The experience thus being accumulated will be of value in future work on language identification. This process also draws attention to another area where improvement can be made: currently, the different parts of ISO 639 (639-1, 639- 2 etc.) have different registration authorities. Bringing all parts of the standard together under a single registration authority would have benefits for ongoing revisions and for transparency and is therefore an important goal. Another important goal is to ensure that linguists are able to provide input to three parts of ISO 639 currently being developed:

• ISO 639-5 a proposed set of codes for identifying groupings above the level of the single language,

• ISO 639-6 a proposed set of codes for identifying linguistic entities below the level of the single language,

• ISO 639-4 will provide an account of the principles on which the various codings rest. Australia is represented in ISO by Standards Australia, and this body has observer status in relation to ISO Technical Committee 37 which is responsible for the 639 group of standards. A group of interested scholars in Australia constitute an informal reference group for these issues (ARGILaRe: http://users.monash.edu.au/~smusgrav/ARGILaRe/) and this group is establishing ways to provide expert input. These include the establishment of a mirror committee for TC37 under the ambit of Standards Australia, ongoing involvement with international projects and endeavours, and the potential formation of a Working Group within the Research Data Alliance framework. The goal of improving access to language resources should be one which unites various research communities and therefore we are optimistic that such endeavours can and will produce valuable results.

David Nathan *On defining the reach of digital language archives: audiences, discovery, delivery, access, accessibility, and feedback*

Over the last decade, and with the help of digital media and technologies, archives (with the focus here on archives for endangered and minority languages) have extended their focus from preservation to also becoming facilities for dissemination. Their innovations have largely been on 'discovery': firstly by encouraging digitisation and inclusion of analogue and obscure materials, and by partnership with funding institutions to support the creation of

new, 'born digital' language resources; and secondly through online provision of language resources via web catalogues driven by standardised metadata and in some cases providing enhanced discovery through web portals aggregating the holdings of multiple archives. These advances have increased the visibility, relevance and authority of archives for language-related disciplines and for language-speaker communities. This paper considers a broader set of parameters describing the 'reach' of archives, where 'reach' includes (a) archives' understanding of their key audiences in order to provide appropriate services for them, e.g. identifying a range of relevant audiences, their languages of access, their varied technological and information literacies, interface design and usability; (b) discovery, drawing on the understandings of audiences in order to help them browse, navigate, search, identify and select their items of interest; (c) delivery, i.e. making available selected resources according to users' preferences whether by download, view-in-browser, through apps or other means; (d) access management such that resource delivery follows depositors' and communities' preferences, and where users have ways of applying for and negotiating for access; (e) information accessibility, where the actual desired content is accessible to users, whether in terms of contextualisation or appropriate complexity, language, or modality; and finally (f) feedback channels, where users can utilise the archive to provide feedback to depositors or to enhance deposits with user-generated content. Through considering how a number of archives are providing such services, we can see their transition from repositories of memory to facilities for fostering participation and understanding.

Jennifer Post *Reconstructing, Reinterpreting, and Repatriating Musical Instrument Data in Ethnomusicological Archives*

In-depth ethnomusicological research on musical instrument production and use is surprisingly scant. At the same time, field data compiled by ethnomusicologists since the mid-twentieth century, now housed in archives and personal collections around the world, demonstrate that ethnographers amassed considerable information on musical instrument production and use. The John Blacking collection, housed at the University of Western Australia and at Queen's University in Belfast, holds audio, visual, and manuscript field data on music collected by Blacking in South Africa, Zambia, and Uganda in the 1950s and 1960s. Among the instrument data are drawings and diagrams, information on woods and other building materials, makers' names, and descriptions and demonstrations of performance practices and social beliefs related to specific instruments. This data can be used to engage more fully with contemporary ethnomusicological, museum, and archival work as we consider the value of cultural and social information, and the materials themselves, to communities in which the documented musical instruments were produced and played. In this study I use selected John Blacking archival data to discuss how

fieldwork documents can be used effectively in the repatriation process to offer musical, social, and ecological knowledge to communities. New interest in organology encourages scholars to reach beyond descriptive, historical, and object-centered views to embrace greater engagement with the social life of musical instruments. Applied to repatriation, the practical and interpretive information offers knowledge to the communities from which instruments and instrument information have been drawn, and it provides opportunities for data sharing among all communities impacted by cultural and material loss.

Martin Raymond *ScriptSource: making information on the world's scripts and languages accessible*

Although there is plenty of script and language information on the web, there has been a need for a site to present the information authoritatively and clearly, making it easier to understand the often complex relationships between scripts, characters and languages. ScriptSource has been designed to meet that need and to answer questions such as: 'Which scripts can be used to write that language?', or, 'Which writing systems use this Unicode character?'. The site allows registered users to contribute information in the form of entries, which are moderated. ScriptSource imports character data from Unicode and locale data from the CLDR (Common Locale Data Repository). Language data is imported from the Ethnologue, and ScriptSource has individual pages for around 7000 languages, each of which has links to the corresponding language pages on several linguistic websites. Some language pages include additional links to sites, such as the 'Aboriginal Languages of Australia' site, and those offering relevant fonts and keyboards. This session will cover some of the needs ScriptSource has been designed to meet, and will explain the invaluable data association mechanism it uses to link information to scripts, characters and languages. The language documentation features of ScriptSource will be explained, including its facilities for the entry of exemplar character lists and phonemic data.

Martin Raymond *Why ScriptSource makes a good hub for language documentation*

The ScriptSource website is designed to be an authoritative source of information on the world's scripts, characters and languages, and the relationships between them. The site presents substantial core data, including more than 200 script pages, 100,000 character pages and 7,000 language pages. Each language page on ScriptSource contains information about which scripts can be used to write the language, with links to the corresponding language pages on at least five other linguistic websites. Software entries provide details of fonts and keyboards which work for the language. Registered users can contribute information in the form of entries, which are moderated. The demonstration will reveal the wealth of information to be found on

ScriptSource, and will show how language information can easily be entered, and associated with other information on the site, to make ScriptSource a valuable aid to language documentation. An interface to enter the character lists for a language is already in place and will be included in the demo; a phonemic data interface is currently under development. An increasing number of links to other language sites, such as the Aboriginal Languages of Australia site, underline the usefulness of ScriptSource as a hub for language documentation.

Paweł Rutkowski, Joanna Łacheta, Piotr Mostowski, Joanna Filipczak and Sylwia Łozińska *The Corpus of Polish Sign Language (PJM): Methodology, Procedures and Impact*

Polish Sign Language (polski język migowy, usually abbreviated as PJM) is a natural visual-spatial language used by the Polish Deaf community. It emerged around 1817, with the foundation of the first school for the deaf in Poland. Up until recently, the hearing linguistic community in Poland devoted very little attention to PJM. The aim of this paper is to present a new large scale research project aimed at documenting PJM. Its main goal is to create an extensive and representative corpus of video material that will further form the basis of detailed grammatical, lexical and cultural analyses. The PJM corpus project was launched in 2012 and its first phase will conclude in 2015. The underlying idea is to compile a collection of video clips showing Deaf people using PJM in a variety of different contexts. The first phase of the project will involve approximately 100 informants. As of May 2013, more than 70 people have already been filmed. When the project is completed, some 500 hours of footage will be available for research purposes. The PJM corpus is diversified geographically, covering more than 10 Polish cities with significant Deaf populations. The group of signers participating in the project is well balanced in terms of age and gender. Data is collected exclusively from signers who either have deaf parents or have used PJM since early school age. They come from different social and educational backgrounds (respective sociological metadata is an integral part of the corpus). Recording sessions always involve two signers and a Deaf moderator. The procedure of data collection is based on an extensive list of tasks to be performed by the two informants. Typically, the signers are asked to react to certain visual stimuli, e.g. by describing a scene, naming an object, (re-)telling a story, or explaining something to their partner. The elicitation materials include pictures, videos, graphs, comic strips etc., with as little reference to written Polish as possible. All the necessary instructions are given in sign language exclusively; they have been pre-recorded and, like the elicitation materials, are presented to the participants on computer screens. The participants are also requested to discuss a number of topics pertaining to the Deaf. Additionally, they are given some time for free conversation (they are aware of being filmed but no specific task is assigned to them). The latter two parts of the recording session scenario are aimed at collecting spontaneous and

naturalistic data. When designing the above procedures, we took into account the challenges and problems encountered in similar projects conducted for other languages, in particular for German Sign Language (DGS), Sign Language of the Netherlands (NGT) and Australian Sign Language (Auslan). For instance, we attempted to make use of elicitation materials that had proved successful in the other projects. The raw material obtained in the recording sessions is further tokenized, lemmatized, annotated, glossed and translated using the iLex software developed at the University of Hamburg. The annotation conventions we employ have been designed especially for the purposes of PJM. The aim of the present paper is to give a detailed overview of the above procedures and show sample clips extracted from the PJM corpus in order to illustrate the most important advantages and disadvantages of the methodological choices that we have made. We also want to emphasize the societal role of this project in the signing community of Poland (as it is the first-ever attempt at collecting an extensive archive of the language and culture of the Polish Deaf).

Guillaume Segerer and Sébastien Flavier*The RefLex project : documenting and exploring lexical resources in Africa*

The RefLex project aims at testing a set of fundamental hypotheses concerning the structure and the evolution of African languages that are often mentioned in the literature, but whose validity was never demonstrated on an empirical basis. These hypotheses share the peculiarity that they can only be tested by means of a quantitative approach, which in turn presupposes the existence of a comprehensive documentation. The more than 2,200 languages spoken in Africa are characterized by great typological diversity, but also display some common characteristics, on each level of linguistic analysis, that go beyond the linguistic phyla and areas. So far, it has never been possible to conduct an in-depth study of these characteristics (e.g., logophoric pronouns, labiovelar consonants, etc.), due mainly to a lack of available data on the majority of African languages. Reflex solves this problem by fully exploiting the existing lexical documentation, which is in fact much larger than the grammatical documentation and yet often ignored in especially typological studies. One of the goals of RefLex is to make the scattered and hard to find lexical documentation available to interested researchers. Indeed, the lexical corpus of African languages, which is available on line for the whole scientific community, gives immediate access to a considerable wealth of data (as to june 2013, 460,000 lexical units for more than 370 languages, but we expect more than 1,000,000 entries within the next two years, representing 1,000 languages). This corpus will allow dramatic progress in several domains: typology, phylogeny, lexical semantics, lexical spread, areal linguistics. RefLex will be the largest online comparative database worldwide. Moreover, the database will be different from other existing databases at two crucial levels: (i) the possibility to have a direct online access to the original documents which are

the basis of the digital data, which makes this corpus a true reference corpus, allowing corrections, checking, argued feedback, replication and even falsifications; (ii) a library of computational tools for the scientific use of the data, designed to facilitate research, retrieval and comparisons. The RefLex project thus conforms to the emerging domain of quantitative approaches to complex linguistic issues. It represents one of the very few projects based on data coming from various languages and the only one to enable easy manipulations of and experiments with the data itself. A set of statistical tools makes it possible to measure all kinds of combinatory distributions, including, but not restricted to, phonological correlations. An other bunch of tools is dedicated to phonological and lexical reconstruction, enabling the management of cognate sets and correspondance sets. Our talk will present the project, the existing tools, as well as their future developments.

Jozsef Szakos and Ulrike Glavitsch *Like a "Swiss knife which cuts in two directions": On the development and use of SpeechIndexer as a documentation and teaching tool*

SpeechIndexer has been developed for language documentation and learning at the ETH in Zurich. Its original goal was to help access the archive recordings of endangered Austronesian languages of Taiwan. In later years, it was more and more applied to preparing teaching materials not only for Austronesian languages, but to help organize and retrieve authentic speech materials of other modern languages. Starting from the precise indexing and retrieval of single speakers, it has grown into a small but complex tool which can deal with multiple speakers and indexing of overlapping speech segments. There have been several challenges:
(1) it was necessary to find an optimal coding of the speech participants in recordings and a retrieval mechanism of participants' speech peculiarities and (2) we have to deal with overlaps in the speech of participants. Overlaps are unnoticed by the system, since the pause-finding algorithm looks for silence breaks for suggesting phrase units. A speech extract where multiple speakers are talking simultaneously is segmented into a single speech segment if there are no obvious pauses. A mechanism that filters out the various speakers' voices that can be individually indexed is the long-term goal. A way of marking speaker overlaps is the realistic short- or medium-term goal. Our presentation introduces the solution we developed for the encoding scheme of speech participants and the respective retrieval of their speech characteristics. Furthermore, we show the marking system we have devised for overlaps. Further disambiguation research is needed to find out the respective types of overlaps in individual speech acts. An indexing possibility is to be provided still, and we are further working on this problem. Finally, we demonstrate on some Formosan Indigenous Languages materials, where SpeechIndexer is used as a tool to build textbooks, how the multiple speakers' linear indexing is one

more powerful asset to choose this software for concordancing speech corpora and integrating them into advanced learners' textbooks and multimedia materials. In teaching dialogues by authentic recordings (dialogues, discussions, drama, theater) selectively speaker's voices can be silence and the learner can practice, playing that part. Documentation and language teaching are kept at a distance by academic and educational institutions. We still believe, however that they are but two sides of the same coin, they all deal with the same authentic language materials, therefore the SpeechIndexer software can help by simultaneously satisfying both needs.

Nick Thieberger *Using the TEI to encode manuscripts of Australian languages*

This paper will discuss the value of using the Text Encoding Initiative's schema (TEI) for a set of manuscript vocabularies of Australian Indigenous languages collected by Daisy Bates in the early 1900s. I will first outline the method used to type the text from manuscript images and then contrast the effort required to render the vocabularies as encoded text with the simpler method of placing page images online (as PARADISEC did, for example, with the Capell papers). I will assess the tools available for markup of the corpus and show that the encoded version affords more research outputs than does the simple rendering of an image, and has benefits for the broader community, including speakers of the languages recorded. In addition, the project provides the National Library of Australia with an enriched description of the Bates vocabulary collection. Once the data structures have been tested by users there is the potential to crowdsource annotation of as yet untranscribed handwritten sections of the work.

Sally Treloyn and Rona Googninda Charles *Repatriation and innovation: the impact of archival recordings on endangered dance-song traditions and ethnomusicological research*

For some time, ethnomusicologists working in Australian Aboriginal communities have repatriated and disseminated audio and video recordings from archival and personal collections to cultural heritage communities as a primary fieldwork method. Increasingly researchers are documenting these processes and are considering the complexities of repatriation and dissemination, and their role in supporting creative innovation and in sustaining performance traditions. As such, while we consider the contexts in which archival materials influence and may be used to innovate endangered song traditions, we might likewise consider ways in which the process of returning materials influences and innovates fieldwork and research. This paper will outline the materials and processes of repatriation involved in the Australian Research Council project 'Strategies for Preserving and Sustaining Endangered Song and Dance in the modern world: the Mowanjum and Fitzroy River Valley

communities of WA'. The paper will present perspectives from both cultural heritage stakeholders and researchers on the role of repatriation of archival materials in: fostering partnerships between researchers and communities; in supporting the capacity of local organizations; in supporting intergenerational engagement around dance-song knowledge; and, in better understanding the intersections and tensions between traditional systems of knowledge management and dissemination, local community archives, and national archives.

Peter Withers *New Developments in Arbil*

This talk will introduce Arbil which is a tool for managing metadata that describes research data, such as audio or video files, allowing research data files to be easily searched both before and after they are archived. Arbil has been developed at The Language Archive at MPIPL (Author, 2012) and was originally designed for the DOBES community to replace the IMDI Editor. The core needs expressed by this group was viewing and editing the metadata when in the field and being able to edit more than one metadata file at once. Indeed, Arbil is fully functional offline, provides tabular editing, and for robustness stores only text metadata files. For moving metadata and associated resources into an LAT archive, the structure is exported from Arbil and then uploaded into LAMUS (Broeder et al., 2006). Arbil was originally designed to support IMDI metadata (Broeder and Wittenburg, 2006). This format has been in use for many years, and it covers most needs with a number of set fields, but also may confuse researchers and slow down the workflow with so many fields to fill in. This issue has been addressed by CLARIN (Va´radi et al., 2008). CLARIN provides flexible metadata fields, allowing a custom profile to be designed for each project only the relevant metadata fields need to be offered to the end user, greatly simplifying the process of creating metadata. Arbil has now been updated to support both IMDI and Clarin metadata formats. Because of the flexible design of Arbil, some of its components such as the metadata table and tree have been utilised in KinOath Kinship Archiver (Author, 2011). This application builds on the core functions of Arbil, onto which it adds an XML database to provide fast searches. Also, a plugin layer has been introduced in KinOath which has been migrated back into Arbil. Another project that is in the prototype stage is a web based search similar to the search in Arbil. These changes are being combined together as a search plugin for Arbil which is in development that will allow much more powerful searches to be available without compromising the original design of the application.

References

Author. 2012. Metadata Management with Arbil. In Proceedings of the Eighth International Conference On Language Resources And Evaluation (LREC 2012) Satellite Workshops, pages 72–75. Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Pro

ceedin gs.pdf Author. 2011. KinOath, Kinship Software Beta Stage of Development. Talk presented at Atelier d'initiation au traitement informatique de la parenté. salle 3, RdC, bât. Le France. 20111216.
D. Broeder and P. Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. International Journal of Metadata, Semantics and Ontologies, 1(2), pages 119–132.
T. Váradi, S. Krauwer, P. Wittenburg, M. Wynne, and K. Koskenniemi. 2008. Clarin: Common language resources and technology infrastructure. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pages 1244–1248, Marrakech. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf.
D. Broeder, A. Claus, F. Offenga, R. Skiba, P. Trilsbeek, and P. Wittenburg. 2006. LAMUS : the Language Archive Management and Upload System. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pages 2291–2294, Genoa. European Language Resources Association (ELRA). www.latmpi.eu/papers/papers2006/lamuspaperfinal2.pdf.

Peter Withers *KinOath Kinship Archiver*

This talk will introduce a new tool for Humanities research, in particular Ethnology, Linguistics, Law, History, but also Genetics and Archiving. This tool is KinOath Kinship Archiver which is an application for collecting and analysing kinship data. It is designed to be flexible and culturally nonspecific, which is important to prevent extraneous concepts being imposed onto the data being recorded. The kinship data can be linked to external resources such as archive data. Graphical representation of the data is a key feature, it produces publishable quality diagrams that can be exported to SVG, PDF and JPG formats. Data can be imported from GEDCOM, CSV and TIP files. Data can be exported into CSV format, with additional formats becoming available as plugins. KinOath provides very flexible data fields for each individual / entity recorded in the kinship data, this is combined with customisable relation types, customisable symbols and customisable kin types. This means, for example, that any number of genders and kinship relations can be defined and represented on a diagram. The most common format, GEDCOM (Family History Department, 1999), can be imported into KinOath. However this GEDCOM format exhibits cultural specificities because it has a predetermined set of kinship types, genders and initiation ceremonies. We know that there is a wider array of kinship types (e.g. suckling relations (Altorki, 1980)) and genders (e.g. the Mahū of Hawai'i (Matzner, 2001)). There are also initiation ceremonies beyond the Christian and Jewish ceremonies that are predefined in GEDCOM. However once this data is imported, all the flexibility of KinOath will be available. KinOath has project based diagrams and freeform diagrams.

Freeform diagrams are like a quick sketch; while project diagrams each have a database of kinship data which can be shared across multiple diagrams. Project based diagrams also allow kin type string queries, such that individuals to be found based on their relations to others. Individuals in a project diagram can be duplicated and merged, which can be useful, for example, in correcting data, or merging multiple data sets where some individuals overlap. In freeform diagrams kin terms can be defined with kin type strings and shown on the diagram, organised in groups, imported and exported. In the future it will be possible to overlay these kin terms onto project diagrams. In order to perform statistical analysis, the kinship data for each project or freeform diagram can be exported for use in R or SPSS. This combined with queries based on kin types and other search parameters, provides great potential in the analysis of both the kin data and the archive data that has been recorded. The intended users of Kinoath are any researchers that collect data in a context of social relations. Kinship data is often not systematically included in the metadata of archives, however these kin relations provide a context that enriches that archived data. KinOath is in active development and new features are regularly being added. The plugin framework that KinOath shares with Arbil has made it possible for external developers to add features. The various versions and the manual are available at: http://tla.mpi.nl/tools/tlatools/kinoath/

REFERENCES

Family History Department of The Church of Jesus Christ of Latterday Saints, 1999, THE GEDCOM STANDARD DRAFT Release 5.5.1 http://www.phpgedview.net/ged5515.pdf
Matzner, Andrew. 2001. *'O au no keia: voices from Hawaii's Mahu and transgender communities*. Bloomington, Indiana: Xlibris.
Altorki, Soraya. 1980. MilkKinship in Arab Society: An Unexplored Problem in the Ethnography of Marriage. *Ethnology* 19(2): 233244.