Sustainable Data from Digital Research Conference

Abstracts
In order of presentation

Monday Monday Monday Monday Monday

Stephen Ramsay

Found: Data, Textuality, and the Digital Humanities

Computational processes generate lists: lists of numbers, lists of words, lists of coordinates, lists of properties. We transform these lists into more exalted forms -- visualizations, maps, information systems, software tools -- but the list remains the fundamental data structure of computing, from which most other structures are derived. Whenever we treat the world as data, we are nearly always creating lists. But what sort of *texts* are these, and can we consider them the same way that we consider other texts within the humanities? In this paper, I offer some meditations on the nature of lists, and suggest that it is the paucity of information they provide -- and the ways in which that paucity licenses narrative and explanation -- that allows us to imagine computational representations as texts that can play a fruitful role in the wider context of humanistic inquiry.

Sebastian Drude and Paul Trilsbeek

The 'Language Archiving Technology' solutions for sustainable data from digital research

Since the late 1990'ies, the technical group at the Max-Planck-Institute for Psycholinguistics has worked on solutions for several of the questions addressed in this paradisec-meeting, in particular, how to guarantee long-time-availability of digital research data for future research. The support for the well-known DOBES (Documentation of Endangered Languages) programme has greatly inspired and advanced this work, and lead to the ongoing development of a whole suite of tools for annotating, cataloguing and archiving multi-media data. At the core of the LAT tools is the IMDI metadata schema, now being integrated into a larger network of digital resources in the European CLARIN project. The multi-media annotator ELAN (with its web-based cousin ANNEX) is now well known not only among documentary linguists. Other tools such as the lexical database tool LEXUS, the related knowledge-space builder VICOS and others are not yet widely used. With further development and integration with other tools they also have the potential for being useful tools for representing non-time-related linguistic data. We aim at present an overview of the solutions, both achieved and in development, for creating and exploiting sustainable digital data, in particular in the area of documenting languages and cultures, and their interfaces with related other developments.

Alexander Borkowski and Andrea Schalley

Going beyond archiving – a collaborative tool for typological research

The work described in this paper aims to outline some of the design aspects for a collaborative tool for typological research. This tool is designed to allow for the collation, from multiple contributors, of linguistic examples and their analysis with regards to an open set of variation dimensions of both onomasiological and semasiological nature. The resulting knowledge base combines linguistically relevant categories of human conceptualisation (e.g. in-group, such as ethnic or family group, categories) together with their linguistic coding (e.g. in gender affixes, verbal agreement), all based on actual linguistic examples from diverse natural languages as its underlying data-driven foundation. The system is based on Semantic Web technology and hence can be queried in a flexible way that allows for combining any variation dimensions within a query (e.g. it allows to answer questions such as which languages exhibit joint attention marking by way of verbal suffixing). We will focus on design aspects relating to sustainable data. How

can sustainable data for such a project be delimited? Surely, this encompasses commonly accepted aspects such as standards conformity, longevity, and accessibility, which we will address in the paper. Additionally and in particular, however, we will argue that user orientation and involvement is a critical factor. Following on from this, the tool is designed in a way that it (i) does not require linguistic users to be trained extensively in system usage, (ii) allows linguists to deploy their standard methods of data entry (e.g. interlinear glossing), and (iii) provides contributors with immediate integration of their own with previously entered data and access to the resulting analysis (i.e. querying) and research potential. The paper will roughly be structured as follows: We will describe the background and aims of the project, and contextualise it in relation to other similar projects. We will then concentrate on how sustainability is addressed, discussing a number of different facets of sustainability. This includes data storage formats, user interface and workflow modelling, knowledge base design, and system features (in particular system output). We will also outline some problems that have arisen so far and close with an outlook on future development.

Colleen Fitzgerald

Investigating Connected Speech from Tohono O'odham Digitized Legacy Data

Archival and legacy resources provide rich material for linguistic investigation, provided such materials are accessible. The growing number of digital tools offers prospects for investigating new research questions. Working with such resources and considering usability, particularly with community members, leads to new insights with old data. In this paper, I talk about how exactly this has happened for the Tohono O'odham language by digitizing two legacy resources: an out-of-print dictionary (Mathiot 1973) and fieldnotes from the late Kenneth Hale. This paper focuses on new linguistic investigations possible because of having digitized these materials. Tohono O'odham is a Native American language primarily spoken in the southwestern United States. The number of speakers has been declining, with an estimated 8,000 to 10,000 speakers. While there has been considerable linguistic research done on the Tohono O'odham language, including preliminary descriptions (Dolores 1913, Mason 1950, Hale 1959, Saxton 1963, 1982, Mathiot 1973, Zepeda 1988), most research has focused on syntax (i.e., Hale 1975, Fitzgerald 2003) phonology (i.e. Hill and Zepeda 1992, Fitzgerald 1997, 2002), and morphology (i.e., Zepeda 1984, 1987, Hill and Zepeda 1991, 1998). Relatively little has been done on connected speech, particularly important for an endangered language, with benefits to revitalization and second language instruction. In fact, there are unanalyzed resources in terms of connected speech in the Mathiot dictionary and the Hale field notebooks. Mathiot noted deletions, assimilations and other attested surface forms, with annotations indicating the citation form. The Hale notebooks were recorded in 1961 from two Tohono O'odham speakers (different dialects) and a third speaker of the Pima variety, with accompanying recordings. While Hale's transcriptions are in citation form, he indicates pauses and high-level phrase groupings and thus gives indications of prosodic phrasing. His transcriptions are also contextualized, whereas Mathiot's connected speech transcriptions are given out of their narrative context. The O'odham digitized materials have considerable potential for data-mining, as well as practical uses for revitalization and maintenance. I address how each set of materials presents challenges, including how to best represent aspects of connected discourse in tools such as FLEx, to consider what elements are helpful for revitalization and language teaching, and to package information for second language learners. Exploring these implications is useful for other revitalization and research teams; it offers ideas about what type of documentation will have future use, as well as challenges in balancing connected speech and citation forms in standardized formats like dictionaries.

Simon Musgrave and John Hajek

ERA and sustainable data

We are now entering the second round of assessment under the Excellence in Research for Australia (ERA) model. A number of writers have drawn attention to the problems inherent in this model for the humanities and social sciences (Cooper and Poletti 2011, Dobson 2011, Genoni and Haddow 2009) but what has not received attention (as far as we are aware) is the impact this model is likely to have on the type of work being highlighted and encouraged in this meeting. ERA relies to a large extent on journal rankings and as a result of its first iteration, many academics are experiencing pressure to direct their publications to highly-ranked journals. We suggest that this pressure is already disadvantageous to innovative work utilising digital data. Prestigious journals are, by their nature, conservative institutions and, at least in the humanities, are unlikely to encourage new models for disseminating results. Whilst journals such as Science and Nature routinely host supporting materials for papers on their websites, such practices are uncommon in the humanities. We are aware of a single journal in our field (Language Documentation and Conservation) which is highly ranked in the ERA process (it has an A ranking) and also encourages such publication. New journals which are published online and other alternative modes of disseminating scholarly work will inevitably have to wait some time to achieve any recognition on the ranking lists; indeed as the lists aim to maintain a proportion of journals at each level, it will be very hard for new publications to achieve a high ranking as that must be at the expense of an established publication. Thus the ERA model will tend to discourage innovative modes of publication. Additionally, the model gives no recognition to the idea which underlies much of the work presented in this forum, that making data widely available to colleagues is an inherently worthy activity. The experience of the British assessment exercise on which ERA is based was that researchers were placed under considerable pressure to ensure that their limited research time was geared to producing outputs which would be visible and valuable for assessment. Producing, curating and sharing sustainable data are activities which will struggle to meet these criteria.

Toby Burrows

Sharing Humanities Data for E-Research: Conceptual and Technical Issues

The humanities, as defined by the Australian Academy of the Humanities, encompass the following disciplines: Archaeology; Asian Studies; Classical Studies; English; European Languages and Cultures; History; Linguistics; Philosophy, Religion and the History of Ideas; Cultural and Communication Studies; the Arts. Researchers in some of these fields employ quantitative and qualitative methodologies similar to those used in the sciences and social sciences, but most research in the humanities is perceived as distinctive and different from research in other fields, both in its methodologies and in its approach to data. Archiving and sharing humanities data for reuse by other researchers is crucial in the development and application of e-research in the humanities. There has been considerable debate about the applicability of e-research in the humanities, particularly around the relevance of programmes to digitize source materials on a large scale. Conceptualized and designed properly, however, a humanities data archive can provide the platform on which data-intensive e-research can be based, and to which e-research processes and tools can be applied.This paper looks at the distinctive characteristics of humanities data, and examines how various models of the humanities research process help in understanding the meaning of 'data' in the humanities. It reviews existing services and approaches to building data archives and e-research services for the humanities, and the assumptions they make about the nature of data. It also analyses some conceptual and technical frameworks which could serve as the basis for future developments, focusing particularly on the place of Linked Open Data in building large-scale humanities e-research environments.

Monday

Monday

Monday

Monday

Tuesday

Craig Bellamy

Dual levels of significance in Australian historical data: the case for equilibrium

This presentation will examine a number of digitised historical corpuses available to Australian researchers to illuminate their ongoing significance within digital scholarship. Many of the corpuses are available through the National Library, state and national archives, and various institutional repositories. Through contextualisation with seminal digital projects in classics and other historical disciplines, the author will examine the significance of the corpuses in both their historical setting and importance for digital scholarship. It is the contention of the author that many digital corpuses, whist emanating from significant recordings of the Australian past, have a less-significant presence online. And inversely, digital corpuses that may not record a significant historical event may, in fact, be highly significant to digital scholarship (both now and in the future). It is these 'levels of significance' between content and form that are often unbalanced in many digital projects, which may impede their longer term sustainability. Historical significance, it its many layers, is a vital determinant of sustainability. And historical significance is also largely a component of scholarly interpretation and the underlying factors that either impede or promote this. The author, through examining a number of case studies, will suggest the properties of an ideal model.

Rebecca Kippen, Janet McCalman and Sandra Silcot, and Len Smith

Roundtable on the ethics of making publicly available historical data 'more' public through linkage and database construction

Chair: Rebecca Kippen, University of Melbourne

Speakers: Janet McCalman, University of Melbourne; Sandra Silcot, University of Melbourne; Len Smith, The Australian National University

Over recent decades there has been a burgeoning of individual-level population data available from historical records, including censuses, and birth, death and marriage registers. These can be linked to other historical material to form rich prosopographical demographic datasets; that is, individual life and family histories synthesised from a variety of sources for an entire population to enable the study of that population.

The creation and analysis of these datasets raises ethical issues around individual and familial privacy. Although the data included are often technically publicly available, their linkage and inclusion in databases gives them a public profile they were unlikely to have in their former homes of archives, registers or libraries.

The speakers will discuss what responsibilities researchers have in linking and analysing historical data on individuals and families, what rights current individuals and families have over the use of their ancestral histories, and what safeguards, if any, should be put in place.

The speakers are key researchers on the 'Founders and Survivors' project, a multi-university, multi-disciplinary study tracing and analysing the life courses and genealogies of Tasmania's population from convict colonisation to World War One and beyond.

Tuesday 13th December

Catherine Ingram with Wu Meifang, Wu Pinxian, Wu Xuegui and Wu Zhicheng

Debating 'fair use' of archived recordings of minority music from the mountains of southwestern China

This paper describes and analyses public discussions within Kam (in Chinese, Dong ?) minority communities in rural southwestern China concerning potential wider online access to one major, recently established archive of Kam song and other cultural recordings. The future of much of the music-making featured in these recordings is uncertain, and this is one reason why the

creation of this sustainable digital archive (with PARADISEC) has been enthusiastically supported within Kam communities. However, many issues are currently influencing Kam debate over wider access to this important collection. Using quotations from Kam people's public discussions concerning potential 'fair use' access to this collection (with most discussions also documented on video during 2011 and to be uploaded into the same Kam archive under discussion), this paper summarises and explores Kam people's varying understandings of and views towards the proposed 'fair use' access agreement. As such, it also offers a valuable first-hand insight into custodians' own responses to archival policy and practice. Kam peoples' discussions are framed within the broader analysis I (first author, [name removed]) present of the current cultural, political, and socio-economic dynamics influential regarding both Kam minority communities in particular and recordings of traditional musics from the Asia-Pacific region in general (especially the high-profile court cases involving world-popular bands *Deep Forest* (Zemp 1996) and *Enigma* (Guy 2002, Tan forthcoming)). The paper concludes by examining points of convergence and divergence between Kam attitudes and Western archival requirements, and makes initial attempts to suggest how these might be accommodated within agreements permitting some access to the archive beyond that of the depositor (first author, [name removed]) and members of Kam communities themselves.References cited:Guy, Nancy. 2002. 'Trafficking in Taiwan Aboriginal voices.' In Sjoerd R. Jaarsma (ed.) *Handle with care: Ownership and control of ethnographic materials*. Pittsburgh: University of Pittsburgh Press, 195-209.Tan, Shzr Ee. forthcoming. *Beyond 'Innocence': An ecosystem of Aboriginal song in Taiwan*. Farnham, UK &Burlington, VT: Ashgate.Zemp, Hugo. 1996. 'The/an Ethnomusicologist and the record business.' *Yearbook for traditional music* 28: 36-55.

Stephen Morey

Documentation of traditional songs and ritual texts: issues for the production of sustainable Data

It is well established that the archiving of materials from endangered languages needs to be not just the archiving of recordings, but also a rich metadata, including, wherever possible, transcriptions, translations, and glossing of the meaningful elements in the languages which would otherwise be lost. All linguistic transcriptions and analysis face complex issues of transcription; there are always alternate ways to represent language transcriptions, such as whether certain grammatical elements should be treated as separate phonological elements, words or particles, or treated as affixes or clitics. In many cases these alternate analyses are in the purview of the linguist; with speakers of the language more or less agreeing on what the form is. With traditional songs and ritual texts, whether in oral or written form, there can be alternate analyses depending on the consultants that the linguist is working with, and these analyses can change over time. For example, when listening back to recordings of traditional ritual / sung texts, consultants sometimes interpret something different on the recording from that which is clearly audible. Which version should be transcribed? Which version is correct? The issue becomes much more complicated when the interpretation of the meaning of such texts is undertaken. And in traditional societies, the interpretations of these materials may have, and did, change over time. So how is this going to be dealt and fit in with the archivist's intention to make 'permanent' records, records that don't change over time? The idea of a 'permanent' record wouldn't have been possible in traditional societies where reanalysis and meaning changing was ongoing. Using examples from Singpho sagas (*Hka yawng ningkin*, the 'water flowing song), the Tangsa ritual songs (*Wihu Qhyoe*, the song for the earth mother) and Ahom ritual manuscripts (*Ming Mvng Lung Phai*, the text for calling back the tutelary spirit of the country), we will demonstrate and discuss these issues.

John Olstad

The Digital Nehan Songbook Project

The Digital Nehan Songbook project seeks to add native language songs to the broader project of describing and documenting Nehan, an Oceanic language spoken on Nissan Island of the Autonomous Region of Bougainville, Papua New Guinea. Like virtually all other current documentation projects, the songs and accompanying texts will be made available to the Nehan community. The logistics of making "data will be made available" a meaningful concept for the Nehan community is the topic of this paper. The language is spoken on an atoll where electricity is especially scarce and therefore the community has little access to computers and no access to the internet. However, small devices such as Chinese .mp4 players are found throughout the island and some residents even own laptops which are charged by petrol generator or solar battery. Any multimedia data should be maximally small in size and easy to share and playback. Therefore, this project looks to leave the community with a nice presentation of the songbook that is appropriate considering resource limits. The solution employed in this case is to make the standard archive-quality general recordings of the Nehan songbook but also leave behind SMIL presentations of .mp3 audio for those with laptops and .lrc (lyrics file) accompanied .mp3 files for synchronised text presentations on .mp4 players. The results are highly portable multimedia with minimal file sizes. I will discuss workflow, give an overview of common software that can be used to view SMIL presentations and the types of devices that currently support .lrc coded audio and finally, report the overall success of the project.

Rosey Billington, Simon Musgrave and John Hajek

Creating a digital wordlist for Lopit: a case study in time and motion

A vast range of techniques and technologies is now available to those working in language documentation. With so many options for the collection, storage and presentation of data, it is essential to find and maintain reproducible approaches that are focused on the accessibility and preservation of language materials, but which are also manageable within the limits of available time and resources. Here, we detail the process of producing a digitised wordlist for Lopit, a Nilo-Saharan language from South Sudan, as a useful case study of the precise resource requirements of what is ostensibly a relatively simple but still important task. The entirety of the process is covered, from the initial planning stages through to the creation of a 200 -item digital wordlist with embedded audio files, in presentation and archival form. We address the motivations for this project, both in terms of the digital wordlist format and the reasons why Lopit was selected, and how these relate to the more general motivations behind language documentation. In terms of the data collection, we explain the selection and composition of the particular wordlist, the procedure of working on this list with Lopit speakers in Melbourne up to, during and after recording sessions, and also give details of the technical considerations for the audio recordings. Our attention then shifts to methods of managing and working with the data, as well as associated hardware and software considerations, before we explain how the final wordlist is produced with embedded audio for presentation and archival form respectively. Throughout the process, our choices are informed by emerging best practices for language documentation, while mindful of the reality of being constrained by limited time and available resources. We also emphasise the need for the materials produced to be accessible and usable by researchers and non-researchers alike. We include some discussion of how materials like digital wordlists may be useful, and the potential directions for future research on Lopit. A combined timeline and tasklist of the process are presented in conjunction with our discussion and recommendations, with the hope that this information will be useful to others in making realistic and best-practice plans for this type of language documentation.

Tuesday Tuesday Wednesday Wednesday Wednesday Wednesday

Jane Hunter

Assessing the Value of Semantic Annotation Services for 3D Museum Artefacts

This paper describes the 3DSA (3D Semantic Annotation) system developed at the University of Queensland that enables users to attach tags/annotations to points, surface regions or segments of 3D digital artefacts. The 3DSA system is based on a common interoperable annotation model (the Open Annotation Collaboration (OAC) model) and uses ontology-based tags to support further semantic annotation and reasoning across digital heritage collections. By using this common model, we enable annotations to be re-used, migrated and shared - across annotation clients and across different 3D and 2.5D digital representations of a single object. Such flexibility, extensibility and interoperability are essential if cultural institutions are to interact with wide audiences that comprise users with different IT skills, client capabilities and bandwidths. This paper describes the design and functionality of the 3DSA system and evaluates it in the context of capturing community-generated tags and annotations for cultural heritage artefacts in the custody of the UQ Anthropology Museum.

Jennifer Green, Gail Woods and Ben Foley

Looking at language: appropriate design for sign resources in remote Australian Indigenous communities

Sign languages, or *iltyem-iltyem angkety,* are in daily use in Arandic speaking communities of Central Australia. They are a form of communication used alongside other semiotic systems, including speech, gesture and drawing practices. Whereas sign languages used in deaf communities operate without any connection to speech, these 'alternate' handsign languages are used in various contexts by people who also use spoken language. They are culturally valued and highly endangered, yet there has been little or no systematic documentation of Arandic sign since Kendon (1988). In this paper we describe a pilot program to record Arandic sign languages, conducted by a community language team, funded by the Maintenance of Indigenous Languages and Records (MILR) program and by the Endangered Languages Documentation Program (ELDP), and auspiced by the Batchelor Institute (BIITE). Research into various aspects of multimodal communication brings with it many theoretical and practical challenges. New technologies and the ever-expanding potentials of data annotation systems create a plethora of choices and huge volumes of recorded material. Whereas the use of film in language documentation has recently become *de rigueur,* at least in some circles, it is often only as an adjunct to studies of spoken language. When the visual is foregrounded, as it is in sign and gesture research, additional layers of complexity are added that impact on all aspects of the documentation process. How, for example, do we balance the desire for naturalistic visual data with the need for visually 'clean' images? What lessons can linguists learn from ethnocinematographers (Dimmendaal 2010)? What kinds of resources will benefit the community and a range of users (scholarly, archival, educational etc), as well as satisfying community aspirations for medium and long-term engagement with their audio-visual language materials? How do we ensure that our methodologies are robust enough to allow comparisons between primary sign language corpora and alternate sign language ones?
We discuss these issues and various others encountered in our research, including our field methodologies, annotation of film data, community consultations and ethical considerations, and issues that have arisen in designing an interactive sign language website for use as a teaching/learning resource in Arandic schools. Although the creation and management of digital archives for primary sign languages have been documented before (see Johnston &Schembri 2006), 'alternate' sign languages have received little attention.

Adam Schembri, Trevor Johnston, Jordan Fenlon, Kearsy Cormier and Ramas Rentelis

Challenges in lemmatising signed language digital video corpora: the measure of lexical frequency in Australian and British signed languages

Digital video archives of Auslan (Australian sign language) and BSL (British Sign Language) are slowly being transformed into machine-readable linguistic corpora. Each archive (Auslan 2004-2008, BSL 2008-2001) consists of data collected from deaf native and near-native signers. The datasets are being annotated using ELAN software. The majority of the video data will be made accessible online (with some limits to access for sensitive data). In this presentation, we report on the on-going studies of lexical frequency in these two signed languages—63,436 sign tokens produced in 360 clips by 109 participants in the currently annotated Auslan dataset, and 25,000 sign tokens from the corpus conversation data in the BSL dataset (500 signs each from 50 participants). Preliminary results signs indicate that between 65% and 60% of the Auslan and BSL data respectively consist of signs from the core lexicon (i.e. those signs which are highly conventionalised in form and meaning across contexts, (see Johnston, 2011, Johnston &Schembri, 1999, 2010). The next two largest categories are pointing signs (12% and 23% respectively) and signs from outside the core lexicon (i.e., gestures and sequences of enactment or 'constructed action') (6.5% and 9% respectively). The remaining number of tokens consists of fingerspelled signs (5% in both datasets), depicting constructions (i.e., depicting verbs of location, motion and/or handling, 11% and 3% respectively), and sign names (0.2 and 0.3% respectively). We discuss some of the challenges creating a lemmatised corpus of a sign language, including difficulties in differentiating core from non-core signs and sign from gesture, as well as how our work informs both sign language documentation and description specifically and linguistic theory more generally. Johnston, T. (2011). Lexical frequency in sign languages. *Journal of Deaf Studies and Deaf Education*. Johnston, T., &Schembri, A. (1999). On defining lexeme in a signed language. *Sign Language and Linguistics, 2*(2), 115-185. Johnston, T., &Schembri, A. (2010). Variation, lexicalization and grammaticalization in signed languages. *Langage et Société, 131*, 19-35.

Anthony Jukes

Culture documentation as linguistic stimulus

Along with the decline in smaller languages around the world, it is well-known that various 'traditional' or locally-based cultural and economic practices are declining or changing rapidly. This paper will show how well-filmed short videos of endangered cultural practices can be used for eliciting procedural/cultural narratives as linguistic data, as well as providing visually appealing material for ethnography, culture documentation, and cultural/eco tourism. By recording narrations as a separate soundtrack (cued by the visual stimulus) researchers are able to collect explanations by different speakers representing different age groups, genders, dialects, or in different languages from different regions or even different countries. Taking traditional usage of the sugar palm in Sulawesi, Indonesia as a test case, I demonstrate data collected in a representative sample of languages, and discuss the technical challenges of a truly multilingual multimedia corpus.

Margaret Carew

Jurra is best: Metadata design for a range of outputs from legacy recordings

This paper describes recent work with Gun-nartpa language material recorded on cassette tape in the Maningrida region in 1993-4. Returning in 2010 after a long absence I have commenced working in collaboration with Gun-nartpa speakers to digitise, log, document, repatriate and archive recordings made by elders now deceased. At the outset there were 3 key aims: (i) to repatriate the recordings in real terms, by making sure that family members have opportunities to listen to and engage with the recordings, (ii) to ensure that the recordings are archived and accessible for family members in the future and (iii) to provide a well formed corpus of recorded material for future purposes. This paper considers some of the challenges regarding access to

digital material for many in this community, and the need to provide a range of outputs according to different levels of technological capacity. For many, jurra 'paper' is still the preferred way to interact with repatriated stories, photographs and memorabilia. Notwithstanding this preference, technological uptake is rapidly increasing, and robust metadata design will enable speakers to access language materials across a range of platforms.

Myfany Turpin and Margaret Carew

'On-line' resources for off-line communities

In small communities in Central Australia, many Aboriginal language speakers do not have access to computers or internet. In such contexts, how then do we ensure research materials are 'on-line' to the people with whom we record and work? This paper considers how my Arandic song research, conducted in such small communities, can be made accessible and 'pertinent' to community members. Pertinence is 'relevant for the language community's aims and efforts for their language' (Nathan 2006). I consider how audio recordings are used and how song books are used and then discuss a collaboration with Batchelor Institute (BIITE) that aims to create an audio book of an Alyawarr women's song series. Data storage, management and retrieval play a role in how pertinent and quickly audio recordings can be made available to community stakeholders. Community requests are often for a particular set of songs, or by a particular person. These songs are embedded in larger recordings and may span many recordings. The use of iTunes, not just as a management tool but also as an annotation tool has proven efficient. A methodology for extracting songs as separate files whilst maintaining their link to the original recording, then importing and annotating in itunes, enables the exact material requested to be retrieved and burnt onto CD within a short time. These CDs are listened to in private and on some occasions played on a ghetto blaster as 'back-up' for singers at a ceremony where only a few people know the words. A new genre of expression that has emerged through a language revival program has been the local creation of songbooks. These children's songs are typed up, printed, illustrated and spiral-bound. They are used regularly in classrooms and the pictures stimulate further discussion. Although there are audio recordings of these songs, it is the books that are used as a tool for teaching the lyrics and their meanings. Furthermore, the audio and the song book are detached together. The Alyawarr song series publication aims to fulfill both needs. The book will include descriptions and images of what the songs refer to, as well as of the ceremonial designs and dancing. An accompanying audio CD will enable it to be used in other contexts. We also consider sound printing, where audio can be accessed by placing an Audio Reader over the text, as a means of reliably linking sound with text and images.

Gabrielle Gardiner, Alex Byrne, Kirsten Thorpe and Elizabeth Mulhollann

Getting it Right from the Start

Setting parameters and calibrating digital research tools at the outset of humanities research can pay dividends in facilitating the preservation of data and facilitating its reuse. This paper explores approaches to both quantitative and qualitative research with particular reference to the Indigenous research data which is managed via the Aboriginal and Torres Strait Data Archive (ATSIDA). This paper will look at the technical, ethical and logistical considerations of establishing a digital data archive. In addition, it will explore the importance of developing and maintaining relationships when establishing a trusted digital data archive, including the researched, the researchers and the technical teams. It will also examine the development of skills across the stakeholder groups and how this strengthens the relationships of the key stakeholders.

Katie Butler

Community-based website building: The Language Documentation Training Center's approach to mentor-mentee partnership

Since its founding in 2004, the Language Documentation Training Center (LDTC) at the University of Hawai'i at Mānoa has helped more than 80 participants publish web sites documenting various aspects of their native languages and cultures. LDTC's approach to language documentation is built on a mentor-mentee partnership, which is an ongoing successful model because of linguistics students who volunteer their time to teach weekly workshops and because of the international student community who consistently show interest in language documentation. This paper will provide an overview of the LDTC program and will argue in favor of community-based web site building, allowing speakers to author their own work and acquire basic skills for language documentation in the process.

Funded by departmental and university grants, LDTC holds eight 2-hour workshops in a semester which are offered free of charge for interested community members. UH linguistics graduate students volunteer to teach the weekly workshops and are partnered with one or two participants who they mentor throughout the semester. The workshops cover topics such as orthography, basic linguistic analysis, creating dictionaries using Lexique Pro, and basic web design. Advanced workshops are also offered to continuing participants who are interested in more advanced tools for linguistic analysis and web design.

Creating web sites focusing on individual languages is beneficial to both the mentor (UH linguistics graduate student) and the mentee (language consultant) as they collect both written and recorded data – such as wordlists, songs, and stories – which can be published to the web at the mentee's discretion. Some participants choose to pursue the advanced workshops to further develop their web site by adding more linguistic content as well as aesthetic design features. Beginning participants design their web sites using a WYSIWYG editor, and the student mentors provide assistance when needed. At the end of a semester, the LDTC student directors collect the web site materials from each participant and upload the pages to the university's open-access digital repository Scholar Space. The LDTC site (www.ling.hawaii.edu/~uhdoc) provides links to individual sites hosted in Scholar Space, while some participants opt to purchase their own web domain.

These web sites give participants an opportunity to showcase their language and culture to the broader online community, which in many cases, has led linguistics and anthropologist field workers as well as minority community leaders to contact past LDTC participants about their sites. This model of LDTC has proven to be rewarding for those involved, and it is the Center's hope that other universities or groups that have the means can adopt similar practices that will benefit their own communities.

Domenico Fiormonte and Desmond Schmidt

Digital representation and the use of shared texts: the case of a theatrical prompt book

The digitisation of cultural textual artefacts is focussed on conservation in digital archives and on the information retrieval tools for which they are designed. Many such artefacts, however, exist only in the form of discontinuous, unstable and stratified objects whose representation within this model presents problems for humanists and technologists alike. A less canonical vision of the past reveals that variation and instability are the constitutive elements of culture and their transmission is characterised by interaction and contamination. The digital representation thus needs an ethnographic perspective, in a cultural context interwoven with relations, exchanges and alterations. The typescript of the comedy by the author, who was also actor and director, the Roman Ettore Petrolini, which we examine here, exhibits the very characteristics that sum up

this challenge. The comedy Peppe er Pollo (Peppe the Fool) was composed in the early 1920s and performed until the end of the 1930s, in standard Italian or in the Roman dialect. The Italianisations and alterations to the substance of the text probably by actors and other agents are very numerous. What we are dealing with is a network of written and spoken traces, which cannot be linked back to a single authorial intention.
How to capture this complex stratification by different agents and how to facilitate interactions with users, readers, students and scholars is obviously a great technical challenge. We extend our earlier work on Multi-Version-Documents or MVDs for recording complex textual variation to the meta-textual properties editors typically use to structure or interpret the text. Instead of using embedded markup to record this information we represent it externally, so breaking down the humanistic information into separate categories of text, markup and images of the typescript. The markup, like the text, exists in many versions, and both are stored in separate MVDs so that the correct markup and text can be retrieved for each version. The two are then combined in the browser with links to the images of the relevant pages. This separation allows different sets of markup, e.g. structural vs. interpretative markup, to be freely mixed or interchanged, so increasing collaboration between scholars and facilitating information reuse. Conformance to embedded XML standards such as TEI (Text Encoding Initiative) is thus forced into an outer layer of the software system, as an import/export format, rather than integrated into its design.

Nick Thieberger and Rachel Nordlinger

Online presentation of media and text. Foundations for verification of analyses.

Schroeter and Thieberger (2006) reported on the development of EOPAS, a method for presenting interlinear text together with the media it transcribes, using timecodes to access points in the media correlated to corresponding chunks of text. The EOPAS model was completely rewritten in 2010 and is now an open-source tool with a working model based at the University of Melbourne. In this paper we describe the new model, illustrate its use and the workflow from transcription and interlinear annotation through to presentation. We then discuss the uses to which data in this format have been put and to which they could be put in future.

Tim Murray, Penny Cook and Conal Tuohy

Archaeological database development: the people and place project

This project, funded by ANDS in 2011, will build the platform for a national database of historical archaeological collections, excavated sites and the people connected to those objects and places; to be called the Australian Historical Archaeology Database.
Each year archaeologists (both academic and private consultants) excavate tens of thousands of artefacts from historical archaeological sites across Australia. While some states (e.g. Victoria) require catalogues to be prepared in a standard format, the majority of catalogue data are stored in small, standalone spreadsheets or custom-built databases, and few are made freely available. There is no central register of these individual datasets, and many significant collections are simply unknown to archaeological researchers.
Between 2001 and 2004, the ARC-Linkage project 'Exploring the Archaeology of the Modern City' (EAMC), led by Murray, created two research databases <http://www.latrobe.edu.au/amc/> which offered, for the first time, a central database of 700,000 artefacts from multiple historical archaeological sites and a companion dataset of historical occupancy data relating to 2200 individuals who occupied those sites. While extensive in their data content, the databases themselves too limited in their structure and require significant design input to make them truly effective tools for managing and sharing historical archaeological data.

Wednesday Wednesday Wednesday Wednesday

This project will federate the two EAMC databases, undertake additional data auditing, and seek new datasets from the private, public and tertiary sectors. It will enable researchers to access a vast dataset that is currently unavailable to them, and provide the platform for future datasets to be made freely available in a standardised and timely fashion. Our paper reports progress and will conclude by discussing several significant challenges:- accessing and integrating privately funded research data created for purposes of public compliance and 'future research';
- integrating data from different archival sources, ie artefacts and documents, through the fundamental lynchpins of people and place;
- the social/IP challenges of creating a platform for information exchange (online research community) that effectively collates multiple recording systems while respectfully presenting divergent interpretations of those data.

Kerry Kilner and Roger Osborne

Bringing Research and Researchers to Light: Current and Emerging Challenges for Discipline-based Knowledge Resources

Australian literary studies have, in the past decade, been greatly assisted by AustLit: The Australian Literature Resource (www.austlit.edu.au), a multi-institutional collaboration between researchers, librarians and software designers from ten universities and the National Library of Australia. Under the leadership of The University of Queensland, this collaboration has produced a web-based research environment that supports a wide range of projects and publications across a diverse array of fields in Australian literary and narrative cultures while also becoming a key resource for teaching and general information. AustLit has consistently worked to integrate the research output of associated projects and is currently planning to expand its position in the community with a new open access and open contribution model. A major innovation in data management and maintenance, the AustLit Research Community[1] structure supports the study of Australian literary and story-making cultures by providing a web-based environment where segments of these cultures can be explored and presented as distinct topics within a larger knowledge framework. Scholars are able to build datasets, annotate, analyse and present that data in a range of ways, and publish scholarly interpretations of their findings in the form of peer reviewed articles. The incorporation of these research-rich datasets into AustLit contributes to an overarching goal of building a comprehensive database of information about Australian writers, writing and print culture more broadly. With a recent decision to move from the current access model as a subscription service, available to relatively few users, to an open access and open contributions model incorporating content produced by a network of volunteers, AustLit is now facing a significant new challenge. The Aus-e-Lit Project[2] has delivered innovative tools and services that will enable AustLit users to engage more directly with AustLit data and to contribute to a Research Commons with collaborative annotations and richly described collections of internet resources. This paper will report on the implications that these innovations bring to current and future research practices. It will consider the successes and challenges that AustLit faces with its aim to be the definitive virtual research environment and information resource for Australian literary, print, and narrative culture, not only for scholars in the field but for students of all levels and the general public.

[1] See www.austlit.edu.au/ResearchCommunities

[2] The Aus-e-Lit project is funded from 2008 - 2011 by the National Collaborative Research Infrastructure Strategy (NCRIS) Platforms for Collaboration, through the National eResearch Architecture Taskforce (NeAT), and by the University of Queensland.