On the reach of digital language archives

David Nathan

Dept Zoology, University of Oxford

djn@soas.ac.uk

www.dnathan.com

Abstract. Over the last decade, through digital media and technologies, archives for endangered and minority languages have extended their activities from preservation to dissemination. While the dominant focus has been on 'discovery', this paper embraces a broader term 'reach' and identifies ten components: (1) *acquisition*; (2) *audiences*; (3) *discovery*; (4) *delivery*; (5) *access management*; (6) *information accessibility*; (7) *promotion*; (8) *communication ecology*; (9) *feedback channels*; and (10) *temporal reach*. Through considering how archives are approaching these, we can see that innovative archives are making a transition from being repositories of memory to being facilities for fostering participation and knowledge sharing.

Keywords: archives, access, audiences. users, social media

Introduction[1]

The aim of this paper is to extend previous work on archival 'access and accessibility' (Nathan 2013) in order to make initial suggestions towards a set of criteria for thinking about archives' 'reach' – their multifaceted capacity to successfully provide language resources to those who can gain value from them. Several of our archives now think of themselves as publishers (Holton, this volume, Nathan 2011b), which leads naturally to thinking about intended audiences and the appropriateness and usability of the archives' materials and services.

The origins of this theme can be traced to the Open Archive Information Systems (OAIS) project initiated by the Consultative Committee for Space Data Systems in the 1990s (CCSDS 2012, OAIS 2012; the CCSDS committee currently has 11 members, including NASA, the European Space Agency, and similar agencies from Canada, China, Japan, Russia and several European countries). The committee's context was a need to deal with massively accruing digital data from space programs, at the same time as preservation strategies were diverging, or worse:

> Problems had often stemmed from terms—such as archives/archiving or metadata—that were used so widely and for so many different purposes that it was difficult to determine if they were being used in the same way by different actors. The combination of pressing need, available

[1] I would like to thank two anonymous reviewers and Peter Austin, Stephen Bird and Birgit Hellwig for valuable comments and corrections to this paper and the presentation on which it is based. However, I am solely responsible for all remaining errors and provocations.

> expertise, and inconsistent language meant the time was ripe for developing a reference model that could codify and support greater consistency in discussions of digital archives[2]

(Lee 2010: 4021). Recognition of these wider problems, and the goal of establishing a "common framework of terms and concepts" rather than specific designs or implementations (CCDSD 2012: iii, 1-3) led to their activity and impact reaching far beyond the scope of space data to "become a fundamental component of digital archive research and development in a variety of disciplines" (Lee 2010: 4020).

The OAIS Reference Model recognises, in addition to long-term preservation, the importance of data dissemination and availability, and archives' accountability to their users and stakeholders. These concepts are expressed in relation to 'data consumers', and in particular *designated communities*:

> [a] special class of Consumers is the Designated Community. The Designated Community is the set of Consumers who should be able to understand the preserved information. ... [i.e. information expressed] in a form that is understandable using the recipient's Knowledge Base. The Designated Community, and its associated Knowledge Base, for whom the information is being preserved by the Archive is defined by that Archive, and that Knowledge Base will, as described below, change over time. The definition of Designated Community may be subject to agreement with funders and other stakeholders.

(CCSDS 2012: 2-3).

Figure 1 illustrates two aspects of the OAIS model that I wish to expand on in this paper. Firstly, the term 'designated communities' highlights the importance of archives being explicit about *who* they serve and in turn *how* they do so; but while many archives pay homage to the OAIS model (Nathan 2011a),[3] few actually make identifying, understanding, and appropriately serving audiences a significant part of their scientific endeavour (see below 2. *audiences*, 4. *delivery,* 5. *access management,* 6. *information accessibility,* and 9. *feedback channels*). Secondly, notice the essentially linear progression from depositor ('producer') to archive and then to consumers/users – an architecture now superseded by the today's potent combination of ethically-based community inclusion in research and current social-networking technologies that enable wider participation (see below 2. *audiences*; 3. *discovery*; 5. *access management*; 7. *promotion*; 8. *communication ecology*; and 9. *feedback channels*).

---

[2] Some readers will recognise some of these problems as still remaining to be solved – or perhaps being recapitulated – for the archiving of language documentation.

[3] See also http://dobes.mpi.nl/meetings/aab-meeting-report-nov-05-v2.pdf and http://www.robertmunro.com/research/munro05elar.ppt. The Data Seal of Approval evaluative scheme (for details, see below), for example, requires approved archives to have "technical infrastructure [which] explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS" – see Guideline #13 of the DSA Guidelines at http://datasealofapproval.org/en/information/guidelines/. Interestingly, this guideline itself seems to flout the OAIS Model's focus on concepts and process architecture rather than infrastructure engineering (cf Lee 2010: 4025).

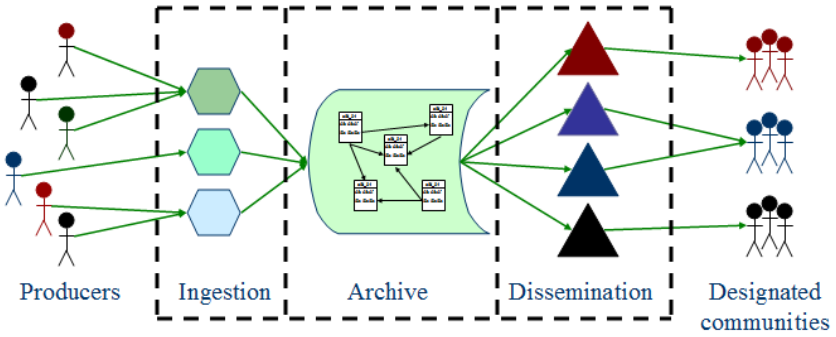


Figure 1. The OAIS model proposes 3 types of packages: ingestion, archival, and dissemination. There can be multiple dissemination packages to serve the archive's "designated communities."

I have borrowed the term 'reach' from Chang (2010) who uses it as a subordinate category in her 'TAPS' grid of evaluative criteria for archives.[4] There are several other evaluative systems that digital archives can use to claim and demonstrate conformance to standards and good practices, including the NINCH guide (NINCH 2002), DRAMBORA,[5] and the Data Seal of Approval.[6] While these are largely focussed on policies. strategies, resources, and technologies for digital preservation, TAPS also included a criterion addressing access and relevance to the archives' intended audiences, recalling the OAIS architectural principle devoted to identifying, understanding, and serving users. The components of 'reach', described below, should be seen as complementary to these existing evaluative schemes.

The body of this paper unpacks 'reach' into a set of ten criteria, illustrating them by examples from some of the DELAMAN archives:[8]

(1) *acquisition,* the archive's collection policies and its acquisition processes and resources;

(2) 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;

(3) *discovery*, drawing on the understandings of audiences in order to help them browse, navigate, search, identify and select their items of interest;

[4] Chang calls her checklist 'TAPS' which is an acronym for Target, Access, Preservation, and Sustainability.

[5] Digital Repository Audit Method based on Risk Assessment. http://www.repositoryaudit.eu [accessed 1 April 2014].

[6] http://www.datasealofapproval.org/en [accessed 1 April 2014].

[8] The archives mentioned in this paper are: AILLA (Archive of the Indigenous Languages of Latin America), ANLA (Alaska Native Language Archive), ELAR (Endangered Languages Archive), DoBeS (Dokumentation Bedrohter Sprachen), and Paradisec (Pacific and Regional Archive for Digital Sources in Endangered Cultures). See http://www.delaman.org/members for details. For DELAMAN, see http://www.delaman.org [accessed 1 April 2014].

(4) *delivery*, i.e. making available selected resources according to users' preferences whether by download, view-in-browser, through apps or other means;

(5) *access management* such that resource delivery follows depositors' and communities' preferences, and where users have ways of applying for and negotiating for access;

(6) *information accessibility*, where the actual desired content is accessible to users, whether in terms of contextualisation or appropriate complexity, language, or modality;

(7) *promotion*; raising the profile of archive deposits and activities, and bringing 'outreach' versions to the intended (or new) audiences;

(8) *communication ecology*; the place of archives' core activities within growing media and informational environments;

(9) *feedback channels*, where users can utilise the archive to provide feedback to depositors or to enhance deposits with user-generated content; and

(10) *temporal reach,* where long term preservation seems to be at odds with today's 'short-termism' of funders and the (apparent) ephemerality of digital media.

Through considering how archives are providing such services, we can see a transition from being repositories of memory to being facilities for fostering participation and understanding.

The ten components of reach

1. Acquisition

Users are drawn to archives when they expect to find resources relevant to their needs. The clarity of an archive's collection and acquisition policies (Conathan 2011: 240) and the vigour with which it seeks new materials will thus draw users, increase usage, and provide regular update topics for announcements (which can be disseminated through the archive's 'information ecology', see below).

The Paradisec archive[9], for example, actively invites and seeks out legacy analogue materials that are vulnerable or valuable[10], thereby increasing its coverage and relevance to users. Acquisition for the ELAR and DoBeS archives[11], by contrast, is largely driven by

[9] See http://paradisec.org.au [accessed 15 April 2014].

[10] See, for example, http://www.paradisec.org.au/blog/2014/04/paradisec-stats-for-2014 [accessed 25 April 2014].

[11] ELAR = The Endangered Languages Archive at http://elar-archive.org; DoBeS = the DOBES archive at http://dobes.mpi.nl.

associated grant-giving – as of early 2014, 90% of ELAR's incoming materials were from ELDP grantees.

2. Audiences

If the mission of archives is preserving and disseminating resources, then audiences are their *sine qua non*. We can think of audiences as being the sum of all individuals that access collections over their entire lifespan, or as aggregated 'types' based on certain shared criteria (such as 'researchers', 'community members' and suchlike). We can alternatively think of audiences as being those using archives today, or those in the (possibly distant) future who discover and access materials, if the archives have fulfilled their preservation role (Woodbury 2014:1, Holton 2013).[12]

Whether thinking of individuals with varied motivations and literacies, or groups who have particular preferences or constraints (e.g. language and other skills, availability of computers etc), effective reach will take into account whether the archives provide suitable content versions and appropriate ways of searching, browsing, viewing and downloading (see *4. Delivery* and *6. Information accessibility* below for more on different methods of delivery and alternative versions of content, also OAIS 2012, Nathan 2006).

How well do archives know their audiences? Audiences are fundamental to what archives do, and archives should take a scientific approach to defining, researching, describing, serving, and reporting about them. Yet, it appears that some language archives take a peremptory approach to audiences, sometimes in contrast to the careful attention they pay to technical issues. Schwiertz (2012: 126) for example, describes DoBeS archive's attempt to address the limitations of its navigational interface:

> When considering the exploitation of language documentation data contained in language archives, three major user groups can be identified: The speaker community, the scientific community, i.e. linguists and scholars of related disciplines, and the general public. Each of these user groups has different interests and different needs, all of which are hardly satisfied by the IMDI-tree representation of the DoBeS archive. For the community users, community portals have been created in some projects. ... we have [also] created a general portal to the DoBeS archive.[13]

[12] Some might include other stakeholders such as funders and host institutions as audiences as well – and increasingly those who require reports about research, e.g. Australians report on their archive deposits for research evaluation in the ERA system. However, for the purposes of this paper, I do not include these categories; while those stakeholders might be those that we are required to 'play to' to sustain our existence, they are not our *reason d'etre*. Funding sources come and go according to fashion or particular funders' strategies, but archives' collections have enduring value.

[13] Schwiertz is the latest of several authors to write about user groups in this way. Wittenburg (2002:36) writes: "Besides the linguists, ethnologists and other researchers we see interests from school and university educators, journalists, and especially from the indigenous people themselves." Farrer and Langendoen (2003:97) arbitrarily identify linguists, indigenous communities, and language learners as groups who will gain

This looks like an admirable advance, but we might ask whether it is sufficient to simply proclaim the reality of these 'major user groups'? Are there other yet undiscovered user groups? What research took place? What shared properties define these 'groups'? How is the archive collecting and reporting evidence about usage by these groups; what counts as serving them, and how well are they being served? How is the archive improving its methods and services based on its growing understanding of these putative groups?

ELAR requires users to register and create a basic profile. Answers to the profile question asking registrants to describe their connection with endangered languages inform the archive about the proportion of its users who are community members, researchers, and professionals in particular disciplines, and about their affiliations, motivations, heritages, interests, and language activities.

It is easier for archives with more specific areal coverage and targeted collection policies to be transparent about the users they say they serve. For example, the Archive of Indigenous Languages of Latin America[14] offers its interface in Spanish which is a *lingua franca* for the region it serves. It is not feasible for international archives like ELAR or DoBeS to provide interfaces to serve all their audiences, however, at the level of the individual deposit, depositors can be encouraged to provide metadata and descriptive information in the subject language of each deposit, or a relevant *lingua franca*. For example, the Movima deposit in DoBeS has metadata and descriptive material in Spanish.[15] Shenkai Zhang, ELAR depositor of Pinjiang love songs[16] edited her deposit's home page to provide contextual information in Chinese to help facilitate access to the Pinjiang community from which these songs come (see Figure 2). See also below under *6. Information accessibility* for Eli Timan's ELAR collection[17] which forgoes analytical linguistic content to provide what Timan, a community member himself, understands that his community wants: transliteration in Arabic, translation into English, and pictures drawn by the story teller.[18]

Considering language choice in the context of audiences highlights the fact that by typically presenting services in a given language (usually English), archives are either making a (probably covert) assumption that English is a *lingua franca* for their audiences or else simply imposing English as a condition of using the archive.

Other audience-related factors include what modalities people would like to access materials in, their computer and literacy styles and preferences, and what computer

---

from web access to linguistic resources. They urge data producers to adopt ontology and the 'semantic web' which would seem to have limited benefit to most of these groups.

[14] See http://www.ailla.utexas.org [accessed 1 April 2014].

[15] See http://dobes.mpi.nl/projects/movima [accessed 13 August 2014].

[16] See http://elar.soas.ac.uk/deposit/0079 [accessed 1 April 2014].

[17] See http://elar.soas.ac.uk/deposit/timan2008jewishiraqi [accessed 1 April 2014].

[18] See these materials at http://jewsofiraq.com [accessed 1 April 2014].

hardware, software and connectivity they have available. Without appropriate research, archives may be insufficiently aware of these factors, or even whether factors pattern with groups or vary more according to the individual user.[19]

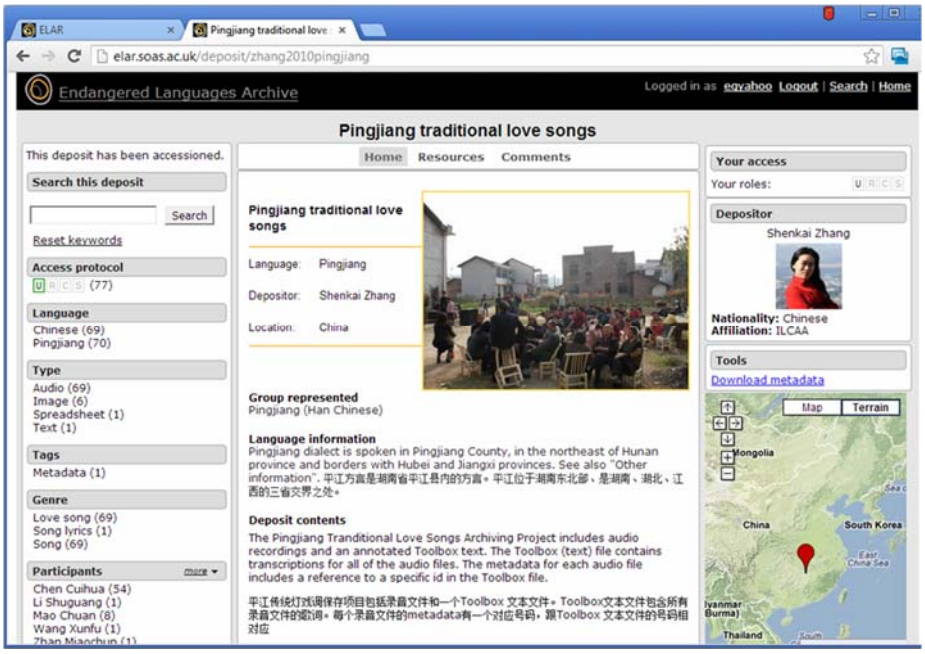


Figure 2. Shenkai Zhang has edited her ELAR deposit to add Chinese text to aid access by language community members.

3. Discovery

Archives have a long-standing tradition of providing methods for helping users to find materials; in archive-speak this is usually called resource discovery (Bird & Simons 2003). There are a variety of methods, from search over cataloguing metadata to additional finding aids produced by curators. Discovery strategies (and users' expectations) are shifted in the digital domain. On one hand, discovery is facilitated by the ability of computers to support search over large amounts of text material. On the other hand, language archives increasingly contain large amounts of media (audio, images, video) which are generally as opaque to computers as uncatalogued objects were in a traditional archive, so metadata (labelling and description of content) are as crucial as ever for resource discovery.

Debates about the extent to which metadata categories need to be standardised have largely given way to concern for encouraging depositors to provide as broad and deep metadata and metadocumentation as possible (Austin 2013), since these represent the

[19] It might be objected that research of these factors might "run into the same language and accessibility issues" which this paper identifies as obstacles to 'reach' (I am grateful to an anonymous reviewer for raising this). While it is beyond the scope of the paper to suggest a full research program, I would suggest that much might be learned if a small fraction of the time and intellectual rigour applied to language documentation and analysis were applied to scientific investigation of a community's preferences, skills and receptiveness to various kinds of language materials. Such research could even be recognised as part of documentation methodology (Nathan and Fang 2014:53).

unique, irreplaceable knowledge that only depositors are likely to possess, and are the keys to carrying a profound understanding of the materials into the future, for future users and usages. Rich metadata, when combined with multiple languages and well-designed interfaces to facilitate search and browse, increase an archive's reach to a greater range of users.

Archives need to understand audiences in order to provide a range of ways for them to search, browse, and navigate effectively to materials they are interested in. While we may make generalisations about real or imagined user groups, it would seem a good starting point for online catalogues to take best advantage of known and effective digital genres. ELAR designed its catalogue to use some of the contemporary visual and interactive methods of social networking applications (e.g. Facebook), a decision that has been validated by various fieldworkers reports that many language communities have recently and rapidly acquired access to the internet with predominant use of social apps on mobile devices.

Providing discovery mechanisms means more than presenting users with search screens allowing them to search 'thin' metadata (Nathan & Austin 2005). This is especially important for endangered languages, where language names can vary widely due to spelling variations or by being expressed in different languages or as exonyms or endonyms, where not all of the 'target audience' (but see *2. Audiences* above) are likely to have relevant literacies, where the materials tend to be at the edges of mainstream knowledge rather than the centre, and where certain users are simply fishing about out of interest rather than being focused on finding particular linguistic material. Thus it is important to provide ways to discover what is available in the archive through browsing. Browsing, such as illustrated in Figure 3, enables users to recognise and select items, even randomly.

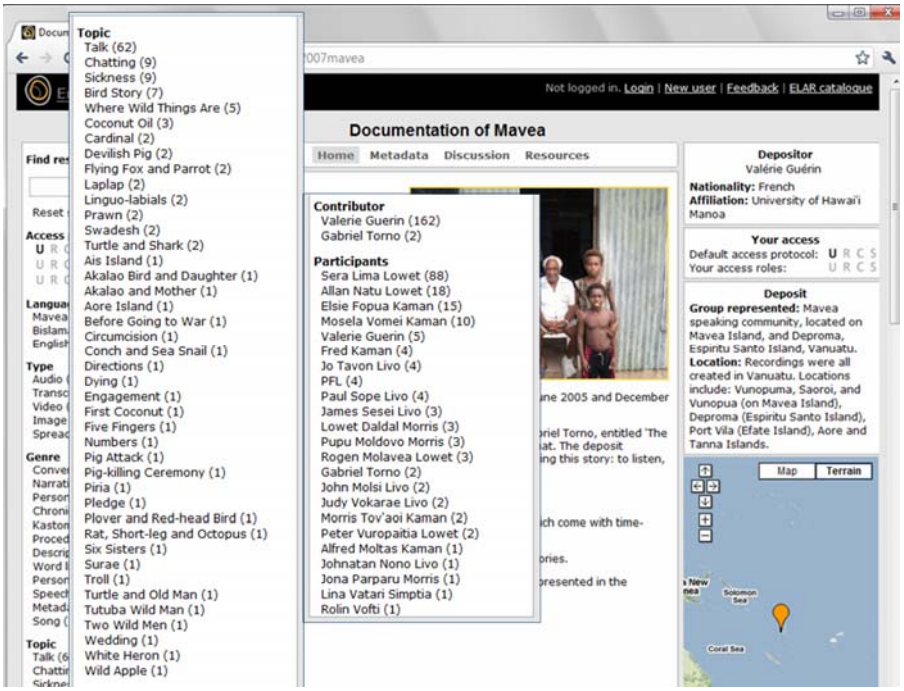


Figure 3. Showing the range of terms available in faceted browse of Valerie Guerin's ELAR deposit (the scrolled sections have been pasted onto this image). See http://elar.soas.ac.uk/deposit/0015.

Many archives now also provide maps to enable discovery by browsing according to location. This has many advantages; it lessens dependence on traditional literacies, it encourages serendipitous discovery, it better supports people who 'think visually', and it conveys additional information such as proximity and clustering of materials and likely landform/environment information.

Archives can also join with others by 'federating' their discovery mechanisms, i.e. sharing and pooling some or all of their metadata so that users can search or browse a larger virtual collection without having to know (at least initially) where a given resource is actually located (Broeder et al 2008). The best known example in the language documentation field is the Open Languages Archive Community (OLAC) catalogue.[20]

Archives' choices – whether explicit or not – about metadata and interfaces control users' ability to discover materials they are looking for, and/or discover materials they were not previously aware of but prove to be interesting or valuable to them. Constraining discovery strategies to structured search via standard, English, academic-centric categories and pre-defined ontologies can limit the reach of archives.

4. Delivery

This criterion is concerned with how a resource, typically a file, is actually delivered to a user. Whether the resource file is text, audio or video, it may be offered for download, or it may be shown directly in the browser or in some kind of browser-embedded player (e.g. a media-player plug-in). The best option for a user will depend on their purposes, skills, devices, software, and internet connection. Viewing a video sample in the browser may be preferable because downloading the whole file would entail a high data cost, especially on mobile; another user might download a video file but not know how to play it. On the other hand, some users will want to view or work with the video or view it later offline.

Consider the choices available for viewing media annotation files produced using ELAN software.[21] Data files produced by ELAN are encoded as opaque XML structures which can only effectively be viewed using bespoke software (typically, the ELAN software itself). Those who want to work with the detail of an ELAN file will likely have ELAN installed and will want to download the ELAN file.[22] On the other hand, those who do not have ELAN (or the correct version of it) installed, who are not interested in technical annotations, or who do not have the skills, time or motivation to find and install software, would rather simply view some version of the material in their browser. To serve them,

[20] See http://search.language-archives.org/index.html [accessed 14 April 2014].

[21] See http://tla.mpi.nl/tools/tla-tools/elan [accessed 1 April 2014].

[22] In fact, such users will also need certain configuration files to display the ELAN file as its producer intended. Peter Austin reports that the situation is even worse with Toolbox files because without the correct .typ and .lng control files none of the 'aligned interlinear text' will actually align, with the result that the informational links between content on adjacent lines will be lost to all but linguists afflicted with Interlinear Glossing Blindness (see blog post at http://www.paradisec.org.au/blog/2012/04/hammers-and-nails).

DoBeS (aka The Language Archive, the authors of ELAN) created a browser plug-in 'Annex' (Annotation Explorer; see Figure 4).[23] Other software developers have also created ELAN content viewers – see *6. Information accessibility*.

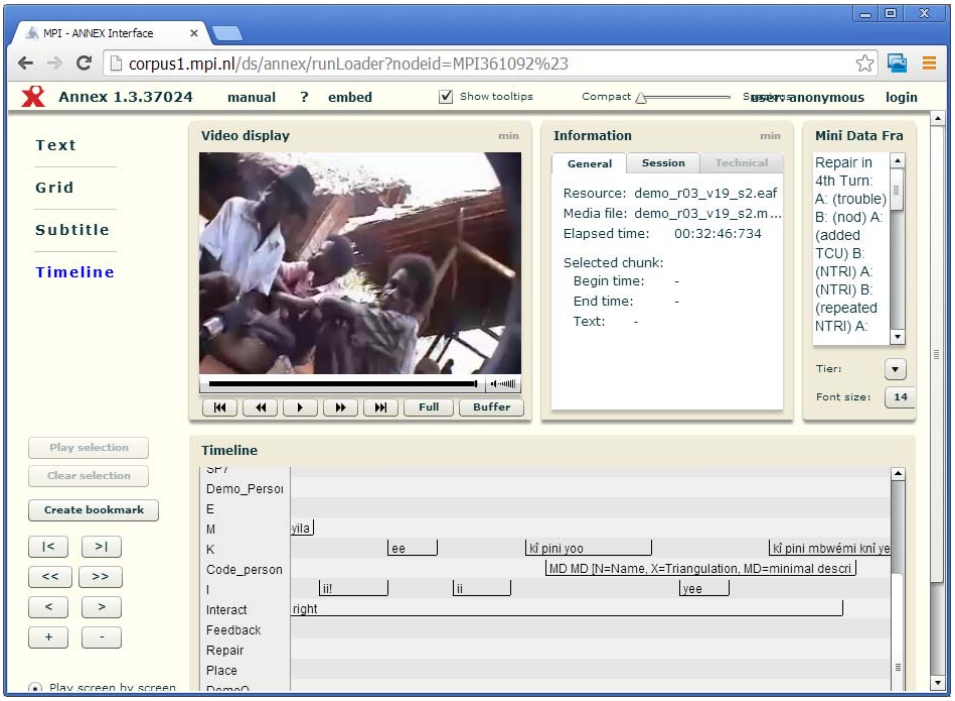


Figure 4. Annex runs as a browser plugin, requiring no software installation.

Archives can extend the reach of their resources by providing different ways of delivering them. This is going to become increasingly important as more non-western communities catch up with, if not overtake, western modes of using the internet, often in different ways, such as solely through mobile devices.

5. Access management

Today, matters of privacy and control of personal information are of increasing general concern. Such concerns are amplified in the case of recordings of endangered languages. Endangered language communities and their speakers are typically under various pressures and deprivations that are often also contributing causes to the decline of their languages and cultures. These difficulties are amplified by the methodologies of documentary linguistics, which most highly values the recording of spontaneous and conversational speech. As the contexts in which languages are spoken decrease (which is what primarily drives endangerment), people tend to use their languages more and more to speak of private, local, sensitive and secret matters. So the primary data of documentary linguistics maximizes the likelihood of it including content that can cause embarrassment or harm to the recorded speakers. As a result, it is broadly agreed among endangered languages documenters and archives that they need to collect, preserve and disseminate materials in accordance with the wishes of the information providers and their communities (Rice

[23] See http://tla.mpi.nl/tools/tla-tools/annex [accessed 1 April 2014].

2011, Austin & Grenoble 2007).[24]

The ELAR archive developed an approach to access management that locates it within a larger framework called *access protocol*. This term refers to the sum of processes extending from the beginning of documentation activity, e.g. starting when a documenter seeks informed consent from speakers, and then collects metadata on the rights and sensitivities associated with documentation materials, through to the mechanisms for dynamically providing, restricting, or negotiating about access to archived materials. It involves careful attention to how the interface represents and guides users around both accessible and controlled-access materials,[25] and it includes methods for negotiating about access, and detailed reporting to depositors and others. It is described in detail in Nathan 2010.

Respect of privacy and control of personal information impose legal as well as ethical obligations. Therefore it is important that an archive's policies and mechanisms for safeguarding access, and its methods for processing and deciding access applications, are transparent, accountable, and ethically and legally sound.

6. Information accessibility

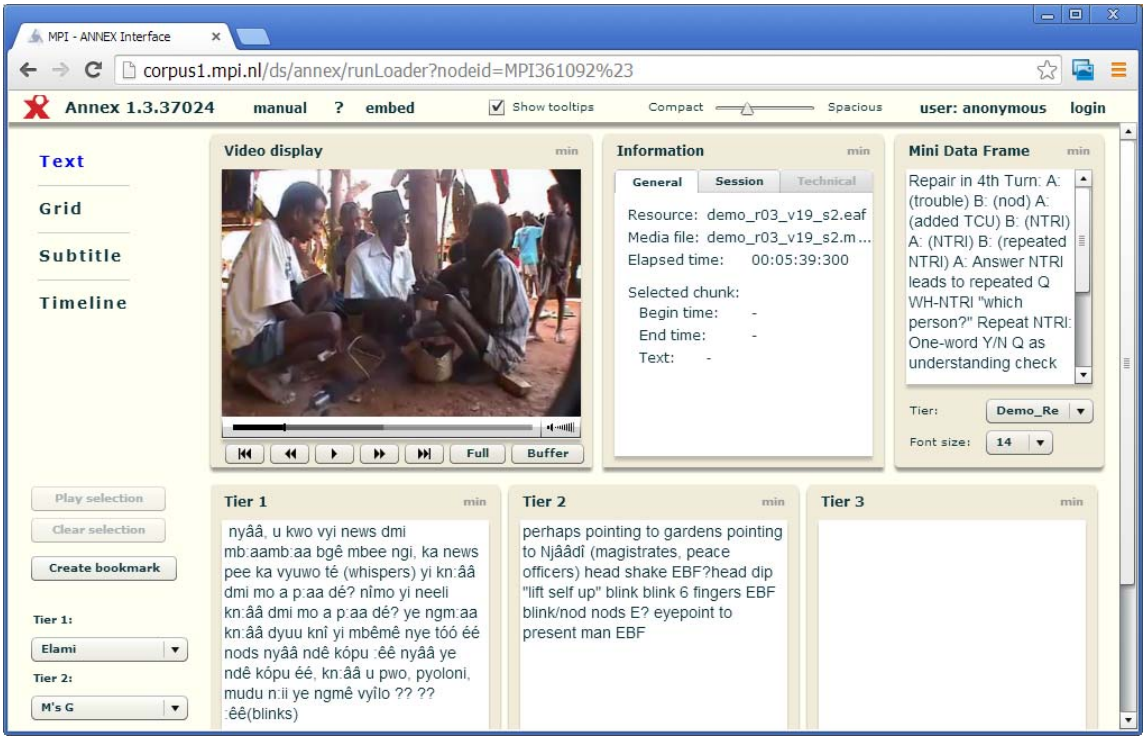


Figure 5. Annex displaying a simpler view of an ELAN file.

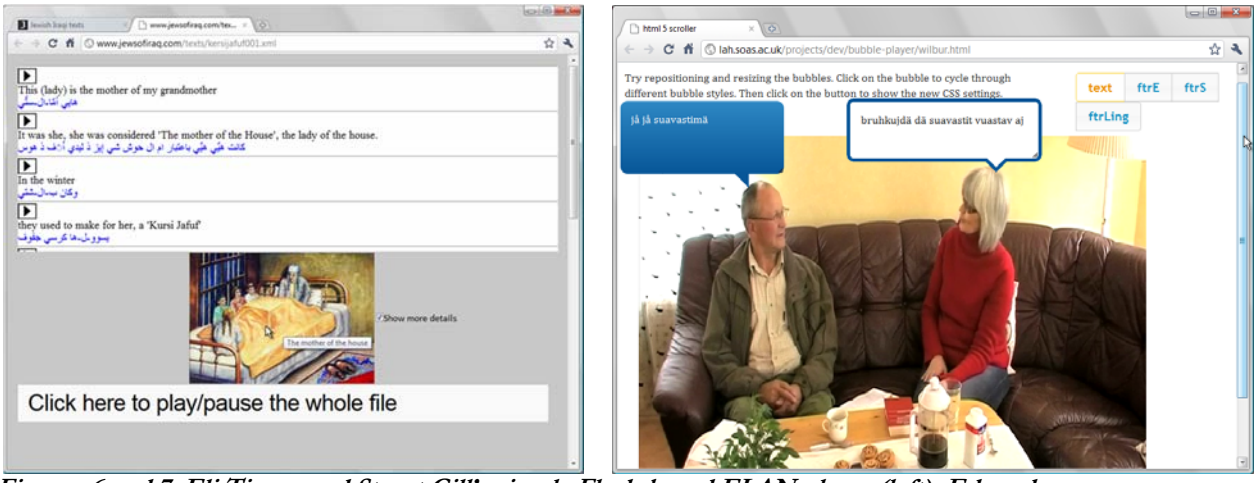
A user might be delivered a resource (see *4. Delivery* above), perhaps after negotiating

[24] This stance has recently come under pressure from funders campaigning for their variant of 'Open Access'.

[25] Prior to 2014, ELAR's catalogue interface provided coloured labels clearly showing any user which resources s/he could and could not access. Helpfully for users, these included navigational controls, which enabled users to restrict a search or browse to only those materials that they could (or could not) access. Following a funder's campaign for 'Open Access', ELAR staff felt pressured to modify this system.

access to it (see *5. Access management* above), but there remains the question: how accessible is the actual *content* of that resource? Consider again the case of an ELAN file, discussed above; it might have a wealth of linguistic detail, but for some users that detail can obscure a more simple experience or bit of information they are after. To help, Annex provides alternative views of the data, for example by showing simpler text versions of the content (see Figure 5).

Figures 6 and 7 show two simpler applications that provide alternative renderings of the information contained in an ELAN file. Figure 6 shows an example from Eli Timan, who documented Jewish Iraqi (a dialect of Arabic), and worked with Stuart McGill to develop a Flash app that runs in a browser.[26] The app draws data from an ELAN file but bypasses the more complex ELAN software to show a simple display that synchronises the audio with scrolling Arabic orthography and English translation. According to Timan, himself a member of the Jewish Iraqi community, this makes the relevant information more accessible to his target audience for the materials.[27]



Figures 6 and 7. Eli Timan and Stuart Gill's simple Flash-based ELAN player (left). Edward Garrett's speech bubble player (right).[28]

Another more adventurous example, developed by Edward Garrett using HTML5, is a speech bubble player.[29] This player selects and pulls data from an ELAN file, and presents it in a familiar comic-book style. A user can 'play' with the speech bubbles, manipulating the interface not only in terms of the linguistic data but in terms of how the display is

[26] The app is similar to Christopher Cox's CuPED; see http://sweet.artsrn.ualberta.ca/cdcox/cuped [accessed 1 April 2014].

[27] See an example at http://jewsofiraq.com/texts/shlomo_kuwaity1.xml#shlomo_kuwaity1.008 [accessed 30 April 2014].

[28] The speakers, conversing in Pite Saami, are Henning Rankvist (left) and Elsy Rankvist (right). From an ELAR collection deposited by Joshua Wilbur, *Pite Saami: documenting the language and culture* http://elar.soas.ac.uk/deposit/wilbur2009pitesaami [accessed 17 August 2012].

[29] See http://lah.soas.ac.uk/projects/dev/bubble-player/wilbur.html [accessed 30 April 2014].

composed and experienced (a 'thick interface' in the terms of Nathan 2006).

I recount here an interesting audience reaction to demonstration of the speech bubble player during the presentation of the talk on which this paper is based. When Garrett's speech bubble player was being demonstrated in morpheme-by-morpheme mode (representing speech content as interlinearised/glossed) several audience members burst into laughter. It took me a few moments to realise that what was amusing them was the dissonance between on the one hand watching video of people in informal conversation and with their speech visualised in speech bubbles, while on the other hand the *content* of their speech was rendered as analytical morpheme-by-morpheme stuff. The friendly video/speech bubble view clashed with the 'technical' interlinearisations. Oddly, perhaps, I had never before seen anybody respond this way despite many years of viewing materials together with others in purely 'technical' contexts such as ELAN. This audience response suggests a challenge to the way we routinely render language events as de-contextualised and a-social without a second thought as to the transformation that we have imposed.

Although in both cases illustrated above the original data file is an ELAN file, the same principle of multiple content-rendering, as per the OAIS model, applies across many types of files and content. For example, a video can be provided with subtitles in a variety of languages (or with varied levels of detail in the transcription or annotation, cf Jukes 2011); a text file could be presented as a print-ready PDF document or in very large font to aid the vision-impaired or elderly. An audio file could be represented spatially and labelled by keywords or images representing topics being spoken about so that a user can easily navigate to sections of interest. Such considerations raise questions about the resources required to produce multiple dissemination versions, and it is an index of the infancy of our field that it is not at all clear whether the onus lies with the archive itself, with the producers/depositors, or even the eventual consumers. In favour of the onus falling on the archive is the OAIS Reference Model, which assigns to archives decisions about 'designated communities' and thus the materials appropriate to them (in addition, an archive can potentially amortise investment in methods across multiple deposits). On the other hand, producers/depositors are most likely to know best about the nature of the materials and their key user communities, and they may have other motivations for reworking materials. Leaving the burden to the eventual consumers is the default but inexplicitly stated scenario for most present archives.

7. Promotion

Archives can increase their reach by raising awareness of their services and activities amongst both existing and new audiences. Up till recently, the activities of endangered languages archives have mainly been disseminated within linguistics and related fields through conferences, workshops, articles, and websites. A few endangered-language-related projects have managed to receive significant mainstream press attention, including the Endangered Languages Alliance[31] (whose stories have been picked up by the New York

[31] See http://elalliance.org.

Times), the (Google-initiated) Endangered Languages Project[32] (which briefly made news in several major newspapers), the Living Tongues Institute[33] (funded and promoted by National Geographic), and the World Oral Literature Project[34] (whose Director, Mark Turin, has appeared in BBC documentaries). However, these are not archives, which raises questions of whether archives are generally too absorbed in their curatorial, preservation or technical services, whether the term 'archive' turns off users, and whether archives should partner with more 'sexy' and outgoing projects like those mentioned or with institutions experienced in outreach such as the British Library or the Smithsonian Institution.

Nevertheless, archive activities can draw wider interest. For example, 'Endangered Languages Week', an outreach event originally initiated by the Endangered Languages Archive and the Endangered Languages Academic Programme at HRELP[35], drew up to 1,000 students, staff and visitors annually to events targeted at a wide range of disciplines and the wider public. In some years, a parallel event was run at other institutions, and during its lifetime from 2007 to 2013, HRELP's Endangered Languages Week came to be seen as a fixture in the calendar for those interested in languages broadly.[36]

There are other opportunities for raising awareness and usage of our archives amongst students and particular language communities. Recently Adam Schembri, depositor (together with Trevor Johnston) of the AUSLAN corpus in ELAR[37], posted a series of Facebook messages about the corpus, and following those posts the staff at ELAR noticed an increase in the rate of archive user registrations and archive accesses. Gary Holton (2012) reports a similar upsurge in community interest as a result of online communications about archive materials.

Joshua Wilbur widened awareness of and access to his Pite Saami materials deposited with ELAR[38] by working with local archives in Sweden to encourage and help them hold language materials so Saami community members can more easily access them (Wilbur 2014).

An archive may decide to locally promote particular deposits in order to attract users. For example the DoBeS archive entry page prominently features attractive videos, thus literally

---

[32] See http://www.endangeredlanguages.com.

[33] See http://www.livingtongues.org.

[34] See http://www.oralliterature.org.

[35] The Hans Rausing Endangered Languages Project at SOAS, which was originally established with three components: the Academic Programme, the Documentation (funding) Programme, and the Archive Programme.

[36] For more details about ELW, see http://www.hrelp.org/events and the annual reports at http://www.hrelp.org/publications/newsletter.

[37] See http:// elar.soas.ac.uk/deposit/0001 [accessed 25 April 2014].

[38] See http://elar.soas.ac.uk/deposit/0053 [accessed 1 April 2014].

promoting the featured deposits.[39] ELAR sponsored a short series of blog posts by postgraduate intern Zander Zambas titled *Meet an endangered language* each of which offers thematic discussion and walk-through of the deposit highlights.[40] Archives could also cross-promote their holdings, for example by listing 'interesting' deposits in partner archives, or by systematic efforts to cross-reference related holdings across archives.

8. Communication ecology

As expressed so well by the title of the 2008 conference of the International Association of Sound and Audiovisual Archives: 'No Archive is an Island'. Archive exist as institutions and services within an interconnected network of communication and interaction types: conferences, workshops, publications, posters, mailing lists, social media (Facebook, Twitter etc), blogs, podcasts, and other events such as training and outreach events. And of course archives can be linked together, through common portals such as OLAC[41], or by placing links in deposits to relevant deposits in other archives (Steven Bird, pc). These all provide possibilities for disseminating information about archives and their collections, and for interaction and exchange.[42] These channels are complementary and mutually reinforcing: Melissa Terras (2012) has shown through experiments with social media that using the right combination of blogging and Twitter – with the right timing ('timing is everything') – it is possible to increase the number of article downloads by up to 11 times.

9. Feedback channels

Archives can implement additional channels to facilitate communication with and between themselves, depositors and users. ELAR provides depositors with detailed real-time information on who has accessed their materials. Reports from depositors and communities confirm that this enhances their trust in the archive. For example, the Warm Springs community (Oregon, USA) has language materials deposited in ELAR with access restricted to 'Community only'.[43] Community members reported their relief on seeing ELAR's access reports explicitly showing zero downloads. In other cases, depositors worried about rampant downloading are reassured on seeing that access to their deposits seems to be moderate, and can be more willing to relax access restrictions.

ELAR implemented an innovative feedback channel for negotiating access to restricted materials (see also *5. Access management* above). Called the 'Subscription system', this system caters to depositors who are willing to share access to materials but only under the

[39] See http://dobes.mpi.nl [accessed 1 April 2014].

[40] See http://elar-archive.org/blog/category/elar-collections/meet-an-endangered-language [accessed 1 April 2014].

[41] See http://www.language-archives.org [accessed 1 March 2014].

[42] As well as to identify new sources of materials for collections.

[43] These were produced with linguist Nariyo Kono; see http://elar.soas.ac.uk/deposit/0066 [accessed 25 April 2014].

condition of express permission, so that they can be aware of access and usage of their data. ELAR conducted research (Nathan 2010) and found a very salient preference for this condition, with the proportion of items under Subscriber access ('S') varying between 25 and 50% over time. The system places a link next to all S-labelled items. Users can click on the link to bring up a dialogue box where they can send a request message to the depositor. In turn, the depositor is notified and supplied with the user's request message and the user's profile information; based on these the depositor can grant or deny access, or send a message back to the user (or both). The system has proved to be a very effective solution both for satisfying depositors' preference for 'need to know' and for delegating access management to those in the best position to handle it. Furthermore it has proved to be a fertile channel for exchange of information, as depositors and users discover the value of reciprocal exchange of information around the topic of the language materials. Although a limited implementation, this transformation of the archive from being a static repository to being a living platform for building and conducting relationships *around* language materials could eventually be extended to include communication around all deposits, involving exchange between various constellations of depositors, users, and language speakers.

Many archives work with depositors and provide feedback about their materials during the depositing/curating process; in this way the archive is 'reaching' future users through its contribution to the content, organisation, and properties of the deposit itself.[44]

While ELAR's subscription system (described above) enables users to negotiate directly with depositors about access to materials, a richer feedback channel between them could result in more effective usage of those materials. Users of data – and especially less experienced users such as students – can benefit from ongoing access to documenters so that the latter can provide methodological guidance or warnings about the limitations of the materials (indeed a free exchange may lead to fruitful collaboration between them). While in general scientific data can be utilised in its own terms, language documentation materials often consist of recordings and other material captured in complex situations that are only partially understood, and where the descriptive aspects can be limited, preliminary, and under revision. In addition, such materials are often unique, with little contextualising, corroborating or cross-referencing literature. While general archive principles encourage depositors to provide metadata and metadocumentation (Austin 2013) to ensure that data is understood and used appropriately, there remain many methodological limitations that can be ameliorated by connecting users and depositors.

10. Temporal reach

Reach across time is conventionally assumed to be archives' mission. However, this can no longer be taken for granted as funds become harder to get, host institutions look for short-

[44] This process makes explicit the influence of archives on the records they preserve and hence on the representation of the world they attempt to record, recalling Jacques Derrida's *Archive Fever* (1995) and its proposal – or accusation – that archives manipulate and construct the historical record through their policies and practices (Nathan 2012). Perhaps the main issue for us is whether archives wield this influence in a transparent, collaborative and scientific way.

term returns, and even the concept of 'archive' becomes crowded out by the proliferation of digital services that appear to converge with what archives do, especially as archives also increasingly portray themselves as publishers (cf. Holton 2014, Nathan 2011b) or software engineers (Koenig et al 2009).

Gary Holton (2013) has pointed out that the value of archives value can be realised through serendipitous discovery in the (perhaps distant) future, and is not calculable in terms of inputs and outputs, impact, or other contemporary evaluative measures. In his example, Eyak materials, after lying unused for some 40 years in the Alaska Native Language Archive, were 'discovered' and suddenly received much attention and use by the community; they went rapidly from zero to 100% reach after 40 years of archival dormancy. Gary has pointed out that in the digital domain, and given today's popular engagement with ephemeral digital data, it is all too easy to delete, revise and substitute – all actions which can dilute or distort the historical record.

Conclusion

This paper has listed a provisional set of 10 criteria that, taken together, could be used to describe an archive's reach. As a coda, I would like to add that they are not proposed as measuring yardsticks or evaluative criteria. That is because that kind of quantitative or box-ticking approach does not take into account the concept of *value*.

As archives struggle to justify their existence to host institutions and funders, they find themselves citing facts and numbers: being a member of this or that body, having X terabytes of data and Y deposits/files/hours (Dobrin et al 2007). While archives might well be proud of some of their numbers, (although cf. Woodbury 2014:2 who honestly discloses disappointingly low access and usage) they also need to work out ways to detect and describe the value found in archive usages. Such information would not only tell us more about the reach of an archive (for example, if a teacher amplifies the dissemination of archive holdings by creating classroom teaching material from them) but also about the *significance* and *meaning* of the materials to those who access them. Endangered languages archives have an important responsibility as custodians of the resources contributed by communities, documenters, and funders, and so any efforts they make to increase their reach will amplify the efforts of all.

References

Austin, Peter K. 2013 Language documentation and meta-documentation. In Sarah Ogilvie and Mari Jones (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*. Cambridge: Cambridge University Press.

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

Bird, Steven & Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. In *Language*, Volume 79. 557-582.

Broeder, Daan, David Nathan, Sven Strömqvist, & Remco van Veenendaal. 2008. Building a federation of Language Resource Repositories: The DAM-LR project and its continuation within CLARIN. In *Proceedings of LREC 2008*. Marrakech, Morocco, 28-30 May 2008.

CCDSD (The Consultative Committee for Space Data Systems). 2012. Reference Model for an Open Archival Information System (OAIS): Recommended Practice, Issue 2 (June 2012). Washington DC: CCSDS Secretariat [Online at http://public.ccsds.org/publications/archive/650x0m2.pdf, accessed 13 Aug 2014]

Chang, Debbie. 2010. TAPS: Checklist for Responsible Archiving of Digital Language Resources. MA thesis Graduate Institute of Applied Linguistics.

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

Derrida, Jacques. 1995. *Mal d'archive: une impression freudienne.* Paris: Éditions Galilée

Dobrin, Lise, Peter K. Austin & David Nathan. 2007. Dying to be counted: commodification of endangered languages in documentary linguistics. In Peter Austin, Oliver Bond and David Nathan (eds.) *Proceedings of the Conference on Language Documentation and Linguistic Theory*. 59-68

Farrar, Scott & Terry Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *Glot International* 7(3). Malden MA: Blackwell. 97-100.

Holton, Gary. 2013. Thanks for not throwing that away: How archival data unexpectedly inform the linguistic and ethnographic record. Paper presented at *Research, records and responsibility (RRR): Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)*. University of Melbourne, 2 December 2013.

Gary Holton. 2014. Mediating language documentation. In David Nathan & Peter K. Austin (eds.) *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*. London: SOAS. pp. 37-52 [Available online at http://www.elpublishing.org/PID/136]

Jukes, Anthony. 2011. Culture documentation and linguistic stimulus. In Nick Thieberger, Linda Barwick, Rosey Billington and Jill Vaughan (eds.) *Sustainable data from digital research*. Melbourne: University of Melbourne. 49-65

Koenig, Alexander, Jacquelijn Ringersma & Paul Trilsbeek. 2009. The Language Archiving Technology domain. In Vetulani, Zygmunt (ed.) *Human Language Technologies as a Challenge for Computer Science and Linguistics*. 295-299.

Lee, Chrstopher A. 2010. Open Archival Information System (OAIS) Reference Model. In *Encyclopedia of Library and Information Sciences, Third Edition*. [Available online at http://ils.unc.edu/callee/p4020-lee.pdf, accessed 13 Aug 2014]

Nathan, David. 2006. Thick interfaces: mobilising language documentation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.) *Essentials of Language Documentation.* (Trends in Linguistics. Studies and Monographs 178). Berlin: Mouton de Gruyter. 363-379.

Nathan, David. 2010. Archives 2.0 for Endangered Languages: from Disk Space to MySpace. *International Journal of Humanities and Arts Computing.* 4(1–2). Edinburgh: Edinburgh University Press. 111-124.

Nathan, David. 2011. Digital archiving. In Peter K. Austin and Julia Sallabank (eds.) *Handbook of Language Documentation.* Cambridge: CUP. 255-273.

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

Nathan, David. 2012. Archive fever: making languages contagious, or textually transmitted disease? Paper presented at *Charting Vanishing Voices: A Collaborative Workshop to Map Endangered Oral Cultures.* University of Cambridge, 30 June 2012.

Nathan, David. 2013. Access and accessibility at ELAR, a social networking archive for endangered languages documentation. In Mark Turin, Claire Wheeler & Eleanor Wilkinson (eds.) *Oral Literature in the Digital Age: Archiving Orality and Connecting with Communities.* Cambridge: Open Book Publishers. 21-40.

Nathan, David & Meili Fang. 2014 Re-imagining Documentary Linguistics as a Revitalization-driven Practice. In M. Jones & S. Ogilvie (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization.* Cambridge: Cambridge University Press. 42-55.

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

NINCH (Humanities Advanced Technology and Information Institute, University of Glasgow & National Initiative for a Networked Cultural Heritage). 2002. The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials. http://www.ninch.org/guide.pdf [accessed 25 April 2014]

OAIS. 2012. Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf [accessed 25 April 2014]

Rice, Keren. 2011. Ethical Issues in Linguistic Fieldwork. In Nicholas Thieberger (ed.) *The Oxford Handbook of Linguistic Fieldwork.* Oxford: OUP. 407-429.

Schwiertz, Gabriele. 2012. Online presentation and accessibility of endangered languages data: The General Portal to the DoBeS Archive. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds.) *Language Documentation and Conservation, Special Publication 3: Potentials of Language Documentation: Methods, Analyses, and Utilization.* Honolulu: University of Hawai'i Press. 126-128.

Terras, Melissa. 2012. Using Social Media to promote your own Open Access research. Presentation at Open Access Week 2012: Opening Research and Data. 22 Oct 2012, Birkbeck College. http://www.slideshare.net/NeilStewartCity/melissa-terras-using-social-media-to-promote-your-own-open-access-research [accessed 5 May 2014].

Wilbur, Joshua. 2014. Archiving for the community: Engaging local archives in language documentation projects. In David Nathan and Peter Austin (eds.) *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*. London: SOAS. 85-102.

Wittenburg, Peter, Ulrike Mosel and Adrienne Dwyer. 2002 Methods of Language Documentation in the DOBES project. In *Proceedings of LREC* 2002. 34-42 [Online at http://www.mpi.nl/lrec/2002/papers/lrec-pap-02b-dobes-talk-final.pdf, accessed 1 April 2014]

Wittenburg, Peter. 2004. The DOBES Programme and its Contribution to Standardization and Revitalization. Presented at the *Linguapax Forum on Language Diversity, Sustainability and Peace*. 20-23 May 2004. Barcelona: Linguapax.

Woodbury, Anthony. 2014. Archives and audiences: toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter Austin (eds.) *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*. London: SOAS. 19-36.