Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 7*. Editor: Peter K. Austin

Language documentation and archiving: from disk space to MySpace

DAVID NATHAN

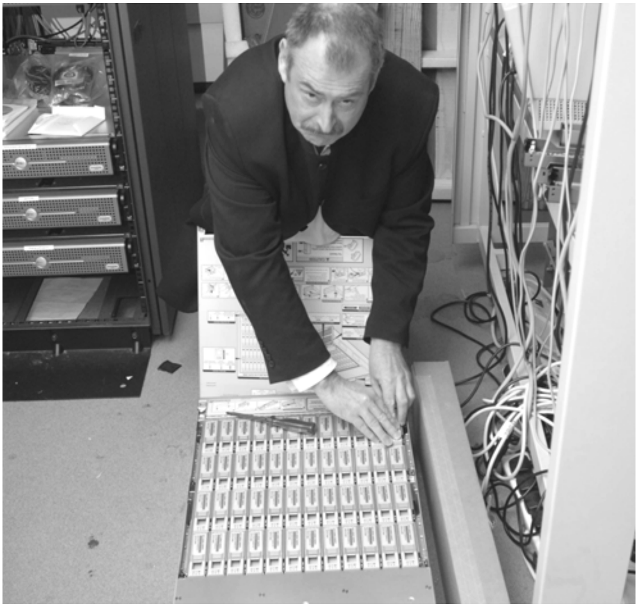
Archiving and language documentation: from disk space to MySpace

David Nathan

1. Introduction

What do you think of when you hear the word ‘archive’? Maybe you think of aisles of dusty filing cabinets on an industrial scale. Or maybe you think of something more high-tech, like the new 48 terabyte disk array unit recently installed at the Endangered Languages Archive (ELAR) at SOAS, and shown in Figure 1.

Figure 1. Tom Castle commissioning ‘the Numbat’, ELAR’s main 48TB storage unit.



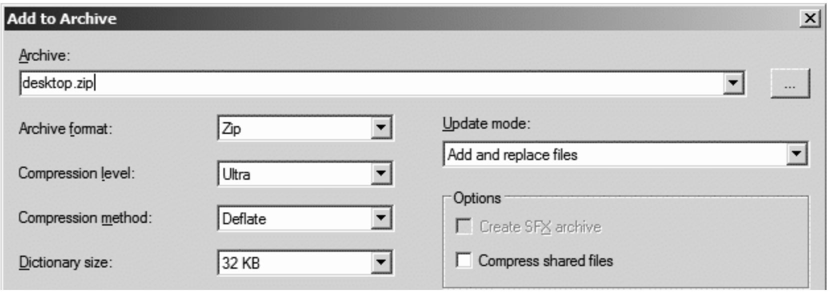
David Nathan 2010 Archiving and language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 7*, 172-208. London: SOAS.

Or maybe you think about that thing you call your own archive which is a pile of CDs of all your data that you have lying around under your bed, as in Figure 2.

Figure 2. A pile of CDs: does your personal archive look like this?

Or maybe it is some mysterious thing that your computer does to you sometimes: it pops up and says ‘archive bit set’. Or someone sends you something – a zip file, for example – and it mentions something about archives, as in Figure 3.

Figure 3. Your computer has a mysterious predilection for archiving.



Add to Archive

Archive:
desktop.zip

Archive format: Zip

Compression level: Ultra

Compression method: Deflate

Dictionary size: 32 KB

Update mode: Add and replace files

Options

□ Create SFX archive

□ Compress shared files

Maybe you think about one of those or maybe all of those or none of them.

What is a language archive then? One answer is that it is the sum of all the horrific problems we have to face as archivists. What horrors do I mean?

- the horror of receiving a stack of 99 disks from a depositor, each one of which we have to feed into a machine, wait for it to read, notice that many reads failed, find out where they failed, log all that, go back to the depositor, ask them to resend that disk or the broken files;
- the horror of videos, occupying many, many gigabytes of our precious disk space which look like absolutely no use at all. One of my archive colleagues received the most outstanding example: a video, 5 minutes long, of an empty chair. No-one is sitting **on** the chair, and no-one is speaking or even visible;
- the horror of receiving data in unusable formats that require a lot of manual work to make preservable; and
- the horror of maintaining complex equipment. All equipment fails eventually, and we at ELAR have certainly had our share of equipment failures leaving us only one or two **more** disasters away from losing data. Of course, we are professional horror managers so that has never happened!

2. Digital archiving

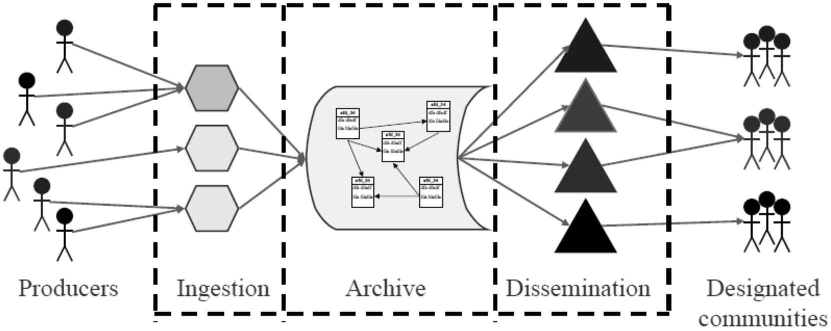
The Endangered Languages Archive at SOAS is responding to the needs of digital archiving in language documentation and description by exploiting social networking technologies to redefine the archive as a forum or a platform for data providers and users to negotiate about, and to exchange, data.

More typically, an archive has been defined as ‘a trusted repository created and maintained by an institution with a demonstrated commitment to permanence and the long-term preservation of archived resources’ (Johnson 2004:142). Documenters of endangered languages want to entrust their materials to a facility that will not only reliably preserve it, but also respect and implement any access conditions or restrictions that are applied. Usually those capacities require an institution that has a commitment to the preservation of resources and which is accountable to its depositors and other stakeholders.

The key word here is **commitment**, that is, commitment to the long term preservation and management of the materials. Any such archive should have policies and processes for acquiring materials, for cataloguing them, preserving them, disseminating them, and then making sure they can live through the various changes in digital technologies that might make files no longer usable as computing systems change. It is a great simplification to think of the archive as a collection of materials that users may (or may not) be able to access or download.

Figure 4 shows presents the model of a digital archive developed by the Open Archive Information Systems (OAIS) project which was initiated by NASA, the American space programme, who were probably the first people to encounter the problem having to organise and store mountains of digital data (OAIS 2002). This model has been very influential and most digital archives follow its main principles.

Figure 4: OAIS view of digital archives. Notice the range of dissemination objects to cater for various 'designated communities'.



On the left in Figure 4 is the community of producers which may include researchers working on language documentation and description. Inside the dotted box is where the archive lives. The model divides this box into three functions:

1. the **ingestion** stage, which sounds horribly anatomical but just means that the data enters the digital domain at the archive;
2. the **archive** storage, and all the supporting processes such as backup and file format conversion;
3. the **dissemination** stage, which is the start of the 'output' side of the archive where materials are distributed to users, possibly by providing alternative distribution-friendly formats of the resources.

Finally – and what can be described as the most important contribution of the OAIS model – is the identification of the 'designated communities' that gain access to resources. It represents the realisation that archives cannot present materials that will serve and satisfy everybody. Just like publishing or dissemination anywhere else, facilities have to be geared to serve particular audiences. In turn, archives have to be able to identify and understand the needs and capabilities of those communities in order to be able to serve them effectively. As discussed below, we identify various such communities: researchers, language communities, the general public, and so on.

2.1 Archiving of language materials

Archiving of language materials means preparing them in structured, well-documented, and complete form. Typically, there is some data, such as an audio recording, and then some accompanying and associated knowledge added by the documenter, often assisted and informed by the language speakers (see Nathan 2010b). The documenter has to understand, inscribe and encode that knowledge somehow, by describing, transcribing, annotating, illustrating, and/or marking up – all ways of giving form to the knowledge. If all that is complete, and the methodologies and conventions explicitly documented, then the package of resources is ready for archiving.

Over recent years the field of language documentation has become rather confused about the relationship between data, data preparation, data formats and archiving. Often, archiving considerations have driven what language documenters do in terms of their processing of data, their methods and software (their so-called ‘tools’), and their formats. That archive-driven approach (which I have called ‘archivism’; see Nathan 2004 and Dobrin et al. 2007) is something that I criticise quite strongly. Good data management and judicious use of standards are part of any research area, especially one which deals with such unique and precious data, much of which is abstract and symbolic (except where audio or video recordings are considered to be data) and therefore quite amenable to encoding (compared, for example, to biology where the objects of description are ‘real’ and physical, not symbolic creations of human minds and culture).

What we do as archives should be less about defining documentation project methods and outcomes and more about supporting other functions that I discuss in section 3.12 below, like building relationships and providing a platform for relationships and transactions between the information providers and the information users.

In other words, archiving is far from being just back up. Neither is it just dissemination or publication, such as putting some materials up on a website. Nor does it define good linguistic practices. What the archivist should want is resources that are worth long term preservation (in their own terms), and which are feasible to preserve. I hope we are moving the documentation field in the direction where researchers are already creating those kinds of resources.

I would like to use a (made up!) example involving a former prime minister of the UK, Winston Churchill. Imagine going to the Churchill archives where you might find his famous pipes in a drawer. On no account would the archivist have gone to Churchill before he died and asked him to arrange his pipes so they would look nicely organised in the archive. Traditional archiving – and in a sense what we are getting back to now after some distractions over the last 10 years – focuses on the intake, preservation and dissemination of materials and does not try to determine what the materials are, let alone wrap its tentacles around the methodologies of the field that generated the materials.

The following is what a language archive can offer:

- **security** – keep electronic materials safe;
- **preservation** – keep them safe for a long time;
- **discovery** – help others to find out about deposited materials. Also help depositors to find out about who is interested in their materials and how other people have used them;
- **protocol** – manage all the issues surrounding sensitivities and restrictions;
- **sharing** – or dissemination, facilitating other people's use of the materials;
- **acknowledgement** – create citable identifiers so that resources can be referenced;
- **mobilisation** – adapting materials and putting them to work, for example in language support and revitalisation activities. This chapter does not have much to say about this aspect, however, language archives, because they often have relevant technical skills, are able to help in the creation of usable language materials for speaker communities
- **quality** and **standards** – researching and then informing clients about the nature and formats of materials that best guarantee preservation. Archives spend a considerable amount of their resources on training, offering advice, and providing feedback.

There are many kinds of language archives and researchers who plan to archive materials should find out which is relevant for their needs[1]:

- local archives, serving their particular community and, like the archive for the Squamish Nation in Canada, not serving outsiders, because they do not have the resources or they want full control over and the privacy of their own community's materials;
- regional archives like the Archive of the Indigenous Languages of Latin America (AILLA), which accepts deposits on Central and South American languages, or PARADISEC in south-east Australia, which is primarily interested in materials from Pacific languages and cultures;
- archives of international scope such as the DoBeS archive located at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, or ELAR at SOAS[2].

Some archives are associated with research institutes like the one at the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS)[3] in Canberra, Australia, or the Alaskan Native Language Centre Archive.[4] Some, like ELAR or DoBeS[5] have a distinct advantage in that they are closely coupled with a granting body (Volkswagen Foundation and ELDP, respectively). This creates a much stronger partnership with documenters who receive a research grant and then go on to be depositors throughout and beyond the life span of their funded projects. Another dimension to check out is whether the archive is digital only, like ELAR, or can offer physical preservation (and perhaps restoration) of analogue tapes and manuscripts.

Who are the users, or the designated communities mentioned earlier? For the DoBeS archive their main users are the depositors. For ELAR, this is also the case, at last initially (although I believe it will change radically in the near future, due to the developments described at the end of this chapter). Depositors want to work with the archives to deposit materials, access

[1] Note that research funders may expect applicants to have contacted an archive and discussed their archiving plans before submitting an application (see Austin 2010b).

[2] See the website of the Digital Endangered Languages And Musics Archive Network (www.delaman.org) for a list of more archives that specialise in endangered languages.

[3] See http://www.aiatsis.gov.au

[4] See http://www.uaf.edu/anla

[5] See http://www.mpi.nl/DOBES/

materials that they may have lost or not have with them, and to update materials.

Members of language communities could be significant, if not the largest, potential users of archive materials. Anecdotal reports suggest that up to 95% of those accessing the Berkeley Language Center archive collection of California Native American language materials, for example, are community members. Similarly, the archives and library of AIATSIS have seen a strong shift in the proportion of people using the library from non-Aboriginal researchers to Aboriginal people who were researching their ancestry and culture (including language) in order to strengthen their claims for land rights.

There are also other researchers, potentially from a variety of disciplines (see Himmelmann's (1998) exhortation to document for a wide range of users and future usages), who may be interested in archived documentation materials. There are other potential audiences for documentation and indeed other stakeholders, including catalytic people like educationalists and policy developers (see Sallabank 2010), who often only need to be convinced that there are resources available for a language in order for them to open up their purse strings and help to foster language programmes and other developments.

Journalists may also wish to access archived materials, especially when they want to write stories about 'the last words of language X', and so on. The wider public, many of whom may have positive, or at least benign, intellectual interests in endangered languages, may wish to use materials to further their understanding of the subject, or possibly to find cute-sounding indigenous words to name their new boat.

There also exist various archive networks and bodies that archives like ELAR are connected to, and, in fact, much of the formative influence on our thinking and on our technologies has come from the libraries area. The D-LIB initiative (http://www.dlib.org/) has been particularly important, as has OAI (Open Archives Initiative), OAIS Open Archival Information Systems (initiated by NASA) and the Open Language Archives Community (OLAC).

More recently there are a couple of groups who are or have been influential in the ways our small but vigorous community of endangered languages archives are working. One is the Digital Endangered Languages and Musics Archives Network (DELAMAN) which has an annual meeting and has been involved in issues including training, pooling resources for some common operations, such as a shared portal for searching, and establishing citation standards so that researchers can begin to have a way for their corpus creation and development work to be recognised. Figure 5 presents an example of DELAMAN's initial recommendations for citing materials that are in our archives, either at the collection level or for individual files.

Figure 5: Examples of citations from Heidi Johnson of AILLA

> *Collection*:
>
> Sherzer, Joel. 'Kuna Collection.' The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: audio, text, image. Access: 0% restricted.
>
> *File/resource*:
>
> Sherzer, Joel (Researcher). (1970). 'Report of a curing specialist.' Kuna Collection. Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Type: transcription&translation. Media: text. Access: public. Resource ID: CUK001R001.

Language archiving is different from other kinds of archiving, and it is difficult. In fact we might say that archiving language is impossible. After all, what is a language? We cannot describe its scope or boundaries. An important thing to remember is that unlike so many other disciplines whose data are conventionalised, e.g. for book publishing we know what ISBNs are, and we know what authors are, with language data and most especially endangered languages data, many of the aspects of particular languages and projects and the way their data is encoded are either unique to that language situation or are perhaps yet unknown. Given the estimates about how many languages there are in the world and how few of them have been documented, it is perhaps rather premature for us to be deciding on such things as standardised sets of morphological glossing terms, for example.

2.2 Archiving of *endangered* language materials

Language archiving gives rise to a paradox because while on the one hand we would like to see standards and comparability across different researchers and disciplines and usages, on the other hand the very nature of the language documentation field demands the recognition of uniqueness and idiosyncrasy across different language situations and different archive resources, for the following reasons:

- languages, cultures, communities, individuals, and projects are all extremely different;
- fieldworkers are often quite an unusual if not eccentric group of people;
- the genres for the field are in flux. While some are stabilising, e.g. the documentation 'bundle' consisting of a video or audio file plus associated .eaf transcription file created using the ELAN tool, in general the genres of the field are not really settled. This makes it

difficult for archive staff to fully manage materials. Below I discuss some of the kinds of strategies we are adopting to deal with this;

- sensitivities and restrictions – languages are endangered because people are under pressure or suffering in various ways. This quite naturally means that language materials are associated with sensitivities and restrictions, which in turn are part of the documentation field. That is amplified even more for archives which have become points of access or distribution.

3. The Endangered Languages Archive (ELAR)

ELAR is one of three programmes within the Hans Rausing Endangered Languages Project. The others are the Endangered Languages Academic Programme (ELAP), and the Endangered Languages Documentation Programme (ELDP) which offers research grants.

ELAR has a staff of three: myself (the archivist), Edward Garrett (the software developer), and Tom Castle (the technician).[6] From time to time we employ research assistants as well. We are involved in developing policies, preservation infrastructure and facilities, and our ongoing activities include curating, cataloguing and dissemination, training, providing advice, and materials development and publishing.

We currently hold about 70 deposits and a total volume of about 8 TB (terabytes), with a great deal of documentation material flowing in for deposit. Our main providers (the Producers of Figure 4) are the ELDP grantees. Our main mission is to archive the materials that are generated as a result of ELDP funding, however, to some extent we can also archive any digital materials for endangered languages. We expect the volume of deposits to nearly double over the next 18 months because materials tend to come in roughly 6 to 18 months after the end of funded projects, and many of these projects have finished during the last year or so. Figure 6 shows ELAR's relative holding of various data/media types for a representative sample of deposits.

[6] We also have access to a fraction of the Faculty of Languages and Cultures technician, Bernard Howard.

Figure 6: ELAR's relative holding of various data/media types

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

ELAR data types for a 10% sample of holdings, late 2008

data type by volume (MB) and number of files, sorted by volume

Figure 6 shows that audio comprises far and away the greatest number of files, and the largest data volume. There are a large number of images as well as many text files in a range of formats. What is particularly interesting are the top two lines showing that although we have almost eight times the number of audio files compared to video, the video volume is two-thirds that of the audio. What this means is that storage space for video is a major issue. The value and methodological problems associated with video are a controversial issue (see Ashmore 2008, and the debate between Nathan 2007 and McConvell 2007 and Wittenburg 2007) but as the use of video by researchers takes off, which is happening now, and as High Definition video (with larger file sizes) becomes commonplace, then holding, preserving and delivering video will become a crucial factor for digital language archives.

It is interesting to compare ELAR's activity profile with how a digital language archive operated only 15 years ago. I used to run a small archive at AIATSIS called the Aboriginal Studies Electronic Data Archive (ASEDA). Although small, it was one of the first digital language archives, and continued to run independently until 2009. It was founded by Nick Thieberger in the early 1990s, based on the model of the Oxford Text Archive. ASEDA's mission was more or less to serve as a backup for researchers, or to hold materials so that they were safe. At that time, most materials were backed up and transferred on floppy disks; even CD disks and writers were prohibitively expensive. Also many linguists then (even more than now) used Macintosh computers, which seemed to be prone to technical problems (much less now since the development of OS-X by Apple). In other words, the ASEDA deposits were backups of otherwise vulnerable data, and they consisted entirely of textual material – lexicons, grammars and texts.

It is interesting to see how much things have changed in the past 15 years. The modalities of the data have changed radically. As Figure 6 shows we now hold audio and video media as the predominant medium of documentation. Nowadays we have an information environment that is much more developed and standardised (e.g. with availability of 'rock solid' data coding methods such as Unicode and XML, and a few widely accepted conventional metadata schemas). Today, we are cataloguing and disseminating materials via the world wide web, and our storage methods have also changed radically. For ASEDA when we wanted to have more back-ups we bought more Macs or magneto optical storage drives (the equivalent of the later minidisk technology; it is still around today but only in niche areas), whereas now we use professional self-monitoring disk-based mass data storage systems with overnight tape backups, more or less the same technologies as those used by a university, a big company or a bank. We are much clearer about our function as providing long term preservation of significant materials, not merely backup of vulnerable data. However, perhaps the single biggest change is that archives like DoBeS and ELAR have expanded their influence on and relationships with the linguistic community to such an extent that they are involved in many stages of the documentation process, especially in providing training, advice, and software resources.

3.1 Why digital?

ELAR is a specifically *digital* archive, although we do occasionally digitise analogue materials such as audio cassette or video tapes and we provide support for people who are willing to come to ELAR and do their own digitising. But why digital? If there were a 'god of archiving', he or she would probably not choose digital as the most robust method of preservation. While digital form is clearly unsurpassed for supporting the transmission, modification and combining of materials, it is inherently fragile and costly as a method of long-term storage. There is only one critical, i.e. unavoidable, reason for using digital form for archiving, and that is for media. The only way we can make perfect copies of things, and therefore be able to carry them forward into the future, regardless of the changes and degradations in their physical carriers, is to have them in symbolic form. Compare the situation with analogue materials, such as cassette or VHS tapes: after about three generations of copying the quality is very poor.

The digital principle is familiar to us as linguists, and we rely on it all the time; our phonological, morphological, and lexical representations are all digital because they use discrete symbols (e.g. a sound is either [g] or [k], a word is either ‘dog’ or ‘dock’). For computers, the choice is either 0 or 1. It is now clearly understood that if we want to preserve audio, for example, the only way to do it is to digitise it. We cannot preserve the tapes indefinitely. Good cassettes will last perhaps 30 years, but there is no way that we can do what we need to do, which is to preserve recordings of the world’s languages for 50 or 100 years and yet further beyond without using digital technologies.

Analogue is real stuff, and if you copy a tape you are making a real thing cause a change to another real thing. However, this is just not something that can be perfect in the real physical world. Actually it is only for the sake of the **content** of media that digital form is absolutely crucial. Once encoded symbolically, we would actually be better off carving the bits of a media file (the 0’s and 1’s) character by character into stone. More seriously, it is said that the very best means of long-term preservation is to print barcodes on microfilm. Under good conditions, including temperature and atmospheric control, such microfilm should last up to 1,000 years. But we are not likely to do that, at least not right now. Using today’s technologies, we can copy and transmit data with zero loss, thus ensuring safe preservation through making redundant (backup) copies. There are also the rest of the functions needed by our discipline and our culture, all of which remain more practical in the digital domain: cataloguing, sharing, disseminating, transmitting, broadcasting, modifying, reusing, combining, etc..

In some ways the digital medium, as we know it today, is the worst possible solution for long-term storage, because of the need for electricity to keep disks spinning, air conditioning running, and so on. There are currently interesting changes in technology such as solid state storage perhaps becoming available at a suitable scale in about five years that may reduce this need. There are huge costs in digitising materials, setting up infrastructure and then maintaining, upgrading and replacing it. At ELAR we have found that we have needed both strategy and luck to get the infrastructure right. Less than 5 years ago, we paid about £30,000 for 8 terabytes of data storage, buying items that were parallel to SOAS IT department’s equipment in order to reduce incompatibility problems. It was said to be good data storage (by its sales people!) but actually it regularly failed (and tested our backup capabilities to the full!). Last year, we replaced it with a unit that can store 48 terabytes, which has operated faultlessly, and cost £8,000, amounting to only 5% of the original unit’s price per unit of storage. At least with the new equipment, we made a major purchase at the right time, just after its price had reduced by 50%

over less than 6 months. There are very few products for which costs reduce so radically. It is probably just as well, because the demand for storing video material, which averages about 10 times the size of audio per hour, is soaring as more and more documenters turn to recording video.

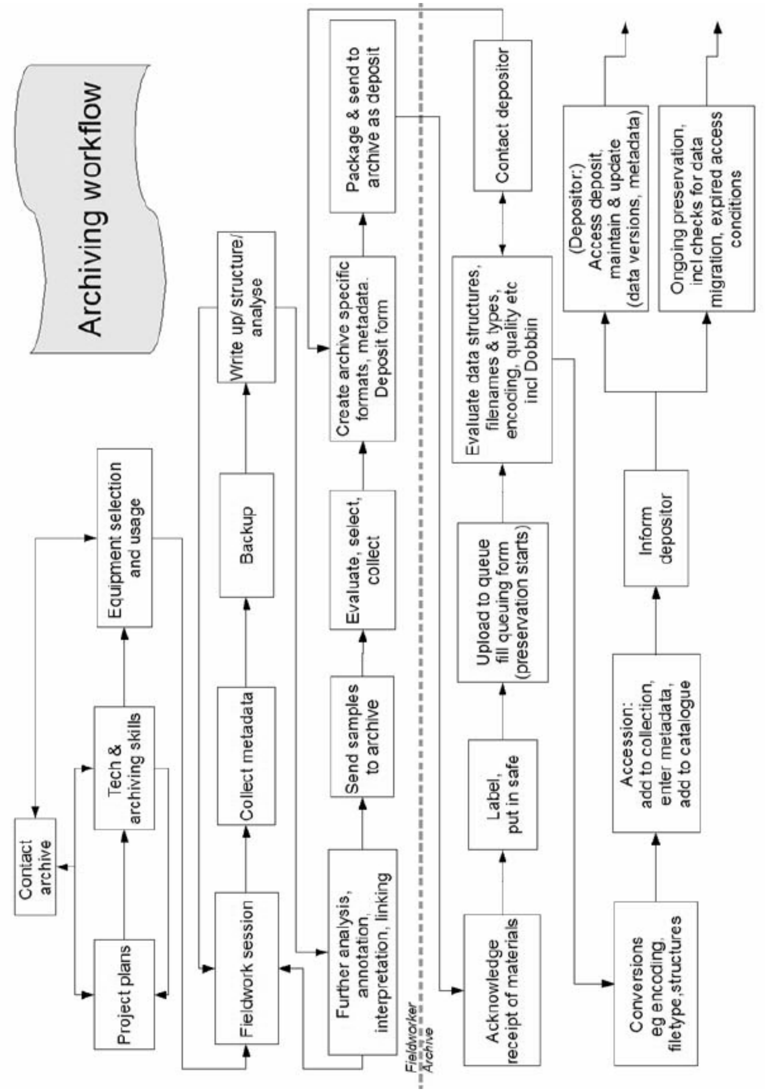
Some issues we face are more complex and are inherent to the digital medium. Successful preservation depends on the use of appropriate file and data formats, and the documenters' ability to use the right tools and techniques to provide these formats. As archives we need to provide the human resources to monitor this material, to convert it, and, as mentioned above, to bring along the documentation communities we work with through training and advice. It is well known that documenters should avoid proprietary formats that can only be created, manipulated and viewed by particular software, such as Microsoft Word. But less noticed is that many resources, even if their file format is open, can only be viewed or experienced using certain software. That is the case for ELAN files (in .eaf format), for example. Because they are XML-based they can be liberated into other formats, but ELAN is needed to experience the tiers and other functions the software provides. Fortunately we can archive ELAN because it is free, open-source and made by our archive colleagues at MPI-Nijmegen who are not restricting it. But what about a FileMaker Pro file? What about an old version of Microsoft Word or Works? We need to make sure that data depends on the least numbers of layers of encoding and software, along the lines suggested in Bird & Simons (2003). However, digital data will always depend on some interpreting agent to be meaningful, and thus, just as for human languages, can become endangered or extinct.

3.2 The archiving workflow from the depositors' perspective

Many depositors are somewhat mystified or even frightened about archiving their data. This is thoroughly understandable, given that they have devoted perhaps years of intense personal effort to the materials, and they have a special familiarity with them. And then, from various technical quarters, they are beset by exhortations to 'Best Practice', 'archive quality', prescribed and proscribed formats, and a range of inconsistent policies from different archives. Archiving might easily feel almost like giving up a child for adoption.

Nevertheless, a hallmark of today's archiving is that documenters and archives are increasingly working together. How do we interact with documenters? Figure 7 is a semi-serious illustration of the variety of types of interaction, including initial discussions about equipment (often even prior to formulating a grant application – see Austin 2010a), participating

Figure 7. A semi-serious flow diagram of interactions between documenter and archive. Activities primarily in the hands of documenters are above the dotted line.


Archiving workflow
Contact archive
Project plans
Tech & archiving skills
Equipment selection and usage
Fieldwork session
Collect metadata
Backup
Write up/ structure/ analyse
Further analysis, annotation, interpretation, linking
Send samples to archive
Evaluate, select, collect
Create archive specific formats, metadata. Deposit form
Package & send to archive as deposit
Fieldworker
Archive
Acknowledge receipt of materials
Label, put in safe
Upload to queue fill queuing form (preservation starts)
Evaluate data structures, filenames & types, encoding, quality etc incl Dobbin
Contact depositor
Conversions eg encoding, filetype,structures
Accession: add to collection, enter metadata, add to catalogue
Inform depositor
(Depositor:) Access deposit, maintain & update (data versions, metadata)
Ongoing preservation, incl checks for data migration, expired access conditions

in training workshops, to providing feedback on materials, collaboration in the conversion and improvement of materials, and managing access to them. I call the diagram ‘semi-serious’ because it was originally conceived as a comment on an archive-centric view of documentation and casts most of the documentation process within the purview of archiving. Needless to say, this is not the view that ELAR really holds: we see ourselves as technical facilitators and as responsible for functions complementary to language documentation, such as preservation and dissemination (see Dobrin et al. 2007).

For the documenter, the ‘main game’ might be the third row (fieldwork session; collect metadata; backup; writeup/structure/analyse). We are increasingly encouraging documenters to send samples of their work to the archive at an early state of their project. We have been able to help many of them through this idea of ‘send a little and send it early’, because we have been able to flag problems such as a microphone that does not match a recorder, or a problem with a format or the way that the documenter is encoding or structuring their data. The result is a win-win situation: we are able to help the documenters, and in the long run it helps us to make materials preservable and to disseminate the relevant knowledge and skills.

Below the dotted line is what the archive focuses on, although some of those activities are shared or even deferred to depositors under our new Web 2.0 model, discussed at the end of this chapter.

To summarise, as an archive we are involved in:

- grant formulation and application;
- various communications, questions, advice;
- training;
- archiving services (curation, conversion, preservation, dissemination etc); and
- ongoing management of materials.

We thus participate in ongoing relationships with our depositors. Archive depositors are no longer expected to be people who turn up one day with a basket full of tapes which they drop like a stork delivering a baby and then fly away forever.

3.3 ELAR feedback

As part of our policy of encouraging potential depositors to send samples for evaluation, we developed a template for providing feedback. For text materials we comment where appropriate under the following headings:

1. document type
2. document format/layout/data structures
3. character/language representation
4. linking/references
5. consistency

For audio and video files we comment on:

- document type/format
- resolution
- quality
- editing
- length
- annotation/transcription
- consistency

And in general, we comment on:

- file naming
- data volume
- delivery
- consistency

Figure 8 contains an excerpt from a feedback form (suitably anonymised) in order to show the kinds of feedback we give.

Figure 8: Excerpt from feedback to a depositor on a data sample

Document format/layout/data structures:

- use of typography (size, underlining, bold, spaces etc) to make headings and other structures is weak – at least Styles should be used (with utter consistency).
- MS Word tables to represent interlinear data is reasonably appropriate, although would need to be converted later.
- is it clear from this document, or somewhere else, where to look up codes etc, such as the speaker initials?
- while the language is consistently labelled in the interlinear section, it is identified only by the alternation in font in the first section.

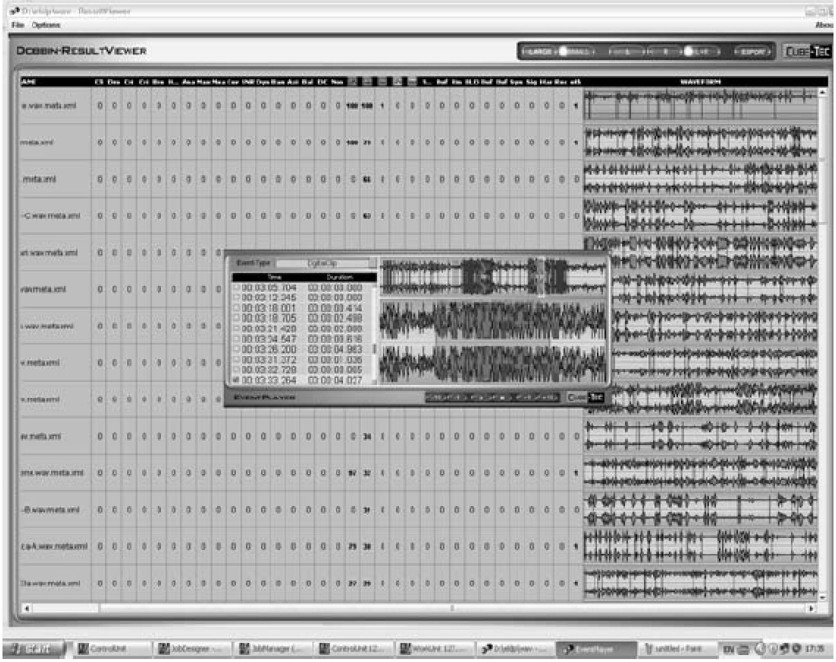
Audio quality:

- gr_amic.wav – quality good.
- gr_amid.wav – quality reasonable, but background hiss is too loud in proportion to the signal. Was this was part of your original recording (on what equipment?) or was it introduced by digitisation, in which case it would be a good idea to try re-digitising.
- gr_amie.wav – quality quite good. Stereo separation of voices is nice.
- gr_amif.wav – suffers a number of faults, including severe clipping (overmodulation), background noise, microphone physical handling, and poor acoustic representation (probably due to poor microphone and/or recorder?).

The feedback shown in Figure 8 was profitable for the depositor, the archive, and future users of the data. It turned out the background noise (hiss) was a result of the depositor's digitisation of his minidisk original, and in subsequent communication I was able to suggest that re-digitisation of the minidisk would make a significant improvement, and it turned out to be exactly so. If we had just taken the finalised data (as if dropped off by the stork), it would not have been discovered that the noise was not in the original recording; by building relationships with depositors and sharing our expertise, things can be better for everybody.

Although we are committed to encouraging the best possible audio quality (see Nathan 2010b), we receive far more audio than we could possibly listen to (the average deposit has around 30-40 hours of audio). To help deal with this, we have purchased some specialised software (called Dobbin) which can work through a set of audio files and give a report, summarising the audio properties and flagging any particular errors. Figure 9 shows the result of one such batch run, where the problems are indicated by highlighting of the relevant points in the waveform representation:

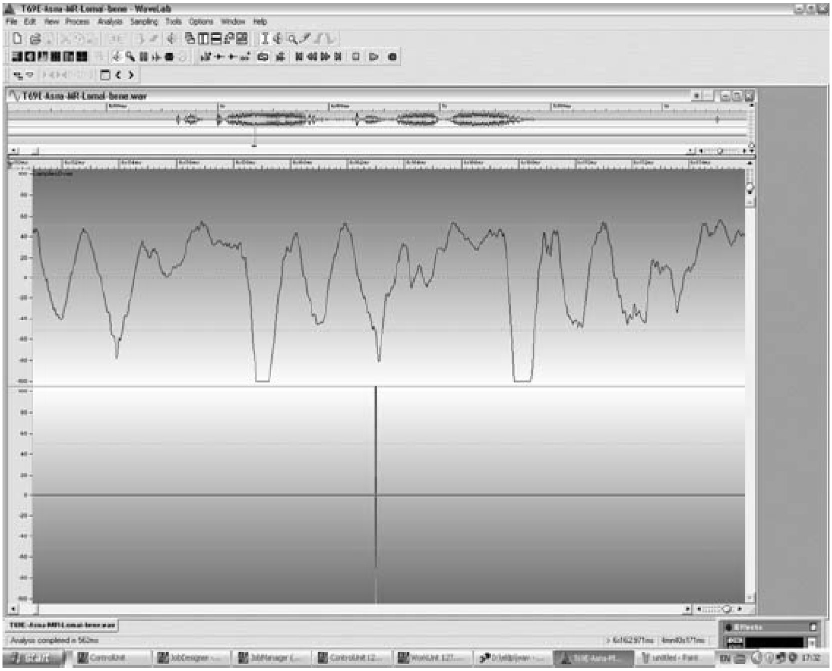
Figure 9: Dobbin report, showing audio evaluation summary and highlighting problem areas in the waveforms.



Time	Duration
□ 00:03:05.704	00:00:00.000
□ 00:03:12.345	00:00:00.000
□ 00:03:18.001	00:00:00.414
□ 00:03:18.705	00:00:02.488
□ 00:03:21.420	00:00:02.088
□ 00:03:24.547	00:00:00.616
□ 00:03:26.200	00:00:04.983
□ 00:03:31.372	00:00:01.036
□ 00:03:32.728	00:00:00.065
☑ 00:03:33.264	00:00:04.037

Clicking on any of the highlighted areas opens an editor where we can inspect and diagnose the problem. Figure 10 shows one such example, where Dobbin has identified audio clipping (gross distortion as a result of the audio source being too loud or the input volume set too high). The problem might be one, like the minidisk example mentioned above, that can be addressed. If it turns out to be in the original recording, although it is probably too late to do anything about it, we can still notate the problem as metadata associated with the deposit.

Figure 10: Dobbin has searched through hundreds of audio files and found various problems, including this example of clipping.


T69E-Asna-bR-Lomal-bene - WaveLab
File Edit View Process Analysis Sampling Tools Options Window Help
T69E-Asna-bR-Lomal-bene.wav

Analysis completed in 562ms

To summarise, ELAR aims to assist depositors in the following ways:

- providing training at various venues;
- providing advice, both general (e.g. that on our website) and specific;
- preserving deposited materials;
- promising to implement access restrictions, etc. (see section 3.9 below);
- achieving the best possible documentation quality, through a philosophy of distribution of labour: it may not be best to expect depositors to convert their data to XML or some other portable format (see below on file and data formats); they may not have the skills, and the result may damage their data. It is best for the archive and the depositor together to find out what skills each has and, where appropriate, for the archive to do the data conversions;
- gradually working towards providing web-based deposit management, which will allow depositors to update materials, send new versions, make corrections and additions, etc. (see 3.12 below);

- occasionally providing equipment and services, and sometimes, on a case-by-case basis, developing resources, such as websites, videos, or multimedia.

3.4 The object of archiving

Archivists tend to think about archiving 'objects' rather than files, partly in continuity with traditional physical archives where collections of objects are held together with information about how to interpret the objects and the relationships between them. Similarly, in the case of digital data, there are relationships among individual files, and metadata and other interpretive items have scope over a range of files. There may well be units intermediate in structure between an individual file and the whole deposit (e.g. a combination of an audio recording and its transcript). Archivists like to refer to these related packages of files as 'bundles'. Such bundles and their identity, structure, and content should be made quite explicit through associated metadata. Some bundles have an implicit existence through simple strategies such as putting items together in directories, or naming the components with the same filename root (e.g. 'gr_ogon.wav' and 'gr_ogon.eaf'). This may work for the researcher while they are putting together the data and working on it personally, but it is liable to be misunderstood or broken as soon as the data is moved to a different location. It is therefore important to explicitly document such structures and local conventions in some kind of metadata table, or, if the system is simple, in a 'readme' file which plainly explains the conventions. ELAR's new cataloguing system will provide a dynamic online method for creating and describing bundles.

Individual files for archiving at ELAR could be any of the following types:

1. **media** – sound, video;
2. **graphics** – photographs (e.g. of consultants, the language speaking settings, objects described or discussed), diagrams (of the recording environment and its location when the recording was made – see Nathan 2010b), sketch maps, scans (of notebooks or local materials or manuscripts). Graphics currently tend to be under-appreciated: photography and diagrams are an effective use of fieldwork resources, compared, say, to video;
3. **text** – fieldnotes, transcriptions, translations, grammars, description, analysis;
4. **structured data** – aligned and annotated transcriptions, databases, lexica;
5. **metadata** – structured, standardised contextual and interpretive information about the materials.

3.5 Data quality and formats

As mentioned above, most data-related issues are properly part of documentation goals and digital linguistic data management, rather than archiving **per se**. There are now few data-related issues that are archive-specific. The digital domain has compressed the effects of time such that what makes data preservable in the long term is not very much different from what we should be doing on a day-to-day basis to make our data portable, in the sense of Bird & Simons (2003). Unfortunately, teaching curricula and documenter practices are generally still considerably behind and need to catch up on these issues. Our broad and shared goal of documenting languages well means that we must find the best 'division of labour' at any given time between education and training curricula, documenter's responsibilities, and archive services.

Bird & Simons (2003) discuss in some detail how to prepare data so that it is robust and 'portable', i.e. complete, explicit, documented, preservable, transferable, accessible, adaptable, and not technology-specific. Of course, documentary materials should also be appropriate, accurate, and useful for the intended users (Nathan 2006).

3.6 Archive specific criteria for deposited materials

The criteria that are particular to archives are:

- materials for deposit should conform to the collection policy of the archive (see above);
- materials for deposit should be fully and explicitly explicated so that users well into the future can understand and use the materials (see 3.8 below); and
- materials are selected.

It is really important to select materials for deposit. There is no a priori reason why a particular piece of audio or video recording (reflecting perhaps the point at which the recording device was turned on rather than anything else), or any particular note that the documenter made, is definitively part of a collection that should be carried into the future. Some materials may distract or even detract from a collection. Archiving definitely does not mean sending in a dump of a hard disk, or the folder that contains everything from a project.

One depositor wrote to ELAR asking 'how much space do you allocate me for my video?' I replied that as the depositor, he needed to make the selection. He repeated his original plea, on the basis that he had a lot of video indeed. However, the proper answer is that the depositor should state the criteria for what makes a good documentation resource and then apply those criteria to selecting video (or indeed any other material); and if it turns out that the criteria (linguistic, documentation or other criteria) indicate that all the materials are relevant, the archive will take all of them. If the criteria say that none of them are relevant we would not take any of the materials. Perhaps this depositor just wanted to be told ‘you can send 40 GB’, but we do not archive endangered languages by the megabyte or kilo.

Some depositors have baulked at the idea of editing audio or video. In some cases this is due to a naïve view that recording captures an actual reality that is rendered untrue or fake by any intervention. In fact, most things researchers do in their academic life are forms of editing and/or selecting. In their linguistic work, they selected, labelled, transformed/processed/edited, summarised, added/corrected/expanded, made links, made or assumed relationships between ‘whole’ and units, invented labels/IDs etc., and imposed formats. When researchers transcribe or annotate, when they choose examples to illustrate generalisations, or when they make decisions to ignore certain things in the audio (e.g. coughing or paralinguistic behaviour) they make selections among which things to pay attention to and which to ignore. It is inconsistent to assume that media is sacrosanct. What is more important in any of these cases is to make clear in the meta-documentation (the metadata that accompanies the documentation) what was selected, on what principles, and with what consequences, if any.

3.7 File organisation in deposits

ELAR does not require data for deposit to have any particular organisation, as long as the files, their names, and their organisation into directories are all rational and consistent in terms of the collection’s own logic. The DoBeS archive, by contrast, has a vision for their collection (the IMDI-corpus) where all deposits are combined as a single united corpus, through which a user can navigate seamlessly. ELAR has taken a less prescriptive stance, because we acknowledge the diversity of depositors’ materials and working styles, and we feel that it is probably premature to believe that we already know the best way to organise language documentations.

To illustrate how some depositors have arranged their materials, the following are some examples. In example (1) the top-level directory ‘IPF10011-Disk3-Story-WulaTuki-LunarEclipse’, contains metadata ‘IMDI_3.0.xsd’ and various other files such as an audio transcription ‘WulaTuki_LunarEclipse.eaf’. This is a simple but effective structure.[7]

[7] There are, however, some comments that could be made about the files and their names; see the discussion questions at the end of this chapter.

(1) IPF987-Disk3-Story-WulaTuki-LunarEclipse [directory, contains the following files:]

```
        IMDI_3.0.xsd
        WulaTuki_LunarEclipse.eaf
        WulaTuki_LunarEclipse.imdi
        WulaTuki_LunarEclipse.imdi.backup
        WulaTuki_LunarEclipse.pfs
        WulaTuki_LunarEclipse.txt
        WulaTuki_LunarEclipse.wav
```

In example (2) the top-level folder contains a file explaining the deposit's labelling system in narrative form. It describes how the depositor has made tables, and what data is in each column. This is good practice, as a form of meta-documentation; a user only has to know basic English in order to be able to understand the arrangement of data in that deposit (although ideally the MS Word file would be in plain text format).

```
(2)     [top level directory, contains the following files:]
        labelling-system.doc
        AngryD-Bsi [directory, contains the following files:]
            AngryD-Bsi.pdf
            AngryD-Bsi.wav
            AngryD-Bsi.doc
```

Example (3) takes a similar approach, but contains additional metadata of various types in the top level directory, including a grid of typical OLAC-style metadata (Overview metadata FTG0025.xls), a legend for the glossing codes used in transcriptions (ELAN transcription key FTG0025.pdf) and some additional read-me notes to the archivist (archivist_notes.txt).

```
(3)     [top level directory, contains the following files:]
        archivist_notes.txt
        ELAN transcription key FTG0025.pdf
        Overview metadata FTG0025.xls
        Kay07-aud [directory, contains the following files:]
            Kay07-aud-jul03a.wav
            Kay07-aud-jul03b.wav
            Kay07-aud-jul03c.wav
```

In all three examples, the depositor has used the technique of having all the related files in the same directory, as well as having the filename root either the same (e.g. 'WulaTuki_LunarEclipse', 'AngryD-Bsi') or partially the same ('Kay07-aud-jul03' + a/b/c). These are, of course, implicit ways of creating bundles of related or interdependent files – the strategy should be described at the top level and followed consistently throughout the deposit. All of these examples also name the implicit bundle's containing folder in some related way, although only AngryD-Bsi does this in a rigorous way. From the archivist's point of view, having this redundancy, i.e. representing bundles or relationships not just in one but in two or even three overlapping ways, is not a bad thing. However, expressing relationships just once and/or completely explicitly would be much better. An ideal deposit would explain the organisational principles in a metadata file, and would also explicitly, consistently and completely list all the bundles and their parts in an inventory/catalogue file.

3.8 Metadata

Metadata is the additional information about data that enables the management, identification, retrieval and understanding of that data. The metadata should explain not only the provenance of the data (e.g. names and details of people recorded), but also the methods used in collecting and representing it. Consider, for example, glossing conventions: using ERG might work fine for a language documenter, but what does it mean to a community member in China? In other words, materials are not only incomplete but seriously flawed if they do not have sufficient metadata, because it is quite possible that they are only understandable by you and no-one else.

Another way to think of metadata is as meta-documentation, the documentation of your data itself, and the conditions (linguistic, social, physical, technical, historical, biographical) under which it was produced. Such meta-documentation should be as rich and appropriate as the documentary materials themselves (see also Austin 2010b:28).

Thus it can be seen that metadata reflects the knowledge and the practices of the discipline and of the individuals undertaking the work, and in doing so, metadata defines and constrains audiences and usages for data. Since metadata enables, or fails to enable, understanding, then it actually controls who can use the materials, and for what purposes. This sometimes leads to bald contradictions; for example, some documenters say 'I'm going to do documentation which will be really useful for the community', but later analysis of the resultant materials, especially the metadata, reveals that the linguist has paid scant attention to documenting the materials themselves in a way such that they are actually understandable and usable by the community (cf. Nathan and Fang 2009:137-139).

Metadata is not unique to documentary linguistics data collections, but the goals of documentation itself heighten the importance of metadata. We know that documentation is data-focused, and that it is supposed to serve multiple audiences; this is the formulation in Himmelmann 1998 that has been constantly repeated in the language documentation literature. But if we **do** want multiple audiences to understand our documentations, we are going to have to work harder on our metadata. This does not necessarily mean learning and doing a lot of technical stuff; it might just mean sitting down and writing a few paragraphs about our assumptions.

There are some widely-used metadata standards, such as OLAC (Open Language Archives Community), IMDI ('ISLE Metadata Initiative', associated with DoBeS), and EAD (Encoded Archival Descriptions). OLAC in particular has been very influential. It proposes a minimal, and by most accounts inadequate, set of attributes to be described, but inherits from its design template Dublin Core (a set of categories defined by libraries to describe their electronic resources) the elegant heuristic that it is designed to be so easy that there is no excuse not to do it. ELAR has created its own set of metadata attributes and is implementing them as part of our online catalogue system. Currently, our deposit form[8] captures deposit-wide overview and discovery metadata, and Edward Garrett is developing an online system to allow depositors and archive staff to add and modify the overview as well as file-level metadata via a standard web browser.

At ELAR, we do not currently oblige depositors to create any particular format of metadata, except for the deposit-wide categories that are included in the deposit form. We took the initial stance that metadata is relative to each project, its goals, its language community, the consultants and other team-members. And each depositor has particular styles and preferences for data management that influence the richness of the metadata that they are actually able to produce. In thus allowing depositors to be more creative with their metadata formats and content we have found that different researchers and projects can give rise to quite different metadata. Given that our goal is to maximise the amount and quality of metadata, we now have some evidence that flexibility is at least as important than standards (see 3.10). Currently, we only insist that our depositors send their metadata in portable formats (Bird & Simons 2003) such as spreadsheets or tables, and that they think carefully about its structure and content (see 3.9).

A lot of depositors are apprehensive about preparing metadata. It seems to be the greatest single impediment to carrying out the deposit process. There are two 'good news' items regarding this. First, the difficulties are understandable, because depositors have had to deal with mixed messages

[8] http://www.hrelp.org/archive/depositors/depositform/index.html

from leaders in documentary linguistics and from archives, and in some cases they have been forced to deal with obligatory but rather impenetrable systems for writing up metadata. Secondly, preparing metadata is probably not as hard as many believe it to be.

The bad news, however, is that if you are considering depositing data in an archive, you should have created your metadata already, because metadata is part of managing any data-bearing project. The fact that many researchers have been unaware of the importance of metadata as an integral part of a data management strategy has led to a systemic but incorrect association of metadata creation with preparation for archiving. In turn, then, the anxieties associated with 'data separation' (see 3.2) are projected onto the process of creating the metadata for the deposit.

3.9 Metadata content

Typically, three main classes of metadata are recognised (Nathan & Austin 2004):

- descriptive metadata
- administrative metadata
- preservation metadata

For example, **descriptive metadata** (about the whole deposit, or any relevant part of it) would be expected to contain information in at least the following categories:

- title, description, subject, summary
- keywords
- the language and its community
- contributors of all types and roles
- location
- dates
- any other information about the content of the deposit

Administrative metadata should help the archive manage the data as well as identify the researcher/depositor and their work context over the long term:

- depositor's affiliation, date of birth, nationality
- project details including funding and hosting institutions
- copyright, IP rights and other stakeholdings

- details of other archived copies elsewhere
- modifications and update status
- details of accession agreement
- source or provenance (where complex or different from that described in descriptive metadata)
- access protocols (see below)

Preservation metadata includes information relevant to the physical provenance and the ongoing physical preservation of the materials, such as:

- original carrier media (if relevant)
- formats, sizes
- any particular software requirements
- history of handling and format conversions throughout the resource's lifespan

As an example of the last point, it might be important to know, for example, the original format of an audio file. Perhaps the documenter made the mistake of recording in MP3 and then heard that the archive prefers WAV. If the would-be depositor then proceeded to convert MP3 to WAV before depositing, the archive would not know this bit of history. While the conversion would not restore any of the information lost on the original compression to MP3, or make the audio better in any way, it puts the material in jeopardy for the future because: (a) there would be no explanation for certain missing bands of frequencies; and (b) there can be interactions between different compression formats, and that could occur if someone disseminates the audio via another compression format in the future. Even if the depositor did not follow good practice and recorded in MP3 initially, and then compounded the error by converting the files to WAV, they can at least atone for their sins by providing metadata telling the archive what they did.

The preceding example is somewhat simplified because MP3 is a standard and open format which could be satisfactorily archived. The situation is different when proprietary compressions such as WMA or ATRAC have been used, in which case there is a strong justification for conversion to WAV, although the importance of documenting the conversion remains as strong.

Ideally, depositors should also provide **file level metadata**, which contains information such as the following:

- for media: duration, file size, MIME type, content type
- for text: font, character set, encoding, format, markup
- for images: captions, links to associated files

Remember that the metadata is itself the resource that enables search, navigation and access to the materials. Some resources, such as audio, video and images, that are likely to be of obvious interest and greater accessibility to community members, would ideally have their metadata, captions etc. provided also in the community language (and/or in the contact or dominant or national language).

Some metadata is used to bundle resources into packages of files that are meant to function or to be used together. IMDI, for example, uses the concept of a 'session' which bundles together an audio and/or video file, an annotation, an IMDI file which glues them together and documents the session, and possibly other files as well. The approach can be generalised, using, for example, some of the strategies described in Section 3.7 above (with clear and appropriate explanation of the conventions used, of course).

Bundles or sessions are really just a special case of linking files or resources. This is currently a very much underused strategy. For example a photograph of a particular language consultant should be able to be connected to all the audio, video, transcriptions, annotations, and other materials such as kinship information, in which that consultant plays a role. It is not very difficult to provide the links in principle, as long as all the metadata is explicit and unambiguous, preferably supplied in a format such as relational tables (a properly designed database, or spreadsheets) or XML. The key to such a linking strategy is to remember that in providing linked data and metadata you are providing the resources upon which a searchable, browsable, user-friendly interface or system would enable the traversal of links. You are not likely to be providing that interface yourself, so you can happily defer the issue of how the links are actually implemented to the archive, or some later development. The important thing is that you provide the information that constitutes the knowledge underlying the link, for you might be the only person in a position to put names to faces, as well as all the other categories that have been discussed earlier.

There are other kinds of metadata that are often overlooked, especially those which make resources accessible to community members, and/or are useful for language maintenance or revitalisation. These could be answers to questions such as: where are the songs? which ones are for kids? where are the segments where the grandparents were talking? where are the likely teaching and learning materials? It could be argued that it is not entirely ethical for researchers to spend hundreds of hours making interlinear transcriptions, without providing simple metadata to enable access to the more community- or pedagogically-oriented content (Nathan & Fang 2009).

Finally, there is the area that archivists call 'access protocol', which concerns addressing sensitivities about data through formulating and implementing access restrictions. This is an area where ELAR has placed a significant emphasis and attempted to play an innovative role, by aligning access metadata categories and values (and the processes for implementing them) with the particularities and intricacies of endangered languages documentation and its data. Archives which use an approach to access control of one-choice (open or closed) and one-stop (define access conditions once and permanently at time of deposit) cannot take into account:

- the shift to disseminable digital media which potentially identifies individuals;
- the ethical and emotional factors often associated with documentation data;
- the differentiation of access, i.e. different formulations of access and restrictions for different groups and individuals; and
- the changeability of protocol over time, as personal, political and other conditions change within the community and outside it.

Access protocol seems to be inherently and intimately connected with the field of language documentation. Documentation focuses on recorded (primary) data, which means that in principle that there are more people involved (more 'human subjects'), there are more genres, and quite likely less researcher knowledge about the conditions under which the data is collected (e.g. compared to standard research data collection). Ethical approaches emphasising community participation (see Austin 2010a) mean that speakers and consultants have more awareness about the documentation activity and more input to shaping its process and products. Furthermore, the potential for subsequent mobilisation (and combinations) of resources in support of language strengthening activities amplifies the issues of ownership and intellectual property.

3.10 On data, standards and tools

There are many sources in the language documentation literature that extol the value of adhering to 'standards', and indeed many processes and technologies depend completely on people following the relevant standards, whether they be railway gauges, temperature measurements, web page coding, or audio file formats. Some linguistic standards are implicit, such as three-line interlinear glossing (see Schultze-Bernd 2006; this is implied in linguistics texts and courses, rather than being prescribed in the way that we are urged by some to use particular file or metadata formats[9]). Currently, ELAR prefers to encourage well-designed and well-managed data, explicitly documented and provided with rich metadata, rather than to impose particular standards. Of course, good data management generally implies perspicacious and standardised representations, such as Unicode encoding for characters, and interoperable data formats such as plain text, tabular and XML-based data. For further information about the file formats ELAR recommends, see the depositors' page at http://www.hrelp.org/archive/depositors.

I do not see the function of an archiving chapter as just to dictate a set of 'correct' formats and practices. What is correct and appropriate is relative to particular contexts, goals, current technologies, and target audiences. Formats and technical factors change over time, although some, such as Unicode, XML and WAV have settled within the last 10 years or so.

It is also worth remembering that so-called software 'tools' such as ELAN and Toolbox are not actually tools in the normal sense. A hammer is a tool, but it does not tell you what sort of house you should build. However, ELAN imposes assumptions about what the user can and should do and how the resultant data can be used. Toolbox is prescriptive about the typology of the languages it can represent and its (in)ability to integrate media, etc. On the other hand, what I would call **real** linguistic tools, e.g. minimal pairs, are conceptual ones, not computer software. The same applies to data management tools, such as data modelling for XML and relational representations; these are devices for notating conceptual exploration, rather than prescriptions of software or standards.

3.11 How does the deposit process work?

As noted above, ELAR's main constituency consists of ELDP grantees but we also take deposits from anyone who has suitable digital documentation of endangered languages, with a preference for materials that will be available on open access as long as the depositor has the rights to deposit the materials. A deposit could be as small as one file: a minimum deposit would be one file, some metadata or inventory for it, and a deposit form (deposit forms are available online at http://www.hrelp.org/archive/depositors/). The act of depositing does not have to be a singular event; depositors can submit some parts of their collection, and then add to them or update them later. This 'ongoing

[9] The Leipzig glossing rules (see Bond 2010:250) are a **recommendation** not a prescribed standard.

archiving' approach suits the workflow of documentation, where audio and video files are usually ready earlier and not likely to be further changed, while transcriptions and annotations are likely to be incremental in both quantity and quality (e.g. as more material is transcribed, and the documenter's understanding of the language increases or their analysis changes).

Delivery of the materials to the archive can occur in a number of ways. Currently, the most frequent method is via portable external hard disks. Many researcher have a spare one, perhaps an older one of smaller capacity. Some grant applicants now include in their budgets the cost of an additional hard disk for assembling and sending their deposit (cf. Austin 2010a). Portable hard disks can be easily posted or sent by courier, and after ELAR has copied off the data, we can send them back. Of the many we have sent and received over the past several years, so far not a single one has been damaged or has failed. Most recently, we have purchased several such disks as a little 'fleet' that we can send out to those who do not have a spare disk to send us. We adopted this strategy in particular to discourage people from sending us DVDs and CDs.

At ELAR we have found CDs, and especially DVDs, to be unreliable. Approximately one in ten DVDs is unreadable or partly unreadable. In addition, they are simply not a rational means of delivering larger volumes of data. At the supply end, researchers somehow have to make their data fit into 600 MB or 4GB chunks, leading to arbitrary re-organisation of data and confusion at the receiving end when the archive tries to reconstruct what the depositor initially intended (if indeed we receive any information at all about how the files have been distributed across the disks).

Those processes, together with burning the disks at the supply end and feeding them in to a DVD drive at the receiving end, create a lot of unproductive work for depositors and for ELAR staff. Only four years ago, a depositor sent us a stack of exactly 99 disks, but fortunately that is unlikely to occur again.

In some cases, conferences and similar events provide an opportunity for the depositor to meet with the archivist or representative and hand over a disk or arrange for the archivist to copy the large media files. The depositor can then email the deposit form and the more compact text-based files such as transcriptions.

Email can also be used, especially for sending text materials and media samples (edited down to one or two minutes) for evaluation. In the future, ELAR will provide a direct web-based upload facility.

Late in 2009, ELAR received its the first deposit delivered via an SDHC (flash memory) card. That development was made possible by the increase in capacity and decrease in price of flash memory. It was an exciting moment that encapsulated the radical changes in data storage that will dramatically change the way we work. For example, flash memory can now be bought for less than 1 pound per gigabyte, which is cheaper than previous forms of media carrier (cassette, minidisc, DAT), meaning that memory cards holding recordings should no longer be re-used but should be labelled and filed as effective means of additional backup.

3.12 Recent developments at ELAR

ELAR's online catalogue system is currently in development and the first phases, the deposit catalogue listings, have been made public. Delay in completing the data access components have actually worked in our (and our users') favour, as several developments in the dynamics of web-based interaction have only recently come to fruition. Web 2.0, or 'social networking', has arrived. In the form of websites such as Facebook and MySpace, a large number of people have become fully accustomed to managing interaction with those who they designate as their friends. The social model implemented by these sites is based on establishing and maintaining relationships that confer access rights, which is just like access protocols for archived deposits.

We surveyed the access conditions selected by ELAR depositors between 2005 to 2009. As shown in Figure 11, the deposit form offers several options, which could be summarised as 'open access', 'restricted access', 'access on a case-by-case request', or 'no access'. Our survey found that the majority of depositors opted for access on a case-by-case request (their second preference was for describing or enumerating the groups or individuals to be given access). Although their first preference might seem counterintuitive because they are obliging themselves to answer each individual request for access to their deposit materials, it exhibits their appreciation of the sensitivity of materials and the fact that access is a relative matter that depends on several factors, but especially on the identities and the purposes of those requesting access. We took this as strong evidence in favour of developing a social networking approach to archive access management.

Figure 11: Main part of ELAR depositors' form, protocol (access conditions) section

P1. Anyone □

Any person may view/listen to or receive a digital copy of any part of the deposit

P2. Certain people or groups

Choose any combination of P2A, P2B, and P2C:

P2A Research community members

What level of access (choose one only)?

P2A1. They can receive a digital copy of requested material □

P2A2. They can view/listen but cannot receive a digital copy □

P2B. Language community members

See below regarding identifying members

What level of access (choose one only)?

P2B1. They can receive a digital copy of requested material □

P2B2. They can view/listen but cannot receive a digital copy □

P2C. Particular named people or bodies □

See below regarding identifying people/bodies

P3. Depositor is asked permission for each request

You will be contacted and asked for permission on each request.

How do you want to be contacted?

P3A. Requester is given address to contact you directly □

P3B. ELAR will relay requests to you □

P4. Only the depositor has access □

Persons other than the depositor will not be able to request access.

In 2010 ELAR will release its data access system, which is a heavily customised open-source content management system (Drupal) based on PHP, MySQL and JavaScript. Just as in a social networking site like Facebook, users will be able to state their credentials and apply to the depositor to access restricted materials (which corresponds to 'I want to be your Facebook friend'). The advantages extend beyond the flexibility this brings for both depositors and users, and people will be able to have whatever dialogue is necessary. This system is going to fully implement our policies of respecting sensitivities and restrictions, while at the same time containing ELAR's administrative workload by delegating much of the activity to the depositors themselves, just as they expressed a preference for doing. For more details about this new model for archiving, see Nathan (2010a).

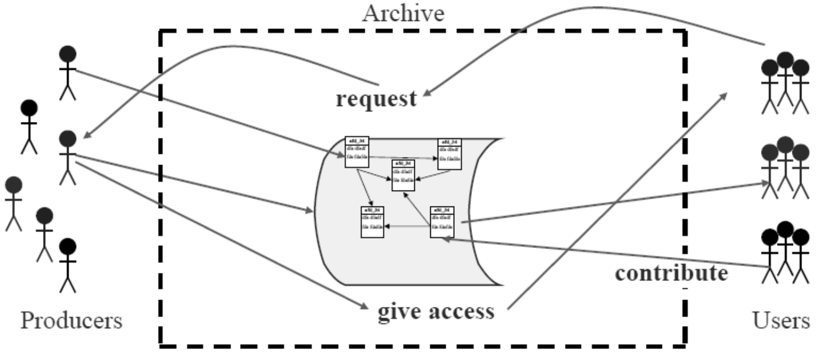
4. Conclusions: archiving for the future

Romaine (2006) has noted that intergenerational transmission may soon be supplanted by institutional learning for many endangered languages. In the longer term, however, documentary corpora and the archives that hold them will become the key vectors of transmission for many endangered and extinct languages. Therefore, the theory and practice of documentation, and the methodologies and capabilities of language archives, will play a crucial role in the future states of many human languages.

Just as documentation itself has found an ethical and community-oriented footing (see Austin 2010a), language archives need to redefine themselves. At ELAR, we believe that we exist in a time when digital preservation practices have rapidly matured and can now be subsumed to an understanding that we must function as the hosts of an important component of human heritage. Management of non-preservation functions will be largely handed over to depositors and users. Tomorrow's digital language archiving is not about technology but about relationships and commitments.

The OAIS model shown in Figure 4 is replaced by the one shown in Figure 12, where the archive becomes predominantly a forum for developing and conducting relationships and data exchange between producers and users of language documentation.

Figure 12: Archiving redefined as the platform for the conduct of relationships and data exchange.



References

Austin, Peter K. 2010a. Applying for a language documentation research grant. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 285-299. London: SOAS.

Austin, Peter K. 2010b. Current issues in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 12-33. London: SOAS.

Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3): 557-582.

Bond, Oliver. 2010. Language Documentation and typology. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7,* 238-261. London: SOAS.

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

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1), 161-195.

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

McConvell, Patrick. 2007. Video - A linguist's view (A reply to David Nathan). *Language Archives Newsletter*, 10: 2-3. [online at http://www.mpi.nl/LAN/issues/lan_10.pdf]

Nathan, David. 2004. Documentary linguistics: alarm bells and whistles? Seminar presentation, SOAS. 23 November 2004.

Nathan, David. 2006. Proficient, permanent, or pertinent: aiming for sustainability. In Linda Barwick & Tom Honeyman (eds.) *Sustainable data from digital sources: from creation to archive and back.* 57-68. Sydney: Sydney University Press.

Nathan, David. 2007. Digital video in documentation and archiving. *Language Archives Newsletter*, 9: 3-4. [on line at http://www.mpi.nl/LAN/issues/lan_09.pdf]

Nathan, David. 2010a. Archives 2.0 for endangered languages: from disk space to MySpace. *International Journal of Humanities and Arts Computing,* Volume 4 (Special issue).

Nathan, David. 2010b. Sound and unsound practices in documentary linguistics: towards an epistemology for audio. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7,* 262-284. London: SOAS.

Nathan, David & Peter K. Austin. 2004. Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 2*, 179-187. London: SOAS.

Nathan, David & Meili Fang. 2009. Language documentation and pedagogy for endangered languages: a mutual revitalisation. In Peter K. Austin (ed.) *Language Documentation and Description. Volume 6*, 132-160. London: SOAS.

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

Romaine, Suzanne. 2006. Plenary lecture at Georgetown University Round Table on Languages and Linguistics, 5 March 2006.

Sallabank, Julia. 2010. Language documentation and language policy. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7,* 144-171. London: SOAS.

Schulze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of Language Documentation*, 213-251. Berlin: Mouton de Gruyter.

Wittenburg, Peter. 2007. Video - A technologist's view (A reply to David Nathan). *Language Archives Newsletter*, 10: 3-5. [on line at http://www.mpi.nl/LAN/issues/lan_10.pdf]

Woodbury, Tony. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description*, *Volume 1*, 35-51. London: SOAS.

Discussion questions

1. Look at the directory name and filenames in example (1) of section 3.7. Why do you think the depositor has chosen these names? Do you think they are the best names for this purpose? Do all these files need to be archived?

2. Who should decide what is to be archived? What criteria could be applied to help make the selection?

3. Is archiving enough? What other means of dissemination/distribution might be useful, and how do these relate to archiving?

4. As stated in the chapter, ELAR is going to ask depositors to play a major and ongoing role in managing their deposits. What tasks do you think this will involve? Do you foresee any problems?

5. Have you thought of setting up your own personal data archive, now or in the future? If you do so, what issues would you have to think about?