Archiving and Language Documentation: from Diskspace to MySpace
David Nathan

Archiving

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 our new 48 terabyte disk array unit shown in Figure 1:

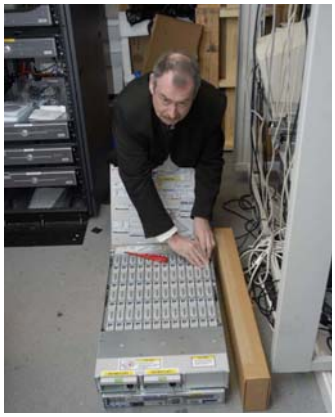

Figure 1. ELAR’s Mr Thomas Castle commissioning “the Numbat”, our main 48TB storage unit.

Or maybe you think about that thing you call your own archive which is that 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.

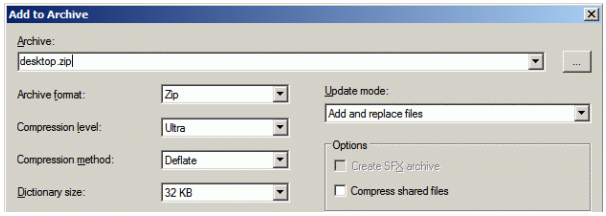

Figure 3. Your computer has mysterious predilection for archiving.

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

What is a language archive then? Well, one answer is that it is the sum of all the horrific problems we have to face. Maybe I can just blame somebody else for the description if not the problems themselves. See Figure 4 which pictures Doug Whalen, one of the key figures in the endangered languages field.

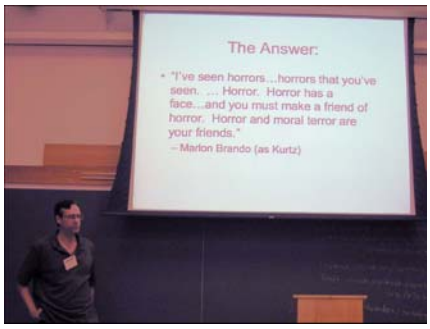

Figure 4. Doug Whalen presages the horrors of archiving.

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 in to 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 of a chair, 5 minutes long. 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 (which I possibly shouldn't really be mentioning!). All equipment fails eventually, and we've certainly had our share of equipment failures leaving us only one or two *more* disasters away from losing data. Of course, we're professional horror managers so that's never happened!

Digital archiving

The Endangered Languages Archive at SOAS is responding to the needs of digital archiving in our field 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 classically, an archive has been defined as a trusted repository. You as a documenter of an endangered language want to entrust your materials to a facility that will not only reliably preserve it but also to respect and implement any access conditions or restrictions that you apply. And usually those capacities are going to 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 there 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 5 shows how this simple archivist's brain works. It presents the model developed by the Open Archive Information Systems (OAIS) project which 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 of follow its main principles.

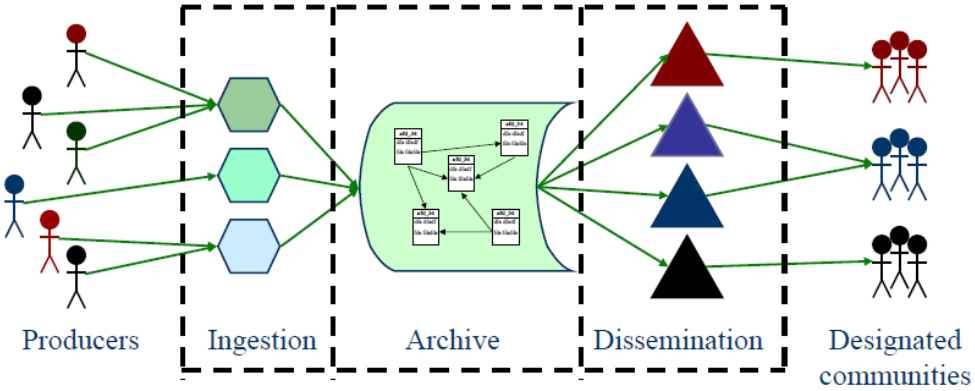

Figure 5: OAIS view of digital archives. Notice the range of dissemination objects to cater for various "designated communities".

Figure 5 shows the community of producers and maybe that includes people like you. Inside this dotted box is where we, the archive, live. The model divides it into three functions. First, we have got the ingestion stage – horribly anatomical though it sounds, it just means that the data enters the digital domain at the archive. In the middle is the archive storage and all the supporting processes such as backup - and then at the start of the "output" end is an important part of the architecture that is making dissemination possible, 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 you 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 you will see, later, we identify various such communities: research communities, language communities, general public and so on.

Archiving of language materials

So what is archiving of language materials? It 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 So the documenter has to understand,

inscribe and encode that knowledge somehow, by describing, transcribing, annotating, illustrating, marking up - all ways of giving form to that 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 our field 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. And that archive-driven approach is something that I criticise quite strongly. Good data management and judicious use of standards are part of any research area - especially ours 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 towards the end of this lecture, things like building relationships and providing a platform for relationships and transactions between the information providers and the information users.

So archiving is far from being just back up. Neither is it just dissemination or publication – throwing something up on a website. Neither does it define good linguistic practice. 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 you are already creating those kinds of resources.

I would like to use a (made up!) example involving a former prime minister of this country, Winston Churchill. Imagine going to the Churchill archives; you might find his famous pipes in a drawer. Now there is no way the archivist went to Mr. Churchill before he died and asked him to arrange his pipes so they would look nicely arranged 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.

So the following is what we say a language archive can offer, as we summarise it in our project flyer:

- Security – keep electronic materials safe
- Preservation – keep them safe for a long time
- Discovery – help others to find out about your deposited materials. And also to help you to find out about who is interested in your materials and how other people have used your materials
- Protocol – all the issues surrounding sensitivities and restrictions
- Sharing – or dissemination – facilitating other people to use the materials
- Acknowledgement – create citable acknowledgement
- Mobilisation – adapting materials and putting them to work, for example in language support and revitalisation activities. I will not say much about that in this lecture;

however, language archives, because they often have relevant technical skills, are able to help in the creation of usable language materials for language communities
- Quality and standards – researching and then informing our clients about the nature and formats of materials that best guarantee preservation. We spend a considerable amount of our resources on training, offering advice and providing feedback.

There are many kinds of language archives and if you are in the position where you are going to archive materials, it is probably worth finding out which is relevant for you.

- There are some which are local, serving their local community and they may, for example, like the archive for the Squamish Nation in North America, not serve outsiders, because they do not have the resources or they want the control over and the privacy of their own community's resources.
- There are regional archives like the Archive of the Indigenous Languages of Latin America (AILLA), which is interested in Latin and South America, and PARADISEC in South East Australia, which is primarily interested in materials from the Pacific.
- Finally, there are archives of international scope such as DoBeS and ELAR.
- See the DELAMAN website (www.delaman.org) for a list of more archives that specialise in endangered languages.

Some archives are associated with research institutes like the one at AIATSIS[1] in Canberra or the Alaskan Native Language Archive.[2] Some, like ELAR or DoBeS[3] have a very distinct advantage in that they are closely coupled with a granting body which creates a much stronger partnership with documenters who go on to be depositors throughout and beyond the life span of their projects. Another dimension to check out is whether the archive is a digital only one like ELAR, or can offer physical preservation (and perhaps restoration) of tapes and manuscripts.

Who are the users or the designated communities as were mentioned earlier? The DoBeS people are clear that their main users are depositors. For ELAR, this will also be 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 lecture). Depositors want to work with the archives to deposit materials, access materials that they may have lost or not have with them, update materials.

But we should not forget that language communities could be significant, if not the largest, potential users of archive materials. We have heard anecdotal reports that up to 95% of those accessing the Berkeley Language Center collection, for example, are community members. I have certainly seen something like this in Australia at AIATSIS – not actually in the archive but in the library – when native title legislation changed, suddenly the proportion of people using the library shifted strongly 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.

[1] See http://www.aiatsis.gov.au
[2] See http://www.uaf.edu/anla
[3] See http://www.mpi.nl/DOBES/

There are also other researchers. You know from Himmelmann's (1998) exhortation to document for a variety of the other disciplines and future usages that there are other audiences for documentation and indeed other stakeholders. I think it is easy to forget some categories of people that are extremely important, for example catalytic people like educationalists who often only need to be convinced that there are resources for a language in order for them to open up their purse strings and help to foster language programmes in schools.

Then of course, you have journalists. We are plagued by journalists who want to have stories about the last speaker of languages and so on. And the wider public, many of whom have very benign or intellectual interests and some of whom are looking for cute indigenous words for their new boat.

There are various archive networks and bodies so we have not just sprung out of nowhere and sitting alone. 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 really important for us. Others include OAI (Open Archives Initiative), OAIS Open Archival Information Systems (initiated by NASA space agency) and the Open Language Archives Community (OLAC).

More recently there are a couple of groups who are or have been influential in the way our small but vigorous community of endangered languages archives are working. One of them is the Digital Endangered Languages and Musics Archives Network (DELAMAN). It has an annual meeting and has been involved in issues such as training, getting archives to pool their resources for some common operations, such as a shared portal for searching, and establishing citation standards so that you as researchers can start to have a way to get your corpus or data work recognised. The following example (Figure 6) is not meant to be definitive but just to give you an idea of our initial recommendations for citing materials that are in our archives, either at the collection level or individual files.

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.

Figure 6. Examples of citations sent by Heidi Johnson of AILLA:

Language archiving is different and it is difficult. In fact we might say that archiving language is impossible, a stupid enterprise. 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, 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 is either unique to that language situation or is perhaps yet unknown. Given the estimations about how many languages are in the world and how few of them have been documented, it is perhaps rather premature for us to be told that we have to use certain sets of morphological glossing terms, for example.

Archiving of *endangered* language materials

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

- languages, cultures, communities, individuals, projects are all extremely different.
- fieldworkers are often quite an unusual if not eccentric bunch of people.
- the genres for our field. While some are stabilising, for example, a video or audio plus ELAN file, in general the genres of our field are not really settled. This makes it difficult for archive staff to fully manage materials and you will see later the kind of strategies that we are adopting to deal with this.
- sensitivities and restrictions – languages are endangered, because people are under pressures or suffering in various ways. So this quite naturally means that language materials are associated with sensitivities and restrictions so these in turn are part of our field and that is amplified even more for archives which have becomes points of access or distribution.

The Endangered Languages Archive (ELAR)

Our archive, ELAR, is one of three programmes of the Hans Rausing Endangered Languages Project. The others are the academic programme, headed by Peter Austin, and ELDP the granting programme, currently headed by Peter Sells.

ELAR has a staff of three: myself (archivist), Ed Garrett (software developer), and Tom Castle (technician).[4] From time to time we employ research assistants as well. We are involved in developing policies, preservation infrastructure, cataloguing and dissemination, facilities, training, advice, materials development and publishing.

We currently hold about 70 deposits and a total volume of about 8 TB (terabytes), with a lot of materials flowing in. Our main providers are the ELDP grantees. Our main mission is to archive the materials that are generated as a result of ELDP funding. But to some extent we can also archive any digital materials for endangered languages. We expect the volume to nearly double over the next eighteen months because materials tend to come in from six to eighteen months after the end of the funded projects and many of these projects have finished over the last year or so. Figure 7 shows 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

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

data type by

[4] And also a fraction of the faculty technician, Bernard Howard.

eaf	33	176	*volume (MB) and number of files, sorted by volume*
text	32	781	
lex	9	29	
trs	5	246	
xls	1	19	
imdi	1	26	

Figure 7: ELAR's relative holding of various data/media types

You might be wondering what kind of materials we have. Figure 7 provides statistics for a representative sample showing that we have far and away audio as the greatest number of files. There are a large number of images as well as the aggregation of all the text formats. What is particularly interesting, I think, is the top two lines showing you how we have almost ten times the number of audio files compared to video but their volume is comparable. What this means is that for us space for video is a major issue. I could go into long discussions about video and its value and associated methodological problems but I will try to not do that here. But you can imagine that as the use of video takes off - which is what is happening now - and as High Definition video (which has larger file sizes) becomes commonplace, then holding, preserving and delivering video will be, in a sense, a crucial factor for us.

It is interesting to compare ELAR's activity profile with how a digital language operated only fifteen years ago. I used to run a small archive at AIATSIS in Canberra called ASEDA – the Aboriginal Studies Electronic Data Archive. Although small, it was one of the first digital language archives, and continues to run today (although perhaps not much longer). It was founded by Nick Thieberger in the early 1990s, based on the model of the Oxford Text Archive.

Its mission was more or less as a backup 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. Ad so many linguists then (even more than now) used Macintosh which seemed to be prone to problems (much less now since Apple's OSX). So the big picture was that materials were vulnerable at any moment in time. These materials themselves consisted entirely of textual materials – lexicons, grammars and texts.

It is interesting to look back on how much things have changed. The modalities of the data have changed radically. As you saw (in Figure 7) we now hold audio and video media as the predominant genre of documentation. Nowadays we have an information environment that is much more developed and standardised (e.g. availability of "rock solid" data coding methods such as Unicode and XML, and a few widely accepted conventional metadata schemas). We are cataloguing and disseminating materials via the World Wide Web. Our storage methods have also changed radically. In ASEDA days when we wanted to have more back-ups what we did was buy more Macs or magneto optical storage drives (the equivalent of the later minidisk technology - which 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 as a University, a big company or even a bank uses. We are much clearer about our function as providing long term preservation of significant materials, not merely backup of vulnerable materials. 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.

Why digital?

ELAR is a specifically *digital* archive, although we do occasionally digitise analogue materials such as 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/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. It turns out that there is only one critical, i.e. unavoidable reason for using digital form, and that is for media. The only way we can make perfect copies of things - and therefore to carry them forward into the future, regardless of the physical 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 really poor.

The digital principle is familiar to us as linguists; we rely on it all the time – our phonological principles, our morphological principles, our lexical principles; these are all digital because they use discrete symbols (e.g. a sound is either [p] or [b], a word is either 'dog' or 'dock'). For computers, the choice is either 0 or 1. So 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. Good cassettes will last maybe 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.

Analogue is real stuff, and if you copy a tape you are making a real thing cause a change to another real thing. And that 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, you would actually be better off carving your ELAN transcriptions character by character into stone. More seriously, it is said that the very best means of preservation is to print barcodes on microfilm. Under good preservation conditions, including temperature and atmospheric control, that should last up to 1000 years. Barcodes for the symbols, put onto microfilm – that is a good medium. But we are not likely to do that, at least right now. Using today's technologies, we can copy and transmit data with zero loss. And then there is the rest of the functions needed by our discipline and our culture – all the practical realities of cataloguing, sharing, disseminating, transmitting, broadcasting, modifying, reusing, combining, etc. – all of which are much more possible in the digital domain.

In some ways the digital medium, as we know it today, is the worst possible solution for long-term storage, because you have to put electricity into it every moment to keep those disks spinning, to keep the air conditioning running, etc., although there are currently interesting changes in technology such that we are looking at solid state storage perhaps being available at a suitable scale in about 5 years. And there are huge costs in digitising materials, setting up infrastructure and then maintaining, upgrading and replacing it. At ELAR we have found that you need both strategy and luck to get the infrastructure right. If you have just committed to spend £10,000 or even £100,000 and then find that the

technology/price ration changes dramatically a month after that, then you will feel rather disappointed. It happened, in a way, to us. Less than 5 years ago, we paid about £30,000 for 8 terabytes of data storage. We bought items that were parallel to SOAS' 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 - which amounts to only 5% of the original unit's price per unit storage. At least with the new unit, 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 the costs change 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 shooting 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 to use the right tools and techniques to provide these formats. We as archives 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 materials (in ".eaf" format), for example – because they're XML-based they can be liberated into other formats, but you need ELAN to experience the tiers and other functions of ELAN files. Fortunately we can archive ELAN because it is free, open-source and made by our archive comrades at MPI 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 work to make sure that data depends on the least numbers of layers of encoding and software, along the lines suggested in Bird and 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.

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 work 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 a child up for adoption.

Nevertheless, a hallmark of today's archiving is that documenters and archives are increasingly working together. How do we interact with documenters? Diagram 8 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), participating 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)

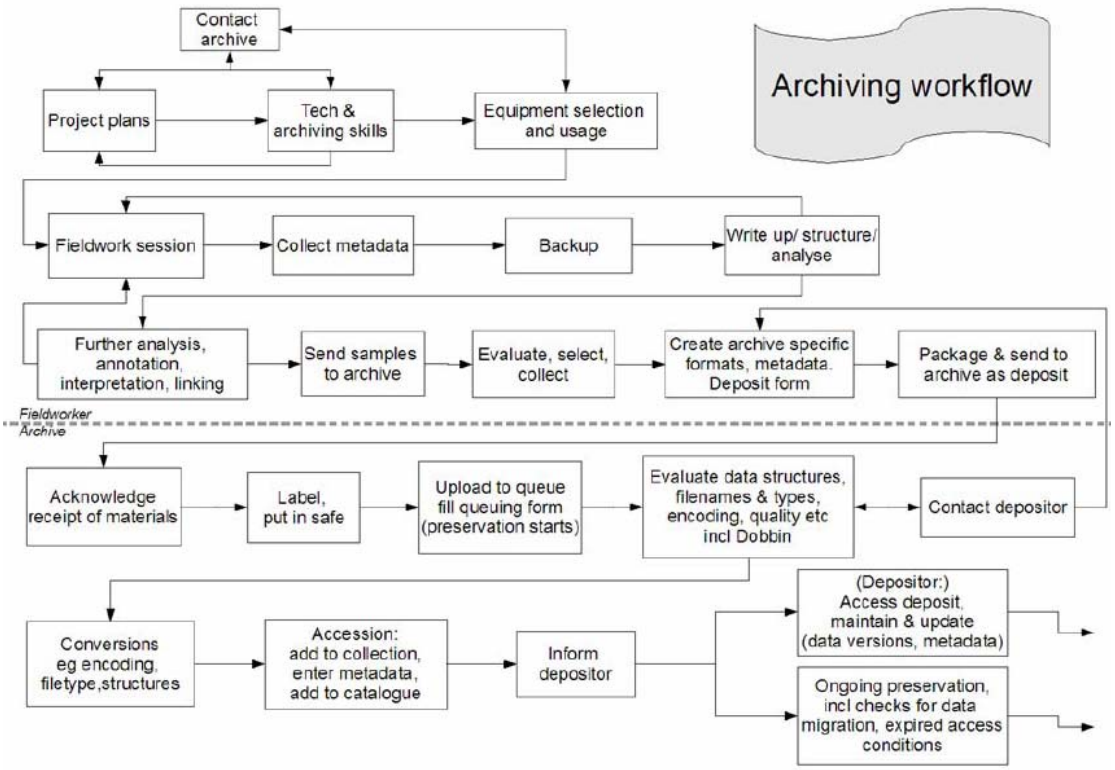

Figure 8. A semi-serious flow diagram of interactions between documenter and archive. Activities primarily in the hands of documenters are above the dotted line.

For the documenter, the "main game" might be the third row (fieldwork session; collect metadata; backup; writeup/structure/analyse). What we are increasingly encouraging is for documenters send samples to the archive. 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 that a microphone does not match a recorder, or that there is some problem with a format or the way that the documenter is encoding or structuring her 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 model, which is discussed at the end of this article.

To summarise, as an archive we are involved in:

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

thus participating 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 fly away forever.

ELAR Feedback

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

- Document type
- Document format/layout/data structures
- Character/language representation
- Linking/references
- 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

In order to give you an idea of the kinds of feedback we give, Figure 9 contains an excerpt from one such feedback form (suitably anonymised):

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 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?).

Figure 9: Excerpt from feedback to depositor on data sample

The case of the feedback in Figure 9 was profitable for the depositor, the archive, and future users of the data. It turned out the background noise (hiss) that I pointed out 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 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, we receive far more audio than we could possibly listen to (the average deposit seems to have around 30-40 hours of audio). To help deal with this, we have 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 10 shows the result of one such batch run, where the problems are indicated by highlighting of the relevant points in the waveform representation:

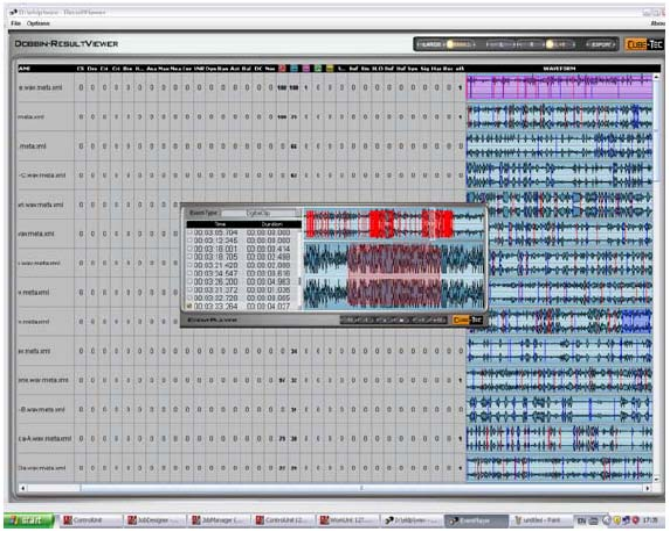

Figure 10: Dobbin report, showing audio evaluation summary and highlighting of problem areas on the waveforms.

Clicking on any of the problem areas opens an editor where we can inspect and diagnose the problem. Figure 11 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 record the problem as metadata associated with the deposit.

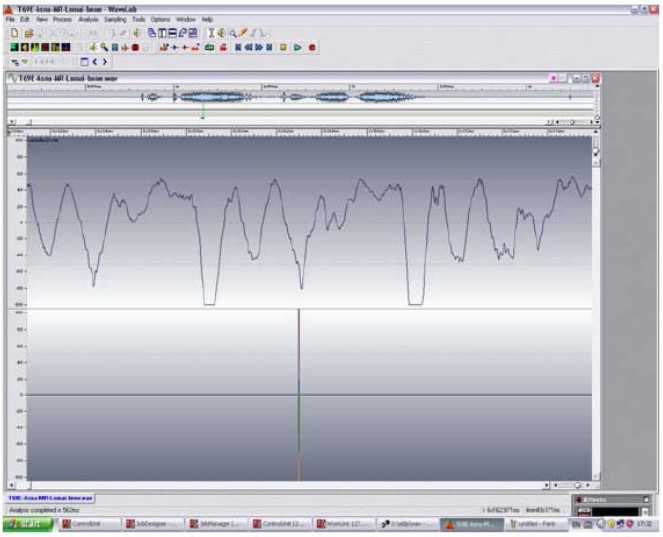
Figure 11: Dobbin has searched through hundreds of audio files and found various problems, including this example of clipping.

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

- we provide training at various venues
- we provide advice, both general (e.g. that on our website) and specific
- we preserve your deposited materials
- we promise to implement your access restrictions, etc. (see section °° below)

- to achieve the best possible documentation ¬quality, through a distribution of labour philosophy: it may not be best to expect you to convert your data to XML or some other portable format (see below on file and data formats); you may not have the skills, and the result may be damage to your data. It is best for us together to find out what your skills are and, where appropriate, we would do the data conversions
- we are gradually working towards providing web-based deposit management, which will allow people to update materials, send new versions, make corrections and additions, etc. (see final section)
- occasionally we can provide some equipment and services, and sometimes, on a case by case basis, we develop resources, such as websites, videos, or multimedia

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 amongst individual files, metadata and other interpretive items have scope over a range of files, and there may well be units intermediate in structure between an individual file and the whole deposit (for example, an audio recording and its transcript). We archivists like to refer to these related packages of files as 'bundles'. Such bundles - their identity, structure, and content - should be made quite explicit through associated metadata. Some bundles have 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 he/she is putting together the data and working on it personally, but is liable to be misunderstood or broken as soon as the data is moved to a different location, so it is important to explicitly document such structures and local conventions in some kind of metadata table, or, if the system is simple, in a simple "readme" file which plainly explains the conventions. ELAR's new cataloguing system (available later this year) is going to provide a dynamic online method for creating and describing bundles.

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

- media – sound, video
- graphics – photographs (of consultants, the language speaking settings, objects descried or discussed), diagrams (of the recording environment, sketch maps, of objects described), 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)
- text – fieldnotes, transcriptions, translations, grammars, description, analysis
- structured data – aligned and annotated transcriptions, databases, lexica
- metadata – structured, standardised contextual and interpretive information about the materials.

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 and Simons (2003). Unfortunately, teaching curricula and documenter practices are generally still considerably behind and need to catch up. Our broad and shared goal of documenting languages well means that we must find the best 'division of labour' at any one time between education and training curricula, documenter's responsibilities, and archive services.

Please refer to Bird and Simons (2003) 'Seven dimensions of portability for language documentation and description', which discusses 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).

Archive specific criteria for deposited materials

The criteria that are distinct to archives are that:

- materials for deposit 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 Metadata sections below); and
- materials are selected.

It is really important to select materials. There is no reason why the point at which the audio or video recorder was turned on, or any particular note that you made, is definitively part of a collection that should be carried into the future. Some materials may distract or even detract from a collection. So archiving definitely does not mean sending a dump of your hard disk, or sending the folder that contains everything from your project.

One depositor wrote 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 answer is that the depositor states criteria for what makes a good documentation resource and then applies 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 your materials are relevant, we'll take all of them; if they say that none of them are relevant we won't take any of them. Perhaps this depositor just wanted to be told 'you can send 40 GB', but we do not archive endangered languages by the kilo.

Some depositors have baulked at the idea of editing audio or video. This is in some cases due to a naïve view that recording captures an actual reality that is rendered untrue or fake by any intervention. In fact, most things you do in academic life are forms of editing and/or selecting. In your linguistic work, you selected, labelled, transformed/processed/edited, summarised, added/corrected/expanded, made links, made or assumed relationships between 'whole' and units, invented labels/IDs/scope/etc., and imposed formats. When you transcribed or annotated, when you chose examples to illustrate generalisations, or when you make decisions to ignore certain things in the audio (e.g. coughing or paralinguistic behaviour) you made selections amongst which things to pay attention to and which to ignore. It's 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, with what consequences, if any.

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. DoBeS, by contrast, has a vision for their collection (the IMDI-corpus) where all deposits are united as one single united corpus, through which a user can navigate seamlessly. ELAR has taken a less proscriptive 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, 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.[5]

(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, with such-and-such 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.

(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 to glossing codes used in transcriptions (ELAN transcription key FTG0025.pdf) and some additional read-me notes to the archivist (archivist_notes.txt).

[5] There are, however, some comments that could be made about the files and their names; see the tutorial questions.

(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 same (e.g. “WulaTuki_LunarEclipse”, “AngryD-Bsi”) or partially the same (“Kay07-aud-jul03” + a/b/c) filename root. 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 in not just one but 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 explicitly, consistently and completely list all the bundles and their parts in an inventory/catalogue file.

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 you, but what does it mean to a community member in China? In other words, your materials are not only incomplete but seriously flawed if they do not have sufficient metadata, because they are quite possibly understandable only by you.

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 - and which should be as rich and appropriate as the documentary materials themselves.

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 potential documenters say ‘I’m going to do documentation and this is going to be really useful for the community’, but a later view over 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 (Nathan and Fang 2009).

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 that Himmelmann gave us and has been constantly repeated (Himmelmann 1998, Austin 2010). But if we *do* want multiple audiences to understand our documentations, we are going to have to work a bit 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', from 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[6] captures deposit-wide overview and discovery metadata, and Ed Garrett is developing the online system to allow depositors and archive staff to add and modify the overview as well as file-level metadata via 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/projects can result in quite different metadata. So given that our goal is to maximise the amount and quality of metadata. we now have some evidence that flexibility is more important than standards. Currently, we are asking our depositors to send their metadata in portable formats (Bird & Simons 2003), such as spreadsheets or tables, and to think carefully about the content and structure of the content (see next section).

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 from leaders in documentary linguistics and from archives, and in some cases 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

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

associated with “data separation” (see above) are projected onto the process of creating the metadata for the deposit.

Metadata content

Typically, three main classes of metadata are recognised:

- 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 to 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
- 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 you had made the mistake of recording in MP3 and then heard that the archive prefers WAV. If you then proceeded to convert your MP3 to WAV before depositing, the archive would not know this bit of history. However, 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, the conversion 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, in the case that someone delivers the audio via another compression format in the future. Therefore, even if you did

the terrible thing of making recordings in MP3 initially, and then compounded your errors by converting them to WAV, you can at least atone for your sins by providing metadata telling us what you did.[7]

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. So 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 contact/dominant/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 the Section **File organisation** 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. However, it is not very difficult to provide the links in principle, as long as all the metadata is explicit and unambiguous, preferably provided in a format such as relational tables (properly designed database or spreadsheets), ·ML. The key to such 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 that 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 we have discussed earlier.

There are other kinds of metadata that are often overlooked, especially those which make resources accessible to community members, and/or which are useful for language

---

7

maintenance or revitalisation, 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.

Finally, there is the area we 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 a one-choice (open or closed) and a one-stop (define access conditions once and permanently at time of deposit) approach to access control cannot take into account (i) the shift to disseminable digital media which potentially identifies individuals; (ii) the ethical and emotional factors often associated with documentation data; (iii) the differentiation of access, i.e. different formulations of access and restrictions for different groups and individuals; and (iv) the changeability of protocol over time, as personal, political and other conditions change in the community.

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 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.

On data, standards and tools

There are many sources that extol the value of adhering to ‘standards’, and indeed many processes and technologies depend completely on people following the relevant standards, whether railway gauges, temperature measurements, web page coding, or audio file formats. Some linguistic standards are implicit, such as three-line interlinear glossing (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). Currently, ELAR is keen on encouraging well-designed and managed data, explicitly documented and provided with rich metadata, rather than imposing 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 we recommend, see our depositors’ page at http://www.hrelp.org/archive/depositors.

I do not see the function of an archiving lecture 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 worth remembering that so-called software 'tools' such as ELAN 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 language 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 software. The same applies to data management tools, such as data modelling for XML and relational representations, which are conceptual matters of exploration, rather than the prescriptions of software or standards.

How does the deposit process work?

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 are on open access and as long as the depositor has the rights to deposit the materials. A deposit could be as small as one file: as a minimum deposit we require 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 deposit does not have to be a singular event - you can deposit some parts of your collection, and then add to them or update them later. This "ongoing 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 his/her analysis changes.

Delivery of the materials to the archive can take place in a number of ways. Currently, the most frequent method is currently using portable external hard disks. Many people 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. Portable hard disks can be easily posted, and after we have copied off the data, we can post them back. Of the many we have sent and received 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 (see below).

We have found CDs, and especially DVDs, to be unreliable. Approximately one in ten DVDs is unreadable or partly unreadable. As well as that, they are simply not a rational means of delivering larger volumes of data. At the supply end, you somehow have to make your data fit into 600 MB or 4GB chunks, leading to arbitrary re-organisation of data and confusion at the receiving end when we try to reconstruct what the depositor initially intended (if 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 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 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 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.

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. Any delay in completing the data access components has 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 protocol for archived deposits.

We surveyed the access conditions selected by ELAR depositors from 2005 to 2009. As shown in Figure 12, the deposit form offers several options, which could be summarised as open access, restricted access, access on 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, 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 using a social networking approach to archive access management.

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.

Figure 12: Main part of ELAR depositors' form, protocol (access conditions) section

This year 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, web 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 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 (in press).

Conclusions: archiving for the future

Suzanne Romaine has noted that intergenerational transmission may be supplanted by institutional learning for many endangered languages (Romaine 2006). In the longer term, documentations 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 play a crucial role in the future states of human languages.

Just as documentation itself has found an ethical and community-oriented footing, 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 5 is replaced by the one shown in Figure 13, where the archive becomes predominantly a forum for developing and conducting relationships and data exchange between producers and users.

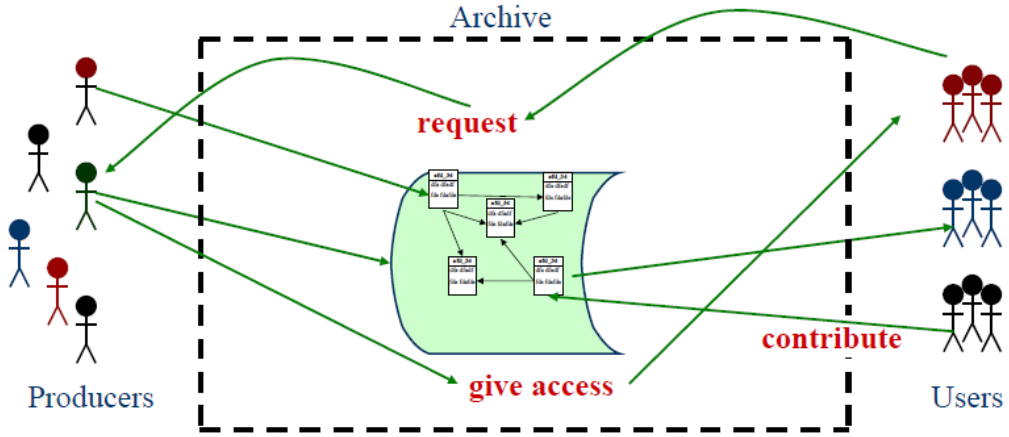

Figure 13: Archiving redefined as the platform for the conduct of relationships and data exchange.

References

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

Dobrin, Lise, Peter Austin, & David Nathan. 2007. Dying to be counted: commodification of endangered languages in documentary linguistics. In Peter 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.

Nathan, David. In press. Archives 2.0 for Endangered Languages: from Disk Space to MySpace. In *International Journal of Humanities and Arts Computing*, Volume 4 (Special issue), 2010.

Nathan, David. 2006. Proficient, Permanent, or Pertinent: Aiming for Sustainability. In Linda Barwick and Tom Honeyman (eds) *Sustainable data from Digital Sources: from creation to archive and back*. Sydney: Sydney University Press, pp 57-68.

Nathan, David and Meili Fang. 2009. "Language documentation and pedagogy for endangered languages: a mutual revitalisation." In Peter Austin (ed) *Language Documentation and Description*. Vol 6. 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.
Woodbury, Tony. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description*, vol. 1, 35-51. London: SOAS.

Tutorial questions

1. Look at the directory name and filenames in example (1) of **File organisation in deposits**. 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 lecture, 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?