Digital archives: essential elements in the workflow for endangered languages documentation

David Nathan, SOAS

1. Introduction

One of the developments associated with the increased attention to language documentation is the establishment of specialised digital archives that provide key contributions to endangered languages documentation and revitalisation.

This paper reflects the perspective and initial experience of the Endangered Languages Archive (ELAR) at SOAS, outlining interactions between linguists and modern digital archives in order to show that archives are now essential participants in the workflow of documentation, and to ask whether the degree of overlap between documentation and archiving is sustainable.

ELAR has been operating since 2005, and is an archive principally in the sense of Johnson 2004:142:

> a trusted repository created and maintained by an institution with a demonstrated commitment to permanence and the long term preservation of archived resources

ELAR joins a number of archives with similar goals and also concerned with endangered languages, such as DoBeS (www.mpi.nl/DOBES), AILLA (www.ailla.utexas.org) and PARADISEC (www.paradisec.org.au). However, as part of the Hans Rausing Endangered Languages Project (HRELP), ELAR is unique because it works in close collaboration with the two other HRELP programmes – the Academic Programme (ELAP), and the Documentation Programme (ELDP). ELDP is an endangered languages field research funding agency that awards about US$2 million per year across the world, and it is through its collaboration with ELDP that ELAR's activities reach out in time and space.

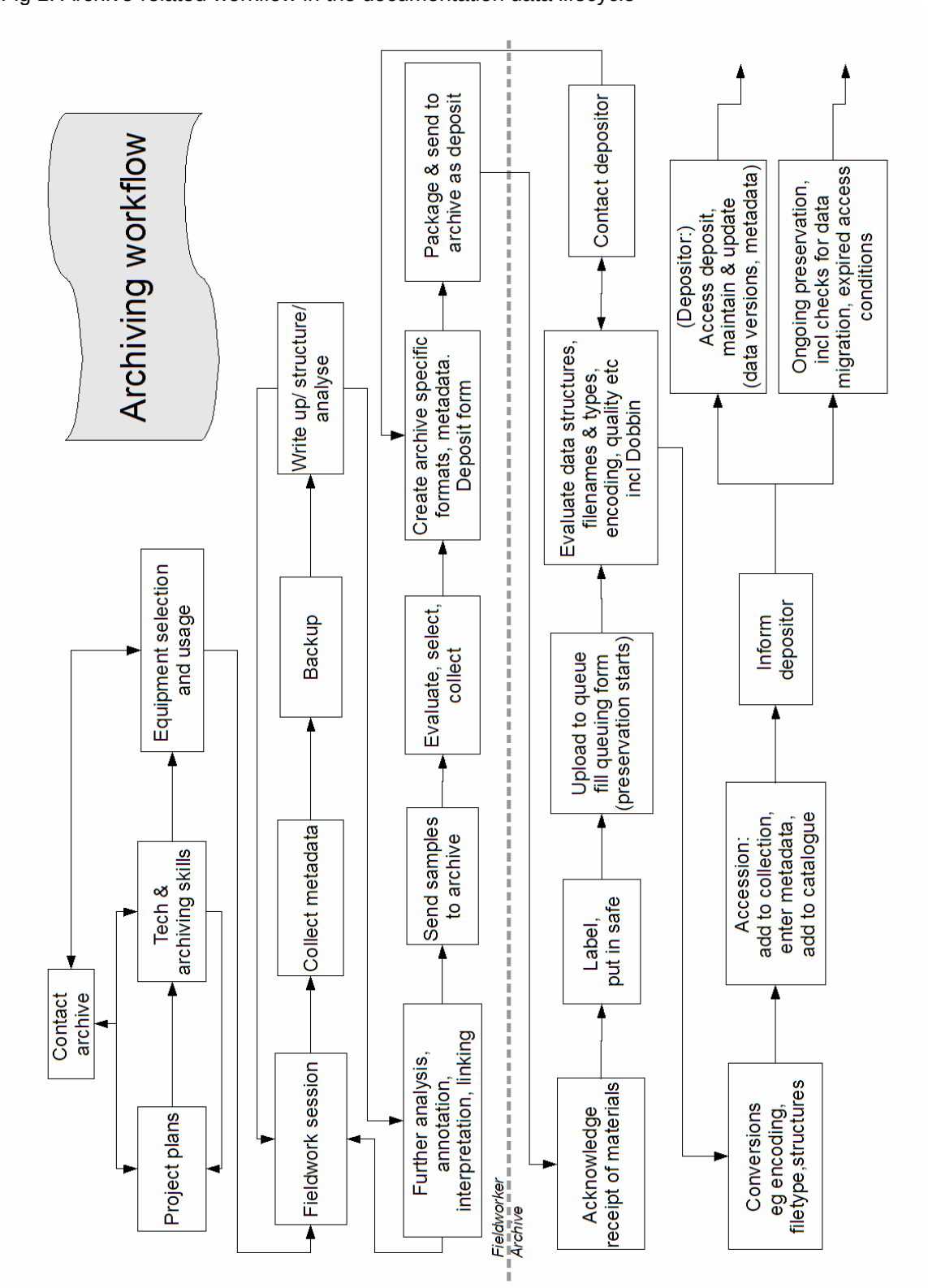
2. Interactions between documenters and archive

The following sections describe the course of a typical funded documentation project, from grant application to archive deposit, focusing on potential points of interaction between documenter and archive, based around a generalised workflow shown in Figure 1. In the top section of Figure 1 (above the dotted line), data is in the hands of the documenter, while in the area below the line it is managed by the archive.

Grant applicants have to think about archiving aspects from the outset because archiving obligations and suggested methods are built into ELDP grant conditions. ELDP application forms and guidelines have had significant input from archive staff relating to both archiving specifically and to various other technical recommendations. In addition, the ELAR archivist advises on preferences for equipment to be funded, checks applications and gives comments and recommendations on equipment and methodologies, and co-ordinates the technical training of new grantees. The archivist's recommendations may result in the applicant being requested to amend their proposal, with this process frequently taking the form of email exchanges of requests, explanations and information. In parallel, many applicants contact ELAR directly to put proposals or ask questions about particular equipment or

methodologies.

Fig 1. Archive-related workflow in the documentation data lifecycle



Once grants are awarded, holders are often requested to contact the archivist to discuss plans for collecting, preparing, and archiving data. At ELAR we provide guidelines, advice and services but we do require particular schedules, workflows, software or formats (see discussion in Sections 6 to 9).

3. ELDP training courses

The next point of contact between grantees and the archive is likely to be the training course that we run annually at SOAS for most new grantees. The course (see e.g. www.hrelp.org/events/workshops/eldp2006_6/) is held at ELAR and covers a variety of topics in documentation, with a focus on those less likely to have been a part of the participants' formal training, including recording, archiving, data management and technical topics as well as wider issues such as ethics and intellectual property. While we try to provide as much information as possible, we emphasise awareness of principles and methods above particular skills or proficiency in particular software tools. Here is a typical topic grid (for ELDP training 2007):

Fig 2. ELDP Training course topics

- Grantee projects sharing
- Language documentation
- Audio: principles, digital audio, practical, evaluation
- Transcription principles & practical
- Mobilising data for communities
- ELAN
- Archiving
- Administering your grant
- Consultation & elicitation
- Video: video in documentation, videography, practical, editing & evaluation
- Data management principles & practical
- Field practical topics (e.g. solar power)
- Advice "clinic"
- Ethics & IP

For participants who look to archiving and specific technologies to provide a complete and prescribed workflow, this approach can be disappointing; however, in general we receive very good participant evaluation of the courses.

4. In the field

Once in the field, documenters are typically involved in a cycle that runs from recording and elicitation sessions to write-up, transcription, analysis; in turn feeding into questions that inform further sessions with consultants – see Figure 1. Densely interwoven in this cycle are many processes and application of skills, including recording techniques, electricity supply management, care of media carriers, data formulation and media formats.

5. Archiving process

At some point depositors start working explicitly towards archiving their materials. For some, archiving concerns may have already considerably shaped how they have created their data, while for others, progression to preparing for archiving represents a departure from their normal way of working. In either case, we prefer to receive representative samples for evaluation and the opportunity to offer advice (in any case, it is useful to gain a sense of how documenters are working). So far, about half of ELAR's depositors have sent such samples, although it is still early in ELAR's operations and we hope this figure will increase. What the samples show is more or less the complete range from those using solely "traditional" or print-oriented methodologies (such as Microsoft Word documents) to those who produce using such methods and then convert to preferred formats (see Section 7), and a small number working entirely within recent archive-friendly formats such as XML.

At the top of the archivist's priority list is metadata. Metadata is important for archives because it is crucial for preservation (e.g. metadata about file types, data conventions, and about people who need to be consulted for permissions), for cataloguing (so the archive knows what it holds and can inform others), and access (for appropriate acknowledgement, access control). Above all, metadata covers areas typically least addressed in the preparation of linguistic data – explicit documentation of the provenance, methodology, conventions, context, and permissions associated with materials. Researchers have long recognised metadata in the guise of bibliographic data in the publishing context, helped by centuries-old conventions and the infrastructures provided by publishers and libraries. The "disconnect" for linguists is that previously data alone has not typically been disseminated, and linguists have conventions only for incorporating data within publications, such as 3-line interlinear format.

6. Content analysis of archive queries

Fig 3. Analysis of archive queries by content area

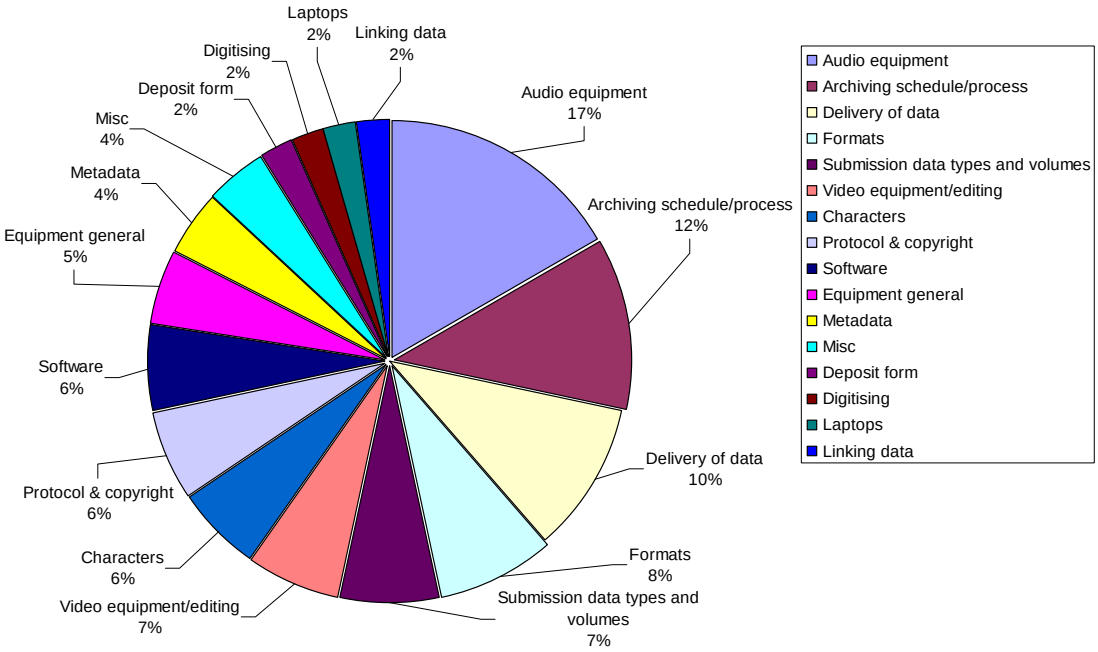


Figure 3 shows an analysis of 150 queries from about 50 grantees (or potential grantees) at various stages of their projects over a period of two years. Note the broad range of topics. Less than 40% of the queries relate to topics that are unambiguously associated with archiving – archiving process, data delivery and volumes, protocols and copyright and deposit form. Fully 40% of all queries relate to audio, video and other equipment.

What follows are some (anonymised) examples illustrating the content of a selection of queries. Grantee J asked about audio recorders. She was concerned that although minidisc recorders have been widely deprecated, use of solid state recorders typically requires a computer also to be available to move data from the flash media cards onto optical disks. This posed a problem in her research area, since the only source of electricity available was at the missionary station, and association with this was felt likely to damage her community relations and hence quality of documentation. We

discussed other powering options, the amount of recording intended and price of flash cards, and concluded that, since she had successfully used minidiscs in the past, using HiMD remained a viable methodology. Subsequently, she informed me that she changed her plan:

> I decided for a Marantz PDM 660. I will bring my old standard minidisc recorder in case there is any problem, but I also got several 2GB cards, an extra battery for my laptop and an extra hard drive. I think I should be able to make it work

Grantee K wrote to enquire about depositing video, asking for a specific space allocation (in gigabytes) to enable him to select his video material. Eventually, I explained that decisions ought not be made on such a basis:

> [we] do not have any policy stating the amount (maximum or minimum) of media material to be deposited. One could apply various criteria to evaluation/selection of what is to be deposited, e.g. audio/video quality, nature/value of interaction/narrative captured, whether or not transcribed/annotated, uniqueness, potential for future products, format/compression level etc. Ideally, this should be done by the depositor (quite possibly with input from the language community), not the archive, since it is the depositor who best understands the nature of the material and the language and community context. Therefore, we expect researchers to have some methodology, and understanding of the role of video in their projects, in order to be able to state and use some relevant criteria. Whatever meets those criteria, whether some or all, or none, of the material, surely defines the best selection for submission to an archive. I do not feel it is adequate on either our or your part to state what should be important aspects of linguistic documentation in terms of sheer data volumes

Grantee L sent data document samples for evaluation and feedback. Amongst other analysis, we reported a frequently encountered problem with character representation and fonts. L's documents, which included 3 languages/scripts, were created as MS Word documents requiring 2 additional fonts. This is guaranteed to cause problems for preservation – at the very least the fonts need to be archived with the document. However, unless the document is very explicitly labelled, a user may not know exactly what the document is supposed to look like (and therefore what to do) when faced with a sprinkling of "empty box" characters, or, worse, a jumble of readable but incoherent characters; the user may not even realise that they need to locate and install fonts. However, as described by Bird and Simons (2003), the fundamental problem for long-term preservation is not the nature of the font so much as that the information that signals the shift of language is merely the assignment of a font. Although of course *some* font is always needed, many researchers are still using older-style interceptive fonts which simply use the font's graphics to re-represent characters that are essentially just Roman characters. Therefore, if the document is transmitted or converted, then the *language* information is easily lost, perhaps without even any overt indication that anything has been lost. The "best practice" way of dealing with this problem is to use Unicode, although this is new to many linguists and is not guaranteed to provide a solution for all languages (Csató & Nathan 2007).

Grantee M had previously grappled with similar problems to grantee L and as a result he had started encoding his data in the archivist's favoured format, XML, by using Filemaker Pro's "export as XML" function. Here is a (modified) snippet:

```
<?xml version="1.0" encoding="UTF-8"?>
<FMPXMLRESULT xmlns="http://www.filemaker.com/fmpxmlresult">
   <PRODUCT BUILD="06/26/2002" NAME="FileMaker Pro" VERSION="6.0v2"/>
   <DATABASE DATEFORMAT="M/d/yyyy" LAYOUT="" NAME="Videos" RECORDS="13"
TIMEFORMAT="h:mm:ss a"/>
   <METADATA>
      <FIELD EMPTYOK="YES" MAXREPEAT="1" NAME="Index name" TYPE="TEXT"/>
      <FIELD EMPTYOK="YES" MAXREPEAT="1" NAME="Image description" TYPE="TEXT"/>
      <FIELD EMPTYOK="YES" MAXREPEAT="1" NAME="Date" TYPE="TEXT"/>
      <FIELD EMPTYOK="YES" MAXREPEAT="1" NAME="Content" TYPE="TEXT"/>
   </METADATA>
   <RESULTSET FOUND="13">
      <ROW MODID="16" RECORDID="40">
         <COL><DATA>Morly Beeta</DATA></COL>
         <COL><DATA>Interview with Morly Beeta</DATA></COL>
         <COL><DATA>Jan/13/05</DATA></COL>
         <COL><DATA>Obu history by Morly Beeta</DATA></COL>
      </ROW>
```

Although this data is preservable, it is weak in *knowledge representation*. Content is represented as rows and columns – the actual data types must be inferred from the "metadata". The Filemaker export is a table-oriented rather a semantic-oriented representation of the data. It is possible that M's move to XML has been at the expense of his control over the data and has led to a loss of information that might have been otherwise provided.

7. Archive format guidelines

The previous two examples focused on data formats. Today's digital archives provide guidelines aimed at encouraging the production of resources that are "portable", as described in "seven dimensions" by Bird and Simons 2003) – content, format, discovery, access, citation, preservation, and rights. These dimensions identify properties that ensure the ability of digital linguistic resources to be preserved, discovered, transmitted, repurposed etc.[1] Some of the dimensions have been the focus of specific projects; for example, ontology projects have focused on the terminological aspect of content (linguistlist.org/emeld/tools/ontology.cfm), and the Open Language Archives Community (OLAC: www.language-archives.org/), addressed discovery by raising awareness of metadata. However, the greatest and widest attention has been paid to format, including markup, the encoding of characters, data structures, and documents, and distinguishing proprietary from open formats.

ELAR's guidelines, published on the depositors' page of our website (http://www.hrelp.org/archive/depositors/) take an ecumenical approach to advice, by pointing depositors to a variety of influential sources. Although, like the authors of "portability", we are more interested in principles than in prescriptions, some depositors do not share this interest and prefer more concrete specifications and specifically prescribed software and workflows. We provide such advice, for example stating a range of recommended formats:

[1] Elsewhere (Nathan 2006b), I have argued that relevance is also a factor for digital preservation.

- sound - WAV
- image - BMP, TIFF, JPEG
- video - MPEG2
- text - plain text, with or without markup
- documents - plain text, PDF or postscript
- structured text - XML, other markup (with description of markup system)
- structured data in commonly available Office formats - ELAR will convert them to archive-suitable formats
- character encoding :
    - preferred encoding is ASCII or Unicode
    - clearly document any other encodings used, e.g. ISO 8859-5
    - discuss with us if you use font substitution to handle non-Roman characters

Note that this list is heterogeneous; it ranges across various layers of format: character encoding, knowledge representation, and document encoding.

There is some confusion about so-called "archival formats". Some of the formats mentioned above, such as WAV, XML, and Unicode are *well suited* for preservation purposes principally because they are open (i.e. one would not need to also archive a copy of any specific software to ensure future access). On the other hand, "archival" is sometimes used as a synonym for correct values of properties such as resolution or compression, despite the fact that for some resources, such as video and images, compressed formats are generally acceptable to archives in recognition of practical real-world constraints. Compression is discouraged, of course, because it typically involves loss of some information and therefore some of its quality.[2] Nevertheless, there are many different ways to lose information; while it may be best for documenters to understand the principles and potential disadvantages of compression, it may be more effective to highlight to documenters the need to monitor, evaluate, and take responsibility for the quality of their materials. The equating of archiving formats with high resolution is even more confusing, since resolutions tend to be either fixed by the equipment used or are scalar with no clear line to be drawn, for example, between 44.1 KHz and 48 KHz, or 200dpi and 300dpi. In fact, the "correct" choice is more likely to depend on what kind of process or product is subsequently involved. Along with these confusions, there is the ever-present danger that documenters come to believe that adopting particular formats or parameters guarantees the quality of the resultant materials.

For many depositors, the gap between guidelines, such as "use explicit means to encode the distinct logical parts of your data", and the concrete means of achieving them, is too great. There is currently only one established approach for bridging this gap – by prescribing specific software, workflow and formats within which data creation takes place. This approach has been used with significant success by the Volkswagen Foundation funded DoBeS project at MPI Nijmegen (www.mpi.nl/DOBES).

[2] This applies to media (audio and video). High rates of lossless compression can be achieved for text. However, certain ways of creating text material, such as using fonts rather than explicit structuring to encode distinct data types, as illustrated in Section 6, could also be regarded as examples of unhelpful compression.

8. Formats and workflow

Another way to think about formats is through their use in particular phases of a resource's lifetime. Johnson (2004:146) and Austin (2006), for example, distinguish formats appropriate for resources in their working, archive, and presentation (dissemination) phases. For example, a grammar might be written in MS Word (working format), archived as XML, and disseminated as PDF or on the web (presentation). However, this schema should be expanded, since it does not take into account (a) ephemeral or informal formats used in additional phases that could be called "raw" and "interchange"; (b) formats do not map simply onto phases – some formats are applicable for multiple phases, either through their expressiveness and robustness (e.g. XML), or through pragmatic concession to the limitations of data storage and transmission (e.g. MPEG); and (c) the three-way distinction does not capture the intricacies of working with multimedia and complex data such as databases.

Fig 4. Example formats for some data types (vertical axis) and work phases (horizontal axis)

	Raw	*Working*	*Interchange*	*Archive*	*Dissemination*
Video	DVI	software-specific	MPEG-2	MPEG-2	MPEG2, AVI, QT
Fieldnotes	Shoebox, Page	Shoebox	FOSF	XML	WWW, print dictionary
Audio	ATRAC	WAV	WAV	BWF	MP3
Complex data	multiple	FM Pro database	RTF, XML	XML	Interactive application
Multi-modal	multiple	multiple	all above	all above	Multimedia application

9. A conversion example

Archives are stuck between a rock and a hard place in relation to the tension between format and preservability. On one hand, accepting data in non-archivable formats such as MS Word places the least burden on many documenters; they can focus on content, not method, thus maximising the continuity and clarity of their personal work patterns, and in turn, encouraging creation of the best quality data. But archives will have to put resources into converting such resources for long term preservation.

On the other hand, imposing format requirements, especially ones that are onerous or little known to documenters, poses several risks: a reduction in quality, errors that require archive intervention, and alienation of some of the community.

ELAR's current approach to this dilemma is to accept a variety of formats, as long as they are either portable (in the sense of Bird and Simons), or *potentially* portable. We acknowledge that there is a diverse range of depositors with different skills, motivations, and constraints. In many cases, we will convert materials at ELAR. The following example (Figure 5) uses data provided by ELDP grantee Dr Alice Taff for the Aleut language of Alaska (www.hrelp.org/grants/projects/index.php?projid=6).

We used Dr Taff's data, deposited in MS Word format, as a case study to investigate the amount of resources that would be needed to make it preservable. Research assistant Lameen Souag analysed the documents and we eventually concluded that the best solution was conversion to XHTML. XHTML has the merits of being robust and well formed (parsable) like XML, making it preservable but at the same time

viewable within ordinary browsers. The latter means that the data is still recognisable to its creator (a considerable benefit, which may not be the case using plain XML), and that no additional work is needed to provide a dissemination format. With minor corrections, regularisation of inconsistencies and the conversion of characters to Unicode, the data is now in preservable form,. In the conversions process, which used a combination of manual and some scripted methods, we were also able to enhance the data; for example, attributes were added to the underlying HTML which explicitly mark the function of various content, such as recorder, recording, speaker, location, etc..

Fig 5. Three views: data converted from documenter's working format to preservation format

5A. Documenters original version (MS Word tables etc)[3]

Language	Unanga{ (Aleut)
Dialect	Nii}u}i{ (Western Aleut)
Speakers	Alice Petrivelli, Vera Snigaroff, Mary Snigaroff, Vivian Koenig
Place recorded	Anchorage, Alaska
Date recorded	Mar. 15, 2005
Recording name	ANC14trk1
Recorded by	Alice Taff, Piama Oleyer
Recording equipment	Marantz CDR300 CD recorder with one flat-filtered, table-mounted cardiod microphone.
Translated/Transcribed by	Simeon L. Snigaroff, December 2005

1	ap	Uqla}ii{, {aaya{, uqla}il agach aliguuta{ a{.
		To take a bath, Steam bath, to take a bath is the one that is Aleut
5	vs	uhmm

5B. Converted preservation version as XHTML, approximate browser view

Language	Unangax̆ (Aleut)
Dialect	Niiĝuĝix̆ (Western Aleut)
Speakers	Alice Petrivelli, Vera Snigaroff, Mary Snigaroff, Vivian Koenig
Place recorded	Anchorage, Alaska
Date recorded	Mar. 15, 2005
Recording name	ANC14trk1
Recorded by	Alice Taff, Piama Oleyer
Recording equipment	Marantz CDR300 CD recorder with one flat-filtered, table-mounted cardioid microphone.
Translated/Transcribed by	Simeon L. Snigaroff, December 2005

1 ap Uqlaĝiix̆, x̆aayax̆, uqlaĝil agach aliguutax̆ax̆.
To take a bath, Steam bath, to take a bath is the one that is Aleut

5C. Converted preservation version as XHTML, source view

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head><title>ANC14trk1</title>
```

[3] In the original, the font Unangam Tunuu has been applied to some of this data.

```
<link href="taff.css" type="text/css" rel="stylesheet"></link>
</head>
<body>
<table class="metadata">
<tr><td>Language</td><td class="language">Unangax̂ (Aleut)</td></tr>
<tr><td>Dialect</td><td class="dialect">Niiĝuĝix̂ (Western Aleut)</td></tr>
<tr><td>Speakers</td><td class="speaker">Alice Petrivelli, Vera Snigaroff,
Mary Snigaroff, Vivian Koenig</td></tr>
<tr><td>Place recorded</td><td class="place">Anchorage, Alaska </td></tr>
<tr><td>Date recorded</td><td class="date">Mar. 15, 2005</td></tr>
<tr><td>Recording name</td><td class="rec_name">ANC14trk1</td></tr>
<tr><td>Recorded by</td><td class="rec_by">Alice Taff, Piama
Oleyer</td></tr>
<tr><td>Recording equipment</td><td class="rec_equip">Marantz CDR300 CD
recorder with one flat-filtered, table-mounted cardioid microphone.
</td></tr>
<tr><td>Translated/Transcribed by</td><td>Simeon L. Snigaroff, December
2005</td></tr>
</table>
<table class="transcript">
<tr><td class="time">1</td><td class="speaker">ap</td><td
class="transcription">Uqlaĝiix̂, x̂aayax̂, uqlaĝil agach aliguutax̂
ax̂.</td></tr>
<tr><td> </td><td> </td><td class="translation">To take a bath,
Steam bath, to take a bath is the one that is Aleut</td></tr>
<tr><td> </td><td> </td><td> </td></tr>
<tr><td class="time">5</td><td class="speaker">vs</td><td
class="transcription">uhmm</td></tr>
```

10. Discussion

This paper has discussed the varied interactions between language documenters and a digital archive, or, more specifically, between language documenters and a range of issues associated with archives.

It has identified a central issue for those working with the diverse range of linguists and others who are documenting endangered languages: how to maximise the amount and quality of documentation while taking into account real-world issues of skills, division of labour, and resource allocation. This inevitably leads to the questions of where lie the essential concerns, and boundaries, of both archiving and language documentation?

The fruitful interaction between documentation and archiving has come about through historical reasons as much as necessity. In some cases, archivists happened to be the ones on the "team" most likely to have or be able to formulate knowledge about topics such as media and IT equipment, and data formats. In other cases, such as DobeS, archives have been instrumental in developing standards and software that have become central to the techniques of documentation.

There has been considerable discussion, following Himmelmann (1998), on the need for language documentation to be contrasted with language description,: potential dangers also lie ahead if documentation does not differentiate its own priorities, skills, processes, and equipment from archiving. If what is distinct about language documentation becomes further subsumed to archiving, then a broader form of archiving could gobble up those parts of documentation that are not identifiably part of linguistic theory or description.

On the other hand, if archives lose their focus on preservation, it will become harder to secure their unique services. Making data preservable does not preserve it. Long term preservation of digital data can be expensive and technically demanding. Although some of the costs are coming down, data volumes are generally increasing, in particular due to the entry of video to the documentary corpus. Outsourcing storage is finally becoming financially feasible, but leaves questions about security and long term stability unanswered (cf. the definition of an archive in Section 1). In addition:

- specialised archives will find it harder to argue for funding to sustain preservation facilities
- preservation will be done by those without the appropriate perspective and skills

One way forward might be to seek parallels and solutions in other disciplines, as well as to distinguish how the nature of documentary linguistics, and its products, create specific needs. If, for example, particular document formats are seen as "belonging" to documentation, researchers are more likely to invest in the relevant skills, and archives will also gain by having clearer definitions of the scope of their tasks.

Other questions also arise: if archiving starts with equipment and data collection methodology, why should it stop at preservation? Since digital archives are an important locus for dissemination, it could equally be argued that they should be involved in mobilisation (Nathan 2006a) – i.e. ensure that the needs of language community members, educators and those engaged in language revitalisation are met.

References

Austin, Peter K. 2006 Data and language documentation. In Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds) *Essentials of Language Documentation*. Berlin: Mouton de Gruyter. Trends in Linguistics. Studies and Monographs 178, pp 87-112.

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

Csató. Éva Á. and David Nathan. 2007. "Multiliteracy, past and present, in the Karaim communities". In Peter K. Austin (ed.) *Language Documentation and Description*, Vol. 4. London: The Hans Rausing Endangered Languages Project, pp 207-230.

Himmelmann, Nikolaus. 1998. 'Documentary and Descriptive Linguistics'. In *Linguistics* 36 (1998), pp 161-95.

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

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

Nathan, David. 2006b. 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.