

Data and language documentation

Peter K. Austin
Linguistics Department, SOAS
4 July 2005

3.1 Introduction¹

The role of data in language documentation is rather different from the way that data is traditionally treated in language description. For description, the main concern is the production of grammars and dictionaries whose primary audience is linguists (Himmelman 1998, Woodbury 2003). In these products language data serves essentially as exemplification and support for the linguist's analysis. It is typically presented as individual example sentences, often without source attribution, and often edited to remove 'irrelevant material'. There may also be a 'sample text' or two in an appendix to the grammar. Language documentation, on the other hand, places data at the centre of its concerns. Woodbury (2003:39) proposes that:

“direct representation of naturally occurring discourse is the primary project, while description and analysis are contingent, emergent byproducts which grow alongside primary documentation but are always changeable and parasitic on it”

For language documentation then, data collection, representation and diffusion is the main research goal with grammars, dictionaries and text collections as secondary, dependent products that annotate and comment on the documentary corpus. The audience for language documentation is also very wide, encompassing not only linguists and researchers from other areas such as anthropology, musicology, or oral history, but also members of the speech community whose language is being documented, as well as other interested people. A significant concern for documentation is archiving, to ensure that materials are in a format for long-term preservation and future use, and that information about intellectual property rights and protocols for access and use are recorded and represented along with the data itself. Important also is 'mobilisation' of materials (Nathan 2005, this volume), ie. generation of resources in

¹ Most of the material presented here has been 'road tested' in lectures at Frankfurt University, Uppsala University, the School of Oriental and African Studies, and the DoBeS summer school; I am grateful for comments and feedback from audiences on these occasions. A proportion of this chapter derives from information on language documentation and guidelines for grant applicants co-written by David Nathan and myself and published on the Hans Rausing Endangered Languages website (see particularly <http://www.hrelp.org/documentation/whatisit/>). I am grateful to David Nathan for permission to incorporate this material into the present chapter, and for his detailed comments on an earlier draft which picked up a number of errors and infelicities. Thanks also to Jost Gippert, Nikolaus Himmelman, Robert Munro, and Peter Wittenburg for suggestions for improvement of earlier presentations. Any remaining errors are solely mine.

Peter K. Austin 2005 'Language Documentation and Your Data' to appear in Nikolaus Himmelman, Ulrike Mosel, and Jost Gippert (eds). *Essentials of language documentation*. Berlin: Mouton de Gruyter.

support of language maintenance and/or learning, especially where the documented languages are endangered and in need of support.

Woodbury (2003:46-47) argues that a good documentation corpus should be:

1. *diverse* — containing samples of language use across a range of genres and socio-cultural contexts, including elicited data
2. *large* — given the storage and manipulation capabilities of modern information and communications technology (ICT), a digital corpus can be extensive and incorporate both media and text
3. *ongoing, distributed, and opportunistic* — data can be added to the corpus from whatever sources that are available and be expanded when new materials become available
4. *transparent* — the corpus should be structured in such a way as to be useable by people other than the researcher(s) who compiled it, including future researchers
5. *preservable, and portable* — prepared in a way that enables it to be archived for long-term preservation and not restricted to use in particular ICT environments
6. *ethical* — collected and analysed with due attention to ethical principles (see Dwyer, this volume) and recording all relevant protocols for access and use.

This means the corpus must be stored digitally and ideally collected digitally.

In this chapter we outline the major processes involved in collecting and representing language data in a documentation framework, briefly discuss the tools that are available to assist with this work, and illustrate some of the products that documentary linguists have developed to present the results of their research. Further technical details about data structures and encoding, tools, archiving, and outputs can be found in other chapters in this volume (see chapters by Gippert, Wittenburg, Nathan).

It is important to emphasise that language documentation is a developing field that has emerged only recently and that is undergoing rapid change in terms of both theory and practice. It can be anticipated that much of what is presented in this chapter will be subject to change and development in coming years.

3.2 Processes in language documentation

Language documentation begins with the development of a project to work with a speech community on a language and can be seen as progressing through a series of stages, some of which are carried out in parallel. In the following we discuss the processes that involve data

collection, processing and storage. These can be identified as follows (see also Wittenburg, this volume):

1. *recording* — of media (audio, video, image) and text
2. *capture* — moving analogue materials to the digital domain
3. *analysis* — transcription, translation, annotation, and notation of metadata
4. *archiving* — creating archival objects, and assigning access and usage rights
5. *mobilisation* — publication, and distribution of the materials in various forms

Note that at time when a documentation project is being developed each of these processes should be considered and relevant procedures included in the project planning. In particular, archiving and mobilisation must be included from the beginning of the project and not left to the end of the project or as an afterthought (see further below).

A crucial aspect that must be kept in mind at all stages is backup.

Backup

It is prudent for any project, and especially one involving digital ICT to develop a regular and effective regime of backing up the project data, ideally on a range of different media (eg. CD-ROM, DVD, flash memory, external hard disk). Backups should be incremental and intended for full recovery, should disaster strike. One widely agreed mantra is LOKSS “lots of copies keeps stuff safe” (see www.lockss.stanford.edu). Remember, it is highly likely that you will lose data at some point in your project work, however a good backup regime will ensure that such loss can be minimised.

3.3 Documentation processes - recording, metadata creation, and capturing

3.3.1 Recording

A good documentation corpus will include audio and/or video materials, ideally recorded in authentic settings and under good conditions. When recording outdoors, if possible attempt to minimise noise from animals, traffic, machinery and electrical equipment, wind and the environment, and non-linguistic activities (eg. children playing in the vicinity). When recording indoors, it is important to keep away from machines and electrical equipment, hard walls (that reflect sound), and windows. For video, it is necessary to consider light conditions, use artificial

lighting and reflectors as appropriate, and to learn some basic filming techniques, ideally from an ethnographic film-maker or relevant textbook.

Note that we are often unaware of and filter out much of the noise and movement around us, however this will appear on your recordings, sometimes over the top of the intended documentary data. There are four ways to check on and reduce unwanted noise:

1. monitor the recording through closed headphones as you make it
2. use a good quality external microphone and never rely on the microphone built into the equipment, especially for video cameras
3. cover the microphone with a wind shield and place it as close to the speaker's mouth as possible, using a boom or shotgun microphone if appropriate
4. reduce all unnecessary movement and sound such as shuffling papers, audience members moving etc.

It is imperative to use good quality equipment (the best you can afford with the project resources available) including good microphones, lighting, headphones, and consumables (tapes, discs, batteries). It is also important to divide up duties and individual researchers should not attempt to do all the recording tasks. It is better to employ and train assistants, ideally interested members of the language community, to help with microphones, recorders, cameras, lights and interaction with the people being recorded.

The choice of recording equipment (DAT, minidisk, solid state, DVD, analogue tape) may be a compromise between quality/cost and convenience and needs to be carefully considered, taking into account such factors as the local climate (DAT recorders are notably unstable in tropic climates, for example), access to electrical power, and portability. Two basic principles however are **never** record in compressed format such as mp3, and **never** record direct to computer hard-disk, as such techniques risk irrecoverable data loss (on sound file formats see below and Wittenburg, this volume). There is good advice about audio and video recording available in textbooks such as Ladefoged 2003 and on the internet (see especially David Nathan's fact sheet on microphones at <http://www.hrelp.org/archive/advice/microphones.html>).

Video recordings have a number of advantages: they are immediate, rich in authenticity, multi-dimensional in context, of great interest to communities, and can be produced independently by members of the community without the researcher in situ. They present several problems, however, including being more difficult to produce, harder to process (transcribe, annotate – see below), difficult to access without a time-aligned transcription, difficult to transfer and store (raw video requires large amounts of storage space), and difficult to preserve in the long term (since there are as yet no universally agreed standards for digital video). There may also be complexities having to do with prohibitions in some communities against viewing the images of dead people appearing in video recordings (necessitating delicate treatment in terms of

access and use restrictions). Note that in some communities making video recordings is not possible for cultural reasons.

Audio recordings are less difficult to produce than video and are relatively simpler to manipulate, store and curate. Audio is also more familiar as a medium and has been in general use by linguists for more than 50 years. Several audio processing software tools exist (see below), and archiving is less problematic than video. Conversely, audio recordings contain less information than video, are difficult to access without time-aligned transcription, and changing formats (both carriers and data formats) make obsolescence a major problem, eg. locating equipment to play the media on. This is especially true of legacy sound recordings (wax cylinders, wire, reel-to-reel tape) but will become increasingly the case for digital media, including DAT recordings and probably minidisk as new machines are introduced by manufacturers and older equipment and carriers are no longer available for purchase.

Before starting fieldwork

It is important to test all your equipment, including cables, connectors, and adaptors **before** you leave for fieldwork. Remember that one missing cable or connector can prejudice an expensive fieldtrip so prepare your equipment before you leave for the field and get professional advice as necessary. Make sample recordings under a range of conditions and check their quality. Transfer the recordings to your computer and be sure you know how to use the relevant processing software and how to burn CD-ROM or DVD backup copies of the data. Check the data on your backups on another computer to make sure that your writer and software are working properly. If in doubt, seek advice.

While making the audio and video recordings it can be useful to take fieldnotes, including rough transcriptions, translations, relevant recording metadata, diagrams, drawings, and notes that can serve as aide memoire for later writing up or checking. Fieldnotes should be written in ball-point pen (not pencil and not washable ink!) on good quality paper (ideally in a bound notebook) using one side of the page only. As soon as possible after the recording session fieldnotes should be checked and elaborated, and transferred to a digital form. It is amazing how rapidly one forgets what abbreviated notes made while recording and interviewing mean².

Digital text has a number of advantages: it is compact, stable, easy to store, access and index, and can express hypertextual relationships (links). There are a large number of tools available to process text data (text editors, word processors, databases, browsers etc), and well established literacy traditions and knowledge of written text in many communities. However, it

² For further suggestions about the role of fieldnotes in documenting languages and cultures it can be useful to look at textbooks on anthropology and ethnography, such as Brewer 2000, Walcott 2004.

is less rich than audio and video as there is always loss of information when ‘reducing language to writing’. Text needs to be connected to richer recordings of speech events through time-aligned transcriptions and hyperlinks (see examples below, and elsewhere in this volume). However, written documentation outputs in the form of books are highly valued in many language communities, and for those where ICT resources are not available or limited will be the ideal form of product from a documentation project.

Labelling and metadata

Whatever the recording medium, it is important to rigorously **label** everything, including tapes, disks, CDs, containers, fieldnote books (number all the pages!) immediately, consistently and uniquely (eg. using date, and sequence number). Write this information with an indelible marker **on** the object itself, since disks and tapes can become separated from their covers. It is also imperative that a proper record of metadata (data about the recorded data, see below), such as speaker name, recording location, dialect etc is made at the same time as the recordings are labelled. You can do this in a notebook or as a computer file (create a structured file using a spreadsheet, database or Word table, whatever is most convenient).

3.3.2 Metadata creation

Metadata is data about data, ie. structured information about events, recordings and data files. It is usually represented as text (but not always, eg. it could be a spoken introduction track on a video or audio recording), but it is a different type of media because it is collected and used differently from other types. Typically metadata is collected and stored according to some formal specification. Metadata is needed for proper description of the data and to enable it to be found and used (see Bird and Simons 2003). There are two main competing international standards for linguistic metadata, that promoted by the Open Language Archives Community (OLAC) and that promoted by the ISLE Metadata Initiative (IMDI), the former being less detailed than the latter. The choice of metadata format should be made in consultation with the archives where the researcher intends to deposit the documentary materials (see Wittenburg’s chapter).

There are several types of metadata:

1. *Cataloguing* — information useful to identify and locate data, eg. language code, file id number, recorder, speaker, place of recording, date of recording etc
2. *Descriptive* — information about the kind of data found in a file, eg. an abstract or summary of file contents, information about the knowledge domain represented
3. *Structural* — for files that are organised in a particular way, a specification of the file structure, eg. that a certain text file is a bilingual dictionary

4. *Technical* — information about the kind of software needed to view a document, details of file format, and preservation data

5. *Administrative* — background information such as a work log (indicating when the files was last saved or backed up), records of intellectual property rights, moral rights, and any access and distribution restrictions imposed by researcher and/or community

Note that information can be metadata for more than one purpose, depending on its nature and use, eg. the identity of the speaker in an audio recording could be relevant for cataloguing purposes and/or also for determining access restrictions.

The following is an example of the different types of metadata associated with a computer file:

Cataloguing	Title: Sasak.dic; Collector: Peter K Austin; Speakers: Yon Mahyuni, Lalu Hasbollah; Language code: SAS
Descriptive	Trilingual Sasak-Indonesian-English dictionary, linked to finderlists, morpheme forms link to Sasak text collection
Structural	Dictionary entries with headword, part of speech, gloss in Bahasa Indonesia and English, cross-references for semantic relations; SIL FOSF record format
Technical	Shoebox 5.0 ASCII text file
Administrative	Open access to all; Last edited version dated 2004-06-25; backup 2004-06-20 on DVD 012

Some linguistically-relevant descriptive metadata that you may wish to use are: speaker (name, gender, age, place of birth, languages spoken, dialect, education level), recorder (name, experience), date of recording, location of recording, duration of recording, type (genre) of materials recorded, transcriber (especially if different from the recorder), date of transcription, location of transcription, location of all digital files, media and text (and location of archive copies).

3.3.3 Capturing

Capture refers to the encoding and transfer of an analogue recording (as on a cassette or reel-to-reel tape) or text written on paper to the digital domain as a computer file. In many cases, modern ICT means that audio and video recordings are ‘born digital’ and can be transferred to

computers without a separate capture process, unless transcoding is involved (see Wittenburg, this volume). When using digital capture software it is important to make sure you use appropriate settings. It is also advisable to transfer fieldnotes from notebooks to computer files, ideally as soon as possible after recording so you don't forget notes, abbreviations and comments. As for recording, it is imperative to name your computer files consistently and clearly, making sure that you should not rely on directory structure to disambiguate file names, eg. if you have a file called fieldnotes1.doc in one directory ("folder") (for year 2004 research, say) and another also called fieldnotes1.doc in another directory (eg. for your 2005 notes) then any loss of directory information will result in confusion between these files. Different naming schemes can be used, but clarity and transparency is the goal – see Johnson 2004 for some suggestions. It is also essential to record the relevant metadata for the data files you create as you make them, ideally in a structured way such as a relational table using standard terminology.

3.4 Processing the materials

3.4.1 Linguistic processing

Processing the documentary materials is a very different operation from recording and capture, and operates on a very different time scale. Thus each minute of audio can take hours to process in terms of transcription and annotation (depending on familiarity with the language and the richness of the annotation), while video is even more labour intensive and requires much more time to process. Video may require cutting and converting to create manageable chunks and file sizes (this is done with computer software³). There are several tools that are useful for transcription and annotation (see below).

Linguistic analysis, that is transcription, translation, and annotation, requires decisions about representation, ie. the levels and types of units. This should make sense within the researcher's chosen framework (theory) and needs to be made clear in the structural metadata that accompanies the relevant files.

There are good reasons for aiming at a degree of standardisation when processing the materials, including transparency, portability and ease of sharing and access (Bird and Simons 2003). Phonetic transcription should follow the conventions of the International Phonetic Association (IPA), and phonemic transcription should be IPA or a regionally-recognized standard. Grammatical annotation tags (i.e. the abbreviated labels for e.g. part of speech categories) should follow general linguistic practice, eg. the recommendations of EURO TYP or E-MELD (including its GOLD ontology), with a list of relevant abbreviations and symbols provided as metadata (for further discussion see Leech and Wilson 1996, Schultze-Bernd, this volume).

³ There are a range of video editing programs, including commercially available software such as Adobe Premiere or freeware such as VirtualDub.

For processed data we need to distinguish between the following:

1. *Character encoding* — how characters are represented, eg. Windows/ANSI, Unicode, UTF-8, Big5, JISC
2. *Data encoding* — how meaningful structures in the data are marked, eg. extensible markup language (XML), Shoebox/Toolbox standard markers, Microsoft Word table
3. *File encoding* — how the data is packaged into a digital file, eg. plain text, Microsoft Word, PDF, Excel spreadsheet
4. *Physical storage medium* — the physical form used to store the file, eg. CD-ROM, minidisk, DAT, hard disk, flash memory stick

As an example, certain documentary materials might be encoded as a hard disk file in plain text Unicode Toolbox format (for further discussion and examples, see Gippert's chapter).

When we consider file encoding it is useful to distinguish between *proprietary formats* and *non-proprietary formats*. A proprietary format is one whose structure is determined and owned by the maker of the software that stores it, eg. Microsoft Word, Excel, Access, FileMaker Pro, or Sony ATRAC (the audio format on mini-disk). As such, this means that the data is not directly accessible, and the format is subject to change (so that attempting to open a file stored in one version of the software with a later version may not always work — see Gippert's chapter for examples). As a result proprietary formats are **not** ideal for long-term storage (ie. the encoding is not portable and reusable). Non-proprietary formats, eg. Unicode plain text, or wav audio, are open and transferrable between hardware and software.

When processing the data it can be useful to distinguish three kinds of contexts each requiring different data formats (see also Johnson 2004):

1. *working* context — the way the data is stored for on-going research work of annotation and analysis
2. *archiving* context — how the materials are to be stored for long term preservation (see below)
3. *presentation* context — the form of the data for distribution and publication

Researchers need to develop ways to *flow* data between contexts, typically by *exporting* the data into some structured format that the software used for other contexts can read (see Thieberger 2004 for some examples). Thus, a common working format for text annotation is Shoebox/Toolbox; this can be exported into rich text format (RTF) to be read by Microsoft Word in order to produce presentation format PDF documents for printing and distribution. The following are examples of the different format types for the three contexts:

	<i>Working</i>	<i>Archiving</i>	<i>Presentation</i>
Text	Word, XLS, FMpro, Shoebox/Toolbox	XML	pdf, html
Audio	wav	wav, bwf	mp3, wma, ra
Video	mpeg2	mpeg2, mpeg4	QuickTime, avi, wmf

As an illustration, the following is a screen shot which shows Shoebox format working context data for the Australian Aboriginal Guwamu language⁴. In the window on the top left is lexical information, on the lower left is elicited sentence data with morpheme-by-morpheme glossing annotation and free translation, on the top right is descriptive metadata about the people involved in the project, and on the bottom right metadata about abbreviations used in the lexical and sentence annotations. Note that the metadata is hypertextually linked to the data in the two lefthand windows, while the lexical root is hypertextually linked from the morpheme field in the sentence window, and the sentence number links from the example field in the lexicon.

⁴ The Guwamu data was collected by the late Stephen A. Wurm in 1955 at Goodooga in Queensland from the late Willy Willis and made available to me for study in 1980. The annotations and glossing are based on Wurm's translations and my analysis of the materials.

The screenshot shows a lexical database interface with four panels:

- guwamu.lex:**
 - bawurra**
 - Vxnum: 161
 - Vg: Gu
 - Vcat: n
 - Vsubcat: n
 - Vgl: k.o.kangaroo
 - Vdef: male red kangaroo
 - Veth: n
 - Vsci: n
 - Vdiscr: used as a generic term for kangaroos
 - Vsyn: n
 - Vant: n
 - Vcf: **gula, gumbarr, dhugandu**
 - Vder: n
 - Vcognt: n
 - Vec: SAW
 - Vsp: WW
 - Veg: Gu206; Gu255
 - Vvnum: n
 - Vdate: 13/Mar/2005
- guwamu.people:**
 - Vd: **SAW**
 - Vname: Stephen Wurm
 - Vrole: **recorder**
 - Vg: n
 - Vnote: collected data on Guwamu in 1955 at Goodooga, Qld, with Willy Willis
 - Vcf: **WW**
 - Vdate: 03/Apr/2005
- guwamu.notes:**

Vsnum	Gu255				nhunga	yilunha	bawurra
Vt	ngaya	banbalguya					
Vm	<i>ngaya</i>	<i>bamba</i>	<i>-gu</i>	<i>-ya</i>	<i>nhunga</i>	<i>yilu</i>	<i>-nha</i>
Vsubcat	pro	vtr	-vtrfl	-proagr	pro	dem	-proinfl
Vgl	1sgnom	spear	-fut	-1sg	3sgacc	this	-acc
Vcat	pro	v	-suff	-suff	pro	dem	-suff
Vxnum	063	088	-012	-028	092	009	-024
Vt	I will spear that red kangaroo						
Vec	SAW						
Vsp	WW						
Vef	Np12As004						
Vnt	pronoun co-occurrence with demonstrative and noun; demonstrative inflected for accusative case						
Vcf	03/Apr/2005						
- guwamu.abbrev:**
 - Vabb: **vtr**
 - Vmg: transitive verb
 - Vtype: **sub-category**
 - Vnt: transitive verbs are a sub-category of verbs; they take a transitive subject argument in ergative case and a transitive object argument in accusative or absolutive case.
 - Vcf: **vi, vdi**
 - Vgr: n
 - Vdate: 03/Apr/2005

A possible presentation form of the illustrated lexical entry is the following:

bawurra *n*

male red kangaroo, *Note:* used as a generic term for kangaroos, *cf.* **gula, gumbarr, dhugandu**, [SAW, WW], e.g. Gu206, Gu255

Note that in the presentation format, typography (eg. italics, bolding, font type, indentation) and dictionary literacy conventions are employed to partially represent the data structure (see Nichols and Sprouse 2003 for other examples). The sentence example can be presented as follows:

ngaya	banbalguya	nhunga	yilunha	bawurra
<i>ngaya</i>	<i>banba-lgu-ya</i>	<i>nhunga</i>	<i>yilu-nha</i>	<i>bawurra</i>
1sgnom	spear-fut-1sg	3sgacc	this-acc	k.o.kangaroo
pro	vtr-suff-suff	pro	dem-suff	n

I will spear this red kangaroo [SAW, WW, Np12As004]

Linguists' conventions (such as the 'Leipzig Glossing Rules' – see <http://www.eva.mpg.de/lingua/files/morpheme.html>) have been established for annotated text so that, as in the given example, horizontal and vertical alignment on the page represents relationships between different types of data⁵.

Lost in the flow

The data structures encoded in these Shoebox files are relatively complex (see the diagram in the Appendix below, and Austin 2005) but the links between the data fields are lost in the process of export to RTF and presentation on the printed page. Note that the links could be captured in a HTML file however, and thus be available to be viewed with a web browser. We discuss archival formats for these examples below.

3.4.2 Tools for linguistic analysis and processing

There are a range of computational resources that facilitate creating, viewing, querying, or otherwise using language data. They include application programs, components, fonts, style sheets, and document type definitions (DTD). Application programs can be classified into two types:

1. *general purpose software* for which the user must design the data structures and can write application programs to manipulate the data and carry out various tasks. Examples are Microsoft Word and Excel, and FileMaker Pro. Such software is powerful and flexible however they store data in a proprietary format which is not optimal for long-term storage and access
2. *specific purpose software* which is designed to be used for particular tasks. Examples of such software in common use by language documenters include: *Transcriber* and *EXMARaLda*

⁵ The Shoebox/Toolbox tool automatically creates the *appearance* of vertical alignment in its interlinear text function, though it actually stores spaces in the data files to do so. Note that it does not store the *relationships* between the aligned information and rather relies on the user's implicit knowledge to interpret these.

(EXtensible MARKup Language for Discourse Annotation – see Schmidt 2004) for time-aligned audio annotations, *Shoebox/Toolbox* for text and lexicon annotations, *Praat* for speech analysis and annotation, *Elan* for audio and video annotation, and *IMDI Browser* for cataloguing and administration metadata.

Some of the specific purpose software is discussed and illustrated elsewhere in this volume.

Other useful software

In addition to the tools mentioned above, there also exist converter programs for transferring data between encoding formats, such as those developed at MPI-Nijmegen for uniting Transcriber and Shoebox encoded files, and converting them to XML for use with Elan. Further information about available programs and computational resources can be found at the E-MELD ‘School of Best Practice’ website and in the list of resources at the back of this volume.

3.5 Archiving

Digital archiving involves the preparation of the recorded/captured data, metadata and processed analysis so that the information it contains is maximally informative and explicitly expressed, encoded for long-term accessibility and safely stored with a reputable organisation that can guarantee long-term curation. A number of digital language and music archives exist; the DELAMAN network created in 2003 links many of them (see resources list). Digital archiving offers opportunities to store data for communities to use, other scholars to access, and for preservation for future generations of community members, the general public and researchers. Note that not all recorded data has to be archived (eg. unprocessed video files) but we should aim to make our materials archiveable, that is, richly structured documentations maximising the possibilities of the digital medium. Archiving must be included as a process in our language documentation project plans, and it is advisable to seek assistance with planning for archiving from an archivist at the beginning of project conception.

Note that archiving is not publication (only those materials prepared for distribution will be published by the archive), nor is it backup (the archive will generally not accept backup copies of files alone but will expect the data and metadata to be explicitly described, often by requesting that deposit forms be completed for each archival object). Archives also commonly have systems in place to manage protocols for intellectual property rights, and for specification of access and usage rights, eg. that a certain archival object is only available to members of the speaker community. The depositor should establish these by discussion and negotiation with the owners, and describe them via metadata and deposit protocols. Data sensitivity is **not** a reason to not archive; it is better to deposit data in an archive with restrictions than not deposit at all. Researchers should also make preparations for assigning their rights into the future by including information in your will and ensuring that your executors understand how to assign them on your death.

3.5.1 Archiving text materials

The preferred format for archiving text materials is eXtensible Markup Language (XML), a document description language used to encode the content of structured documents (see Sperberg-McQueen and Burnard 2002). XML is a subset of SGML (standard generalised markup language) and is used to explicitly describe a domain of knowledge through *markup tags* enclosed in angle brackets (see Gippert's discussion in this volume of a 'play structure' implicit in a published document). Each part of a structured document is described within a defined and logical structure (stored in XML schemas or DTDs 'document type definitions'). XML is a good archival format because XML documents explicitly represent data structure, and are directly readable by humans even if computer software to display the documents is not available.

XML documents are typically created by export from working context materials, rather than being directly written by the researcher, because the process of writing well-structured XML tends to be tedious and error prone (various XML editors exist and these can be used to create documents, to check markup tag syntax (well formedness), to create DTDs, and to ensure that a document complies with a schema or DTD). XML encoded documents can be transformed into various archival and presentation formats by XSLT, extensible stylesheet language transformations. Thus, an XSLT could create a concordance of an annotated text collection, or HTML files for web publication. Archivists can provide advice on possible transformations of XML documents.

The following are two examples of XML encoding. First, consider the structure of a typical bilingual lexicon (such as seen in the Guwamu example presented above)⁶:

1. lexicons contain entries
2. the attributes of entries are: form, category, subcategory, language, meaning specification (and any other additional information such as notes, speaker, recorder, sense relations, sentence examples)
3. meaning specification can be gloss (for morpheme-by-morpheme glossing and finderlist production) and definition
3. cross-references to other lexical entries have a sequential order chosen by the lexiographer
4. cross-references to sentences examples also have a specified sequential order

⁶ The chosen example is deliberately simple in order to present the main concepts here; in practice lexical entries may have much more complex structures and relationships.

Here is the Guwamu sample entry discussed above in XML form, which would be a possible archival representation:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<lexicon>
<entry id="161">
<form>bawurra</form>
<language>Gu</language>
<cat>n</cat>
<subcat>n</subcat>
<gloss>k.o.kangaroo</gloss>
<def>male red kangaroo</def>
<note>used as a generic term for kangaroos</note>
<rec>SAW</rec>
<sp>WW</sp>
<date>13/Mar/2005</date>
<xref>
<cf n="1">gula</cf>
<cf n="2">gumbarr</cf>
<cf n="3">dhugandu</cf>
</xref>
<egref>
<eg n="1">Gu206</eg>
<eg n="2">Gu255</eg>
</egref>
</entry>
</lexicon>
```

If we view this data using XML-aware software such as an XML editor⁷ or web browser such as Mozilla Firefox or the current version of Internet Explorer, the hierarchical relationships between the data entities are displayed, as in:

⁷ A number of commercial and freeware editors are available; the screenshots below show views within the ElfData XML editor (see www.elfdata.com).

```

? xml version="1.0" encoding="ISO-8859-1"
└─ lexicon
  └─ entry id="161"
    └─ form abc bawurra
    └─ language abc Gu
    └─ cat abc n
    └─ subcat abc n
    └─ gloss abc k.o.kangaroo
    └─ def abc male red kangaroo
    └─ note abc used as a generic term for kangaroos
    └─ rec abc SAW
    └─ sp abc WW
    └─ date abc 13/Mar/2005
  └─ xref
    └─ cf n="1" abc gula
    └─ cf n="2" abc gumbarr
    └─ cf n="3" abc dhugandu
  └─ ehref
    └─ eg n="1" abc Gu206
    └─ eg n="2" abc Gu255

```

For an annotated corpus we can set up a structure where:

1. the corpus contains sentences
2. sentence properties are: sentence number, sentence form, sentence gloss, speaker, recorder, sentence source reference, grammatical notes
3. sentences contain words in sequential order
4. word properties are: word form, word gloss
5. words contain morphemes in sequential order⁸

⁸ A simple concatenative item-and-arrangement morphological model is adopted here for purposes of illustration (this is the model assumed by the *Shoobox* software); other morphological models could be used and represented in XML. For further discussion of the

6. morpheme properties are morpheme form, morpheme gloss, morpheme category, morpheme subcategory

Here is an XML representation of the Guwamu sentence shown above. Note that the XML representation makes explicit the sequential order of words in the sentence, and the relationships between elements, eg. word forms and their constituent morphemes, which are purely implicit in typical working format (Shoebox) and presentation format (printed example) which rely on horizontal and vertical alignment on the page or screen to signal the relationships:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<corpus>
<sentence>
<snum>Gu255</snum>
<sform>ngaya banbalguya nhunga yilunha bawurra</sform>
<ft>I will spear this red kangaroo</ft>
<rec>SAW</rec>
<sp>WW</sp>
<ref>Np12As004</ref>
<nt>pronoun co-occurrence with demonstrative and noun; demonstrative inflected
for accusative case</nt>
<date>03/Apr/2005</date>
<word seq="1">
<wform>ngaya</wform>
<wgloss>I</wgloss>
<morpheme id="053" seq="1">
<mform>ngaya</mform>
<cat>pro</cat>
<subcat>pro</subcat>
<gl>lsgnom</gl>
</morpheme>
</word>
<word seq="2">
<wform>banbalguya</wform>
<wgloss>will spear</wgloss>
<morpheme id="088" seq="1">
<mform>banba</mform>
<cat>v</cat>
<subcat>vtr</subcat>
<gl>spear</gl>
</morpheme>
<morpheme id="012" seq="2">
<mform>lgu</mform>
<cat>suff</cat>
<subcat>vinfl</subcat>
<gl>fut</gl>
```

structure of interlinear text and a proposal for representing it in XML using the annotated graph formalism (Bird and Libermann 1999) see Bowe, Hughes and Bird 2003, and Hughes, Bird and Bowe 2003.

```

</morpheme>
<morpheme id="028" seq="3">
<mform>ya</mform>
<cat>suff</cat>
<subcat>proagr</subcat>
<gl>1sg</gl>
</morpheme>
</word>
<word seq="3">
<wform>nhunga</wform>
<wgloss>him</wgloss>
<morpheme id="092" seq="1">
<mform>nhunga</mform>
<cat>pro</cat>
<subcat>pro</subcat>
<gl>3sgacc</gl>
</morpheme>
</word>
<word seq="4">
<wform>yilunha</wform>
<wgloss>this</wgloss>
<morpheme id="009" seq="1">
<mform>yilu</mform>
<cat>dem</cat>
<subcat>dem</subcat>
<gl>this</gl>
</morpheme>
<morpheme id="024" seq="2">
<mform>nha</mform>
<cat>suff</cat>
<subcat>proinfl</subcat>
<gl>acc</gl>
</morpheme>
</word>
<word seq="5">
<wform>bawurra</wform>
<wgloss>kangaroo</wgloss>
<morpheme id="161" seq="1">
<mform>bawurra</mform>
<cat>n</cat>
<subcat>n</subcat>
<gl>k.o.kangaroo</gl>
</morpheme>
</word>
</sentence>
</corpus>

```

Again, we can view this representation using XML-aware software and see its hierarchical structure; firstly in terms of a sentence made up of a sequence of words:

```

? xml version="1.0" encoding="ISO-8859-1"
└─┬─ corpus
  └─┬─ sentence
    └─┬─ snum abc Gu255
      └─┬─ sform abc ngaya banbalguya nhunga yilunha bawurra
        └─┬─ ft abc I will spear this red kangaroo
          └─┬─ rec abc SAW
            └─┬─ sp abc WW
              └─┬─ ref abc Np12As004
                └─┬─ nt abc pronoun co-occurrence with demonstrative and noun; demonstrative inflected for accusative case
                  └─┬─ date abc 03/Apr/2005
                    └─┬─ word seq="1"
                      └─┬─ word seq="2"
                        └─┬─ word seq="3"
                          └─┬─ word seq="4"
                            └─┬─ word seq="5"

```

Now, if we view the information about words in the sentence in detail we see that they consist of one or more morphemes in sequence (notice that the triangle icon on the left margin changes from horizontal to vertical as we move down the hierarchy):

```

? xml version="1.0" encoding="ISO-8859-1"
└─ corpus
  └─ sentence
    └─ snum abc Gu255
    └─ sform abc ngaya banbalguya nhunga yilunha bawurra
    └─ ft abc I will spear this red kangaroo
    └─ rec abc SAW
    └─ sp abc WW
    └─ ref abc Np12As004
    └─ nt abc pronoun co-occurrence with demonstrative and noun; demonstrative inflected for accusative case
    └─ date abc 03/Apr/2005
    └─ word seq="1"
      └─ wform abc ngaya
      └─ wgloss abc I
      └─ morpheme id="053" seq="1"
    └─ word seq="2"
      └─ wform abc banbalguya
      └─ wgloss abc will spear
      └─ morpheme id="088" seq="1"
      └─ morpheme id="012" seq="2"
      └─ morpheme id="028" seq="3"
    └─ word seq="3"
      └─ wform abc nhunga
      └─ wgloss abc him
      └─ morpheme id="092" seq="1"
    └─ word seq="4"
      └─ wform abc yilunha
      └─ wgloss abc this
      └─ morpheme id="009" seq="1"
      └─ morpheme id="024" seq="2"
    └─ word seq="5"
      └─ wform abc bawurra
      └─ wgloss abc kangaroo
      └─ morpheme id="161" seq="1"

```

More on archival format

Note that the information stored in the XML representation is extremely compact but is still readable by humans and the structure can be recovered, even if the software to display the data is missing; this is why XML is a good archival format. For more information on archival encoding see the Text Encoding Initiative (www.tei.org) or the resources websites listed at the end of this book. There are numerous introductory textbooks for XML though none of them explicitly deals with language documentation issues.

3.5.2 Archiving sound and video

The formats for real-time media are subject to rapid technological change and one of the major concerns of archives is to attend to refreshing files ('forward migration') so that they remain readable to the existing equipment. For video there are two internationally agreed compressed formats, namely mpeg2 and mpeg4, however there is no agreement about raw formats which in any case are extremely difficult to store due to the very large file size. For audio recordings, archives generally use uncompressed CD-quality (44kHz, 16 bit) encoded as wav files; some archives also use 48kHz and/or BWF 'broadcast wave format' where metadata is bundled together with the audio. Note that mp3, Realaudio or Windows Media Player formats are all compressed in a way that loses information; they are useful for working and presentation (eg. for publication, on web sites) but are not suitable for archiving.

More on sound archiving

There are a large number of well equipped sound archives around the world, ranging from regional, to national to international coverage. Some, such as the Austrian National Sound Archive have been established for a long time and have extensive experience with material in older 'legacy' formats. The International Association of Sound Archives (www.iasa.org) publishes lots of valuable and up-to-date advice about archiving issues, and the Language Archives Newsletter (www.mpi.nl/LAN) focusses on archiving for linguistic research.

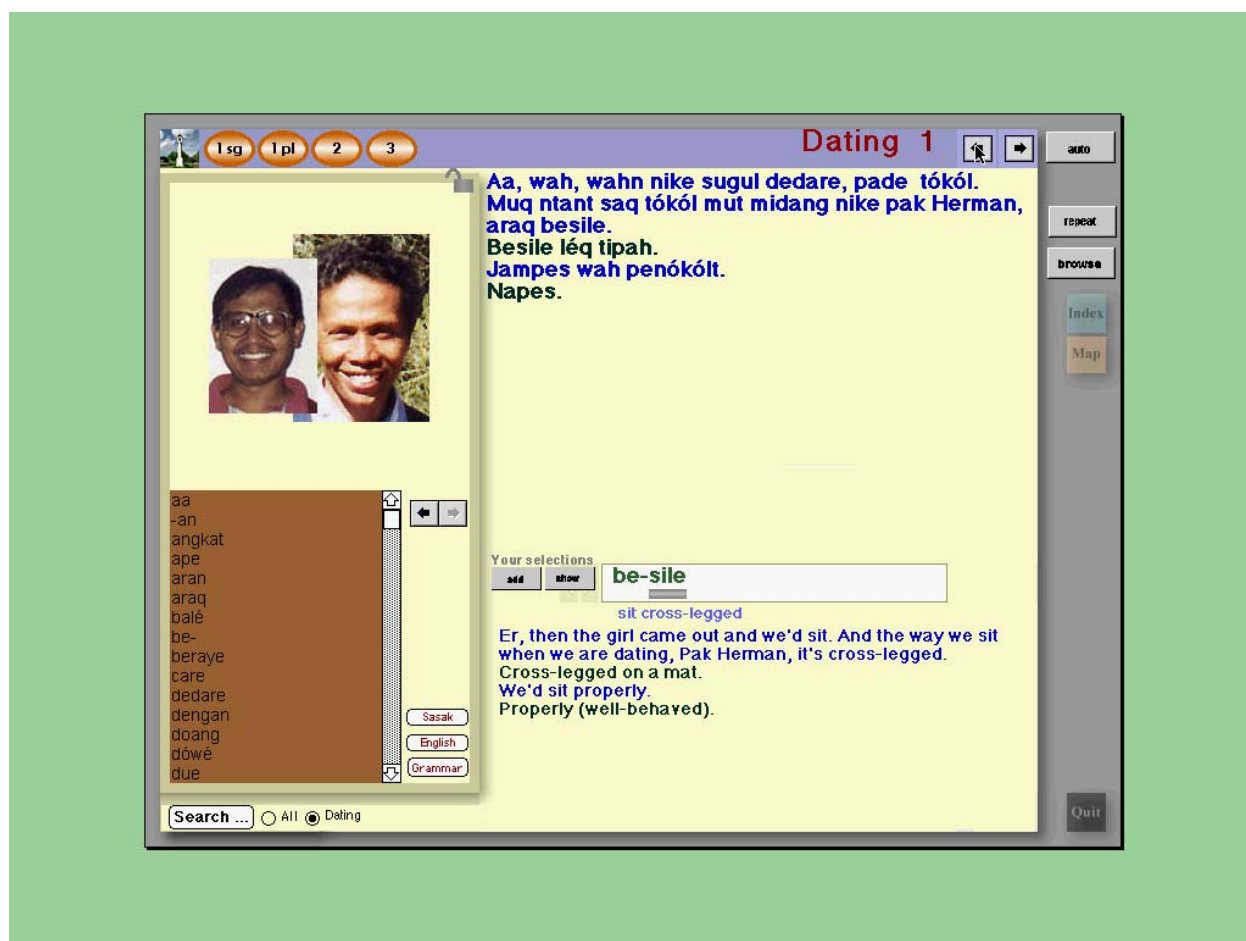
3.6 Presentation, publication and distribution

One of the ways that the presentation, publication and distribution of rich language documentations can be achieved currently is via multimedia which links media, annotations (time-aligned transcriptions, analysis and translations, hyperlinks) and metadata. One such format is linked files (including HTML, mp3 sound clips, Quicktime etc) distributed via the world wide web, but bandwidth can be problem for publication of media files – even small movies of a few minutes in a compressed format can be megabytes in size and take a long time to download via slow connections (the use of video streaming software can partially overcome this limitation). There is also SMIL 'synchronised multimedia integration language' which is an application of XML to encode mixed media, text and image information in a presentation form (see www.w3c.org/TR/2001/REC-smil2020010807/).

For highly complex richly annotated and linked media currently we need to use multimedia platforms such as Macromedia Director, delivered on CD-ROM or DVD as a publication format (see Nathan, this volume). Unfortunately the future of these formats and the carriers is unclear and how we can archive multimedia for the future is also currently problematic. One current major need is good multimedia players and ways for users to interact with the rich documentations; it is necessary to model and design interfaces and access formats for various

audiences. An example of such a format is the *Spoken Karaim* CD described by Csató and Nathan 2003 which presents video and audio recordings with accompanying transcriptions, translations, glosses, lexicon and cultural information, all of which are linked and interactive. The interface enables users to explore their own pathways through the corpus and to search, collect items of interest, backtrack, and interact with the corpus. It has a simple attractive interface that enables maximum interactivity without forcing the user to digest too much information, and has been used for Karaim language support in education, language maintenance and revitalisation (Nathan and Csató 2005).

The following is a screenshot from a CD-ROM of conversational documentary materials in the Sasak language of eastern Indonesia (Austin, Jukes and Nathan 2000) which is based on the Karaim model. The top-left window shows images of the consultants who worked on the corpus, and below it a Sasak lexicon arranged alphabetically (clicking on an entry in the lexicon reveals full details of the individual item in the top left window in place of the images), and on the top right is the Sasak transcription of the conversation (colours indicate the two speakers, their voices can be heard in the left and right channels respectively of the associated time-aligned digital stereo recording). Below the transcription is a small central window displaying morpheme-by-morpheme analysis and gloss for a selected item in the text, and below that a display of the free translation in English of the speaker turns (again colour coded). In the lower bottom left of the display there is a search facility which the user can employ to find occurrences of morphemes or glosses of interest throughout the corpus, and in the top left is a set of buttons that produce pronominal inflected forms of verbs (via a morphological generator) when the user moves them over a selected lexical entry in the top left window (see Nathan, this volume, and Nathan 2000 for further details about the morphological generator developed for the *Spoken Karaim* CD).



3.7 Conclusions

Language documentation is an emerging field that involves recording, analysis, annotation, archiving and publication of rich and complex data. By properly structuring the data representations and planning methods to flow data between different formats and contexts, you can work productively with your materials, as well as publish and distribute them for others and archive your resources to preserve them for the future. It is important that all these aspects of a documentation project be incorporated in its planning and execution, in order to ensure maximally effective and useful documentation.

References

Austin, Peter K. 2005 New documentation from old sources. In Diana Eades, John Lynch and Jeff Siegel (eds) *In honour of Terry Crowley*. Amsterdam: John Benjamins.

- Austin, Peter K., Anthony Jukes and David Nathan 2000 *The Sasak conversation CD*. University of Melbourne.
- Bird, Steven and Mark Liberman 1999 A Formal Framework for Linguistic Annotation. Linguistic Data Consortium, University of Pennsylvania: Technical Report MS-CIS-99-01 Department of Computer and Information Science
- Bird, Steven and Gary Simons. 2003 Seven Dimensions of Portability for Language Documentation and Description *Language* 79:557-582.
- Bow, Catherine Baden Hughes and Steven Bird 2003 Towards a General Model of Interlinear Text. *Proceedings of EMELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings*. LSA Institute: Lansing MI, USA. July 11-13, 2003.
- Brewer, John D. 2000 *Ethnography* Open University Press
- Csató, Eva A. and David Nathan 2003 Multimedia and documentation of endangered languages. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 1*, 73-84. London: SOAS.
- Himmelmann, Nikolaus 1998 Documentary and descriptive linguistics *Linguistics* 36:161-195.
- Hughes, Baden Steven Bird and Catherine Bow 2003 Encoding and Presenting Interlinear Text using XML Technologies. *Proceedings of the Australasian Language Technology Workshop 2003*. Melbourne, Australia. December 10, 2003.
- Johnson, Heidi 2004 Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 140-153. London: SOAS.
- Ladefoged, Peter. 2003 *Phonetic data analysis: an introduction to fieldwork and instrumental phonetics*. Oxford : Blackwell
- Leech, Geoffrey and A. Wilson 1996 Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R, March 1996. Available at: <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>
- Nathan, David 2000 "The Spoken Karaim CD: Sound, text, lexicon and 'Active Morphology' for language learning multimedia". In: Göksel, Aslı; and Kerslake, Celia (eds). *Studies on Turkish and Turkic languages*, 405-413. Wiesbaden: Harrassowitz.
- Nathan, David and Eva Csató 2005 Multimedia: A Community-Oriented Information and Communication Technology. In Anju Saxena (ed.) *Minor Languages of South Asia*.
- Nichols, Johanna and Ronald L. Sprouse 2003 Documenting Lexicons: Chechen and Ingush. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 1*, 99-121. London: SOAS.

- Sperberg-McQueen, C.M. and Lou Burnard (eds.) 2002 *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford
- Schmidt, Thomas 2004 Transcribing and annotating spoken language with EXMARaLDA. *Proceedings of the LREC-Workshop on XML based richly annotated corpora*, 69-74. LREC 2004, International Conference on Language Resources and Evaluation. 29 May 2004, Lisbon, Portugal. Paris: European Language Resources Association.
- Thieberger, Nick 2004 Documentation in practice: Developing a linked media corpus of South Efate. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 169-178. London: SOAS.
- Wolcott, Harry 2004 *The Art of Fieldwork*. 2nd edition. Walnut Creek: Alta Mira.
- Woodbury Anthony 2003 'Defining Language documentation' in Peter K. Austin (ed.) *Language Documentation and Description, Vol 1*, 35-51. SOAS.

Appendix – Guwamu data structures

