CHAPTER 4

LINGUISTIC DATA MANAGEMENT

NICHOLAS THIEBERGER
ANDREA L. BEREZ

4.1 INTRODUCTION[1]

Documenting a language requires the production of records that will persist and be accessible into the future. If we look at the number of language descriptions for which there is no corresponding accessible primary data, it would seem that creating persistent, well-structured, and citable language records has proven to be a considerable barrier to linguists in the past.[2] This is due, in part, to the lack of advice and training in the use of appropriate methods and tools for recording, transcribing, and analyzing linguistic data. This chapter seeks to provide such advice by focusing on the nuts and bolts of the creation and management of sustainable data in the course of linguistic field research, including pre-fieldwork planning and the follow-up archiving, annotation, and analysis of field-collected materials. The approach to data management presented here assumes that it is the

[1] Thanks to Christopher Cox and Toshihide Nakayama for their comments as reviewers of this chapter. Thanks also to Laura Berbusse, Tobias Bloyd, and Aaron Waldrip for helpful suggestions.

[2] See Johnston (1995) and Antworth and Valentine (1998) and the other papers (a little dated but still useful) in Lawler and Aristar-Dry (1998) to see how these topics have been part of linguistic discussion for some time.

professional responsibility of linguists to create long-lasting, archivable primary data and to then situate any subsequent analyses in that data.

We advocate the use of appropriate technologies to record, manage, and annotate linguistic data, in order not only to create good records of the languages we study but to also provide access to the data upon which we make generalizations. Good data management includes the use of specific software tools (e.g. Toolbox,[3] Fieldworks,[4] ELAN[5]), but more importantly centres on an understanding of the nature of linguistic data and the way in which the tools we use can interact with the data. Tools will come and go, but our data must remain accessible into the future.[6] We are primarily aiming our discussion here at academic linguists, but we hope that it will be useful to anyone engaged in the many tasks involved in the documentation of underdescribed languages, which may not involve academically trained linguists at all, or may involve linguists who are from the communities in question, for whom 'the field' is in fact home (cf. Crippen 2009).

If we liken our research endeavor to building a house, we can think of planning and managing our data collection as providing the firm foundations on which a solid house is built—that is, one we can live in for many years, and which can grow as we need further renovations and extensions. These extensions are akin to the different ways we will be able to use or present our data in the future, including text collections, various lexicons, and multimedia representations of the language. Some of these outputs we can envisage now, others we cannot. Our foundations need to be built today in a manner that makes our data perpetually extensible.

If, however, we do not build our foundation correctly, there will be little hope for extending our research, and our data may not even serve our immediate purposes without the need for constant reworking. Consider, for example, a dictionary written in Microsoft Word in which the elements (headwords, definitions, part of speech tags, and so on) are differentiated on the page only through formatting conventions (i.e. the use of bold or italic styles) and are not explicitly marked for their status as elements. One cannot easily convert such a dictionary into multiple versions for diverse audiences (say, a reversal of headword language or a learners' or children's dictionary), nor can one automatically link entries to media and so forth (cf. Black and Simons 2009). This is because the elements in the dictionary, not being explicitly identified by use of structural tags of some kind (e.g. the 'backslash' tags of Toolbox, or perhaps XML tags, see below), are not automatically accessible to computational processes. Equally important are our data descriptions; if we do not take a few moments during our field sessions to write simple descriptions

[3] http//www.sil.org/computing/toolbox

[4] http://www.sil.org/computing/fieldworks

[5] http://www.lat-mpi.eu/tools/elan

[6] See Bird and Simons (2003) for a detailed discussion of the principles underlying the creation of good linguistic data from fieldwork.

about our collections, their very existence may become unknown—not only to ourselves, but to the speaker communities we work with.

This chapter is not concerned with determining how much data needs to be recorded in the course of fieldwork. Instead, we assume that researchers can decide for themselves how much is adequate for a given situation, and we focus on proper principles for managing the data that is collected. It is of no small concern to see colleagues swamped by masses of recordings, photos, and texts, confused by discussions of which standards to follow and unable to keep track of not only the items in their collection, but also the status of those items (e.g. 'Has this video been transcribed?', 'Has this text been interlinearized?', 'Has this collection of photographs been described?'). The result is that within a short time these colleagues cannot fully access their own collections because of insufficient description of the contents, or because the data is tied to now-defunct software, or even because of permanent data loss from lack of a suitable backup procedure.

In a discussion of the general problem of data 'deluge', Borgman (2007: 6) notes: 'The amount of data produced far exceeds the capabilities of manual techniques for data management.' This has become a truism for computational tasks on the web, and it is also true of the comparatively small datasets created by documentary linguists. If we want to move beyond tedious manual techniques for manipulating our data and make use of far more efficient automatic and global operations, we need to learn how to first prepare *well-formed data* in a predictable structure that conforms to existing standards, so that we can then allow computers to do the rest of the work. After all, handling repetitive (and tedious) tasks quickly and accurately is what computers do best.

Publishing and citing the primary data on which an analysis is based is becoming more common and may soon become a requirement for major research grants (cf. Borgman 2007: 240–41) and submission to journals. Until the late 1980s it was difficult to cite primary data in a linguistic analysis (although Heath's 1984 grammar of Nunggubuyu is a valiant attempt at linking examples to textual sources). Since then, however, only a few grammatical descriptions (e.g. Morey 2004; Thieberger 2006) link the descriptions and analyses of linguistic phenomena to examples from a corpus. The methods described in this chapter build citable corpora and encourage the use of cited data in any resulting analysis, a practice that has been part of lexicographic and corpus-based linguistics for some time (cf. Liberman 2009).

We want to be clear, however, that the goal of this chapter is not to drown field linguists in the virtual ocean of stringent and ever-changing specifications of 'best practice' procedures that our discipline sometimes seems to be swimming in. In this context it is important to bear in mind Voltaire's admonition (Arouet 1877) that 'the best is the enemy of the good'. Bloomfield is quoted as saying, 'each of us should take a vow of celibacy, not teach, and dedicate the entirety of our summers to fieldwork and our winters to collating and filing our data, year after year' (Hockett 1962). We suggest it would be better for linguists to become more efficient using the methods we advocate, and to still have a life!

This chapter aims to help find the balance between the desire to record and annotate our data to the highest possible standards, and the reality of the time constraints we all work under. In fact, if we take the time to produce well-formed, reusable data, our linguistic analyses will be more convincing, and our production of language materials will become more efficient. The initial outlay of time and effort required to understand the methods and tools for good data management may seem high, but the investment will pay off in the end.

This chapter is organized as follows. The remainder of §4.1 provides some basics about data and metadata, as well as an overview of the workflow for creating well-formed data. In §4.2 we outline the planning required before setting off on fieldwork, including preparing to use (or not use) technology in the field and making contact with an archive. The data management tasks you will face, including file naming, care for fieldnotes, metadata collection and storage, regular expressions, time-aligned transcription and interlinearization, and the use of lexical databases, are discussed in §4.3. For those who are not so technically oriented, we suggest skipping §4.3.3.1 to §4.3.4. In §4.4 we discuss the broader advantages of creating well-formed field data for linguistics as a discipline, and §4.5 contains concluding remarks.

4.1.1 What is data?

Linguistic data arise from a range of techniques that straddle traditional humanities and scientific methodologies, from field-based observation to corpus data to experimentation. For this chapter, 'data' is considered to be the material that results from fieldwork, which may include: primary observational notes, recordings, and transcripts; derived or secondary material such as lexical databases and annotated texts; and tertiary material, in the form of analytical writing, published collections of texts, or dictionaries. In addition, fieldwork results in elicited data: elicitation is always a part of fieldwork, and many linguists also use questionnaires and experimental field methods (see Majid's Chapter 2 above). All of these types of data and their associated metadata (see below) together make up the larger documentary corpus. Like the house with solid foundations, a documentary corpus grows and changes over time and is continually updated as we collect more recordings, refine our annotations, and publish more analyses.

4.1.2 What is metadata?

Simply stated, *metadata* is one set of data that describes another set of data. In terms of linguistic field data, your metadata is your catalogue of 'what you have'—a list, stored in a relational database or a spreadsheet, of all the important bits of information that describe each recording, transcript, photograph, and notebook in

your corpus. Keeping accurate, up-to-date metadata is crucial, because without it, you very soon lose track of the details of the items in your collection (try to remember the contents of a recording you made last year to see how important a brief description is). You would not know, for instance, if recording *x* predates or postdates recording *y*, or if photograph *p* is of mountain *m* or mountain *n*, or if recording *z* has been transcribed yet, or if WAV file *w* was originally on cassette.

You may think that you can keep track of all of this information in your head—and indeed, you may be able to—but consider what would happen if others wanted to access your collection, perhaps when you are back home after summer fieldwork, or after your retirement or death. Without a catalogue of metadata, there would be no way for anyone else to know what your corpus contains, or the conditions under which it was collected.

Fortunately, there are two widely accepted metadata systems for linguistic material that can guide fieldworkers in which bits of information to collect as part of a catalogue. These are the standards provided by Open Language Archives Community[7] (OLAC) (see below) and the ISLE MetaData Initiative[8] (IMDI). You should not feel limited to just the fields provided by either one. Since you may want to collect types of information that neither system requires, you can build your catalogue so that it can export to a form that can be used by any archive.

4.1.3 Form and content

Critical to the goals of this chapter is the distinction between the *form* of the presentation of data and its *content* (cf. Black and Simons 2009). The content of data should be constructed, or *marked*, in such a way that it is possible to derive many different presentation forms. A simple example of this is shown in Fig. 4.1. Two different kinds of markup are shown: in (a), a form-driven markup like HTML (the encoding used on the World Wide Web), and in (b), a content-driven markup.

In this example the presentation of the data is the same for both, but the mechanism by which this presentation happens is different. In (a), the bold type would be rendered directly from the strictly presentational <b> (bold) tag, whereas in (b), the data would be processed through a stylesheet that assigns different styles to different tags (in this case, everything marked as <sentence>, <header>, and <adj> are bolded). Without a structural identification of a word as being, for example, an adjective, there is no way to quickly change the presentation form of just one kind of element (e.g. so that the adjective *grumpy* is italicized instead of bolded), nor is there any way to search for structurally different elements, e.g. headers only. Imagine having to comb through a 10,000-entry lexicon, entry by entry, just to make a minor format adjustment that could instead be accomplished with a single click or a couple of keystrokes.

[7] http::://www.language-archives.org
[8] http://www.mpi.nl/IMDI/documents/documents.html

(a) Form-driven markup (in HTML)

Marked data:	This is a sentence.
Presentation:	**This is a sentence.**
Marked data:	Chapter1 He was grumpy about being left out.
Presentation:	**Chapter 1** He was **grumpy** about being left out.

(b) Content-driven markup

Marked data:	<sentence>This is a sentence.</sentence>
Presentation:	**This is a sentence.**
Marked data:	<header>Chapter 1</header> He was <adj>grumpy</adj> about being left out.
Presentation:	**Chapter 1** He was **grumpy** about being left out.

Figure 4.1. Form-driven versus content-driven markup

A more relevant example is that of a dictionary project. A single well-structured lexical database can give rise to a number of different presentation formats: on paper, on-line, with linked sound files and images, just as a simple wordlist, with reverse-language indices, and so forth. These many presentations of dictionaries are possible if contents of the elements (lexemes, glosses, definitions, example sentences, etc.) are described with, for example, backslash[9] field markers, rather than just with formatting information. Similarly, texts have logical structures: they usually include an orthographic representation of the spoken words, a morphemic parse, a morpheme gloss, and a free gloss as shown in the example of a Toolbox text fragment in Fig. 4.2, where each line is marked by a code (e.g. \tx = text; \mr = morphemes; \mg = morphemic glosses; \fg = free gloss). Describing or marking the content of these structures in our data allows various presentation formats to be generated for different audiences.

Structuring linguistic data by content allows it to be arranged and rearranged in new ways, providing visualization of relationships within the data. Such visualization allows new generalizations to be made, and also helps ensure consistency in the data. For example, being able to sort a lexicon on various fields puts similar types of information

[9] So-called 'backslash' codes or standard field markers are a simple way to delimit fields in a text document. MDF, or Multi-Dictionary Formatter, is the system used by Toolbox to create dictionaries as styled RTF documents from a structured Toolbox file (cf. http://www.sil.org/computing/shoebox/mdf.html), and it provides a set of over 100 standard 'backslash' codes.

```
\id     061: 005
\aud    AHT-MP-20100305-Session.wav 02: 19.320-02: 21.780
\tx     Ga      łdu'    ben     yii     taghił'aa.
\mr     ga      łdu'    ben     yii     ta-     ghi-    ł-      'aa
\mg     DEM     FOC     lake    in      water   ASP     CLF     linear.extend
\fg     'As for that one (river), it flows into the lake.'
```

Figure 4.2. Structure of interlinear glossed texts

next to each other which can also be useful for creation of topical lists where the lexicon is sorted by semantic field rather than just alphabetically. Mistakes, such as duplicate entries or typos, are more easily located when such sorting is carried out.

4.1.4 Reuse of field data

Reuse of our field recordings and research outputs is central to the methodology advocated here. We need to be able to reuse our own field materials in our analysis of the language, and also ensure that others can access them in the future. It is increasingly common for linguists to deposit primary material with an archive early and then to regularly keep adding to their collection. Once items are archived, any subsequent work—transcribing, annotating, or selecting example sentences for published analyses—can be cited back to an archival file. The archive provides this citability for you, so if you archive your files before you begin your analysis, you allow your research to be embedded within the data (Thieberger 2009) and then to be verified by other researchers with reference to the primary data.

4.1.5 A workflow for well-formed data

Creating well-formed data from linguistic fieldwork entails a workflow like the one in Fig. 4.3. The workflow begins with project planning, which for our purposes includes preparing to use technology in the field and deciding on which file naming and metadata conventions you will use before you make your first recording. (Clearly there is a great deal of non-data-related planning to be done before the first recording can be made, but that is not the focus of this chapter.) After recordings are made and metadata are collected, data must be transcribed and annotated with various software tools, and then used for analysis and in

representation of the language via print or multimedia. Note that depositing materials in an archive is carried out at every phase of the procedure.

Because you will subject your data to a number of processes, there is always a risk that some parts of it may be lost along the way if the outputs of one software tool do not align exactly with the input requirements of the next tool in the workflow. For example, importing a time-aligned transcript from ELAN (see §4.3.5) into Toolbox for interlinearization may result in some extra fields appearing in Toolbox. These fields correspond to the original ELAN data categories, but are not necessary for the Toolbox interlinearization process. You must take care, then, that you do not delete these extra fields so that they will still be present when you re-import the data into ELAN.

Another significant problem that arises from using a variety of tools in a workflow is that the exports of one tool may not be in a format that fits into the import routine of the next tool in the workflow. For example, time-alignment software (e.g. ELAN or Transcriber) typically produces a native XML file which can be exported to a number of interchange formats, but none of these can be opened *directly* in, say, Toolbox or Fieldworks (although more import and export formats are being added to these tools regularly, so this may become less of a problem over

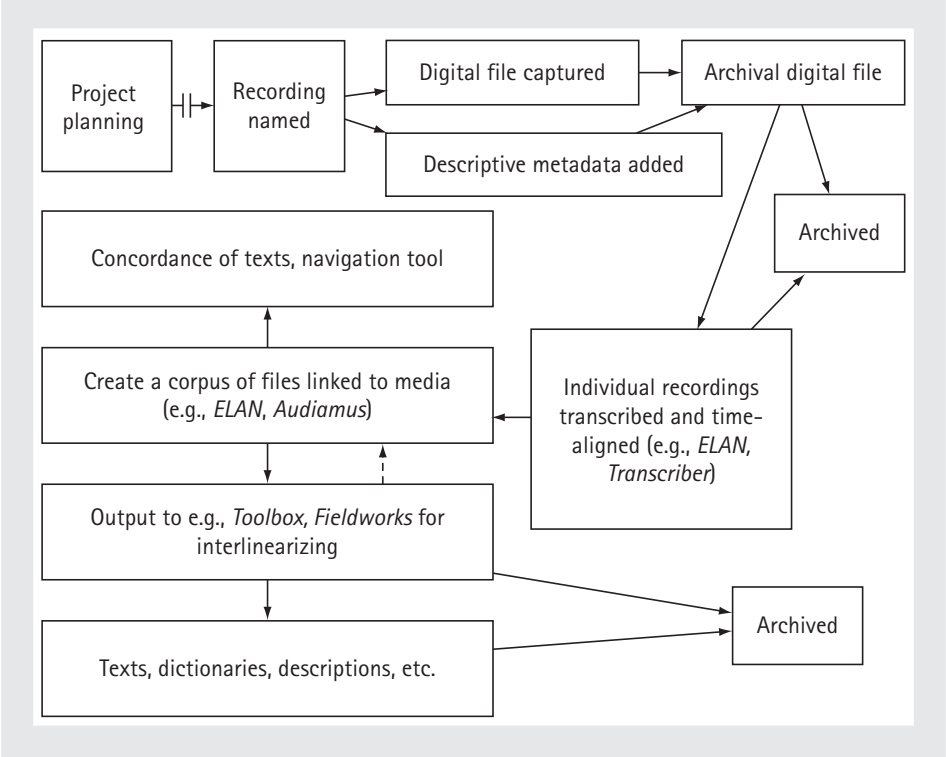

Figure 4.3. Typical workflow resulting in well-formed linguistic data (adapted from Thieberger 2004)

time). A useful way of converting text into new formats is to use regular expression functions as discussed in §4.3.4.

A final potential problem with workflows is that you need to keep track of which parts of your corpus have undergone which steps, with the risk that you may end up with different versions of the same file on your hard disk. If this happens, you may accidentally edit two different versions of the same file and subsequently have to use a 'compare documents' function to keep all of the edits you have made. For this reason it is also helpful to keep *administrative* metadata (which describes your processing of the data) to track items through the steps in your workflow, as well as *descriptive* metadata (descriptions of the data itself) to describe the items.

Care on your part and an awareness of what is lost and what is gained by the use of a particular tool can mitigate all these dangers. Understanding what the various tasks along the workflow involve can make your life easier, and will also determine what tools are applicable at each stage of the workflow.

4.2 Advance planning for data management before you leave for the field

While you are planning the logistical details of your upcoming fieldwork (and see Bowern 2008: 222 or Fox[10] for lists of things to take and do before departing), one component of your preparation should be planning for the management of the data you will be collecting in the field. Some advance planning before you leave will save time and effort later and will ensure you have a good strategy in place when you arrive.

4.2.1 Preparing to use data management technology in the field

Technology has always been a part of fieldwork. Clifford (1990: 63) notes that the typewriter, once used by many anthropologists in the field (he cites Clifford Geertz, Margaret Mead, Colin Turnbull, and Gregory Bateson as examples), allowed researchers to reorder and enhance the day's handwritten notes. He suggests that typing notes dislocates the researcher, who writes for a distant professional

[10] http://www.stanford.edu/~popolvuh/field-checklist.htm

readership, from the field: 'Facing the typewriter each night means engaging these 'others', or alter egos. No wonder the typewriter or the pen or the notebook can sometimes take on a fetishistic aura' (Clifford 1990: 64).

The fetishizing of technology is not surprising, nor is it necessarily a problem. We adopt the best tools for the job at hand, and typewriters have given way to laptops, which are now commonly taken to the field, as are various handheld devices, e.g. PalmPilots (cf. Gibbon 2006), iPods, iPads, and the like. If you can use a computer in your fieldwork situation, it can help you stay organized and allow you to query data that may otherwise have to await your return.

The use of computers for managing field data could be seen as exemplifying the technological distance between the linguist and the people we work with in the field. Carrying costly equipment in a cash-poor community is incongruous and potentially dangerous (if you are likely to provide a target for theft). This is something that you need to seriously consider before departing for fieldwork; trust your knowledge of the field location to guide what you think it is appropriate to take with you (cheap netbooks are quite inconspicuous, and it's not the end of the world if they break). Unlike lugging a typewriter into the field, the laptop, recorder, camera, GPS, and video camera are not so robust and need special care (see Margetts and Margetts, Chapter 1 above, on recording). You may decide that a laptop is not going to survive your field environment, but that you will access a computer in a nearby location where you can organize your material away from the field.

Digital data is very fragile, so avoiding its loss has to be integral to your planning. Ironically, analog cassette tapes are in some sense hardier than digital files: a tape kept in a dry, temperature-stable filing cabinet can usually be played again a decade later, but digital files stored on a hard disk are unlikely to last that long. Also, accidents can happen—your field materials could be burned, flooded, eaten, stolen, or lost, despite your best efforts to keep your data backed up. Backing up your digital files can be done in one of several ways depending on the details of your field situation. Taking an extra external hard drive or two with you and copying your files to them on a regular basis is a good solution (one author who works in a region that is prone to forest fires takes two external hard drives to the field, copies files to both of them daily, and stores one in her cabin and the other in her car). You may also investigate online storage that can be accessed when you have an internet connection, perhaps on a regular visit to a town if that is an option. If you are not bringing a laptop to the field and plan to use multiple memory[11] cards to hold your recordings, a USB 'On-The-Go' (OTG)[12] storage

[11] Memory cards of various kinds are used in many portable devices like digital recorders and cameras.

[12] OTG USB devices are those that can act as hosts to other USB devices and also interact with a computer; in this case, a device to which you can upload files from your recorder's SD card for storage until you can transfer them to a computer.

device is a compact solution for backing up those recordings (but leave the original recordings on the memory cards during your trip so that you have multiple copies, rather than reusing the cards).

Having sufficient power to run the equipment is no small undertaking if there is no electricity supply in your field site. Choosing devices that run on ordinary batteries makes life easier, and chargers that run off car batteries, generators, or solar panels are also an option (see McGill and Salffner n.d. and Honeyman and Robinson 2007 on the use of solar power and portable generators for fieldwork, and also Margetts and Margetts, Chapter 1 above).

4.2.2 Archiving: plan for it now

A major difference between the newer methods in language documentation and those that preceded them is the use of digital media. As pointed out above, digital data has no proven longevity, and format changes can potentially lead to orphaned data that can no longer be read by available software. Therefore, the sooner your data can be stored in a dedicated repository, with trained staff and an institutional commitment to protect and migrate your data, the better. A proper archive will describe your data adequately, assign persistent identifiers, and share data in several locations to ensure its long-term accessibility. Archiving is an essential part of any language documentation work in which a major motivation is to create an enduring record of the language. Unfortunately, your computer is not an archive. An external hard drive, as crucial as it is for day-to-day backup, is not an archive, nor is a website from which you make multimedia files available for download.

Many people think of archiving as the last step in a research program, to be done years after the fieldwork is complete, after everything has been transcribed and mined in service of analysis—essentially, something to be done after we are 'finished' with the data. Here, we advocate the exact opposite: data should be archived immediately and often (subject, of course, to your having resolved any ethical issues that may have arisen about what you recorded, and whether the speakers have agreed to have their recordings archived—see Rice, Chapter 18 below, for further discussion of ethical issues).

It is quite possible—and becoming increasingly common—to archive recordings periodically from the field as they are collected. Then, whenever you finish a transcription, archive it. When you write up an analysis, archive it. As your lexical database grows, archive it. If you make a children's book from your data, archive it. Think of your archived collection as another backup of your entire corpus, only this one is stored in a remote offsite location and maintained by professionals. If the idea of sending your data to another location before it is ready to be made

available to the general public makes you uncomfortable, you can temporarily restrict access to certain parts of your collection.[13]

Establishing a relationship with an archive before you leave for the field ensures that the staff can work with you from the beginning of your project to secure the long-term preservation of your collection. Finding out what kind of metadata and file naming conventions the archive uses is a good idea, because this will guide how you set up your metadata database. Be sure to keep track of the information the archive requires, as well as any extra information that you will find helpful that is not necessarily requested by the archive. You can also make some arrangements for regular deposits of your data from the field, either via direct upload to the archive or by mailing DVDs or a small hard drive.

Example of the use of methods advocated in this chapter

Scenario 1

Having established my field location and the relationships with the people I will be working with, I am now in the process of eliciting and recording information. I set off for a story-telling session with my backpack of equipment (video and still cameras, audio recorder, headphones, tripod, and microphone) and my notebook. The speaker has already completed a consent form and I have taken their photo to include in future presentations of stories they have told. After a cup of tea and a chat they are ready to start. I do my best to ensure that there is no background noise and that the lighting is appropriate, then set up the tripod, video, and audio recorder. I put on the headphones, then start the recording by saying the date, place, and name of the speaker. They then tell their story and I take notes of contextual information that may be relevant, like the direction they are facing, or gestures they use to help in communicating the story. When they are done they may want to retell parts of the story or explain aspects of it, and I will try to record that information as well. Afterwards, I note the name of the file assigned by the recording device. If the location of the recording is not easily described with a place name, then I would also have a geographic reference from a GPS and that would be part of the metadata for the recording.

[13] Temporarily restricting access may also be necessary if your recordings include sections with local gossip or other content that a speaker subsequently requests to have suppressed. Given the problem of 'data deluge', this necessary work can become more and more onerous as the collection grows, and is not a task that current digital tools provide much help in automating. Restricting access to all primary materials contributed to the archive until they can be reviewed and edited offers an immediate solution to issues of preservation and privacy, but it does not necessarily address the bottleneck created by this kind of review, and potentially poses difficulties for later open access. Furthermore, it could be controversial to archive language materials outside of the speaker communities involved in documentation, especially in cases where past research relationships have contributed little of appreciable benefit to the speaker communities involved, or even proven hurtful in the treatment of language materials which may have been removed from those communities. This is a matter that often requires more discussion between all partners in documentation to arrive at a solution that ensures long-term preservation while respecting the concerns of all involved.

I employ a speaker of the language to transcribe recordings, using exercise books to write the language and a translation. These books are then scanned for ease of access and will be typed when I return from the field.

My notebooks are the primary source of information about this recording session, and when I can get to a computer, I enter that information (e.g. speaker name, place and date of recording) into a database, with the key field of the record being the name of the media item. Later, as I proceed with my analysis of the language, the name of the transcript of the media will be entered, or at least, the fact that it has been completed would be checked off in the database. Then, when I extract a text from that transcript for interlinearization, I note the identifier of the text in the database too.

4.3 Managing your data

If you work in a location with electricity, much of your initial data management will be done in the field on a daily basis as you collect recordings, including naming your files appropriately, recording the associated metadata in your database, creating backup copies and preparing files for archiving. If there is no electricity in your field site, you will want to take care of data management at the soonest possible opportunity, either on a visit to a town or immediately upon returning home. Be vigilant and resist the temptation to do it later, as it is too easy to lose track of what you have collected. It is useful to audio record simple details of who is speaking, where, and when at the beginning of the recording session; if you forget, you can add it at the end of the recording.

4.3.1 File naming and persistent identification

File names are critically important, and, while a seemingly trivial matter, the failure to plan properly for file names can wreak havoc with your data organization. Each file that you create has to be named in a way that lets you find it later. The names have to be unique—i.e. you should avoid having one directory with file 1, file 2, file 3, etc., and another directory with exactly the same file names in it. If those files are ever removed from their original directory (and they surely will be), there will be no way to trace their provenance.

You may be tempted to use the default file names assigned by your equipment; for example, your digital recorder may name files with something like STE-001, and your camera may assign names like IMG_0086. As a first step, you can always make note of the default name in your notebook, but be sure to rename it and record the correct metadata to them as soon as you can. These default file names may not be

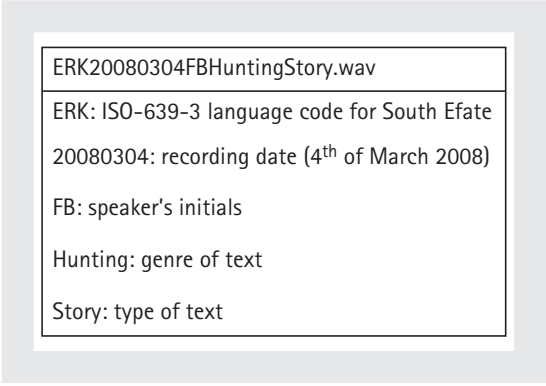

Figure 4.4. Example of semantic file naming

unique, especially if you are using multiple recording devices or if your device recycles file names upon download.[14]

There are several strategies you can choose from when you are deciding on your file naming convention. Some people prefer semantic file names, which carry a great deal of information, as in Fig. 4.4. The file name, shown in the top line, includes several meaningful chunks, including a language code, the date of recording, the speaker's initials, and a brief description of the contents. One advantage of this system is obvious: you can easily see basic information as you look through your directory of files. Another advantage is that your computer will sort your files depending on how you order the information in the file name (in this case, by language first, then by recording date, then by speaker). A disadvantage of this strategy is that the names are rather long and can be difficult to read. Another disadvantage is that this system may not coincide with your archive's naming system. When you deposit your files you will need to assign each file an archive-appropriate name. This strategy for file naming is also a bit redundant: all the information contained in the file name will also be repeated in separate fields in your metadata catalogue. Nevertheless this system is useful for browsing your directories.

Another approach is to include most metadata only in your catalogue, and assign relatively simple unique file names that include only a minimum of meaningful information. An example would be *2011001.wav*, for the first recording made in 2011. This approach is easier to read and a bit less redundant in terms of record-keeping, but the file name may still need to be converted for deposit in an archive.

A third approach is to simply adopt the file naming convention that your archive requires, for example *ALB01-001-A.wav* for an item held in PARADISEC. This file name carries little meaningful information beyond the depositor's initials and the

[14] A temporary solution might be to append the date, such that IMG_0086 could become 20110102IMG_0086 (for 2 January 2011).

order of creation, but it is the least redundant method and will require no conversion upon deposit.

In the end, the choice of a naming convention is up to your preferences, but consistency is key and, more importantly, a file's name must be permanent once it is assigned. Take care with hyphens and underscores (and other non-alphanumeric characters), as these may be 'reserved' characters in some archives and will need to be converted later. To ensure the greatest legibility and persistence of your file names it is still best to use ASCII characters, although Unicode may become more acceptable in the near future. Be consistent in using upper and lower case—for some computer systems upper- and lower-case characters are treated equally, but in others they are not.

The names you assign to your files can be used by the archive and, together with their address, will provide a persistent identifier, i.e. a name that will allow the file to be located now and into the future. Persistent identifiers may include a URL, but remember that websites alone cannot provide persistence of location unless they have adopted conventions like those offered by Digital Object Identifiers[15] or Handles,[16] which are both services that allow permanent names to be applied, via a redirection service, to impermanent addresses.

4.3.2 Handwritten fieldnotes and headnotes in digital language documentation

Since long before the age of digital language documentation, handwritten notes have been integral to linguistic fieldwork. Even in the current age of widespread awareness of the advantages of digital documentation, much day-to-day practice by field linguists is still based on analog procedures. Many linguists consider pen-on-paper to be the best method for preliminary elicitation and analysis; some field sites have no reliable power source, or perhaps using a computer in some locations would be obtrusive. Regardless of your reason for making handwritten notes, it is important to stay organized. You can use paper notebooks of whatever size you find convenient (or, in extreme need, scraps of paper that are conveniently to hand, or even, as in the case of one report, a coconut that was inscribed with a precious example: cf. Crowley 2007: 107). Whenever possible, use acid-free paper and pens with waterproof, archival-quality ink. Researchers working in damp, rainy, or flood-prone areas might try waterproof all-weather notebooks.

In addition to being an obvious source of primary data, written fieldnotes also provide contextual information for recorded media, including the place, date, and participants, as well as any observations you want to note about the recorded performance itself. Much of the process of linguistic discovery is embedded in these

[15] http://www.doi.org/ [16] http://www.handle.net/

notebooks in the form of ethnographic notes, thoughts, and early errors, and in fact the left-to-right, top-to-bottom format of western notebooks itself provides important temporal information about the fieldwork session.

Written notes are further supplemented by 'headnotes' (Ottenberg 1990: 144), which include the memories the researcher has that continue to enrich their analysis over time. Headnotes and written notes are, says Ottenberg, in constant dialog: 'only after their author is dead do written notes become primary, for the headnotes are gone' (p. 147). The ability to access the primary data in digital form facilitates this dialog, bringing new forms of interpretation to bear on the collection over time due to the ease of access provided by files in a digital format.

While the interpretive and creative application of the researcher's experience results in more than can be reduced to the data objects stored in an archive, without those objects the linguistic analysis lacks an authoritative base. Fieldnotes, like recordings, should be scanned as part of your regular backup routine and archived as images. You can also type up your written notes to make them electronically searchable, allowing you to reflect on what you have written, add to it, and formulate new questions.

4.3.3 Your metadata catalogue

Simply amassing a number of digital files will not give you access to the material in them, just as a bookshop with thousands of uncatalogued books does not make information easy to find. The metadata that you provided for items in your own collection as they were created can later be uploaded to an archive's catalogue. These metadata will also make it easier for you to work with your own collection in your continuing analysis of the language.

An up-to-date metadata catalogue is a key component of your documentary corpus. Even if you have chosen a file-naming convention that captures a lot of information, you still need to keep a digital catalogue. Initially you are likely to be creating metadata in your paper notebooks, jotting down the name of a speaker or the object being photographed, the date, the location, and so on, and then keying this information into your catalogue at the earliest opportunity. Ask your archive what metadata they require for deposit, and you can also add other information that is important to your particular situation.

4.3.3.1 Relational databases

A relational database is a good choice for a catalogue, and in this section we discuss how one can be conceptualized (but we refer readers to other sources for the details of using any particular database management system (DBMS), like Filemaker Pro, MS Access, or OpenOffice.org Base). If you do not know how to build a relational database, you can still keep well-organized metadata in a spreadsheet, but you should be aware that a spreadsheet has some limitations. For instance, you will

need to enter identical information multiple times, and there is no easy mechanism for enforcing consistency, whereas a DBMS can constrain entries to a fixed list, using dropdown menus to assist in data entry. For these reasons you may wish to make the leap from a spreadsheet to a relational database when you have time to learn how to use one (Harrington 2009 is an excellent introductory guide), or adapt an existing relational database to suit your needs. As with all digital files, periodically backing up your catalogue to a readable text format is good insurance against losing your catalogue in the event that someday the DBMS software is no longer supported.

Relational databases provide ways of linking related but conceptually different kinds of data, such as information regarding recordings, transcripts, and people that are part of your fieldwork project, as shown in Fig. 4.5. In your database, each *record* relates to an *item* in your collection. An item can be whatever you select it to be, perhaps a recording session that lasted two hours of which there are photos and video and audio, all of which can be summarized in one database record. On the other hand, you may want to list each recording as an item, each with a unique name and with a catalogue record devoted to it. This would allow you to find information at a level lower than the session. Sample catalogue files in various formats are available from archiving projects (see e.g. Arbil,[17] IMDI,[18] PARADISEC[19]) to help you decide. We hope that in the near future we will have access to more user-friendly metadata entry tools such as Saymore[20] and Fieldhelper,[21] which promise to use drag-and-drop functions for metadata entry.

Everyone has different ways of working, but we all need to keep track of some basic kinds of information, such as information about media recordings, about people, about transcripts and texts, and about lexica. A DBMS stores similar kinds of information together in *tables*, and then establishes *relationships* between tables so that you should (in theory) never need to enter the same piece of information twice, thus saving time and eliminating chances of typos. The links between tables in a DBMS work by using *keys*, a unique identifier for each record in the database.

To illustrate this, consider Fig. 4.5. A metadata database will have a table that keeps track of your **recordings**. You will want to store a number of different pieces of information about each one, e.g. the file name, the date recorded, the equipment used, the location of the recording, a summary description of the contents, and the length. At the same time, the DBMS can also establish a unique identifier, or key, for each recording (here, Recording_ID).

To understand how relations between tables work, imagine you also want to keep track of your **transcripts**, which are related to, but separate from, the recordings

[17] http://www.lat-mpi.eu/tools/arbil/
[18] http://www.mpi.nl/IMDI/
[19] http://www.paradisec.org.au/downloads.html
[20] http://saymore.palaso.org/
[21] http://www.fieldhelper.org/

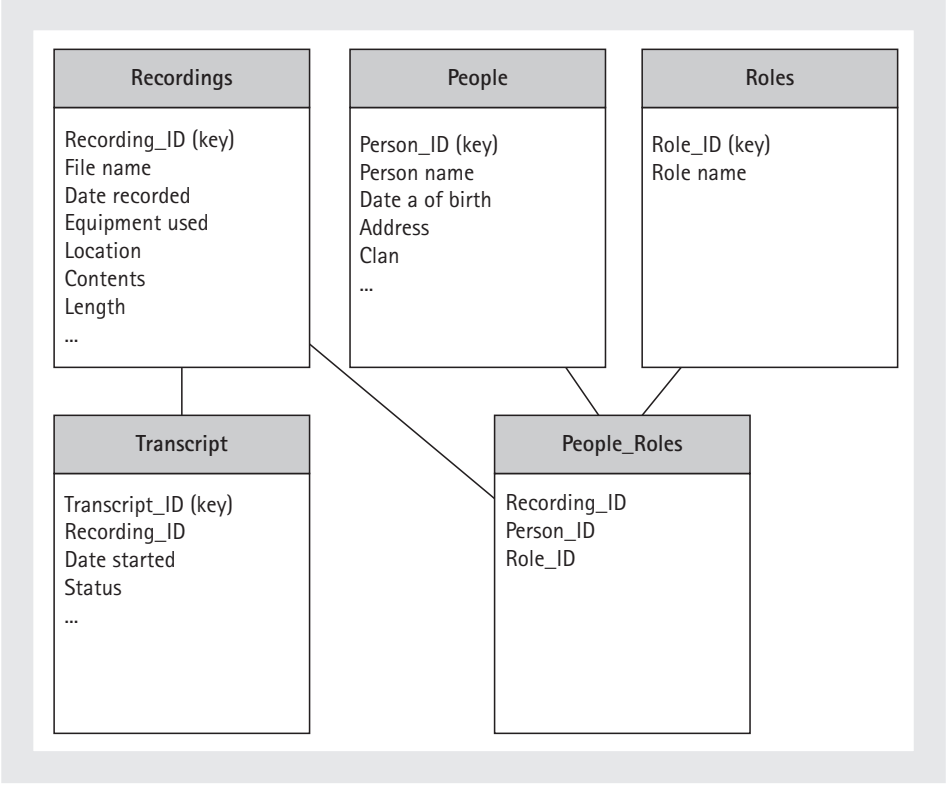

Figure 4.5. A simple relational database for fieldwork metadata

they transcribe. You can keep information about your transcripts, like the start date and the status (in progress, finished, etc.), in another table. Notice in Fig. 4.6 that the transcripts table not only contains the unique identifier, or key, for each transcript, but also references the key of each associated recording. Including in one table the key from another table establishes the relation between the two tables.

A more complex example is that of tracking the various **people** who are associated with the items in your collection, with information like their names, their dates of birth, their clans, etc., and their **roles**. People can have roles as speakers, linguists, transcribers, singers, and so on; furthermore, a single person may hold multiple roles across your collection, and particular roles will most certainly be filled by different people for different items.

Rather than listing all the people associated with a particular recording and their roles within the recordings table, it is more efficient to have a separate table with all of the people involved in the project, and another table listing all the possible roles people can hold. A third table, known as a 'composite table' (or 'link table'), then lists **person–role** pairings (each by their key), and the key of the recording with which each

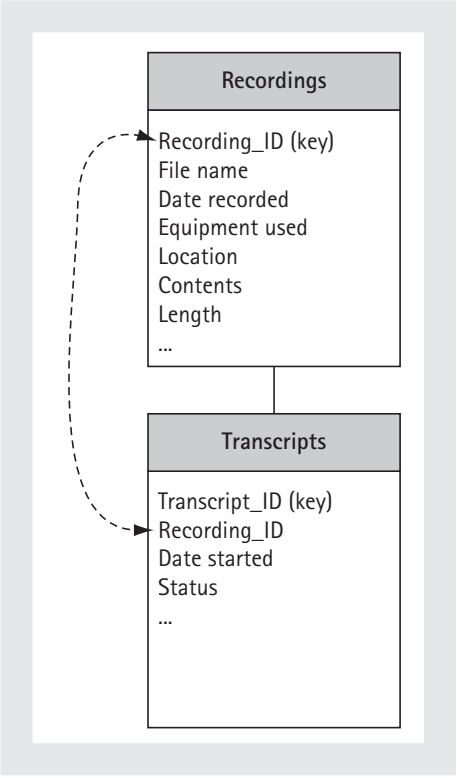

Figure 4.6. Establishing the relation between tables using keys

person–role pairing is associated. This means that the names of speakers will only appear once and so you do not risk entering variations on the same name into your catalogue, thereby making it impossible to search on all references to one person. Using a database can allow you to query your data in complex ways (e.g. 'Who are the speakers for whom I have completed transcripts of their recordings?', 'What are the different roles that person *p* held in my collection between 2005 and 2010?').

Your catalogue could also include tables listing photographs, interlinearized texts, or geographic data. As long as your metadata catalogue conforms to a few basic principles, it should be possible to export textual data from it into in an archive's catalogue. These principles include using appropriate file names, including only one kind of information in each field of your catalogue database, and using standard forms where possible, e.g. ISO standards for dates[22] (YYYY-MM-DD), country names,[23] and language names.[24]

[22] ISO 8601, http://www.iso.org/iso/ue_detail?csnumber=40874
[23] ISO 3166, http://www.iso.org/iso/country_codes.htm
[24] ISO 639, http://www.sil.org/iso639-3/codes.asp

Open Language Archives Community (OLAC) metadata

The most commonly used open-access metadata system is Dublin Core,[25] with a set of fifteen terms that form the basis of most library catalogues, as well as the Open Archives Initiative.[26] If we want our material to be locatable via international search mechanisms, then all we need to do is to prepare our metadata in a form that they can read. Luckily, OLAC provides a ready-made set of terms that can be the minimum or at least the core of a description of items in our collection. OLAC provides services that aggregate catalogues from cooperating archives and create a set of webpages listing what is available, not only in each archive, but also providing a dynamic page showing what is available in all archives for each of the world's languages, updated every eight hours.

There seems to be some confusion over the use of OLAC metadata. It is not, and was never designed to be, an exhaustive set of terms for describing linguistic data. Your catalogue can have much more information in it than just that provided by the OLAC terms. The key consideration is that if you take the OLAC terms into account in your archive's system, then you can participate in a global information system about the world's languages.

4.3.4 Regular expressions

A consequence of a workflow using different software tools is that data needs to be converted from the output format of one tool to the input format of another tool without losing any of the data in the process. Textual conversions can be achieved by using *regular expressions* (also known as 'regex'), which are sophisticated search-and-replace routines. Regular expressions can save you a great deal of time, and it is well worth at least learning what they can do, even if you don't want to learn how to use them yourself. Basic information is easily found on the web,[27] Friedl (2006) is a useful reference, and Gries (2009: 68–99, 105–72) provides good instruction on using regular expressions in linguistic contexts. You can always find someone to help you with regular expressions if you do not feel confident creating them yourself.

Three examples of the power of regular expressions are given below. Fig. 4.7 shows how regular expressions can be used to convert the tab-delimited output from Transcriber software in (a) into the structured text that Toolbox software requires, shown in (b). While the change could be done with a single regular expression, we have split it into two steps for illustrative purposes.

Step 1 finds the first tab stop, then inserts '\as' (the MDF field marker indicating the time code for start of the audio clip) at the beginning of the line and replaces the first tab stop with a carriage return and '\ae' (for the time code of the end of the

[25] http://dublincore.org/

[26] http://www.openarchives.org

[27] http://en.wikipedia.org/wiki/Regular_expression, http://www.regular-expressions.info/, or http://etext.virginia.edu/services/helpsheets/unix/regex.html.

(a) Output from Transcriber (timecodes followed by the relevant text) 78.813 [TAB] 83.083 [TAB] apu motu nigmam upregi.., rutrau wesi go rapolak (b) Converted into Toolbox format \as 78.813 \ae 83.083 \tx apu motu nigmam upregi.., rutrau wesi go		
Step 1	Find: \r([^\t]+)\t *(carriage return followed by any non-tab characters followed by a tab)*	Replace with: \r\\as \1\r\\ae *(carriage return followed by \ as and the characters in parentheses in the find expression followed by a carriage return and \ae)*
	78.813 [TAB] 83.083 [TAB] apu motu nigmam upregi.., rutrau wesi go	\as 78.813 \ae 83.083 [TAB] apu motu nigmam upregi.., rutrau wesi go
Step 2	Find: \r(\\ae [^\t]+)\t *(carriage return followed by \ae and by any non-tab characters followed by a tab)*	Replace with: \r\1\r\\tx *(carriage return followed by the characters in parentheses in the find expression followed by a carriage return and \tx)*
	\ae 83.083 [TAB] apu motu nigmam upregi.., rutrau wesi go	\as 78.813 \ae 83.083 \tx apu motu nigmam upregi.., rutrau wesi go

Figure 4.7. Example of a text (a) and its derived form (b), arrived at by use of regular expression search-and-replace routines

audio clip). The second step finds the next tab stop and replaces it with a carriage return and the MDF text field marker '\tx'.

A second example of the use of regular expressions is the conversion of a dictionary made in a word processor into computationally tractable files, ready to be used in specialized lexicographic software. Fig. 4.8 shows an unstructured Microsoft Word document with a regular pattern of headwords preceded by an asterisk and followed by a definition in italics.[28] Using a regular expression in the

[28] This example is from Clark (2009).

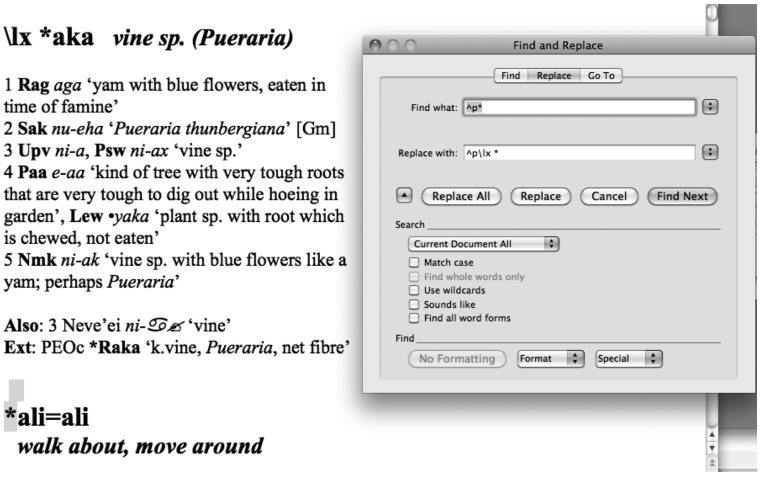

Figure 4.8. Limited regular expression search in MS Word, inserting '\lx' before each headword in a document

\lx *aka
\de vine sp. (Pueraria)

\nt 1 **fb:Rag** *aga* 'yam with blue flowers, eaten in time of famine'
\nt 2 **fb:Sak** *nu-eha* '*Pueraria thunbergiana*' [Gm]
\nt 3 **fb:Upv** *ni-a*, **fb:Psw** *ni-ax* 'vine sp.'
\nt 4 **fb:Paa** *e-aa* 'kind of tree with very tough roots that are very tough to dig out while hoeing in garden', **fb:Lew** •*yaka* 'plant sp. with root which is chewed, not eaten'
\nt 5 **fb:Nmk** *ni-ak* 'vine sp. with blue flowers like a yam; perhaps *Pueraria*'

\nt Also: 3 Neve'ei *ni-* 'vine'
\nt Ext: PEOc ***Raka** 'k.vine, *Pueraria*, net fibre'

Figure 4.9. Second insertion of codes into a document on the way to structuring all elements

Word find-and-replace window locates a carriage return followed by an asterisk and inserts '\lx' as a field marker identifying the headword, as has been done for **aka*.

Eventually, more codes can be inserted, as shown in Fig. 4.9, where the '\de' and '\nt' fields have been added. Ultimately, all elements of the entries will be explicitly coded, rather than relying on formatting to imply structure.

As a third example of the use of regular expressions, imagine you have a corpus for which you need to quantify the occurrence of a particular word token expressed as a proportion of the total number of tokens in the corpus. However, the corpus also includes coding in angle brackets for, say, syntactic roles (e.g. <sub> for subject, <prd> for predicate), which you need to exclude from the text count. Using an ordinary find-and-replace tool will not locate all of these items at once (you would need to search for each of them individually). With a regular expression search you can locate all patterns of the form 'three characters enclosed in angle brackets' with the regular expression of the form < . . . >, where a full stop or period represents 'any character except return', thus looking for three such characters inside angle brackets.

MS Word has a very weak form of regular expression search. There are a number of tools with fully-featured regular expression functions, including OpenOffice.org, Text Wrangler (Mac), EditPad Pro (Windows), and regexxer (Linux). Regular expressions can swap the order of items in a text, and insert and change regular patterns based on shape rather than content.

4.3.5 Annotation of media: time alignment and interlinearization

Transcription and translation of recordings provides another opportunity to create well-structured data. A time-aligned transcription matches sections of media to a textual representation of the contents of the recording, creating links between the two via time codes. There are several advantages to aligning text to media over creating an unaligned transcript (i.e. a transcript created in a word processor). From a research standpoint, time alignment allows access to language as an audiovisual signal, rather than just a written interpretation. You will be able to quickly search and retrieve particular portions of the media to assist you in developing an analysis. Most time alignment software produces an XML file that is both archivable and human-readable, allowing the link between the recording and the transcript to endure beyond the life of the software used to create it. From a presentation standpoint, a time-aligned transcript can be quickly and easily converted into display formats via tools like CuPED,[29] Eopas,[30] and Annex/Trova.[31]

[29] http://sweet.artsrn.ualberta.ca/cdcox/cuped/
[30] http://www.eopas.org
[31] http://www.lat-mpi.eu/tools/annex/

There are several software options available for time alignment designed for documentary linguistic use, including ELAN,[32] Transcriber,[33] and EXMARaLDA.[34] When choosing software, things to look for are:

(i) flexibility—does the tool support multiple speakers, overlapping speech, and hierarchical relationships between tiers for interlinearization, comments, or annotation of gesture?
(ii) interoperability—do the tool's import and export options allow easy conversion between the other tools in your workflow, like a lexical database or a text editor?
(iii) archivability—does the tool produce a well-documented XML file?[35]

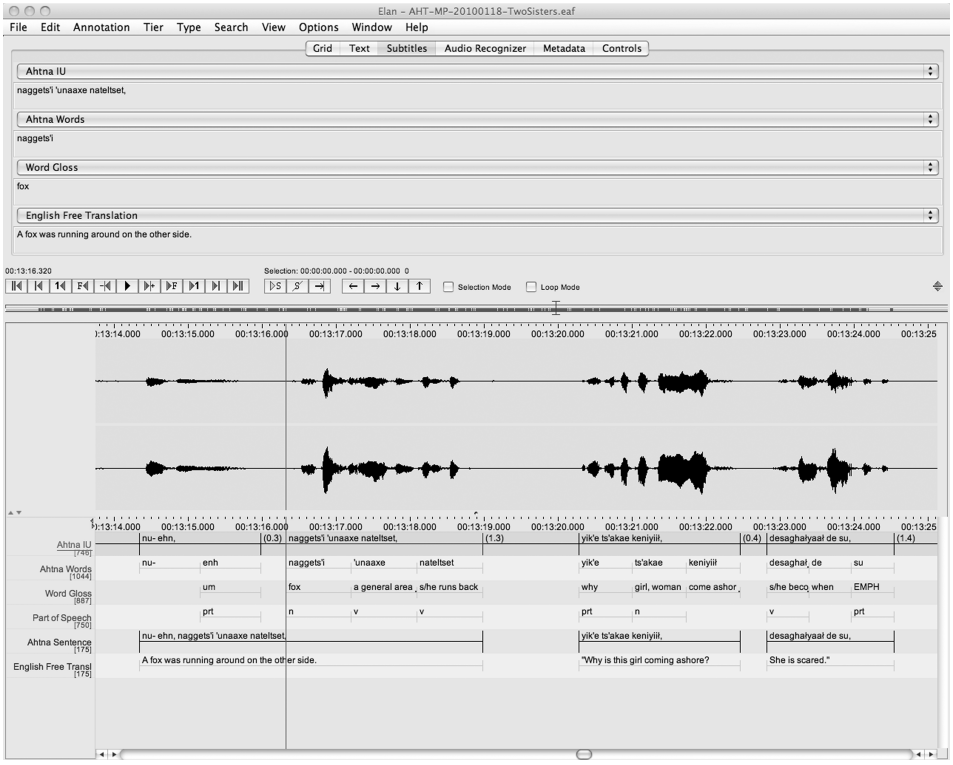

Figure 4.10. Time-aligned text and audio in ELAN, showing several levels of interlinearization: intonation units, words, word glosses, parts of speech, normative sentences, and free translations

[32] http://www.lat-mpi.eu/tools/elan/
[33] http://trans.sourceforge.net/en/presentation.php
[34] http://www.exmaralda.org/en_index.html
[35] It is best to avoid the less desirable commercial transcription software options targeted to the legal and medical professions.

Once you have produced time-aligned transcripts of the recordings, the next step is to further annotate the transcripts as interlinear glossed text (IGT), adding additional information like morphemic parsing and glosses of words and morphemes. Interlinearization tools that are tied to a lexical database (Fieldworks and Toolbox) are advantageous because they lend consistency and speed—entries are constrained to ensure they are consistent with those already in the database, and as the database grows, interlinearization becomes more and more automatic. It is essential, however, that the time codes from the time-aligned transcript be preserved so that the resulting IGT will still relate to the media. You can choose to bring your fully interlinearized text back into your time-aligned transcript (in essence making a 'round trip' with the data), or not. Fig. 4.10 shows a transcript that was time-aligned in ELAN, then

Example of the use of methods advocated in this chapter

Scenario 2

As soon as the sun comes up, my consultant drives over to my cabin in his pickup truck, knocks the snow off his boots, and comes in to sit by the heater and enjoy a cup of coffee and share the latest gossip before getting down to work. We have determined that my cabin is the best place to work, as it is warm, quiet, and free from distractions like the telephone and the television.

Today we are working on transcribing a legacy recording from the 1980s that has been digitized but never transcribed. On my desk are my laptop computer, a pair of speakers, a set of headphones, my notebook and pen, and my digital recorder. I play the entire legacy recording once through, then turn on my recorder to capture the transcription session and play the old narrative sentence by sentence from the ELAN software installed on my laptop. I have already started using ELAN to chunk the recording into intonation units, but as we work today, I use my laptop only for playback, and not for transcription. My consultant repeats the sentences back to me slowly and then provides translations, all of which I quickly scribble in my notebook. We work for a few hours until my consultant tires.

After another cup of coffee, he leaves, and my data management tasks begin. I copy the audio recording I made of the morning's work session onto my laptop, record a bit of post hoc verbal metadata about the session, and then join the two in an audio editor. I give the file an appropriate name, record the metadata in my digital catalogue, prepare a copy for sending to the archive, and back everything up on to external hard drives. I also jot down the relevant metadata on notebook pages containing my transcription. If my catalogue is ever lost, the pages can be rematched to their associated recordings.

Later that afternoon I use the handwritten notes in conjunction with the recording of the morning session to digitally transcribe the legacy text in ELAN. I then export it to Fieldworks, where I simultaneously create consistent multi-tier IGT and populate my growing lexical database. Once IGT is complete, I export the data back to ELAN to create the final time-aligned transcription. The transcript and the Fieldworks database are also prepared for archiving, entered into my metadata catalogue, and backed up locally. I scan my notebook pages monthly, and also mail DVDs of all my archival data to the overseas archive I am working with.

exported to Fieldworks for interlinearization, and then reimported into ELAN, taking advantage of ELAN's hierarchical tier structure for establishing parent–child relationships between the levels of IGT (see Berez 2007; Hellwig and Van Uytvanck 2007).

4.3.6 Lexical databases

Creating a lexical database, rather than writing a dictionary in a word processor, will allow you to build dictionaries of various kinds, including topical lists, reversals or finderlists, simple wordlists or a more fully featured dictionary. All of this will be derived from a single set of lexical data, and this set can also be built from the texts that you interlinearize, with each word in the texts being cross-checked against the lexical database. This is the work that can be done using software like Toolbox or Fieldworks. In the lexical database you can keep track of items associated with the headword, like images, sounds, or videos, and the link to these items can later be developed for whatever delivery mechanism you need to use. For example, you may want to have all headwords in the dictionary spoken and available as media in a web browser. To do this, you would first have prepared a list of headwords that would be read and recorded by a speaker of the language. The same list can then be time-aligned with the media which will give you a citable media reference for each word. These word-by-word references can then be sorted and joined to your lexicon, and the form of the link (e.g. an HTML hyperlink) can be made as needed, using the headword as the key that links the sound and the lexical entry in the database.

4.4 Advantages of well-formed data in analysis, publication, and presentation

Interaction with field recordings via their aligned transcripts allows the transcripts to improve as your understanding of the language grows. As an example, the first author's research language, South Efate, has two phonologically similar pronouns, *ga* '3SG'and *gaŋ* '2SG.POS', that need to be carefully distinguished, because in preverbal position the latter in fact acts as a benefactive. In the early versions of the South Efate transcripts, the distinction between pronouns was not always made correctly because of the difficulty of hearing the final velar nasal in *gaŋ*. As the author's knowledge of the language—and his awareness of the benefactive forms—grew, he was able to easily return to the primary recordings and look for discrepancies in the transcript. By having well-structured data, he was able to improve the transcription and confirm the analysis of benefactives in South Efate.

A well-formed corpus allows us to seek answers to linguistic questions that are difficult to ask when data is limited to what can be expressed on the printed page. Much of language structure that is embodied in phenomena like intonation, tempo, and stress, is observable only acoustically; certain aspects of spontaneous language use like disfluencies, gesture, and speaker overlap are difficult to represent on paper. Even the best transcription systems—for example, ToBI (Beckman and Hirschberg 1994), Discourse Transcription (Du Bois et al. 1992; 1993), Conversation Analysis (Schegloff 2007)—are only substitutes for direct observation. Being able to cite archived, time-aligned examples of the phenomenon you describe in your publications by their permanent handles, down to the level of the clause, word, or even segment, allows others to confirm, challenge, and build on your claims. Well-formed data is a powerful tool for research and publication, allowing linguistics as a discipline to participate in the practice of open data that has long been an ethos of scientific inquiry.[36]

The possibilities for multimedia presentation of language material are enticing, and linguists have long recognized the value of dynamic media in capturing and representing the dynamism of language. When multimedia software was first developed sufficiently, a number of projects like CD-based talking dictionaries immediately took advantage of the ability to link text to audio or video as a pedagogical and presentational tool. Unfortunately, many of these early projects had very limited lives. As the software they were built in aged, they became unplayable. Even worse, in some cases, producing the multimedia CDs was the primary goal of the language project, so the data has become effectively lost.

A better objective is to build well-formed, archival data from fieldwork, and then to derive delivery forms of the media from that base. These delivery forms can be on-line or local or paper-based, but they all benefit from being derived from well-structured underlying data. Time-aligned transcripts, with their links to specific points in their associated media, will allow you to install an iTunes jukebox in a local school (see also Barwick §7.3.4.1 below), or develop a simple Flash presentation online or on CD. Well-structured lexical databases will allow you to deliver targeted topical dictionaries on paper or over the web.

4.5 Conclusion

An often-voiced objection to the types of processes described in this chapter is that they take too much time. It is true that time-aligned transcriptions and interlinearized texts take some effort to construct, but we can liken it to building a house that has strong foundations. Once the initial effort has been made, the resulting house is able

[36] http://en.wikipedia.org/wiki/Open_science_data

to withstand the elements and the passage of time, unlike a shack that is blown over by the first storm. Similarly, a well-constructed set of data will provide ongoing access, allowing you to ask new questions of it and to keep interacting with the primary material for a deeper analysis of the language. There is a great return on this initial investment of effort in creating an accessible and reusable corpus, especially if you will continue working with the same data over a long period of time. Furthermore, there is an ethical responsibility to prepare the data that we have recorded so that it can be located, accessed, and reused by those recorded or their descendants.

It is indisputable that there is more to research on a previously undescribed language than just the recorded data. The headnotes, the associated memories, and contextualization of the research that remain in the head of the researcher continue to inform their analysis. Similarly, the body of material recorded can only ever be a partial glimpse of a language that will continue to change over time. Nevertheless, the records we create will be invaluable for many purposes, and many more than we have planned for in our own research.

Creating good archival data can allow various derived forms that can be improved over time relatively easily. Rather than having to labour over formatting complex documents by hand, automatic processes can be applied to generate diverse outputs periodically as the underlying data is improved or as the analysis develops.

Finally, while the technologies we use will change, the underlying principles described in this chapter should provide some guidance in choosing new tools and methods from those that will appear in future. There is no question that linguistic fieldwork practice needs to change, and that future language descriptions will have to include recorded examples, cited in the analysis, with an underlying corpus that is archived and described in the course of the analysis. By adopting the methods described in this chapter, linguists will be building more detailed and accessible records of the world's languages.

4.6 Sources of further advice

New technologies and methods associated with them are becoming available all the time. To keep up with options for tools or methods to use in your fieldwork, you should subscribe to appropriate mailing lists such as the Resource Network for Linguistic Diversity[37] or read blogs such as Endangered Languages and Cultures.[38]

[37] http://www.rnld.org [38] http://paradisec.org.au/blog

The SOAS Online Resources for Endangered Languages[39] (OREL) links to many useful sites. Another key source of information is the EMELD[40] site. There are also mailing lists specific to each of the tools discussed in this chapter, and this is likely to be the case for any new tools that appear in future. Useful references on topics around fieldwork and data management include Chelliah and de Reuse (2010: 197–225), Austin (2006), and Bowern (2008: especially ch. 4 on data management and archiving, and ch. 13 on working with existing and historical field materials which covers issues not dealt with in this chapter).

[39] http://www.hrelp.org/languages/resources/orel/tech.html

[40] http://emeld.org