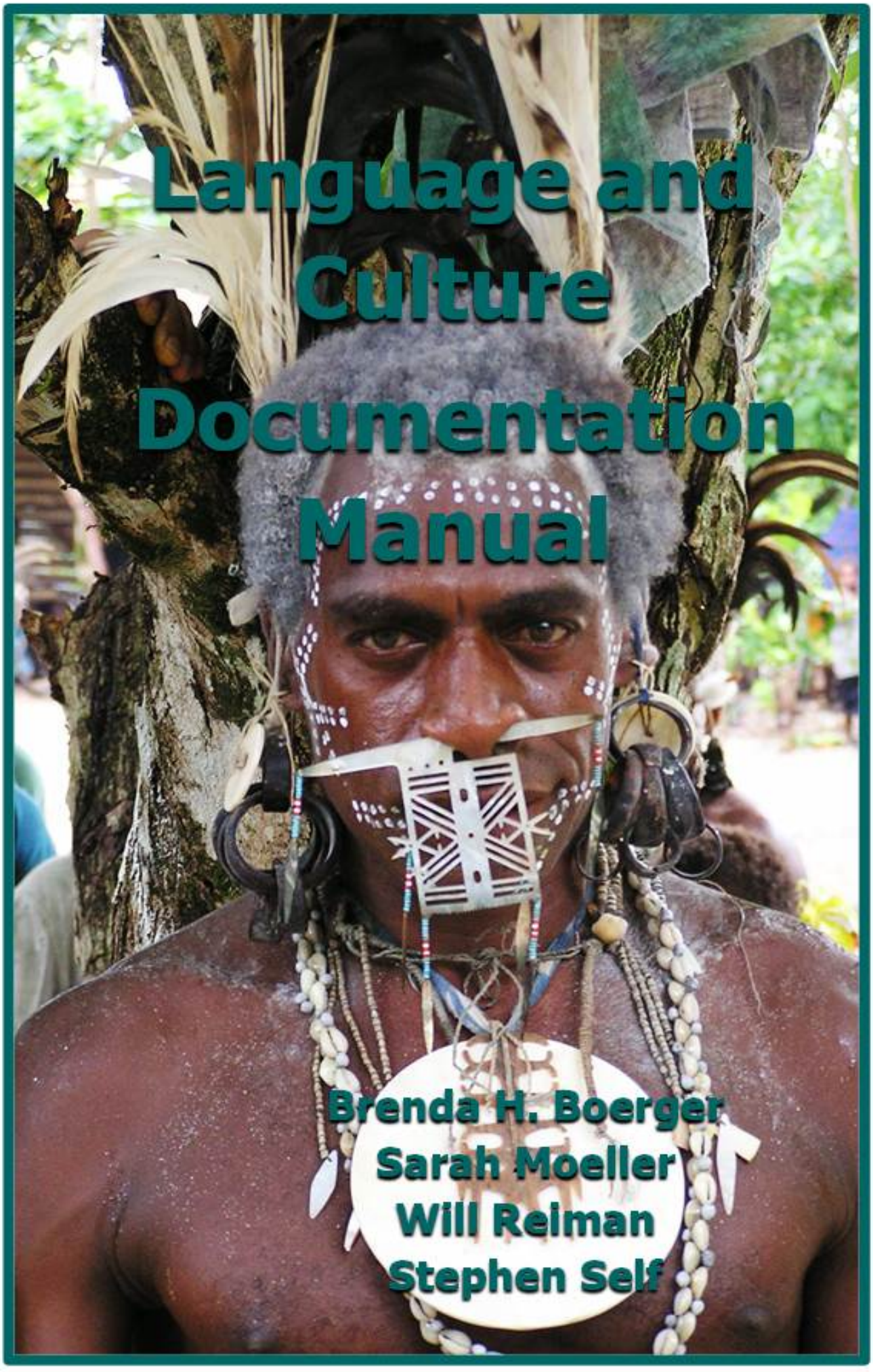
Language and Culture Documentation Manual



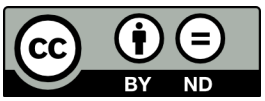
Brenda H. Boerger
Sarah Moeller
Will Reiman
Stephen Self

Language and Culture Documentation Manual

Brenda H. Boerger, Sarah Ruth Moeller, Will Reiman and Stephen Self

This book is for sale at http://leanpub.com/languageandculturedocumentationmanual

This version was published on 2023-05-18



Contents

Chapter 2. What is Language Documentation?

Think of the important events in your life. Think not only of the major, life-changing events like marriage or the birth of a child, but also minor events, like a special family vacation or outing. Did you take pictures of these events? Did you keep those pictures? Did you write anything on the back of your pictures to help you remember when and where they were taken and what you were doing then? Did you organize your photos into albums or scrapbooks with additional keepsakes and notes to help you look back on and relive the important events they depict? If your pictures are all digital, did you store them on your computer in folders that are labeled by when and where they were taken? Did you tag the photos according to the individuals pictured in them? If you have done any of these things, then you already understand quite a lot about the basic principles of language documentation, albeit on a personal level.

Language documentation as external memory

Like photographs, scrapbooks, and photo albums, documentation is a complex system of externalized memory. The capacity of human memory is finite in two senses. Not only does our memory perish with the physical bodies and brains that house it, but even in a single lifetime, none of us is capable of storing—much less retrieving—every piece of relevant information, every detail of every event, every name and every face we have come across. Art, music, literature, the internet: these are all systems of externalized memory designed to fend off mortality for our complex mental lives. Language documentation is another such system.

A brief sampling of definitions of language documentation will serve to illustrate this point. Language documentation has been defined as:

- "concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties" (Gippert, Himmelmann & Mosul 2006:v).
- "a comprehensive record of the linguistic practices characteristic of a given speech community." (Himmelmann 1998:166).
- "the creation, annotation, preservation and dissemination of transparent records of a language." (Woodbury 2011).

What key term do all three of these definitions share? **Record**. Language documentation is about creating a record of the language and culture of a human community that will outlast the individual memories of those creating it. Like a photograph, the records created for language documentation should directly reflect or recapitulate aspects of the original events they record.

PLATO, THOTH, WRITING, ORALITY, AND MEMORY

In his dialogue entitled *Phaedrus*, the Greek Philosopher Plato dramatizes his teacher Socrates' retelling of the myth of the Egyptian god Thoth (also spelled Theuth). The myth recalls an encounter between Thoth and King Thamus of Egypt, in which Thoth reveals that he has created what he expects will be a great gift for humanity: the art of writing. Thoth claims that writing will greatly benefit humanity by providing a sure remedy for faulty memory.

Thamus, however, surprises Thoth by criticizing his gift rather than praising it. Thamus argues that writing will only make human minds weaker by allowing them to get around the practical problem of working hard at remembering and recalling things they wish to hold in mind.

Many, if not most, of the communities where language and culture documentation is taking place have no historical traditions of writing and literacy. These communities embody what are usually referred to as predominantly "**oral cultures**," where information is passed down through stories and songs from one generation to the next. In such environments, we could say that the problem of externalizing memories has been solved not with physical technologies like writing and photography, but through recruiting other minds to store, recall, and retell traditional stories and songs and thus to perpetuate them through vast stretches of time.

Many of the great epic poems of Western literature—*The Iliad* and *Odyssey* in Greek, *Beowulf* in Old English, *The Tain* (or Cattle-raid of Cooley) in Old Irish—started out as traditional oral literature handed down for centuries prior to being reduced to writing.

Language documentation as linguistic and anthropological fieldwork

Language documentation involves numerous component activities. The activity that lies at its heart, however, is working directly with members of speech communities in their home environments to compile an audio-visual record of their linguistic and other cultural practices. Thus, language documentation occupies a central position in a long tradition of linguistic and anthropological fieldwork dating back to the beginning of the twentieth century.

German-American anthropologist Franz Boas (1858-1942) stressed the importance of first-hand contact and fieldwork with **indigenous** communities of interest to scientific research. Both he and his students like Edward Sapir, Ruth Benedict, and Margaret Mead conducted extensive onsite research among numerous indigenous communities throughout the world. Large amounts of recorded data resulted from these early investigations.

In those early days, field recordings were often made on wax cylinders that could be played back on specialized phonographs. Later, reel-to-reel tape recorders were used for field recordings. By the 1950s, the first modern cassette tapes had become the medium of choice. Recording technology has changed considerably in recent decades and will doubtless continue to change rapidly over the

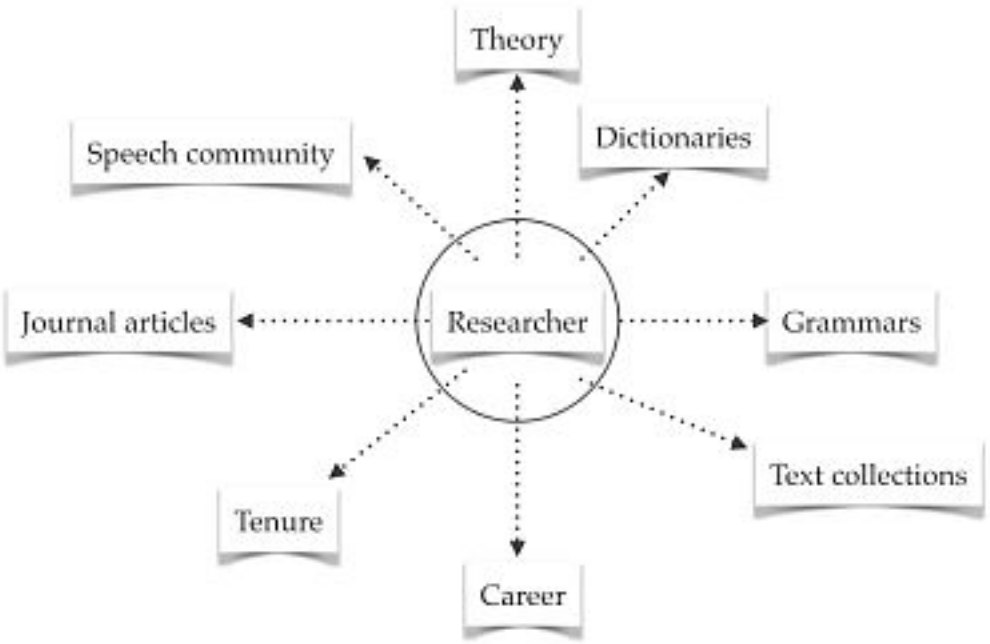
course of this century. However, onsite research with speech communities will continue to form a key component of language documentation. And such research will always involve cultivating long-term relationships with individual speakers and becoming directly involved with them on a personal level.

Language documentation as a shift of focus from earlier fieldwork

Language documentation has benefited from the long tradition of linguistic and ethnographic fieldwork stretching from the early twentieth century. Yet it nonetheless represents a significant shift of focus from earlier conceptions of fieldwork in two important respects.

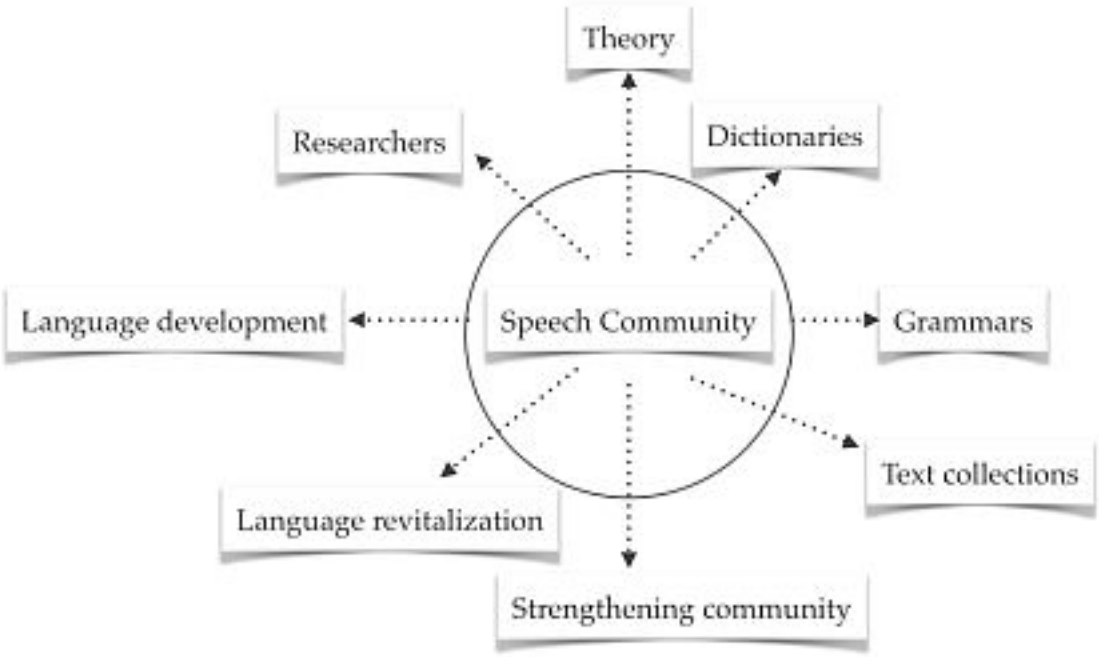
Shift of focus off researcher to speech community

First, language documentation takes its cue from the advocacy research, negotiated fieldwork, and empowerment research models that have gained currency since the 1960s. All three of these stances to linguistic and anthropological fieldwork emphasize the subjects of research—in this case, speech community members—as individual people with their own concerns and goals that should be entertained and met whenever possible. Earlier fieldwork models were researcher-centered, focusing principally on the research aims and career goals of the scientist rather than the community (See Earlier models figure).



Earlier models of fieldwork were researcher-centric

A concentration on social diversity and the value of linguistic diversity form a key shift of emphasis represented in the documentary linguistics movement. This emphasis entails that the documentary linguist remains primarily responsible to the speech communities whose practices are being documented. Language documentation places the speech community and its goals and concerns in the center of attention (See figure below).



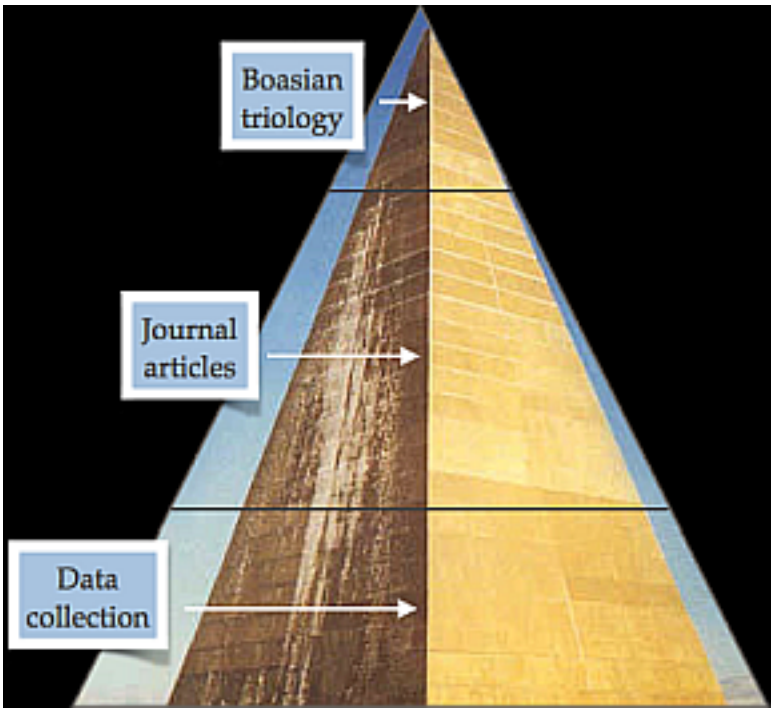
Language documentation focus on the needs and aims of the speech community

Shift of focus from data analysis to data collection

Even though language documentation evolved out of traditional **descriptive linguistics**, its focus has shifted off viewing data collection as just a necessary step toward the ultimate aim of analysis and onto seeing data collection, storage, and management as proper goals in and of themselves.

Since the time of Franz Boas, traditional linguistic and anthropological fieldwork has involved collection of large amounts of linguistic and cultural data from speakers of small languages. The researchers used these data to form conclusions and theories about the languages and their speakers. Later, they would publish their conclusions and theories as a contribution to their fields. The now standard three outputs of traditional descriptive linguistics—a descriptive grammar, a lexicon, and a collection of edited texts—are sometimes referred to as the "Boasian trilogy" in deference to the practice of Boas and his students. The mostly unanalyzed collections of data themselves were not considered important beyond the specific purposes of the researchers who gathered them. Perhaps, following the death of researchers, their collections of primary data would be gathered and stored in museums and archives for historical interest. Yet little thought or care was given to storing the

information in these collections in such a way that it could be used by future researchers or members of speech communities for their own purposes, whatever those may be.

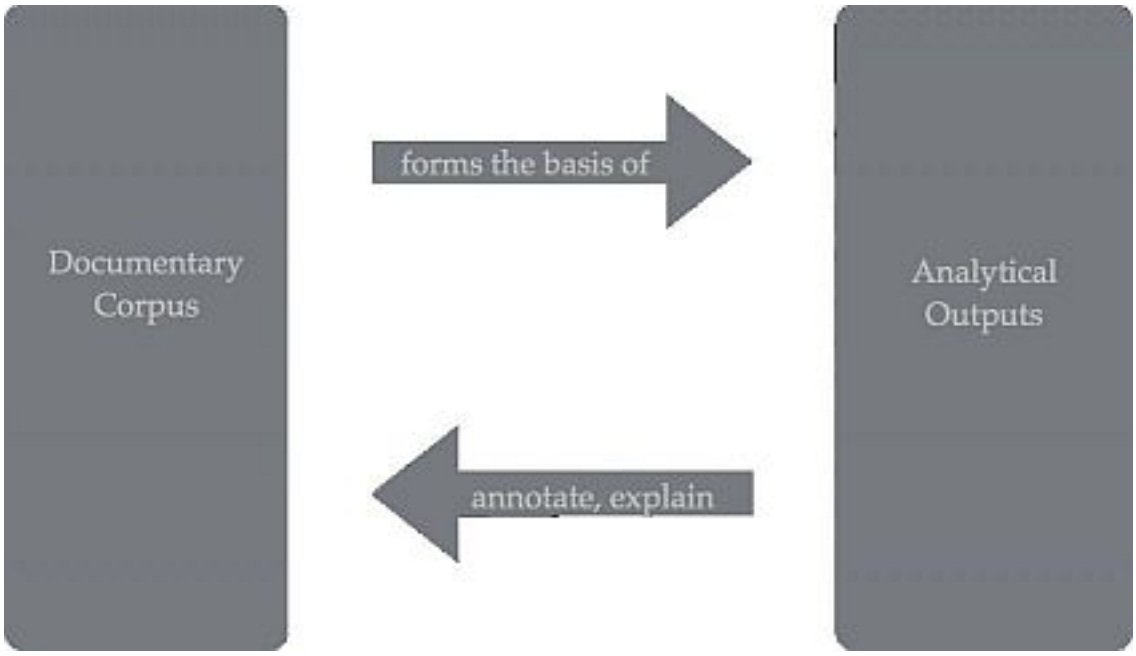


Traditional pyramidal model of descriptive linguitic outputs

Two major early theoretical voices in language documentation, Nikolaus Himmelmann of the University of Köln, Germany, and Anthony Woodbury of the University of Texas at Austin, have argued for a reversal of priorities when it comes to the relationship between data gathering and analysis. In a language and culture documentation project, the data itself—and the processes by which it is gathered, stored, and later accessed—assume the primary focus. Significant consideration is given to ensuring that recorded materials are safely stored and available for the long-term future.

Himmelmann, especially, viewed the separation of linguistic analysis from data collection as necessary in order to provide the collection and presentation of primary data with the theoretical and practical attention they deserve. Since both traditional descriptive linguistics and documentary linguistics involve basic **transcription** and **translation**, it is easy to blur the lines between the two and see compiling data as a component part of descriptive analysis. Yet language documentation goes well beyond merely editing and archiving a researcher's field notes on the way to producing a Boasian trilogy of grammar, dictionary, and collected texts. Language documentation constitutes a methodologically distinct field, characterized by greatly expanded focus on the collection and management of primary linguistic data. This vision of language documentation seeks to reverse the interdependency between documentary and descriptive linguistics. Rather than a pyramid-shaped model that places grammars and dictionaries at the pinnacle as the goals of documentation (Figure 3 traditional pyramid), language documentation views explanatory and analytical materials such as a grammar, dictionary, and collected texts, as informing and annotating the **documentary corpus**

(See figure language and culture documentation views).



Language & culture documentation views explanatory and analytic material as informing and annotating the documentary corpus

The core tasks of language documentation

Language documentation aims at preserving a record of the linguistic practices and traditions of a speech community (Himmelmann 1998:166). This goal is broad and requires a wide range of expertise to fulfill. For this reason, language documentation is a multi-disciplinary or interdisciplinary field, drawing on not just linguistics, but also anthropology, ethnomusicology, art, audiovisual design, recording technology, and more.

Whether a documentation project comprises a single archival submission of language data from a single speech community at one point in time or the totality of interdisciplinary materials archived for that language over a span of time by a researcher or team of researchers, it involves four core tasks.

Compiling

Documentation begins with compiling raw data. These data consist of audio and video recordings of **communicative events**. Himmelmann defines the term communicative event as a sound event or moment of linguistic utterance with as much of its accompanying context as can be captured: location, posture, gestures, **artifacts**, and any other way that such an event can be contextualized. A complete record of communicative events thus requires both audio and video recordings.

In addition to the communicative event itself, a full compilation should also contain **wordlists**. These may comprise morphological paradigms, numbers, measure words, folk taxonomies, or any other types of lists that are culturally appropriate and fill the needs of both researcher(s) and speech community.

The compiling component should also include a third type of material that Himmelmann (1998:169) referred to as **analytic matters**. These elements may be data elicited by the documentary linguist(s) or **analytic discussions** of language matters independent of particular communicative events or wordlists. Such discussions allow researchers to arrive at a better understanding of the language and culture in a given speech community. What Woodbury (2003:42) referred to as a metacorpus –that is, recordings of members of the speech community discussing previously recorded communicative events—might also fall into this category.

Compiling is discussed in more detail in the chapter on capturing quality audio and video recordings.

Annotating

In order to ensure that a documentary corpus fulfils the aim of being truly transparent and useful to those who might wish to consult it in the future, it should also contain what Anthony Woodbury (2007) has called "thick translation." Thick translation is essentially a form of elaborated **annotation** and may include any or all of the following:

- Audio recordings of real-time free translations
- Word-by-word and sentence-by-sentence translations by a mother-tongue speaker
- Linguists' breakdown or **parses** into minimal meaningful elements with an invariant gloss or definition in a language of wider communication
- Drafts of ever more refined literary translations by mother-tongue speakers, target-language speakers, or a collaboration of both
- Formal poetic analyses of the original
- Alternative versions of the same text
- Literary exegeses
- Discussions
- Footnotes

In many ways, Woodbury's notion of thick translation encompasses much of what documentary linguists desire from the annotating component of documentation. This component consists of recordings by native speakers commenting on the communicative events and wordlists in both the vernacular and, crucially, a language of wider communication (LWC). While this component may sound superficially similar to the analytic matters included in the discussion of compiling raw data above, the purpose of thick translation remains distinct. The original raw data may include recordings of speakers discussing previously recorded communicative events; however, the thick translations of those recordings will additionally include translations and parses that will make the materials of use to linguists, other researchers, and heritage members of the speech community not

personally involved in the original recording events. These elements parallel the interlinearization and glossing of written texts done by traditional descriptive linguists.

Annotating is discussed in more details in later chapters.

Metadata

Language documentation should be accompanied by "a thick description of a non-linguistic nature" (Bergqvist 2007b) providing information about the recording situation as well as its nature and contents. This information is known as **metadata** (literally 'about data'). Metadata may be superfluous to the actual linguistic analysis, but due to the long-term perspective of documentary linguistics, it is necessary for the purposes of preservation and management. In addition to its other benefits, metadata also helps ensure continuity within organizations, as personnel move in and out and pass along data to those who come after. In this way, the inclusion of metadata in documentary corpora ensures that language documentation achieves its goal of being transparent, ongoing, and distributed. The practical aspects of documenting metadata are discussed in Part 4.

The National Information Standards Organization[1] (NISO) identifies three basic types of metadata: **descriptive**, **structural**, **administrative**. Some documentary linguists also recognize a fourth type: **technical**.

Descriptive metadata. Descriptive metadata provides a description of the material for the purposes of discovery and identification. Imagine a dusty attic (perhaps your own?) filled with countless sealed boxes and no labels. How would you ever find anything? A documentary corpus will serve no purpose if other interested parties cannot locate it and identify its contents. Descriptive metadata ensures the documentary corpus does not languish in a (digital) dusty attic.

Structural metadata. Structural metadata indicates how the material is put together, what its internal components and divisions are, how it is ordered, and the like. Often a documentary corpus contains disparate items, such as journals, metadata sheets, and image, audio and video files, all related to a single recording event. These need to be joined together. Structural metadata connects all these different elements. Also, if files stored with the corpus are named with standardized codes—a recommended best practice we will discuss in DB chapters—the key to the codes would fall into this category of metadata. Structural metadata indexes the internal structure of the documentary corpus and ensures that it stays together, so that individual elements of the corpus cannot be lost.

Administrative metadata. Administrative metadata provide information to help manage the material, such as when it was created, how it was created, what digital file types and formats it contains, and other technical aspects of its production and management. There are several subsets of administrative metadata. One example is **rights management metadata**, which provide information about the intellectual property rights involved in the material and governs who can access the material. Another example is **preservation metadata**, which contain information required to chronicle changes to the material and preserve access to it through time. This information includes its **provenance**, authenticity, past preservation activity, and the technical specifications of equipment needed to interact with and use it.

[1] www.niso.org

Technical metadata. Some practitioners of language documentation recognize a fourth category: technical metadata. This category includes detailed notes about the type of compiling and annotating equipment used, instrument settings, **sampling rates**, and the like. Others consider this information part of administrative metadata.

Metadata may be likened to the information contained in a bibliographical citation in an academic paper or the information contained in the catalogue entry for that same item in a library. It is a tool for managing information. Thus, it should come as no surprise that in this "Information Age" interest in and concern for metadata and information standards have proliferated. **Metadata standards** (or schemas) ensure that metadata remains consistent across corpora. They prescribe what terms should be as labels and define what kind of information (or values) should be entered under each label (or field). End users should be able to compare the labels used in a given corpus against the universal standard and immediately conclude the content and purpose of the metadata value. Several metadata schemes exist for the description of language resources. Two popular standards used for language documentation are the Open Language Archives Community (OLAC) standard and the ISLE Meta Data Initiative[2] (IMDI) standard.

The Dublin Core Metadata Initiative[3] (DCMI) is a non-partisan, international organization dedicated to metadata design and best practices. The organization's activities focus on the Dublin Core Metadata set which originated in 1995. The Dublin Core Metadata set consists of fifteen standardized elements that help fully identify a given piece of information by categories such as title, creator, subject, etc. The Dublin Core Metadata set and the Dublin Core Metadata Terms (DCMT) lie at the heart of the Open Language Archives Community[4] (OLAC). Founded in 2000, OLAC is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on current best practices for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources. SIL International's Chief Research Officer Gary Simons is one of the coordinators for OLAC and co-author of the OLAC metadata usage guidelines (Simons, Bird & Spanne 2008), along with computational linguist Steven Bird and SIL member Joan Spanne. These scholars have worked hard to propagate high standards for data storage, cataloguing, and retrieval.

The ISLE Meta Data Initiative (IMDI) **metadata schema** was developed specifically for language documentation at the Max Planck Institute for Psycholinguistics[5]. It was designed to describe language resources preserved in various media. This metadata standard allows browsing and searching documentation corpora when using specific software tools developed at Max Planck Institute and used by The Language Archive[6]. IMDI is part of the Component Metadata Infrastructure (CMDI), which combines several metadata schemas, including OLAC. CMDI was initiated by the Common Language Resources and Technology Infrastructure (CLARIN) to enable discovery of language corpora and individual items. CMDI and CLARIN are still under construction but its goal is to allow researchers to use different metadata components in a way that suit particular needs.

[2] www.mpi.nl/IMDI
[3] dublincore.org
[4] www.language-archives.org/OLAC/metadata.html
[5] www.mpi.nl
[6] tla.mpi.nl

Archiving

Finally, the documentary corpus with all its cross-referenced recordings and annotations, and all attendant metadata, must be submitted to a stable, protected, **accessible**, and reputable archive. Until this happens a documentary project cannot be considered complete. All of the meticulous work leading up to this point will amount to nothing if the documentary corpus is not secured in a location that guarantees digital storage for the indefinite long-term. In an archive, the copious amounts of data that may, in some cases, represent the only surviving record of the linguistic and other cultural practices of a given speech community will remain accessible to scholars and members of the local speech community far into the future.

Archiving is discussed in more details in chapters on Finding an Archive and Preparing for Archiving.

The Seven Dimensions of Portability

The data records that result from a language and culture documentation project will be most useful in the long-term if it meets certain standards. Documentary linguists must take these standards into consideration when compiling, annotating, adding metadata to, and archiving the corpus. For language and culture documentation, Helen Dry Aristar and Gary Simons (2006) distinguish between good, better, and best practices. **Good practices** are the acceptable minimum standards. In language and culture documentation, better practices are standards that provide preservable, intelligible, and **accessible**, but not completely interoperable documentation. **Best practices** achieve the best results in a particular field and in documentary linguistics; they are essential to the final goal of a preservable, intelligible, accessible, and interoperable documentary corpus. Being *preservable* means that the media and formats used for recording and storage should be easily stored and usable far into the future. Being *intelligible* means there are **annotations** that make it possible for those who do not know the language to interact with the data. Being *accessible* means that those who want to use the data actually have a way of finding it and obtaining access. And finally, being *interoperable* means that the data is coded in a way that allows it to be the input to multiple means of processing. Even though these are not all consistently attainable, operating under best practices will guide one to the best currently attainable results.

In 2003, linguists Steven Bird and Gary Simons published a seminal article entitled "Seven dimensions of portability for language documentation and description" which addresses difficulties of data access and preservation due to issues of content, format, discovery, access, citation, preservation, and rights. In the article, Simons and Bird note with irony that the proliferation of technologies that has made possible the storage of so much language data has also condemned that data to an endangered status along with the very languages documented. Proprietary software, platforms, and configurations are ever changing, mutually incompatible, and short lived. Since the great excitement of the early 1990s, so much emphasis has been placed on collecting language data that few have stopped to consider the need for long-term data preservation.

To prevent data from becoming part of the "embarrassing level of digital detritus," researchers need to store the fruits of their labor in a format that is open to multiple users across multiple platforms, that makes use of markup that will permit format conversion, storage, and query, and can be rendered for end users in a clear and meaningful way. Meeting these needs is what portability is all about. The seven dimensions of portability are 1) content, 2) format, 3) discovery, 4) access, 5) citation, 6) preservation, and 7) rights. Many of the issues raised in the following sections are discussed in detail in subsequent chapters.

Content

The content of the **documentary corpus** consists of digital files containing data and metadata. The three main areas of concern for these files are coverage, accountability, and terminology.

- **Coverage** – What do you record and how was it elicited?
  To obtain optimal coverage of the possible categories of interest, two steps are critical. First, making rich records of rich interactions, especially in the case of endangered languages or genres. Here "rich" means a sufficient number of linguistic categories (sometimes referred to as linguistic types or genres) are represented and have been commented on (annotated) by speakers of the language. Second, documenting in the metadata the "multimedia linguistic field methods" that were used. For example, if the data was elicited using a questionnaire, the metadata should say which one and how many speakers were interviewed.
- **Accountability** – Can others verify your claims by finding and reviewing the data they are based on?
  Accountability regarding how the data and any scholarly conclusions based on that data are related requires the researcher to provide four things. First, language descriptions must indicate the documented data on which they are based. For example, if a grammar is based on a corpus of texts, the author must indicate which texts. Second, researchers should provide access to the original recording (without segmentation or trimming) whenever texts are transcribed; it is also appropriate to include the segmented recordings. Third, transcriptions should be time-aligned to the recording to facilitate verification. Similarly, when recordings have been significantly edited, the original recordings should be included in the archived corpus as well, to guarantee authenticity of the materials.
- **Terminology** – Have you used a common ontology of terms at all points?
  It is critical that others be able to understand what transcriptions and analyses mean. Therefore, all a) terminology and abbreviations, b) element tags in descriptive markup, and c) symbols used for phonological transcription should be mapped to commonly accepted linguistic or anthropological terms.

Format

The formats of documentary digital files need to be accessible to many users through time. They should minimize any obstacles to the files' ongoing use. That means digital recording and processing

must consider factors such as openness, **encoding**, **markup**, and rendering. See Software and DB chapters for discussion of these issues in more details.

- **Openness** – Are the files in open formats?
  Documentary digital files should be stored in formats that are **open source**, that is, whose specifications are published and non-proprietary. Furthermore, digital formats should be preferred if they have these characteristics: a) they are supported by software tools available from multiple suppliers, b) they can be accessed with free tools rather than commercial tools only, and c) if proprietary formats, they are published formats rather than secret proprietary formats.
- **Encoding** – Will the characters used in the files be understood properly?
  Four steps will ensure that characters are properly understood. First, encoding all characters with **Unicode**. Second, if at all possible, avoiding "Private Use Area" characters, but if they are used then document their use fully in the metadata. Third, describing any 8-bit character encodings. And finally, describing any scheme used to transliterate characters.
- **Markup** – Does the markup system give the highest functionality of the corpus?
  It is important to consider what markup schemas the files use to communicate with the software, so that the documentary corpus can be used by the greatest number of software programs. That means choosing digital formats based on the following six principles. First, descriptive markup are preferred over presentational markup. Second, **XML** (with an accompanying DTD or Schema) is preferred over other schemes of descriptive markup. Third, if the XML DTD or Schema is not a previously archived standard, the standard should be archived, giving a unique identifier (or label or filename) to each version. Alternatively, if a descriptive markup scheme other than XML is used, a document that explains the markup scheme should be prepared and archived. Furthermore, whenever a resource using descriptive markup is archived, the archived resource should be cross-referenced (or linked) to the archived version of the associated markup format's definition. And finally, if punctuation and formatting are used to represent the structure of information, the metadata should document how they are used.
- **Rendering** – Will the fonts and formats chosen be properly rendered?
  Although different type fonts and elegant formatting may enhance the information in text files, the documentary linguist needs to consider how the choices of font or formatting affect the appearance of the files submitted and whether they will be properly rendered in other software programs. For example, if the fonts needed to appropriately render the resource are not commonly available, then the fonts need to be archived along with the resource and cross-referenced to the resource. Likewise, if stylesheets have been used to render the resource, they should be archived. Similarly, it is best practice to provide one or more human-readable versions of the digital information, using presentational markup, such as HTML. Proprietary formats are acceptable for this purpose as long as the primary documentation is stored in a non-proprietary format.

Discovery

Discovery concerns the ability of speech community members, scholars, and interested persons to discover that the data exists and where it is housed. The goal of a documentary corpus is to allow others to discover and use the documentary data.

- **Existence** – Will people be able to find your data in an internet search?
  Two things are necessary to reveal the existence of archived data. First, all language resources should be listed with a repository that has an online portal accessible to online search engines or listed in the OLAC union catalog hosted by LinguistList[7]. And second, any item presented in HTML on the web should include metadata containing keywords and description that can be used by conventional search engines.
- **Relevance** – Will people be able to determine if your documentary corpus is relevant to their interests?
  In addition, to make it possible for people to know if the corpus is relevant for their purposes, it is essential to follow best practice for describing language resources in metadata, especially concerning language identification and data type. This will ensure the highest possibility of discovery by interested users.

Access

Access is concerned with how future users will gain access to the data once they have found it.

- **Scope of Access** – What parts of your corpus are accessible?
  Documentary corpora should be made as fully accessible as possible. This can be done by first, publishing (i.e. archiving) a complete primary documentation, including a process for anyone to obtain the documentation; second, publishing documentation in formats which allow users to manipulate the files in novel ways; and finally, transcribing all recordings in the orthography of the language (if one exists), along with a translation in an international language.
- **Process for Access** – Who can access your corpus? How is access obtained?
  The documentary linguist and/or archive must establish procedures for obtaining permission to access the corpus. This involves several steps. First, the process for access must be documented as part of the metadata for the corpus, including any licenses and charges. Next, all restrictions on access must be included as part of the metadata. Then, for resources not yet distributed over the web, the expected delivery time when a resource will be published should be documented. Similarly, for resources not yet distributed over the web, online surrogates or summaries that are easy for potential users to access and evaluate should be published.
- **Ease of Access** – Is it relatively easy for all potential users to obtain access to your corpus?
  Different users have different needs. The archived files and any products produced must attempt to meet most needs. First, appropriate delivery media for a variety of digital resources must be used. For example, use the web to share small resources, but a CD or DVD for larger

[7] linguistlist.org/projects/olac.cfm

resources. Second, low-bandwidth surrogates for multimedia resources should be provided. For example, publish MP3 files that correspond to large, uncompressed audio **WAV** files. Thirdly, transcriptions and translations of extended recordings must be provided to facilitate access to relevant data segments. And finally, for communities with limited internet access, printed book versions of the relevant recordings should be published and a written list of any multimedia content should be provided in a major language.

Citation

Citation refers to issues regarding bibliographic citations of electronic documentary corpora. Citation involves four key concepts: the ability to cite a resource in a bibliography, the persistence of the electronic resource identifiers, the **immutability** of materials that are cited, and the granularity of what may be cited.

- **Bibliography** – How should the relevant resource be cited?
  There are four parameters to address regarding how to cite resources in documentary corpora. First, the complete bibliographic data must be furnished in the metadata for all resources created. Second, complete citations must be provided for all language resources used. Third, instructions on how to cite electronic resources must be given on the website for a digital archive. And lastly, the metadata record of a resource must be used to document its relationship to other resources.
- **Persistence** – Does your resource have a long-term identifier and a way to obtain the resource?
  Two strategies help assure long-term accessibility for documentary resources. The first is to ensure that resources have a persistent identifier, such as the ISBN system which persistently identify books. In digital corpora, the identifier is often the same as the filename. The second strategy is to ensure that the persistent identifier links to an online instance of the resource, or else to online information about how to obtain the resource.
- **Immutability** – Do all references to your resource refer to the same thing?
  Users need to be sure that a resource is referred to in such a way that no confusion is raised. This is addressed by, first, providing fixed versions of a resource, either by publishing it on a read-only medium, or by submitting it to an archive which ensures immutability; and secondly, by providing thorough **provenance** metadata which among other things, distinguishes multiple versions with a version number or date, and assigning a distinct identifier to each version, so that all changes to the resource are tracked.
- **Granularity** – Is it possible for users to refer to individual components of your corpus?
  Often users will not want to refer to an entire corpus, but to individual elements: recordings, transcriptions, translations, and the like. Therefore, it is essential to do two things. First, provide a formal means by which the components of a resource may be uniquely identified. Second, take special care to avoid the possibility of ambiguity, such as, for example, arises when headwords only are used to identify lexical entries in a wordlist or dictionary, since multiple entries can have the same headword.

Preservation

Preservation is about ensuring that digital resources will be accessible to future generations. This involves three factors: the longevity of the format, the safety of resources from catastrophic loss, and the ongoing migration of resources to current physical and **digital media**.

- **Longevity** – Will your corpus files be available in the long-term future?
  Longevity has multiple sub-points for consideration. To assure longevity, a documentary linguist should: a) deposit the full documentary corpus in a digital archive that can credibly promise long-term preservation and access; b) ensure that the archive satisfies key criteria for a responsible archive; c) digitize any **legacy data** to permit access in the future; d) publish documentation data on the internet so that they can be captured by internet search engines and internet archives; e) transfer digital resources to update storage media every five years or before the existing media becomes obsolete and unsupported; f) archive physical versions of the data (printed versions of documents, etc.); and g) prefer digital file formats that have the best prospect for accessibility far into the future.
- **Safety** – How can you protect the existence of your resources?
  It would be a great loss if a documentary corpus or any of its files became lost and unrecoverable for any reason. To avoid that, first, copies of archived documentation and description should be kept at multiple locations. This follows the **LOCKSS** concept, "Lots of Copies Keeps Stuff Safe." Second, the archive and/or depositor should create a disaster recovery plan, containing procedures for salvaging archived resources in the event of a disaster.
- **Media and migration** – What will keep your files from becoming outdated?
  Several practices prevent files from becoming obsolete or unusable. First, language resources should be stored on digital mass storage systems to make it easy to backup and transfer the files onto upgraded hardware. Similarly, every one to five years, any offline digital files should be transferred to new storage to keep digital files fresh. The frequency of transfer will depend on type of media and where it is being stored. Finally, any language resources that are stored in a proprietary format should be migrated to a new format before the existing format becomes unsupported, approximately every five years.

Rights

- **Terms of Use** - Who can use the resources, for what purposes, and are there restrictions?
  Once individuals have given permission to record them, the documentary linguist needs to be conscientious about observing any restrictions they place on the use of the materials. To that end, the intellectual property rights for each resource must be established and included as part of the archived corpus. Also, it is wise to include a "terms of use" statement for the whole corpus that says what users may and may not do with the resources in it.
- **Benefit** – Will all intended users benefit from the corpus?
  The documentary linguist should strive to obtain as far-reaching permissions as possible since it is difficult to anticipate all uses to which the resources might be put. Explicit permissions should

be obtained for the data to be used for research purposes. Also, an explicit statement should state that the use of the materials, i.e. the archived corpus, is not limited to the researcher/s, agency responsible for collecting the corpus, or the agency funding the work.

- **Sensitivity** – Did anyone say anything they do not want made available?
  It is nearly inevitable that someone will say something that should not be circulated. To protect confidentiality of such resources, the nature of any sensitivity should be documented in detail. To help the archive that will host the resource, the depositor should provide example scenarios of what the speaker does not want to happen. This will help the archive interpret the restrictions and share the resource appropriately.
- **Balance** – How can you make the maximum number of resources available?
  It is important, though, to balance the restrictions with accessibility to as many resources as possible. Therefore, non-sensitive sections of a corpus should be generally accessible while only the sensitive sections are limited. The corpus metadata should determine and stipulate a time after which each sensitive item can be shared, including objective criteria for determining when that time has been reached. As a general principle, for publishing purposes, a researcher should limit exclusive access to the primary documentation for a period of five years following the recording. This allows adequate time for the researcher to derive the most personal benefit, while not unduly delaying access to the data.

Chapter Questions

1. Why is language and culture documentation important in the twenty-first century?
2. List and explain the critical elements of language documentation.
3. How does documentary linguistics differ from and contribute to traditional descriptive linguistics?
4. What are the three types of metadata?

Chapter 13. Software

Language and culture documentation arose with the digital age. Much of the impetus behind the development of documentation as a discipline has been predicated on the possibility to capture and store large amounts of digital data at affordable costs. Primary data in linguistic and anthropological studies have always been important, but the benefit of large data corpora has not, until recently, justified the financial cost of creation and the manpower and time cost to sufficiently process it. Nor, before the advent of the internet, did even the largest archive's relatively limited accessibility allow the full benefit of what corpora existed to be exploited. Language and culture documentation—from compiling, commenting, to metadata creation and archiving—is primarily a digital process. Its motivating vision to preserve data on endangered languages, if not the languages themselves, for future generations of scholars, speech community members, and the wider world—all of it, in the end, is about the creation, manipulation, and preservation of a specific sequence of ones and zeros stored in a bunch of digital files.

A digital process requires different types of digital tools. Chapter 11 on equipment presented guidelines and recommendations for selecting tools that *create* digital language and culture documentation data—digital audio/video recorders, microphones, and the like. After the digital recordings and images are created, the documentary linguist needs another set of tools to *manipulate* and *preserve* (until depositing in an archive) the data. These tools are software programs.

Software programs abound for audio editing, video processing, linguistic analysis, and file management. Even in the rapidly developing world of digital technology, there is no one software tool that fulfills all the necessary tasks of a language and culture documentation project. Yet, with so many programs to choose from it can be difficult to know what combination of software programs will work best for a specific project. This chapter offers general guidelines and questions to ask when selecting software tools. After the general guidelines, four software programs are introduced which fulfill these guidelines. After becoming familiar with the four core programs, documentary linguists will want to explore other tools until they find the ones that suit the specific goals of their projects; the last sections in this chapter has tools and resources that are worth investigating.

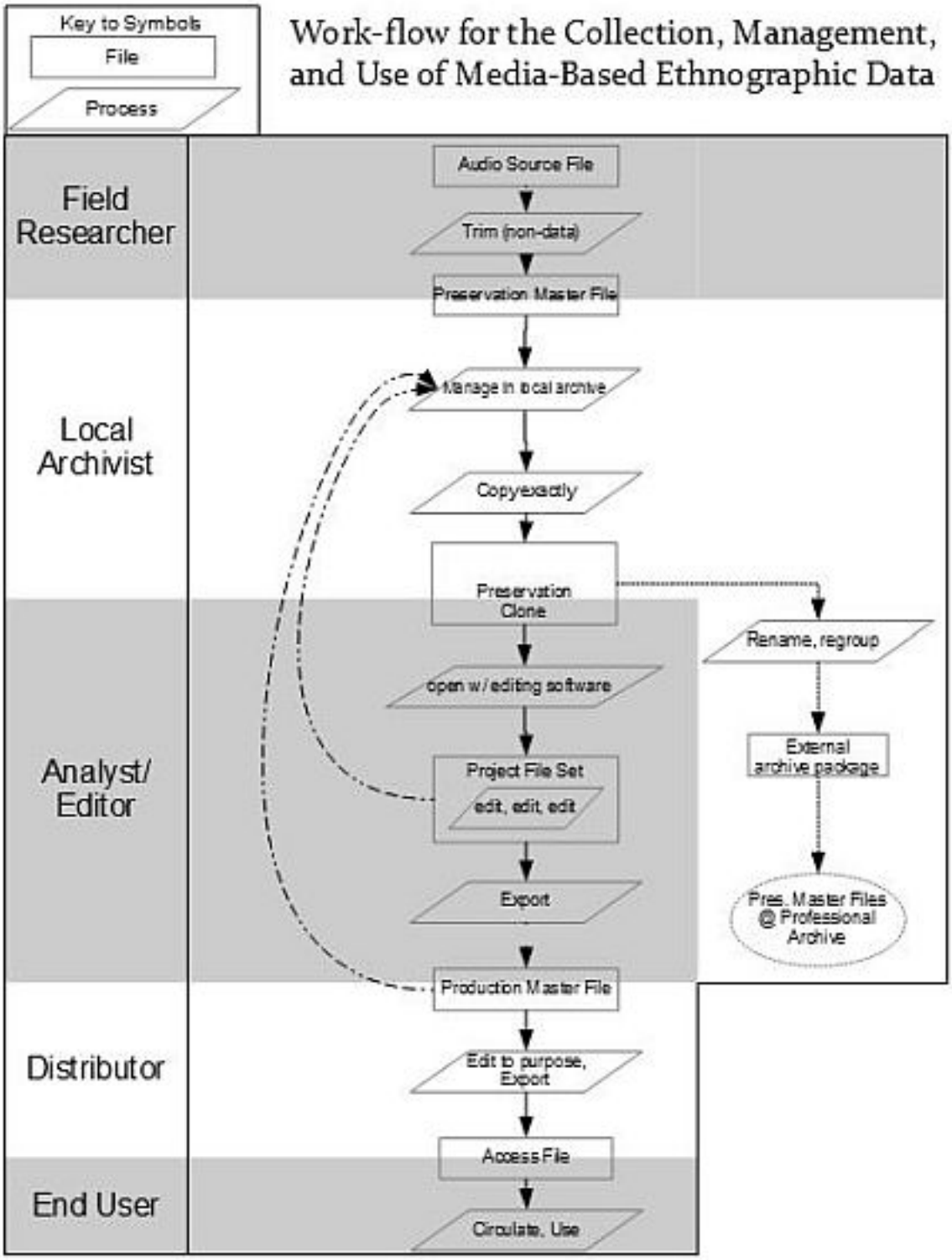
The recommended programs in this chapter by no means constitute an exhaustive list of software used in the documentation or descriptive process. In particular, if a documentary project or speech community members desire to make derivative products based on the documentary data such as storybooks, educational material, or informative webpages, they will want to explore more software options than those described in this manual.

Language and Culture Documentation as Digital Activity

In order to execute a successful documentation project the documentary linguist needs to integrate current attainable best practices from the outset. The documentary corpus needs to be accessible by multiple users through time, minimizing any obstacles to the files ongoing use. That means digital files must be open, have common encoding and markup, and be renderable by multiple software programs.

Workflow of data files

Each digital file that makes up a documentary corpus must travel along a path that starts at recording and journeys to its final destination which is an archive. The Workflow for the Collection, Management, and Use of Media-Based Ethnographic Data is an adaptation of digital file workflows found in *Sound Directions* (Casey and Gordon 2007) and *IASA-TC04* (Bradley 2009). These books are roadmaps for the **analog-to-digital (A/D) conversion** of large-scale media collections, but our workflow chart, in contrast, describes the path for single born-digital files until their long-term archiving. This chart illustrates a thorough treatment of the entire documentation digital process, but it is not a definitive, detailed description. Each documentation project will find that certain details need to be inserted according to the project's specific goals and its chosen archive's requirements. Also, while the original workflows anticipated a team of full-time recording professionals and archivists, it is entirely likely that the four roles spelled out in the chart—Field Researcher, Local Archivist, Analyst/Editor, and Distributor—will have only *you* to fulfill them! Whatever the final digital path for a given project, it should keep in mind the needs of the responsibilities and goals of the roles that are beyond this workflow, that is, the Archivist, the Analyst, and the (derivative) Product Distributor.


Key to Symbols
File
Process
Work-flow for the Collection, Management, and Use of Media-Based Ethnographic Data
Field Researcher
Local Archivist
Analyst/ Editor
Distributor
End User
Audio Source File
Trim (non-data)
Preservation Master File
Manage in local archive
Copy exactly
Preservation Clone
open w/ editing software
Project File Set
edit, edit, edit
Export
Production Master File
Edit to purpose, Export
Access File
Circulate, Use
Rename, regroup
External archive package
Pres. Master Files @ Professional Archive

With that background, let's take a look, step-by-step, at the workflow through the lens of each role in the digital process.

Field Researcher. The field researcher records a communicative event and then processes the resulting source files to be managed in the local archive. Readers may be familiar with the process that recording studios practice with artistic recordings. Artistic recorders labor to enhance their source files with filters, signal **compression**, varied **amplitude**, etc. in order to achieve a desired artistic quality in the recording. The documentary fieldworker has a different objective, and because of this must practice a different process.

Documentation emphasizes the need to preserve sound and images in an unedited form. Even so, some editing must take place. First, most unstaged communicative events have **non-meaningful silence** and noise which should be edited out, or "trimmed". Second, many digital recorders, when set for monophonic recording, will send the mono signal to both channels of a stereo file; this is just the result of how the device is hard-wired to record to mono. Such files should be processed to eliminate one of the redundant channels. Third, the field researcher should rename all files as soon as possible according to the professional archive's specifications, if known. These three things are generally the limit of the "post-processing" that a documentary data file should receive. A smart field researcher considers these steps when planning the field recording schedule.

Local Archivist. The field researcher's recordings of a communicative event, once trimmed and renamed, together with their metadata and annotations are referred to as a **bundle** (or **session**). The local archivist is steward of all bundles.

The master files he receives are never edited, and copied as seldom as possible. Any copy is made exactly, hence called a **Preservation Clone**. The cloning process is managed and possibly automated by the local archivist. Any party wishing to work further with the data starts from a "preservation clone'.

The local archivist ensures the data stays intact and usable, and that data is not distributed in a manner contrary to any rights, permissions, or licensing agreements. The local archivist handles the renaming and regrouping of files that are to be submitted to a professional archive, applying the professional archive's specifications to preservation clone files.

Analyst/Editor. Anyone wanting to analyze the data, or wanting to add **annotations** to it, or wanting to edit the data for publishable products can request a preservation clone from the archivist. After the local archivist grants the distribution request, the analyst/editor can then work on the data with editing software.

Most media-editing software (such as Audacity™ or Sony Vegas™) create a **project file** set, which the analyst/editor may wish to save in case the project needs to be later re-purposed. Since these derived files correlate to language data–and may even generate fresh data–the local archivist may request/require the project file set be added to the local archive.

If the preservation clone files are edited to produce publishable materials for distribution, the final export from the editing software can be called a **Production Master File**. This is what is turned over to the Distributor. Since the production master file often falls in the category of language/culture data, the local archivist may request/require that it be submitted to the archive as well.

Distributor. A distributor makes files accessible for reasons other than long-term preservation. The production master files have a set number of channels and effects, but the distributor may choose to modify these files to make them accessible by the intended end user. This usually means converting the production master file to popular formats that do not meet archival standards.

Note that this workflow has two destinations to its journey. One is the professional archive; language and culture documentation is not accomplished until the data resides there. The other is (local) circulation and use, itself an iterative and developing process. Ongoing use of documentation resources by the speech community and by scholars, whether accessed from local copies or from a professional archive, is destination worth the journey! Employing the right software tools will aid in a successful journey and safe arrival.

Archival standard and other file formats

One section of Bird and Simons' article "Seven dimensions of portability for language documentation and description" addresses the formats of documentary digital files. Bird and Simons propose that recommendations for file format be given according to the type of file (audio, video image, text). This also means that the software chosen for documentation project must output files that satisfy these standards.

The principles of portability, however, do not necessarily apply equally to all digital products of a documentation project. Not all the digital files are destined for long-term preservation. All the various files and copies of files produced during the documentation process fall into three forms. Each form has certain acceptable formats, indicated by the **file extension**. The table below lists the three forms and gives examples for recommended formats for each type of file.

Recommended Formats by Type

Copy	Example File Formats	Type
Source File	PHOTO: TIFF, BMP, JPEG VIDEO: discretionary (MPEG2 is recommended by some archives) AUDIO: **WAV**, 24 bit/48kHz TEXT: TXT (with Unicode character encoding) or .pdf METADATA (and other structured text): XML files	Archival / Working
Preservation Master File	same as Source File	Archival / Working
Preservation Clone	same as Source File	Archival / Working
External Archive Package	to external archive's specifications	Archival
Project File Set	Proprietary	Working
Production Master File	as required by the project	Working
Access File	as needed by end user	Presentation

Archival form. The archival form is the form in which data is stored for long-term access. It requires

the highest resolution of the data and should be the format in which the file was originally created. Unlike popular formats, such as MP3 for audio files, the archival must never **compress** data in order to make file size smaller.

The archival files must be in **open source** formats. That is, the file format's specifications are published and non-proprietary. Also, archival quality prefers file formats that are supported by software tools available from multiple suppliers, can be read with free tools, and if they are proprietary, have published formats, such as Adobe Portable Document Format (PDF), over those with secret proprietary formats, such as Microsoft formats.

Any text or programming characters used in archival quality files must be encoded with **Unicode**. However, if non-Unicode encodings are used, they must be documented fully in the metadata along with any scheme used to transliterate characters. Most modern language software outputs Unicode characters, but this should be double checked.

> UNICODE
>
> Unicode is a computing standard for consistent character encoding. It defines the way that computers should handle various alphabets and writing systems. Before Unicode was established, hundreds of different encoding systems were employed. Each assigned its own code to represent the different written characters on the computer screen. The European Union alone required several different encoding systems to represent all its writing systems. Any two encoding systems might use the same code number for different characters, or use different code numbers for the same character. Unicode provides a unique encoding number for every unique character, no matter what the computer platform, no matter what the software program, no matter what the language.

Markup language refers to the underlying computer code that communicates to the software how it should read and implement the file. It is important to consider what markup language you use, not only so that your corpus can be used by the greatest number of software programs, but also so that it can be understood and used even if the appropriate software is unavailable or obsolete. For that reason, descriptive markup is preferred over presentational markup. Descriptive markup is used to label parts of a file, rather than specifically instruct the computer how to process or present it. In language and culture documentation, the preferred descriptive markup is the very common **XML** schema. XML files can be verified by a Document Type Definition (DTD). Software that output XML should have a published DTD, which can be referenced in the corpus introduction. Generally, naming all software used in the documentation project should provide enough information for future users to find the software's DTD and render the file correctly. If you are computer-savvy enough to create a DTD schema that is not a previously archived standard, be sure to include the DTD as part of the archived corpus. Alternatively, if a descriptive markup scheme other than XML is used, prepare and archive a document that explains the markup scheme. Furthermore, whenever a resource using descriptive markup is archived, reference the resource to the archived version of the definition of the associated markup format. Fortunately, the advantage of descriptive markup is that blocks of information are tagged in such a way that the computer can be programmed to deal with them just

like presentational formats, and because the tags are in English and are specific to the type of data in the file, any literate person with or without a computer science background can make sense of the markup by just looking at the underlying code.

Example XML markup:

<olac xmlns=http://www.language-archives.org/OLAC/0.4>

<creator>Michael Scott</creator>

<format>WAV</format>

<language>English</language>

<contributor>Jonny Yorky</contributor>

All these points can be summarized to one easily memorized rule:

*Archival quality formats must ensure long-term preservation by making **LOTS**.* That is, the software produces formats that are Lossless, Open, Transparent, and Supported by multiple suppliers (Simons 2004).

Working form. The working form is the form in which data is stored while being annotated or edited. Working forms of data can include accompanying files specific to the software, not all of which may be preserved later. For example, Elan, Audacity™, and Flex may create ancillary files (with file extensions such as .typ, .prj, etc.) that are only necessary while working in that program. Unlike video or text files, which must undergo compression, audio and still photo file formats are generally the same as archival quality formats.

Presentation form. The presentation form is the file formats in which data is presented to the public. It is derived from working or archival forms and is often a proprietary format. The data is usually compressed and arranged in a way that makes it easier to deliver and interpret. The presentation form usually falls short of archival standards because it is compressed, reduced, or proprietary, and cannot be expected to endure. Presentation forms are chosen for some other advantage. They can be downloaded more quickly, played easily by mp3 players, or used effectively for CD distribution, radio broadcast, book or article publication, slideshow presentation, and so forth.

Guidelines for selecting software

In 2000, a five-year collaborative project called the Electronic Metastructure for Endangered Languages Data (E-MELD) began to promote common standards for digital language data. One result of that project was E-MELD's online School of Best Practices, which promotes best practices in digital language documentation that have been reviewed by documentary linguists, archivists and language engineers. The School of Best Practices' Toolroom contains a small database of software tools that had been used and reviewed by linguists at the time the project ended. Viewing the database today, the ephemeral nature of software and software recommendations is quickly evident.

However, the School also provided guidelines for choosing software and these guidelines remain valid.

- Prefer software that outputs formats supported by software tools that are available from multiple suppliers (i.e. many different software programs can open and read the output).
- Use software that exports documentation and description in formats that are open source (i.e. whose specifications are published and non-proprietary).
- If you must use proprietary software, prefer software that outputs published proprietary formats, e.g. Adobe Portable Document Format (PDF), over non-published proprietary formats (e.g. most formats produced by Microsoft products).
- Where possible (without sacrificing quality), prefer tools which are free or low-cost over those from commercial suppliers, so that high cost will not prohibit users from accessing the data.
- Prefer software that generates XML (with an accompanying DTD or Schema) over other schemes of descriptive markup.

To these guidelines we add one more:

- Use software that generates formats which do not compress the data in order to make file size smaller, i.e. Waveform Audio File (WAV) instead of MPEG-1 Audio Layer3 (MP3).

All these guidelines can be summarized to one easily memorized rule: Look for software that will ensure long-term preservation by making LOTS. That is, the software produces formats that are Lossless, Open, Transparent, and Supported by multiple suppliers (Simons 2004).

As further guidance, below are some questions to consider when selecting software. Fortunately, you do not necessarily need to understand all the concepts in order to find the answers. Online forums or friends with technological knowledge may be able to guide you to the best software as you think through the issues raised in these questions.

Will this program run on my computer? Every software program has its own recommended operating system configuration. Check your computer's specifications against the software's requirements.

Do I have special computing needs that will limit my software choices (e.g. limited grid power, primarily working offline)? It may be necessary to download a program and test it with extensive data or large files in order to know how much power it requires or whether its offline version offers full features. If your electrical power will be generated primarily by batteries, you want to avoid programs that eat a lot of power. For example, although FLEx provides powerful support for creating written annotations at various levels and in multiple languages, its processing size makes it a poor choice for low-power or low-memory devices. If power is limited, SayMore may be a better choice for simple transcriptions and/or translations.

What do I need in order to manipulate the data (e.g. selecting a specific subset of data, searching for specific words, comparison with data from another source, etc.)? Many software programs support search or filtering features in some limited form, but few, if any, have sophisticated capabilities in all these areas.

What is the software's primary storage format? Some programs store the data in its own database (e.g. FLEx). Others link to the computers files (e.g. ELAN) or folder structure (SayMore). The user should be familiar with the software's saving and backup procedures and prefer a program with simple procedures.

What formats does the software output? The software should have the ability to export plain text data with XML markup. Options for audio should include non-compressed formats such as WAV.

What character formats does it support? The software should provide Unicode (UTF-8) support.

What do I need in order to analyze and process the data in my situation (multiple translation languages, multiple orthographic transcriptions, network collaboration, word-for-word interlinearization, video subtitles)? The program with the most sophisticated features may not be best for your goals. Some programs may have very developed features in one area and limited tools for another. For example, SayMore offers wonderful tools for oral annotation, but none for creating morpheme glossing. Some software programs are suitable for the lone, offline researcher. Others encourage collaborative work.

Will I be teaching others to use this software? The appropriateness of a software tool for any given situation depends not only on the program's capability to accomplish all necessary tasks but also on its "user-friendliness." Some programs (e.g. ELAN) have a steep learning curve. Learning curves can cause frustration for new user but the programs should not be discouragingly daunting to learners with low computer literacy. Some programs present a low to moderate learning curve, but have fewer or less powerful features, which may lead to downloading and learning more programs.

Core software

The software programs widely recommended for language and culture documentation are SayMore, EUDICO Linguistic Annotator (ELAN), Audacity™, and Fieldworks Language Explorer (FLEx). SayMore is designed for compiling and managing a documentary corpus, including metadata records, and for preparing the corpus for archiving. SayMore and Audacity™ are recommended for oral annotations (commenting). SayMore and ELAN support written annotations (commenting). Finally, ELAN and FLEx are excellent tools for taking documentary data to the next step and beginning descriptive analysis. We believe these four programs should form the core of any documentary linguist's software toolkit. You will find yourself returning to them again and again as your own documentation projects progress. These four programs have proven themselves on the field for many years. For this reason, users can realistically depend on continued technical support and regular updated versions.

SayMore

SayMore is the only software program that supports all the core components of language documentation (compiling, commenting, metadata, archiving), if only at a basic level. In addition, SayMore offers at least two things that the other programs do not: database management and a metadata creator that are designed specifically for language documentation.

Download from	software.sil.org[8]
Released by	SIL International. Distributed under Academic Free License[9].
Platforms	Windows only
Description	SayMore makes common language and culture documentation tasks simple and keeps you productive by gathering recordings directly off your camera or audio device, creating folders for each documentary **bundle**, and allowing you to add annotations, enter metadata, record information about participants, and track informed consents. Files created in SayMore output in XML format.
Support & Training	Subscribe to SayMore Google Group and watch SayMore developer John Hatton's presentation. Links to both can be found on the SayMore's website. For training, read Sarah Moeller's (2014) review in *Language Documentation and Conservation*[10], or use the SayMore Tutorial[11] or Canada Institute of Linguistics' (CanIL) tutorial videos[12] (these videos are for an early, beta version, but are still useful) or Dr. Andrea Berez' demo video[13].
Recommended Tasks	All documentation tasks at a basic level, including time-aligned written transcription into one writing system and written translation into one language. Especially recommended for database design and management and metadata.
Pros	Organizes the corpus around bundles (documentary sessions). Automatically renames files, reducing broken links due to typos. Tracks corpus progress, session status, and workflow stages. Exports data in formats supported by ELAN, FLEx, Audacity™, and others. Easy-to-use and attractive interface.
Cons	Does not directly support network collaboration. Does not make clear to new users where data is stored on the computer and how to access and manipulate the files without SayMore. Archiving packages available only for two archives (and affiliates). Status markers and progress charts are designed for BOLD and may not be convenient for other language and culture documentation methodologies.
Learning Curve	Low to Moderate

ELAN

ELAN is notable for its sophisticated annotation features. Also, if an event was recorded with more than one audio or video recorder, ELAN allows you to link up to four media files and work with

[8] https://software.sil.org/saymore/
[9] www.opensource.org/licenses/afl-3.0.php
[10] https://hdl.handle.net/10125/4610
[11] drive.google.com/folderview?id%3D0B2TFy02zqdRwOE9sdEZaNFlpcHM%26usp%3Dsharing
[12] tutorials.canil.ca/?page_id%3D162
[13] http://sustainableheritagenetwork.org/digital-heritage/demonstration-saymore-language-documentation-software-tutorial

them at the same time. Unfortunately, ELAN is not known for having an intuitive user interface but, as testimony to its popularity, user-created ELAN video tutorials and written user guides abound online. Still, it is worth the investment, if feasible, to find and attend an ELAN workshop.

Download from	tla.mpi.nl[14]
Released by	Max Planck Institute for Psycholinguistics and The Language Archive, Nijmegen, The Netherlands
Platforms	Windows, Mac OS X, and Linux
Description	ELAN is a standalone tool for complex annotations on video and audio streams widely used for multimedia annotation. It was designed for the creation of text annotations for audio and video files of language use. Annotations are stored in ELAN's own open EAF format (XML).
Support & Training	An extensive manual and shorter user guide is available on the website. Web searches turn up many user-created helps.
Recommended Tasks	ELAN is essential for creating time-aligned written annotations, especially if more than one orthographic convention or more than one target translation language are needed. It allows any number of annotations levels, even tiers to describe hand motions or facial expressions. ELAN can also be used for descriptive analysis.
Pros	Annotations are grouped on an unlimited number of hierarchically organized tiers that can be independent, aligned, or embedded in relation to each other. Video files can be linked with an audio recording and its transcription.
Cons	Not easy to learn its most powerful features. It can be used for creating morpheme glossing or word-for-word interlinearization, but the setup is discouragingly time-consuming and complicated compared to FLEx.
Learning Curve	Steep

Audacity™

Even though a documentary linguist will find that much audio work can be done in SayMore and ELAN, anyone working with audio files should include Audacity™ in his or her software toolkit. Its user guides are thorough and very helpful. With the possible exception of trimming noise, if Audacity™ is used to edit original documentary recordings, the original, unedited version should be kept and deposited in the archive.

[14] http://tla.mpi.nl/tools/tla-tools/elan

Download from	audacity.sourceforge.net[15]
Released by	Open source. Distributed under the GNU General Public License[16]
Platform(s)	Windows, Mac OS X, Linux
Description	Audacity™ is a free, open source, cross-platform software for recording and editing sounds. Import sound files, edit them, and combine them with other files or new recordings. Export your recordings in many different file formats.
Support & Training	audacity.sourceforge.net/help/documentation[17]
Recommended Tasks	Any audio editing - trimming noise from audio files before exporting to SayMore or ELAN. Pre-segmenting files for oral annotations (see Chapter 12), converting files to smaller MP3 files for CDs or websites, and so on.
Pros	Able to reduce stereo WAV files to mono. Its "Edit Decision List" leaves original files unchanged, making editing very safe. Exports MP3 files for presentation copies. Allows track labeling and keeps tracks and labels synchronized.
Cons	Requires additional download of free LAME Encoder to convert to MP3.
Learning Curve	Low to Moderate

FLEx (Fieldworks Language Explorer)

Strictly speaking, FieldWorks Language Explorer (or FLEx, for short), is not a tool for language documentation. It was designed to assist language description. However, since some linguistic analysis increases the value of language and culture documentation data, and since FLEx makes basic analysis and dictionary building fairly easy, it is an essential tool for any linguist.

Download from	software.sil.org[18]
Released by	SIL International
Platforms	Windows, Linux
Description	FLEx is designed to help field linguists perform many common language documentation and analysis tasks, such as recording lexical information, creating dictionaries, interlinearizing texts, analyzing discourse features, and studying morphology.

[15] audacity.sourceforge.net
[16] audacity.sourceforge.net/about/license
[17] audacity.sourceforge.net/help/documentation
[18] https://software.sil.org/fieldworks/

Support & Training	Subscribe to FLEx Google group (in English and French) and watch demo clips. Both are available on the website. Training guides are available in the Help menu.
Recommended Tasks	Text interlinearization, morpheme analysis, and lexicon building.
Pros	Allows bulk edit of entries. Customizable fields. Can be used with Phonology Assistant (see "Additional Tools" on website) for phonological analysis. Supports some network collaboration.
Cons	Limited number of tiers. Does not provide tiers for phonological information.
Learning Curve	Moderate to steep

A Word about video software

Unfortunately, the current world of video editing has no equivalent to the free, open source Audacity™. When choosing a video processing software program, each documentation team will have to examine its budget, evaluate the team members' previous experience with video, and list what equipment will be available for running the software. This section is meant to help novices orient themselves among the many possible choices.

Inexperienced video editors may want to start with user-friendly, point-and-click interfaces, such as Windows Live Movie Maker for Windows or iMovie for Mac. For the more advanced user, there are more complex (and more expensive) software programs, such as Sony Vegas™, Final Cut Pro, or Adobe Premiere.

For those on a shoestring budget, Ubuntu Studio for Linux is a user-friendly option. Its learning curve is moderate to steep. Another free option with more powerful features but a steeper learning curve is Lightworks (for all platforms).

Since video (unlike audio) should not be archived in its raw format, video conversion and transcoding (see inset below) support is important to have. Most large suites like Sony Vegas™, Final Cut Pro, and Adobe Premiere include conversion and transcoding software. If you prefer free software, Handbrake is a reliable program. However, when it comes to licensing issues some free programs tread in a gray ethical area. Specifically, be wary of free programs using the h264 codec; this codec is licensed and really should not be available for free. It is also worth noting that free transcoding programs may be bundled with malware. Read the reviews, test the program before you commit yourself, if possible, and *caveat emptor* (Latin for 'Let the buyer beware!').

CONVERSION & TRANSCODING

To find out what transcoding is and why it is important for video processing, start by Googling "video

transcoding" and "video transcoding software", or read this website: http://www.jwplayer.com/blog/an-overview-of-audio-and-video-transcoding/ to learn more.

OPEN SOURCE TIP

For those who enjoy testing and finding just the right free video editing program, start by exploring this website: http://open-tube.com/10-awesome-free-and-open-source-video-editors/[a]. This website is aimed at Linux users, but several of the applications reviewed there run on other platforms as well.

[a] http://open-tube.com/10-awesome-free-and-open-source-video-editors/

Expanding your software toolkit

The chapter about recording equipment encourages aspiring fieldworkers to purchase a few essential pieces of equipment and then gradually enlarge their "field kit" according to their needs. In the same way, fieldworkers should download a combination of software tools, starting with those recommended in this chapter, and then gradually expand their software toolkit. Eventually, you may find certain programs are better suited to your workflow. This section includes only a few software programs as a place to begin exploring. As always, keep in mind that as technology constantly changes, the programs may soon be replaced by newer and brighter tools.

Starting points

- For editing and annotating the three data stream types—text, audio, video: Toolbox[19] by SIL International is an older linguistic analysis program that is still used by some linguists and archives.
- For file management: Total Commander[20]
- For metadata creation: ARBIL[21] by Max Planck Institute for Psycholinguistics
- For still photo editing: Irfanview[22] is good for batch processing several media types. Also, Microsoft Photo Tools 2[23] is for editing photo metadata; it is a proprietary Microsoft program, but it is also free.
- For text editing: any decent text editor such as Notebook or WordPad will do nicely.

[19] www-01.sil.org/computing/toolbox/
[20] www.ghisler.com
[21] tla.mpi.nl/tools/tla-tools/arbil
[22] www.irfanview.com
[23] www.microsoft.com/en-us/download/details.aspx?id%3D13518

- For archiving and archiving preparation: LAMUS[24] by Max Planck Institute for Psycholinguistics is designed for IMDI standard metadata only (see Chapter 2). OpenOffice.org[25] has wonderful, free, and open-source tools that can be used for creating metadata spreadsheets in CSV format, tables of contents, and corpus introduction documents.

Exploring Further

- The journal *Language Documentation & Conservation*[26] (LD&C) regularly publishes technology reviews.
- Stanford University recommends this link[27] at SourceForge.net for linguistic fieldwork tools.
- For a brief description of all software released by Max Planck Institute (MPI) for Psycholinguistics, which developed ELAN, see section 7 of Daan Broeder, Han Sloetjes, Paul Trilsbeek, Dieter van Uytvanck, Menzo Windhouwer and Peter Wittenburg's article[28] in *Documenting Endangered Languages: Achievements and Perspectives*.
- The Language Archives at MPI for Psycholinguistics' has a webpage[29] of language software tools, including many developed at MPI.
- SIL International[30] provides links to their computing resources and projects.
- E-MELD's website[31] provides a "small database of software tools that have been used, classified and reviewed by linguists" that was last updated in 2005, but there are no dates on the comments.

Chapter Questions

1. Describe four guidelines for selecting software tools.
2. When selecting software, what questions should you ask yourself? Why are these important?
3. Do some research on a software program of your choice. List 1-3 pros and cons. What tasks of language documentation or context of research would this program best suit?

[24] tla.mpi.nl/tools/tla-tools/lamus
[25] www.openoffice.org
[26] nflrc.hawaii.edu/ldc
[27] fieldling.sourceforge.net
[28] pubman.mpdl.mpg.de/pubman/item/escidoc:1263591:4/component/escidoc:1448958/broeder_ch3_2011.pdf
[29] tla.mpi.nl/tools
[30] software.sil.org
[31] emeld.org/school/toolroom/software/index.cfm