The Besemah Language Documentation Project

Bradley J. McDonnell

September 15, 2012

1 Introduction

The aim of the Besemah Language Documentation Project (BLDP) is to create a comprehensive record of the Besemah language as it is currently spoken in as natural a setting as possible, utilizing current methods in the field of language documentation (Himmelmann, 2006; Woodbury, 2011; Bird & Simons, 2003; Thieberger & Berez, 2012). BLDP currently has a fairly large corpus primarily made up of dialogic spontaneous speech (conversation), but also including traditional and modern narratives, elicited wordlists and sentences, and lexicographical materials. All of these data are enriched with metadata, which include information on the setting, speakers, genre, etc. Other non-linguistic materials that enrich the corpus are photographs and video that show the contexts within which the data was collected.

This report outlines the theory and practice that serves as the foundation of the BLDP, which includes a description of the procedures and practices that I have undertaken (and will continue to undertake) in recording, transcribing, and archiving Besemah language data. The report begins with an introduction to the language itself, describing its linguistic affiliation, environment, previous research, and level of endangerment. The following section outlines the procedures in creating "a lasting, multipurpose record of the language" (Himmelmann, 2006, p. 1), including methodologies for naming and storing files and a principled workflow for collecting, annotating, and archiving linguistic data. This section also includes a rational for choosing the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC; `http://www.paradisec.org.au`). The final section describes some of the intended outcomes of the BLDP.

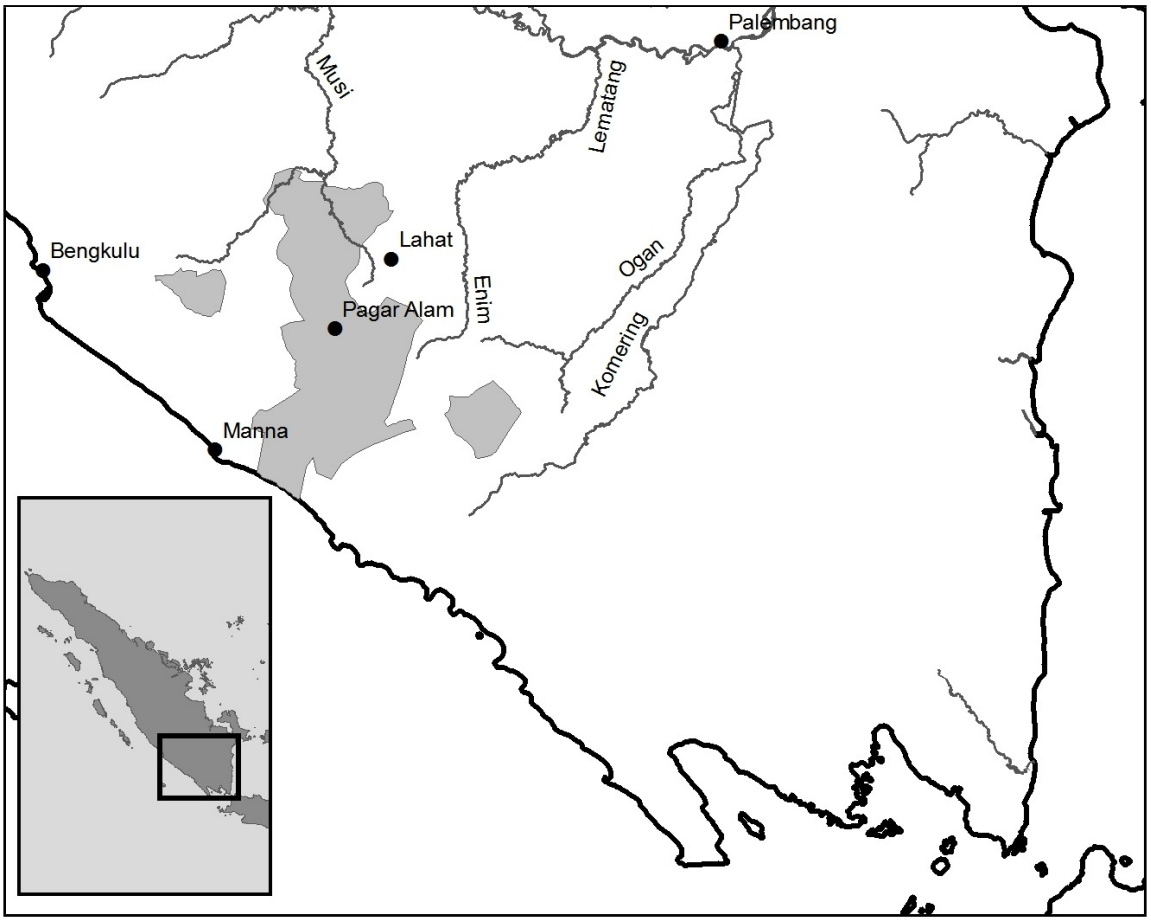
I hope that this report will serve as a roadmap to help others (now and in the future) to navigate the archived materials and see the processes that I took when compiling the material in the archive. Aside from the primary purpose of describing the BLDP, this report will hopefully demonstrate how current best practices in language documentation can be implemented in a language documentation project.

1.1 The Language

Besemah (alternatively, Pasemah) is a little-known Malayic language in the Western Malayo-Polynesian branch of the Austronesian language family. It is spoken by approximately 400,000 people primarily in the highlands, but sporadically throughout the lowlands of

southwest Sumatra (see Figure 1), straddling both South Sumatra and Bengkulu provinces (McDowell, 2007; McDowell & Anderbeck, 2007). Besemah is considered to be a part of a cluster of Malayic isolects that roughly cover the southern half of Bengkulu province as well as the western highlands of South Sumatra province, traditionally referred to as the Middle Malay or Central Malay languages (Brandes, 1884; Voorhoeve, 1955; Adelaar, 1992; McDowell, 2007; McDowell & Anderbeck, 2007; Lewis, 2009; McDonnell, 2009). However, as I have argued in McDonnell (2010), these names are misleading as this cluster of related languages is the southernmost of the Malayic languages of Sumatra and is located on the western half of the island. For these reasons, I propose Southwest Sumatran Malay as the most fitting title for the cluster of Malayic languages that are spoken in southwest Sumatra.

Figure 1: The Besemah Language Map



According to Adelaar & Prentice (1996) and Adelaar (2005), Besemah can be additionally classified along sociolinguistic lines as a vernacular Malay as it is spoken by a traditionally Malay speech community. This distinguishes Besemah from literary historic varieties, such as Old Malay and Classical Malay, and Pigeon-Derived Malay varieties, such as Ambon Malay and Sri Lankan Malay. Furthermore, Besemah is spoken by a rather homogenous community, which makes Besemah somewhat different from Malay-Indonesian varieties such as Palembang Indonesian (see below), Riau Indonesian (Gil, 1994), and Jakarta Indonesian

(Wouk, 1989, 1999). However, like these Malay-Indonesian varieties, Besemah is diglossic or even what might be called polyglossic with Standard Indonesian and Palembang Indonesian, the language of wider communication in South Sumatra. Standard Indonesian is used in all formal situations, including speeches at weddings, funerals, and other cultural events, such as Friday sermons at the mosque. Palembang Indonesian is a koine that came out of the Malay spoken in Palembang, the capital of South Sumatra. Palembang Indonesian is used in the city of Palembang and its subsidiary towns and cities throughout South Sumatra where ethnically Malay, Javanese, Minangkabau, Batak, and Chinese Indonesians congregate. Palembang Indonesian is therefore the medium that Besemah speakers use in interethnic communication. Finally, Besemah is used between Besemah speakers in the home and in everyday village life.

1.2 Level of endangerment

Besemah does not clearly fit into any of the current classifications of language endangerment. On the surface, Besemah might be considered to be a vital language because (1) children are still actively learning the language and (2) there is still a relatively large number of speakers (approximately 400,000) using the language. These are both very good signs for Besemah, but there are a number of other factors that show that Besemah may not be as stable as one might hope. These include various factors, such as: contact languages and multilingualism, language attitudes, and a lack of formal education and standard orthography. Besemah speakers are inundated with other varieties of Malay-Indonesian. As discussed in the previous section, Besemah speakers use Palembang Malay and Standard Indonesian, but also hear other varieties on a daily basis, such as Jakarta Indonesian from the media and closely related Malayic varieties in neighboring communities (i.e., Lintang). Additionally, Besemah shows low prestige and is commonly seen as an uneducated and parochial variety of Malay. This is in direct opposition to educated languages like Standard Indonesian or even English and cosmopolitan varieties like Palembang Malay or Jakarta Indonesian. Although Besemah speakers see some ancestral value in their language, there is no modern value in the language for moving ahead in a globalizing world. Finally, formal education from preschool to high school is conducted entirely in Standard Indonesian, even though it is common to hear teachers and students using Besemah in and out of the classroom. The Indonesian government does allow local languages in the classroom, but this is not employed in the schools of the Besemah highlands.

The effects of these factors are not equally distributed among Besemah speakers. While speaker over sixty years of age are for the most part monolingual, younger speakers under the age of thirty show a high level of bilingualism in Standard Indonesian. This bilingualism and exposure to other Malayic varieties has had various effects on the language. One clear example from McDonnell (2009) is the difference in the vowel system of Besemah speakers of different ages. That is, younger speakers are able to distinguish six vowels that are present in other varieties of Malay-Indonesian, while older speakers can only distinguish the four vowels of Besemah. From my own observations and a cursory look at recorded conversations, there also appear to be other differences in the lexicon and grammar of older and younger speakers.

There is no doubt that these differences grow out of the fact that the Besemah highlands have drastically changed over the last forty years. Elizabeth Fuller Collins accompanied her

husband, anthropologist, William Collins from 1971 to 1973 to the Besemah highlands. In (Collins, 2007, p. 7), she describes the Besemah speaking region as follows.

> We settled on the Pasemah Plateau, a fertile plain below the majestic volcano Gunung Dempo in South Sumatra. At that time there was no electricity or running water, no newspapers, and only one telephone at the post office in the market town of Pagaralam. On our first trip to the highlands cars and trucks had to travel in convoys so one vehicle could be used to help haul another through places where the road had deteriorated to a muddy swamp.

When my own fieldwork began in January 2008, I found a much different situation. Most homes have electricity (albeit somewhat inconsistent), television, gas stoves, and at least one cell phone per house and sometimes more. Cars and trucks travel quickly from Palembang to Bengkulu on the paved single lane roads. Although the Besemah region is still considered somewhat remote, access to outsiders and access to major cities has increased. With these changes in mind, the level of endangerment for Besemah is unclear, but it would be safe to say that it is at least a threatened language. More research on the level of language endangerment in these situation is needed.

2 History of the documentation of Besemah

2.1 Previous research

Besemah, like many of the Malayic varieties of western Indonesia, has received very little attention from linguists; there are virtually no recent publications on Besemah grammar, and McDonnell (2009) represents the only recent study on Besemah phonology. The most significant work on Besemah is the dictionary, short grammar sketch, and texts by the late Dutch government linguist O.L. Helfrich (1904; 1915; 1921; 1927; 1933; **?**). There have also been two surveys on South Sumatran Malay, including both Southwest Sumatran Malay and Musi Malay, which includes the Malayic languages surrounding Palembang in the Musi River basin by Mitani (1980) and more recently by McDowell (2007); McDowell & Anderbeck (2007). Other anthropological work on Besemah has been conducted by Collins (1979; 1998).

Current research is also being collected by the Padang Field Station of the Max Planck Institute for Evolutionary Anthropology (Gil & Litamahuputty, in preparation). Although this work is not part of the BLDP, I am collaborating with the Padang Field Station for the mutual benefit of both projects.

2.2 Besemah corpus

The Besemah corpus includes recorded conversations, narratives, songs and elicited wordlists and sentences all recorded from January 2008 to June 2010. The majority of the corpus is made up of a naturalistic conversations that amount to well over eight hours of recording, four hours of which is transcribed and glossed in both ELAN (EUDICO Linguistic Annotator; `http://www.lat-mpi.eu/tools/elan`) and the Field Linguist's Toolbox (`http://www.sil.org/computing/toolbox/`). The second largest portion of the corpus is the collection of

narratives that include twenty narratives that range from three to ten minutes a piece. These narratives primarily fall under the category of 'fairy tale' or *andai-andai* in Besemah. All of the recordings have participants ranging from 19 years old to approximately 70 years old with varying levels of education and bilingualism in Standard Indonesian. All participants are members of Karang Tanding or villages in close proximity to it. While all participants are native speakers of Besemah, approximately half have never left the Besemah speaking region, while the other half have spent months and in some cases years working or attending university classes outside of the Besemah region.

3 Language documentation methods

This section outlines the methodology that I employ from the time that I ask speakers to record until I deposit the materials in the archive. Before embarking on issues of workflow, metadata, and file naming, I lay the foundation to the methodology with an introduction to the archiving plan.

3.1 The Archiving plan

Archiving is essential to any language documentation project. In fact, every definition of language documentation that I am aware of considers archiving to be a crucial component of the definition (Woodbury, 2011; Himmelmann, 2006). Conathan (2011), for example, states that "All documentation projects should have an archiving plan, and consider the long-term preservation of their records from the outset" (p. 235). Even though I failed to make such an archive plan at the outset of this project in January 2008, all of the Besemah data is archive-ready. This means the materials are in the appropriate formats (i.e., wave audiofiles, XML transcriptions) and have the appropriate metadata to be directly deposited in the archive. This section describes the archiving plan for the BLDP, outlining how the existing data is archived and how data in the future will be archived. Before outlining the procedures for archiving, I will provide my rational for choosing to deposit the Besemah data in PARADISEC.

PARADISEC is an archive that "offers a facility for digital conservation and access for endangered materials from the Pacific region, defined broadly to include Oceania and East and Southeast Asia" (`http://www.paradisec.org.au/`). The archive is run by a consortium of four Australian universities, including: the University of Sydney, the University of Melbourne, the University of New Castle, and the Australia National University. Out of the small handful of archives that I could choose to deposit the Besemah corpus, PARADISEC is particularly appropriate for its regional focus, which includes Southeast Asia. Even though this sets PARADISEC apart from other digital archives for linguistic data, there are a number of other reasons that this archive is the best choice for the Besemah data, which are outlined in the following sections.

3.1.1 Essential components of a digital archive

Simply put, "Archives maintain and provide access to records of enduring value"(Conathan, 2011, p. 236). Assuming that the linguistic data in the Besemah corpus have "enduring

value", the archive serves to ensure that the linguistic data still exists (in a readable, watchable, or hear-able format) in the near and distant future and, equally important, that the linguistic data is accessible to interested scholars and community members. In order to ensure that an archive can achieve these purposes, linguists, communities, and archives have developed several criteria that help to assess any digital archive. For example, Conathan (2011) lists the following seven core archival functions: appraisal, accession, arrangement, description, preservation, access and use. However, Chang (2010) has developed an extremely detailed and clear criteria that are conveniently grouped into four major categories and summarized in the acronym TAPS, which stands for Target, Access, Preservation, and Sustainability. Chang's (2010) TAPS checklist includes a detailed set of questions that help linguists and community members practically assess a digital archive. Since Chang (2010, p. 69-73) has already assessed PARADISEC according to her TAPS checklist, I only provide a short discussion of the four major categories of the TAPS checklist and demonstrate how the archive is the appropriate fit for the Besemah corpus.

Target According to Chang (2010, p. 82), "Target refers to the 'fit' of the archive with regard to the data to be deposited and the needs of the identified designated communities." For the BLDP, this means that the archive must fit both the Besemah data described in section 2.2 and the interests of the Besemah community that produced the data (and their decendants). Chang (2010) lists four criteria under this category, including: the mission statement, submission criteria, designated communities, and ongoing relationship. Based on these four criteria, PARADISEC is certainly suitable for the BLDP and vice-versa. PARADISEC's mission statement includes both a strong commitment to the community and a desire to follow "emerging international standards for digital archiving" (`http://www.paradisec.org.au/home.html`). This means that PARADISEC ensures that Besemah language data will be available for scholars and Besemah community members now and in the future. The submission criteria for PARADISEC specifies that (1) data should come from the Pacific region and (2) data is accepted in various formats based on their type (i.e., text, audio, video, and image). By and large, the existing Besemah data are already in these formats; audio files are in a wave file format and transcriptions are in an XML file format. It is important to note that these formats follow best practices in archiving linguistic data (Bird & Simons, 2003). For the BLDP, the designated communities most crucially include the academic community and the Besemah community, both broadly defined. PARADISEC is interested in providing access to both of these communities, but first targets the language community by "making field recordings available to those recorded and their descendants" (`http://www.paradisec.org.au/about.html`). Finally, PARADISEC is dedicated to an ongoing relationship with the community. In fact, the deposit form requests the contact information of community members (or other parties) who might have rights to the language material deposited (`http://www.paradisec.org.au/PDSCdeposit.pdf`).

Access According to Chang (2010, p. 87), "Access refers to the accessibility and usage of the data and corresponding metadata once materials are deposited." For the BLDP, access is an important issue because it is crucial that the data are not freely available anyone for any reason on the internet. At the same time, I do not want the data to be so restricted

that designated communities are not able to gain access to it. Chang (2010) includes the following criteria under this heading: discoverability, fixed identifiers, reach, and access restrictions. PARADISEC enjoys a high level of discoverability because it is a member of the Open Languages Archive Community (OLAC; `http://www.language-archives.org/`). OLAC is "an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources" (`http://www.language-archives.org/`). It is, therefore, possible for designated communities to search OLAC for materials that are deposited in PARADISEC. In turn, OLAC is discoverable through mainstream search engines (i.e., Google and Yahoo), which further increases the discoverability of PARADISEC records. Fixed identifiers concern the persistent citation of an electronic resource on the internet. Bird & Simons (2003, p.567) summarize the problem of not using fixed identifiers well.

> Often a language resource is available on the web, and it is convenient to identify the resource by means of its UNIFORM RESOURCE LOCATOR (URL) since this may offer the most convenient way to obtain the resource. However, URLs are notorious for their lack of persistence. They ?break? when the resource is moved or when some piece of the supporting infrastructure, such as a database server, ceases to work.

PARADISEC follows this best practice well, as fixed (unique) identifiers are assigned to each item by the curator upon deposit. Furthermore, "PARADISEC identifiers are never reassigned and location independent to facilitate persistence" (`http://paradisec.org.au/naming.html`). Reach is concerned with the ability of the community to obtain the materials from the corpus. PARADISEC is able to make materials available to communities in appropriate formats (i.e., in MP3 format over the internet or on a CD), which depends on the needs of the users (`http://www.paradisec.org.au/services.html`). Currently, many in the Besemah community do not have access to the internet and so may not be in the reach of PARADISEC. However, I expect that this will change in the next couple of decades, so this may not be an issue in the future. Furthermore, many of the Besemah diaspora in major cities, such as Palembang or Jakarta, would have internet access, so the archive may be able to reach them. Finally, access restrictions concerns the policies and procedures that the archive has in place to ensure that the wishes of the community and the linguist are honored. For PARADISEC, metadata are completely searchable through OLAC, but the deposited items can only be accessed by filling out a *Conditions of Access* form (`http://www.paradisec.org.au/PDSCaccess.rtf`). It is, however, the depositor who sets the access conditions on the depositor's form (`http://www.paradisec.org.au/PDSCdeposit.rtf`). For the BLDP, I have chose the option, "Access by permission of depositor only." This means that for someone to access the Besemah data, PARADISEC would need to obtain my permission before providing access to anyone. This allows me to still have control over the data and consult with the Besemah speakers who produced the data about sharing it with others.

Preservation According to Chang (2010, p.96), "Preservation refers to the overall system and technical structures of the archive that ensure materials will be managed in ways that make them available and usable, with their authenticity and integrity intact, far into the

future." This category includes: evidence of long-term planning, preservation strategies, integrity, and authenticity. Evidence of long-term planning is important for any archive, but the essence of Chang's criterion concerns long-term planning for *digital* data. According to (Chang, 2010, p.173), PARADISEC personnel have set up a system that (thus far) needs little maintenance, but the responsibility to maintain this system rests on the shoulders of a small group. However, what is crucially important here is the fact that the archive adopts best practices in preserving digital data. Preservation strategies ensure that "digital resources remain accessible to future generations" (Bird & Simons, 2003, p.567). Because PARADISEC has been using appropriate digital formats from the outset, it has not been necessary to counter problems such as obsolescence of hardware and software. Integrity ensures that all records are complete and unchanged. Chang (2010, p.173) states that PARADISEC uses checksums to maintain the integrity of their records. Finally, authenticity is the criterion that makes sure that every record is, in fact, what the metadata says it is. At the present time, PARADISEC puts its trust in the depositors to make certain that their data are reported accurately.

Sustainability According to Chang (2010, p. 104), "Sustainability refers to the demonstrated organizational robustness of the archive, lending long-range viability to the functions that it performs." This category includes the following criteria: adequate infrastructure, financial sustainability, disaster preparedness, and succession plan. Much like the previous category, a sustainable archive is one that has planned for as many foreseeable circumstances as possible. Chang (2010) demonstrates that PARADISEC demonstrates uncertainty when it comes to adequate infrastructure, financial sustainability, and a succession plan. These are areas that will hopefully improve for PARADISEC over time and, for the most part, is beyond the archive's control. The one area that PARADISEC is prepared for is disaster preparedness. The archive stores copies of the material at different location throughout Australia to safeguard against a natural disaster occurring in one part of the country.

3.2 Metadata

Good (2011) states "Metadata is an essential part of any documentary corpus, and a metadata plan forms an integral part of a general data plan." This section attempts to follow through with Good's suggestion. Metadata, according to Good (2011), is "information describing the constituent resources of a documentary corpus, including, for example, their content, creators, and access restrictions" (p. 226). Metadata for the the Besemah corpus is not merely shaped by me, the linguist; rather, it is shaped by institutions, participants, cultural and linguistic factors. First, the archive plays a major role in shaping the metadata and may in fact serve as the foundation for which categories are initially included in the metadata (Good, 2011, p. 226). Even though this was not the case for the collection of the Besemah metadata from January 2008 to June 2010, PARADISEC's required metadata categories have served as the base for which categories are added into the reformatted metadata that I compiled from April to June 2011. The second entity to shape the metadata are the speakers. For example, early in my fieldwork in the Besemah highlands, I noticed that some Besemah speakers who had moved to major cities such as Jakarta or Palembang tended to use lexical items from the local Malay-Indonesian isolects spoken in those cities. This is the

reason I added the category 'time out of the area' in the metadata. Another entity that has an effect on the metadata is the language itself. Besemah, for example, makes use of *teknonymy*, so that adults with children are rarely referred to by name. It is, therefore, useful to collect the names of the children of each speaker in order to understand the transcriptions. Until now, I have made note of speakers' children's names in the transcription, but now would like to put this in the metadata itself. There are, of course, other entities to shape the record, such as the genre of the recording, but I will not discuss those here. What is important is the recognition that multiple entities shape the metadata, so that metadata for different languages may look different.

For Besemah, metadata is organized into four major sets: recordings, transcripts, people, and archive. The metadata categories for each set are shown in Table 1 below.

Table 1: Metadata Categories

Recordings	**Transcripts**	**People**	**PARADISEC**
Date	Language	SpeakerID	Cassette length
Filename	ISO code	Role	Country
PDSC Filename	File type	Name	Data type
Language	Recording Date	FullName	Date
Language code	Export Date	Code	Dialect
Region/village	Filename	FamilySocialRole	Item
Genre	PDSC Filename	Languages.Description	Language as in source
hasTranscript	Transcript begin	EthnicGroup	Language code
Title	Transcript end	Age	language standard
Researcher	Transcription notes	BirthDate	Language subject
Recorder	Orthography	Sex	Language content
Speaker 1	Region/Village	Education	Media
Speaker 2	Transcribes	Anonymized	Notes
Speaker 3	Version	Contact.Name	Number of cassettes
Speaker 4	Researcher	Contact.Address	Orthography
Speaker 5	Transcription assistant	Contact.Email	Priority
Speaker 6	Transcription checker	Contact.Organisation	Region/village
Notes	Speaker 1	Description	Relation
Original Filename	Speaker 2	Nickname	Rights
(before 2011-05-01)	Speaker 3	Time.away	Role
	Speaker 4	Marital.status	Rtor time
	Speaker 5	Spouse	Track
	Speaker 6	Birthplace	Transcript
			Video length
			Date sent
			to PARADISEC

The metadata for the BLDP are stored in a working format, an archival format, and a presentation format. The working format (for the time being) is a spreadsheet file with three sheets that have the first three sets above on each sheet. Even though, it would be best to use

a relational database (Thieberger & Berez, 2012), it is simply more practical to organize the metadata in a spreadsheet. The archival format is produced by PARADISEC and stored on their servers. In fact, PARADISEC creates this format based on the spreadsheet that I send them. This is the metadata set in the last column of Table 1. In many ways, this format is also a presentation format, as this is format that is visible to the public when searching OLAC. The presentation format is stored in ARBIL (Archive Builder), a software program developed by the Max Planck Institute for Psycholinguistics (`http://www.lat-mpi.eu/tools/arbil`). Using the ISLE Meta Data Initiative (IMDI), ARBIL has the unique ability to be directly associated with ELAN transcription files. Even though I call this the presentation format, ARBIL files are also in an XML archival format. This makes Arbil somewhat redundant, but it does allow me to store a much more detailed level of metadata than the PARADISEC metadata files.

3.3 File naming

There are two file naming conventions that are used for the BLDP. The first convention is for non-archived files used by the researcher, while the second is for the archive and follows the archive's conventions. Both file naming conventions follow Thieberger & Berez (2012). Even though it is not ideal to have two different conventions for naming files, it is a preferred working format for me (as the researcher) to be able to easily access files. With this in mind, it is important when archiving the files to be disciplined to keep track of the different file names in the metadata.

3.3.1 Researcher file naming

The file names have three different component parts, specified in Table 2. Obligatory components are in brackets and optional components are in parenthesis. The first component is the ISO code as set by the Ethnologue (Lewis, 2009); for Besemah the ISO code is PSE. Using the ISO code (as opposed to my initials in the archive format) allows me easily identify the language and differentiate between different languages that I may be documenting. The second component part is the date in YYYYMMDD format. For audiovisual recordings and the ELAN/Toolbox formats that accompany them this is the date that the recording took place (and not the transcription date). However, in the case of the lexicon, this component specifies the date that the lexicon was last exported. It is important to keep this date the same for the recording and its transcription, so that they can easily be matched together. The third component of the file name is the discourse type, which is classified in one of four categories: Conversation (C), Narrative (N), Song (S), or Elicited (E). If more than one recording of the same discourse type is recorded on the same day, then an alphabetic character beginning with A follows the discourse type for each of recordings in the order in which they were recorded. Recordings have an additional (optional) component that specifies whether the recording is the 'original'. As will be explained in the next section, after I make a recording, I append the original recording with audio metadata. To ensure that there are no issues with this process, I think it best to save the original file. This is the file that is appended with the final O.

Table 2: Filenaming Conventions

Format	**Audio/Video**
Structure	[ISO Code]-[Date Created]-[Discourse Type](Order)-(Original)
Examples	PSE-20080520-C.wav
	PSE-20080520-C-O.wav
	PSE-20080520-N.wav
	PSE-20080520-S.wav
Format	**Transcription (ELAN/Toolbox)**
Structure	[ISO Code]-[Date Recorded]-[Discourse Type + (Order)]-[Order Exported]
Examples	PSE-20080520-C-EA.eaf
	PSE-20080520-C-TA.txt
Format	**Lexicon (FLeX)**
Structure	[ISO Code]-[Date Exported]-[Lexicon Label]
Examples	PSE-20080520-L
Format	**Fieldnotes (scanned)**
Structure	[ISO Code]-[Date Created]-[Fieldnote Label]-[Page Order]
Examples	PSE-20080520-FA
	PSE-20080520-FB

3.4 Archive filenaming

PARADISEC uses a persistent (and unique) identifier for each item. The file name consists of two primary components, the *Collection ID* and the *Item ID*. The collection ID is typically the depositor's initials followed by a number, while the Item ID is a three digit number that specifies the order in which the item was deposited. If an item has more than one component item, (i.e., a recording and its transcription), then the file name is followed by an alphabetic character beginning with A. Therefore the structure is as follows:

[Collection ID (*Depositor's initials/number*)] -[Item ID (*Order of deposit*)]-(Associated item)

For example, my first recording is named *BJM01-001-A.wav* and the ELAN transcription for this recording is *BJM01-001-B.eaf*. My second recording then is *BJM01-002-A.wav*. If this recording, for example, did not have any other associated item, then the file would be *BJM01-002.wav*

3.5 Workflow

This section outlines the workflow of the BLDP, which begins when I ask the speaker to record and ends when the recording, transcript, and and any other associated files are archived. Figure 2 illustrates a proposed workflow. It is important to note that this is not the exact workflow that I used to collect the data from January 2008 to June 2010, although it is

very similar. For instance, I did not photograph the setting (upper right-hand corner of the diagram) nor did I immediately archive my recordings after returning to a major city (left-hand side of the diagram). However, I followed almost all other activities as specified on the workflow.

The workflow diagrams specifies three different locations, specified in all caps and within brackets. These locations include: the field recording site, the field site and the research site. The field recording site is the house or space where the recording takes place, while the field site is the village where the research project takes place, which is the village of Karang Tanding for me. The research site refers to the home university, which for me is the University of California, Santa Barbara. The research site could also be a local university located in a major city.

The first step in the workflow is to ask speakers to record, which may happen at the field recording site or at the field site in general. This is when I set up the time and place of the recording as well as possible other participants. After meeting at the recording site, I ask for an informal consent to record and casually explain what I am recording and why. It is important for me not to obtain formal consent at this time as this could interfere with the naturalness of the recording. This is because Besemah speakers are diglossic and so once speakers sense a tinge of formality they are likely to switch to Standard Indonesian or at least intersperse Standard Indonesian throughout their speech. Again, due to diglossia, I leave the field recording site when recording, allowing for a much more natural setting. After a set time (usually an hour or two), I return to the field recording site and photograph the setting and each speaker, collect metadata from each speaker, and obtain formal consent from each speaker. When obtaining informed consent, I make sure that each speaker feels okay about the recording and ask if s/he would like me to delete the recording in its entirety or in part.

Once the recording is finished, I return to my village home and record the audio metadata, which includes the location, date, and the names of the speakers. I then make a copy of the recording and append the metadata to the beginning of the copied recording. I then save these recording to my laptop and make CDs of the recording for each speaker. At the same time, I listen to the recording and judge its authenticity and transcribabilty alongside the overall quality of the recording. I may consult with my transcription assistant to judge the authenticity of the recording. If the recording is subpar in any of these respects, I do not proceed, but do save the recording and input all of the metadata into the database. At my first opportunity to be in a major city, I send a DVD with all of the recordings to PARADISEC for archival and email them a copy of the metadata.

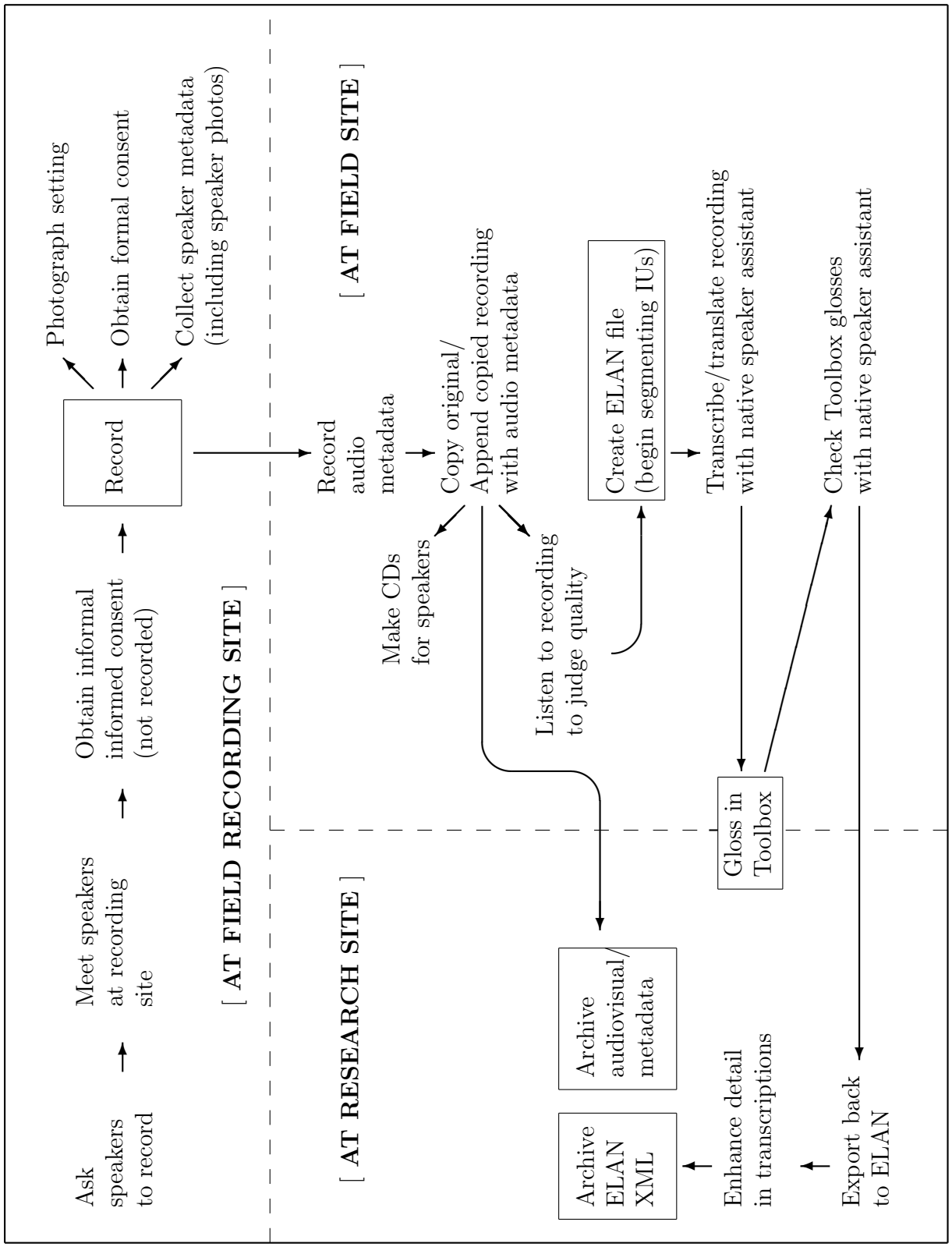
If the recording is of good quality, I begin the long process of transcription by creating an ELAN file. Before each transcription session, I segment the recording into Intonation Units (IUs) to speed up the transcription process. Working with my transcription assistant, we transcribe each IU in Besemah and translate them into Indonesian and English. For each three hour session, we are able to get through two to three minutes of the recording. Once a transcription is finished in ELAN, the ELAN file is exported into Toolbox and I begin to gloss the file. This could occur in either the field site or the research site. Questionable glosses are marked as such and then later checked with the speaker. Once the Toolbox file is glossed it is then exported back into ELAN. Once in ELAN, the file can be enhanced or annotated further by me. This is the stage where I enrich my transcription with elements from Discourse Transcription (Du Bois et al., 1993). Once the transcription is in its final

form, I archive the fully annotated transcription in an XML format.

4 Outcomes

The outcomes of a language documentation (hopefully) are more far-reaching than the linguists who instigates the project intends. This is precisely my hope for the BLDP. However, I do have a some intended outcomes that I hope are built upon the data in the BLDP. These primarily include a comprehensive reference grammar and dictionary of Besemah. However, another less tangible outcome that I hope will grow out of the BLDP are language revitalization efforts by the Besemah community. At this time, however, it is unclear when this will happen and how it might happen.

Figure 2: Workflow Diagram



References

Adelaar, K. Alexander (1992) *Proto-Malayic: The Reconstruction of its Phonology and Parts of its Lexicon and Morphology*. 119, Canberra, A.C.T., Australia: Department of Linguistics, Research School of Pacific Studies, the Australian National University.

Adelaar, K. Alexander (2005) Structural Diversity in the Malayic Subgroup. In *The Austronesian languages of Asia and Madagascar*, K. Alexander Adelaar & Nicholas P. Himmelmann, eds., New York, NY: Routledge Curzon, 202–226.

Adelaar, K.A. & D.J. Prentice (1996) *Malay: Its history, Role and Spread*, vol. II.1. Berlin: Mouton de Gruyter.

Bird, S. & G. Simons (2003) Seven dimensions of portability for language documentation and description. *Language* **79**(3): 557–582.

Brandes, Jan Laurens Andries (1884) *Bijdrage tot de vergelijkende klankleer der Westersche afdeeling van de Maleisch-Polynesische taalfamilie*. Utrecht: Van de Weijer.

Chang, Debbie (2010) *TAPS: Checklist for Responsible Archiving of Digital Language Resources*. Master's thesis, Graduate Institute of Applied Linguistics.

Collins, E.F. (2007) *Indonesia betrayed: how development fails*. Univ of Hawaii Pr.

Collins, W.A. (1979) Besemah concepts: a study of the culture of a people of South Sumatra .

Collins, W.A. (1998) *The guritan of Radin Suane: A study of the Besemah oral epic from South Sumatra*. KITLV Press, Leiden.

Conathan, Lisa (2011) Archiving and Language Documentation. In *The Cambridge Handbook of Endangered Languages*, Peter K. Austin & J. Sallabank, eds., chap. 12, Cambridge: Cambridge University Press, 235–254.

Du Bois, J.W., S. Schuetze-Coburn, S. Cumming, & D. Paolino (1993) Outline of discourse transcription. In *Talking data: Transcription and coding in discourse research*, Jane Anne Edwards & Martin D. Lampert, eds., Hillsdale, NJ: Lawrence Erlbaum Associates, 45–89.

Gil, D. (1994) The structure of Riau Indonesian. *Nordic Journal of Linguistics* **17**(02): 179–200.

Gil, David & Betty Litamahuputty (in preparation) The MPI Southwest Sumatra Corpus., a joint project of the Department of Linguistics, Max Planck Institute for Evolutionary and Bung Hatta University.

Good, Jeff (2011) Data and Language Documentation. In *The Cambridge Handbook of Endangered Languages*, Peter K. Austin & J. Sallabank, eds., chap. 11, Cambridge: Cambridge University Press, 212–234.

Helfrich, Oscar Louis (1904) *Bijdragen tot de kennis van het Midden Maleisch: (Běsěmahsch en Sěrawajsch Dialect)*, *Verhandelingen van het Bataviaasch Genootschap van Kunsten en Wetenschappen*, vol. 53. Batavia: Landsdrukkerij.

Helfrich, Oscar Louis (1915) *Nadere Aanvullingen en Verbeteringen op de Bijdragen tot de kennis van het Midden Maleisch (Běsěmansch en Sěrawajsch Dialect) - Lampongsche dwerghertverhalen*, *Verhandelingen van het Bataviaasch Genootschap van Kunsten en Wetenschappen*, vol. 66. Batavia: Albrecht & Co.

Helfrich, Oscar Louis (1921) *Supplement op de in Deel LXI, 3e en 4e stuk der Verhandelingen Gepubliceerde Nadere Aanvullingen en Verbeteringen op de Bijdragen tot de kennis van het Midden Maleisch (Běsěmansch en Sěrawajsch) (Verschenen in Deel LIII der Verhandelingen)*, *Verhandelingen van het Bataviaasch Genootschap van Kunsten en Wetenschappen*, vol. 63. Batavia: Albrecht & Co.

Helfrich, Oscar Louis (1927) *Nadere Bijdragen tot de kennis van het Midden Maleisch (Běsěmasch en Sěrawajsch Dialect)*, *Verhandelingen van het Bataviaasch Genootschap van Kunsten en Wetenschappen*, vol. 68. 's-Gravenhage: Nederlandsche Boeken Steendrukkerij V/H H. L. Smits.

Helfrich, Oscar Louis (1933) *Bijdragen tot de kennis van het Midden Maleisch (Běsěmahsch en Sěrawajsch Dialect) Supplement op de "Nadere bijdragen" (1927).* 's-Gravenhage: Martinus Nijhoff.

Himmelmann, N.P. (2006) Language documentation: What is it and what is it good for. In *Essentials of language documentation*, Jost Gippert, Nikolaus P. Himmelmann, & Ulrike Mosel, eds., Walter de Gruyter, 1–30.

Lewis, M.Paul (2009) Ethnologue: Languages of the world. URL `http://www.ethnologue.com/`.

McDonnell, Bradley (2009) A Conservative Vowel Phoneme Inventory of Sumatra: The Case of Besemah. *Oceanic Linguistics* **47**(2): 409–432.

McDonnell, Bradley (2010) *Two passives and a paradigm: Exploring undergoer voice in Besemah*. Master's thesis, Arizona State University, Tempe, AZ.

McDowell, Jonathan (2007) The Malays of southern Sumatra: Unity in diversity, the Eleventh International Symposium on Malay/Indonesian Linguistics, Manokwari, Indonesia.

McDowell, Jonathan & Karl Ronald Anderbeck (2007) Unity in diversity: Malayic varieties of Southern Sumatra., sIL, Jakarta.

Mitani, Yasuyuki (1980) Languages of South Sumatra. In *South Sumatra: Man and Agriculture*, Tsubouchi, ed., Kyoto: The Center for Southeast Asian Studies, Kyoto University, 1–16.

Thieberger, Nicholas & Andrea L. Berez (2012) Linguistic Data Management. In *The Oxford Handbook of Linguistic Fieldwork*, Nicholas Thieberger, ed., Oxford: Oxford University Press, 90–118.

Voorhoeve, Petrus (1955) *Critical survey of studies on the languages of Sumatra*, *Koninklijk Instituut voor Taal-, Land- en Volkenkunde Bibliography Series*, vol. 1. The Hague, Netherlands: 's-Gravenhage - Martinus Nijhoff.

Woodbury, Anthony C. (2011) Language Documentation. In *The Cambridge Handbook of Endangered Languages*, Peter K. Austin & J. Sallabank, eds., chap. 9, Cambridge: Cambridge University Press, 159–176.

Wouk, F. (1999) Dialect contact and koineization in Jakarta, Indonesia. *Language sciences* **21**(1): 61–86.

Wouk, Fay (1989) *The Impact of Discourse on Grammar: Verb Morphology in Spoken Jakarta Indonesian*. Ph.D. thesis, UCLA.