For projects interested in documenting, describing or revitalizing languages, especially endangered languages, historically existing materials (whether digital or analog) like tape recordings made in earlier times or written materials collected years or even centuries ago may exist and may represent important sources of information, indeed, in some cases, the only information available. Making use of legacy text material raises many challenges that need to be confronted if we wish to include it in a corpus or to treat it along with other contemporary data. There are practical, technical, contextual, ethical, and political issues that legacy materials raise, and many questions which it may be difficult or even impossible to answer. There are also many opportunities to add value to legacy materials using documentary linguistics methods. We discuss and exemplify these throughout this paper. A conclusion that can be drawn from this work is that creating good metadata and meta-documentation for current written materials can potentially reduce legacy data problems for future researchers compared to the issues that we face today.

Keywords: language documentation, legacy text materials, metadata, meta-documentation, archiving

Preamble
1. Introduction
2. Form, content and context issues
3. Stakeholder issues
4. Conclusions

This paper arises from teaching materials developed for sessions on data management for DocLing training courses at Tokyo University of Foreign Studies and is intended for beginning researchers to alert them to some of the challenges around working with legacy text sources and how documentary linguistics methods can be applied to confront these challenges and add value to such material. In some situations there may be no living speakers of the language recorded in such sources and legacy materials are all that exist; if processed and analysed well they can play an important role in language reclamation and revitalisation (see, for example, Amery 2000; Baldwin et al.)
24 Asian and African Languages and Linguistics 11

2016, 2017; Costa 2003, 2015; Giacon and Lowe 2016). The present paper expands on the training handouts and adds further examples and references to relevant literature.¹

1. Introduction

Over the past 20 years a new area of research on human language called Language Documentation (or Documentary Linguistics) has developed, drawing on ideas from linguistics, anthropology, verbal arts, information science, media and recording arts, ethnoscience and other areas (Austin 2010, 2016; Himmelmann 1998, 2006; Lehmann 2001; Woodbury 2003, 2011). In his seminal article on the field, Himmelmann (1998: 161) presented its main goal as ‘to provide a comprehensive record of the linguistic practices characteristic of a given speech community’. Himmelmann (2006: vi) restated this as concerning ‘the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties’, while Woodbury (2011: 159) gives a similar definition: ‘language documentation is the creation, annotation, preservation and dissemination of transparent records of a language’. This approach emphasizes transparency and multifunctionality, as well as ethical engagement with a wide range of stakeholders, including speech community members. Himmelmann (1998: 161) also argued that language documentation ‘differs fundamentally from... language description [which] aims at the record of a language... as a system of abstract elements, constructions, and rules’. This view is critically discussed in Austin and Grenoble (2005) who argue that such a strong distinction is not useful or desirable.

Most of the theoretical and practical work on language documentation to date has tended to assume implicitly that it deals with audio-visual and textual data collected at the present time, and has not paid much attention to historically existing materials (whether digital or analog) like tape recordings made in earlier times or written materials collected years or even centuries ago (an exception is Himmelmann 2012 which discusses various data types). I will refer to such information sources as legacy material or legacy data in the discussion which follows, and concentrate on written textual materials in particular (what Himmelmann 2012 calls ‘original written document[s]’, which he considers to be raw data²). Pre-existing audio-visual recordings present many challenges to language documenters, including issues of digitisation, transcription, interpretation and contextualisation.³ For example,

---

¹ Thanks to Lyle Campbell, Lisa Conathan, David J. Costa, Lise Dobrin, Andrew Garrett, Anthony Jukes, Susan Smythe Kung, David Nathan, Gabriela Perez-Baez, Anthony C. Woodbury and two anonymous reviewers for discussion and comments on earlier versions. I am also grateful to Lauren Gawne, Cristina Muru, Julia Sallabank and Candide Simard for feedback on a workshop presentation of some of this material at SOAS, 20 February 2017. None of these scholars can be held responsible for errors in the use of their feedback.

² The notion of ‘raw’ here is problematic as some, perhaps much, legacy written material is ‘processed’ in various ways, ranging from transcription and/or translation to full grammatical analysis (see examples in sections 2.1, 2.3 below).

³ Many thanks to Lise Dobrin for discussion of these points.
digitisation problems may include dealing with mould on old tapes or the difficulty of finding machines which are able to play obsolescent formats like Hi8 video or minidisk audio. Specialist organisations such as *Paradisec*⁴, the *Phonogramarchiv* of the Austrian Academy of Sciences⁵ or the *British Library Sound Archive*⁶ are set up to deal with tasks like cleaning the carriers, providing equipment, and digitising such material. Difficulties of transcription and translation can include being able to comprehend the speaker(s) and/or language in the recordings, differences between contemporary language use and that in the legacy recordings, mismatches between transcriptions and translations (if they exist) and what is in the recording (see Mosel 2014), and lack of sufficient background information to be able to make sense of the content and/or context. Fuller discussion of these issues in relation to audio or video recordings is beyond the scope of this paper, and I will concentrate on text materials only in the following.

Legacy text material raises many challenges that need to be confronted if we wish to include it in a documentary corpus or to treat it together with other contemporary data.⁷ There are practical, technical, ethical, and political issues that legacy materials raise, and many questions which it may be difficult or even impossible to answer. There are also many opportunities to add value to legacy materials using documentary linguistics methods. We discuss and exemplify these below. When adding value to legacy materials, it is important to distinguish between adding structure (categories, entities, relationships), adding content, and adding format. Note that explicit and well-structured data (e.g. stored in a database, or marked up in extensible markup language (XML)) can have format added computationally, and also lends itself to repurposing for other uses and/or other users than the immediately intended audience (and so can be multifunctional, as language documentation proposes), see Gippert (2006: 358–361) and section 2.1 below for discussion and exemplification.

To make sense of legacy materials it is essential to explore the socio-cultural and historical context of the documents and their creation, including the biography of the author(s), especially what prior language knowledge and/or study and/or exposure they had, who their teachers/mentors/correspondents were, how long they worked on the language and at what point in their careers, how the work was funded and with what goals, whether there were previous studies of the language or the community that they could have had access to, and so on.⁸ It is also important to explore aspects of the historical period during which the materials were created in terms of the kind and impact of contact between communities, including colonialists, and what descriptive categories and formats would have been known and might have influenced the author(s),

---

⁶ See [https://www.bl.uk/subjects/sound](https://www.bl.uk/subjects/sound), accessed on 2017-04-07 and Copeland (2008).
⁷ See Goddard (1973) for discussion.
⁸ An example of this kind of detailed exploration is Silverstein (2015) on the historical development of Franz Boas’ fieldwork methodology and products.
e.g. traditional grammar based on Latin or Greek models. For further discussion see section 2.4 below.

In the following sections we discuss and exemplify some of the issues to be confronted when working with legacy text materials in terms of formal, analytical, stakeholder and ethical matters.

2. Form, content and context issues

In this section we discuss some of the challenges raised by the form, content, context and analysis of the written documents.

2.1. Issues with the form of the original

Reading and interpreting handwritten texts can be problematic as it can be difficult to deal with cursive writing and especially old writing styles that are no longer in use. The study and interpretation of old handwriting is called paleography, however most of the training material readily available on it focuses on reading medieval or later manuscripts in European languages and there is nothing that I have been able to locate on reading other sources, especially non-Western materials.

An example of challenging linguistic materials is Figure 1 which is an extract from the Diyari grammar of Flierl (1880) where the grammatical explanation uses an archaic German script (the Diyari words are in regular Roman script).

![Fig. 1 Diyari Grammar extract from Flierl (1880), from Stockigt (2016: 80)](image)

Sometimes, documents are written in non-Roman scripts such as ideographic or syllabic characters or phonetic scripts. Interpreting the symbols used in a document

---

9 See also Bowern (2003) for examples of some of the topics covered here. Andrew Garrett (p.c., 2010-02-05) notes that similar issues of interpretation can arise with contemporary ‘born digital’ data, and are not restricted to legacy materials in the sense intended here.

10 For example the online course at http://www.nationalarchives.gov.uk/palaeography/, accessed on 2017-04-09.
and mapping them to a modern form can be problematic. Campbell et al. (2015: 23) describe representational problems in the analysis of documents on Uralic languages:

much documentation was written in inaccessible transcriptions, often written phonetically (not phonemically) in the Finno-Ugric Transcription System, also known as the Uralic Phonetic Alphabet (UPA). It is important to note that the phonemic principle did not come into force until around 1930 or later. While the UPA is established among Uralic scholars, materials originally rendered in close phonetic transcription in this notation are formidable even for scholars accustomed to the UPA, and far worse still, are intimidating to the point of uselessness for speakers of the languages transcribed.

An example of this presented by Campbell et al. (2015: 23) is given in Figure 2.

An example of the transcription is a Saami paradigm for the word for ‘fish’ in Figure 3 from the website of the 3rd Sami Linguistics Symposium:11

---

Dobrin (p.c., 2017-05-11) gives a different example of problematic source materials for the Papuan language Arapesh:

The earliest Arapesh grammar was published in the journal *Anthropos* (Andreas Gerstner’s *Grammatik der AlubänSprache* [Nordküste von Neuguinea]). It used the most bizarre notation for vowels, which I later learned was their standard house style. The only reason I could make sense of it was that it was modelled on the customary Massoretic pointillation of Hebrew.

The huge corpus of materials in the Meskwaki language spoken in Wisconsin, USA, is written in a unique syllabic script called papepipo which ‘has a basic matrix of 48 syllables, comprising four vowels by themselves and compound characters for 11 consonant symbols combined with four vowel symbols’ (Goddard 1996: 117).

These examples illustrate that sources need to be studied closely and that there may be different issues and problems to be faced depending on the particular situation. Interpretation of document orthography and spelling of a given language may require philological and linguistic training to be able to analyse the original and map it to a modern phonetic or phonological representation (see Broadbent 1957; Crowley and Austin 2005; Koch 2011). Austin (2008) discusses problems with the interpretation of William Ridley’s 19th century publications on the Gamilaraay language of New South Wales, Australia. Ridley marks vowels in this language with or without a macron (e.g. ā versus a), however this is ambiguous and can represent vowel length...
and/or stress. Issues that arise in reading and interpreting the corpus of Meskwaki have been discussed in Goddard (1996) and Bear and Thomason (2011) (see also Dahlstrom 2015). Bear and Thomason (2011) note the following:

Meskwaki papepipo is adapted to be wonderfully easy to write. It is much harder to read. Papepipo omits vowel length, the consonant ‘h’, and nearly all punctuation. This means that spoken Meskwaki has eight distinct vowels, but papepipo writes only four vowels; spoken Meskwaki has eleven distinct consonants, but papepipo writes only ten consonants; spoken Meskwaki has 29 distinct consonant clusters, but papepipo writes only 16 consonant clusters. As a result, there is a great deal of educated guesswork involved in transcribing papepipo into fully phonemicized words, phrases, and sentences. The edited and translated text of ‘Skunk and Grizzly Bear’ should be taken provisionally and read critically: there may be mistakes or infelicities in Y[oung] B[ear]’s original papepipo, in L[ucy] T[homason]’s reading of the papepipo, in L[ucy] T[homason]’s choices regarding phonemicization, word breaks, and sentence breaks, and in L[ucy] T[homason]’s choices regarding Meskwaki-to-English translation.

Goddard (p.c., 2017-05-26) notes that in his current work on an edition of a Harmony of the Gospels in Southern Unami (Oklahoma. Delaware; Lenape) produced by a Baptist missionary in 1837–1839:

The orthography is at the same time remarkably modern and seriously underdifferentiated (not distinguishing the long and short consonants), as well as inconsistent in the writing of the numerous vowels (not surprisingly). Reversing the odd choices of letters (‘v’ for /h/; ‘h’ for /ʃ/; etc.) would still only produce an inconsistent mess of a transcription. But the translation is remarkably free and idiomatic, for example supplying enclitics that sounded right where the King James text has, of course, nothing. So here the edition has both the original and the phonemic transcription.

There may also be textual amendments (crossing out, additions), abbreviations, or other obscurities in the document, as in Figure 4 from Gerhard Laves’ fieldnotes on the Bardi language presented in Bowern (2003):

Bowern (2003) notes that Laves abbreviated words that occur frequently in his note (e.g. g. for ginyinggi ‘this, he, she, it’, g.on for ginyinggon ‘and then’). Stephen Wurm’s 1955 fieldnotes of New South Wales Aboriginal languages contain glosses and translations in Hungarian shorthand.12

Careful retranscription of legacy documents into a modern form is necessary if they are to be made more useful and multifunctional. It may also be advisable to link

---

12 AIATSIS archive item 002895A.
Fig. 4 Gerhardt Laves fieldnotes on Bardi, from Bowern (2003).
such transcription to images of the original documents so that readers may confirm the proposed analysis. A publication in print or online that aims to accurately reproduce all significant features in the original manuscript, including spelling and punctuation, abbreviations, deletions, insertions, and other alterations is called a diplomatic edition of a text. Online examples drawing on language documentation principles include colonial Zapotec documents curated by Lillehaugen et al. (2016), the Daisy Bates collection described in Thieberger (2016), and the website of the William Dawes manuscript of the Sydney language of Australia.\footnote{13} The underlying digital representation of the Dawes material uses XML extensible markup language and encodes crossing out, insertions, and amendments, as well as additions by the editors, such as spelling out abbreviations.\footnote{14} The transcription was created by the Dawes project team, with reference to Troy (1994), and then edited by tagging it in XML in order to:\footnote{15}


distinguish various text content structures (such as Sydney language vs English, and with the language content tagged for lemmas etc), person and place names, meanings, and commentaries. In addition, layout structures such as columns are represented.

From this XML source, XSLT transformations generate the content of each of the transcription pages (in two versions, edited and unedited). CSS is used to lay out the text visually on the web page.

An example of the resulting display which resembles a diplomatic edition (what the website somewhat misleadingly calls ‘unedited’) is shown in Figure 5.\footnote{16}

\footnote{13} See http://www.williamdawes.org/, accessed 2017-04-07.
\footnote{14} There is a long tradition in epigraphy of encoding this kind of information and value adding for creation of editions of ancient documents and stone inscriptions. For current approaches using XML see Bodard and Stoyanova (2016), the EpiDoc guidelines (http://www.stoa.org/epidoc/gl/latest/, accessed on 2017-05-29) and the cheat sheet available at http://www.nesc.ac.uk/action/esi/contribution.cfm%3FTitle=964, accessed on 2017-04-09.
\footnote{15} See http://www.williamdawes.org/howtouse.html, accessed on 2017-04-09.
A ‘cleaner’ (so-called ‘edited’) view of this page can be seen in Figure 6.\footnote{See http://www.williamdawes.org/ms/msview.php?image-id=book-b-page-2&edited=true (accessed on 2017-04-09).}

Note that Unicode fonts should be used for the representation in the corpus (see
Gippert 2006 on character encodings in language documentation).

Text material which is implicitly structured, such as by using typography or layout on the page to distinguish analytical categories or kinds of information, can be made more useful by encoding the structure separately from the form. For example, (1) presents a sample entry from the Scherer (1981) English translation of J. G. Reuther’s four volume manuscript Diyari-German dictionary.\(^\text{18}\)

\[(1) \text{Entry for } \textit{banbana} \text{ in Scherer (1981).}\]

5. \textit{banbana} (v) = ‘to stand still; to pause; to stop; to cease; to go no further’, e.g.
   1) \textit{ninkida nau banbana warai} = ‘here he stopped’, i.e. he did not continue on.
   2) With reference to the circulation of blood: \textit{matja kumari banbai} = ‘the [flow of] blood has already ceased’
   3) With reference to water: \textit{matja ngapa banbai} = ‘the water is already stationary’, i.e. has ceased to flow.
   4) With reference to the star of the wise men from the East: \textit{pungani miri ditji waka banbana wonti} = ‘the star stood still above the house’ (in Bethlehem).
   5) With reference to a track: \textit{ninkida paltu banbai} = ‘here the track terminates, comes to an end’
   6) With reference to a sandhill: \textit{dako kajirani banbai} = ‘the sandhill finishes up in the creek’

Nathan (2016) processed a digitized version of Scherer (1981) computationally and manually to create XML files where the different data types are tagged into a proper hierarchical structure, as shown in (2).

\[(2) \text{XML representation for } \textit{banbana} \text{ entry in Scherer (1981).}\text{\(^\text{19}\)}\]

\[
<\text{xml version="1.0"?>
<\text{diyarilexicon version="20170409"}>
<\text{entry label="5" num="5"}>
<\text{lemma}>
<\text{di}>banbana</di>
<\text{pos type="lemma">v</pos>
</\text{lemma}>
<\text{gloss}>to stand still; to pause; to stop; to cease; to go no further</gloss>
\]


\(^{19}\) The presentation here is based on Nathan (2016) with amendments by the present author.
<entry>
<di>ninkida nau banbana warai</di> <eg_gloss>here he stopped, i.e. he did not continue on</eg_gloss>.
</entry>
<entry>
<di>matja kumari banbai</di> <eg_gloss>the [flow of] blood has already ceased</eg_gloss>.
</entry>
<entry>
<di>matja ngapa banbai</di> <eg_gloss>the water is already stationary, i.e. has ceased to flow</eg_gloss>.
</entry>
<entry>
<di>pungani miri ditji waka banbana wonti</di> <eg_gloss>the star stood still above the house (in Bethlehem)</eg_gloss>.
</entry>
<entry>
<di>ninkida paltu banbai</di> <eg_gloss>here the track terminates, comes to an end</eg_gloss>.
</entry>
<entry>
<di>dako kajirani banbai</di> <eg_gloss>the sandhill finishes up in the creek</eg_gloss>.
</entry>
</diyarilexicon>

Such an XML file can now be unambiguously searched for different types of data as well as presented in various formats, such as on a web page using XSLT and CSS to convert the data labels into HTML formatting. Nathan (p.c., 2015-10-02) noted the following issues arose when converting the original document files (which had been created by scanning and optical character recognition (OCR) of a printed copy of Scherer 1981):

- the documents had a massive over-use of quotation marks, for multiple purposes and often redundant or fatuous scare quotes. This made it hard to drive markup from them, even though in a huge number of cases they were the only clue to structure.
- scoping - the original documents had some mixed assumptions about scope, e.g. of entries, pages, footnotes etc. In particular footnotes and footnote references are relative to page units so in many cases they don’t work properly (some are just in-line footnotes, which are OK).
- character data errors - mostly resulting from OCR errors.
- others, such as Scherer’s spelling errors, inconsistent use of all kinds of punctuation and line breaks (e.g. some parsing could be driven by looking for strings like “i.e.” but this appeared as i.e, i.e., ie, i.e. Also, the part-of-speech
formulations are inconsistent and otherwise problematic.

Researchers need to be aware of possible complications such as these when working on adding value coding to documents.

A further problem with the form of legacy text materials may arise from cryptic glossing, or wrong glosses, because the author could not understand their language consultant’s accent or pronunciation, or because the semantics of the source language terms were misunderstood. Crowley and Austin (2005: 60) give the examples in Table 1, drawn from various wordlists of Australian Aboriginal languages.

Table 1 Semantic problems in text materials, from Crowley and Austin (2005).

<table>
<thead>
<tr>
<th>Wordlist meaning</th>
<th>Correct meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Pronunciation problems</strong></td>
<td></td>
</tr>
<tr>
<td>heart</td>
<td>hot</td>
</tr>
<tr>
<td>wet</td>
<td>sweat</td>
</tr>
<tr>
<td>moths</td>
<td>boss</td>
</tr>
<tr>
<td>dung, shit</td>
<td>tongue</td>
</tr>
<tr>
<td><strong>2. Meaning problems</strong></td>
<td></td>
</tr>
<tr>
<td>a. generic versus specific</td>
<td></td>
</tr>
<tr>
<td>grass</td>
<td>vegetation</td>
</tr>
<tr>
<td>boy</td>
<td>uninitiated youth</td>
</tr>
<tr>
<td>beard</td>
<td>hair</td>
</tr>
<tr>
<td>day</td>
<td>now</td>
</tr>
<tr>
<td>thumb</td>
<td>your hand</td>
</tr>
<tr>
<td>girl</td>
<td>female</td>
</tr>
<tr>
<td>b. related word</td>
<td></td>
</tr>
<tr>
<td>thighs</td>
<td>buttocks</td>
</tr>
<tr>
<td>cloud</td>
<td>sky</td>
</tr>
<tr>
<td>woman</td>
<td>wife</td>
</tr>
<tr>
<td>hair</td>
<td>head</td>
</tr>
<tr>
<td>frown</td>
<td>blind</td>
</tr>
<tr>
<td>spider</td>
<td>to bite</td>
</tr>
<tr>
<td>dig</td>
<td>drink</td>
</tr>
</tbody>
</table>

Another issue to pay attention to is that understandings and analyses by the text’s author(s) of the language being recorded can be seen to change over time, and thus different parts of a collection of text material may show different spellings, translations etc. For example, Bowern (2003) mentions that Gerhard Laves began to analyse the
Bardi material he was writing in his fieldnotes as he collected it and made a number of analytical mistakes as a result, that is, later parts of the document does not contain what he actually heard but rather what he thought he heard. Also, Steele (2005: 84) notes in relation to William Dawes’ records of the Sydney language that ‘in order to be in a position to make some assessment of the soundness of an interpretation of a word, expression or sentence provided by Dawes, it is useful to have an idea of at which stage of his language learning an entry was created’.

2.2. Issues with the content of the material

There can be issues of various types that arise due to the content of the original text. For example, some of the content may be inappropriate to discuss in front of particular individuals or groups within a community (for more on sensitivities see section 3.1). Bowern (2003) notes that Gerhard Laves’ Bardi fieldnotes contain names that should not be spoken aloud because they are the same as the name of a close relative of a community member who has passed away and thus subject to a death taboo. Innes (2010) mentions stories in legacy texts that are considered to be ‘dangerous’ by a given community and should not be reproduced or distributed. Bowern (2003) also cites the problem that some of Laves’ fieldnotes contain information about secret male rituals and that she, as a female, should not read them; in this case the challenge is that she, as a trained linguist, is the only person with knowledge to decipher and potentially pronounce particular forms in such texts.

There may also be content that is dated or inappropriate by contemporary standards. For example, Bowern (2003) notes that manuscripts may also contain notes by the original author that were intended as private personal comments on the people who being worked with which other people, such as contemporary relatives, may find offensive. In addition, legacy materials may use ways of referring to indigenous people which were acceptable at the time they were written but which are offensive now. Authors may also have written down personal judgements about the nature of the material collected (e.g. ‘superstitious rubbish’) that do not match modern community judgements (e.g. interesting personal history). Andrew Garrett (p.c., 2010-02-05) notes that:

A lot of excellent examples of inappropriate content come up in early 20th century California fieldwork. People who never thought that any non-scholars, and certainly no Native people, would ever see these field notes, routinely wrote stuff like ‘half-breed’ to summarize genealogy, or commented on fluency levels in ways that people now would find insulting, or even wrote things like ‘a syphilitic’ next to people’s names.

How and indeed if these kinds of content should be represented in the modern corpus will need to be decided (e.g. by censoring them in some way, or placing them in
password-protected files).\textsuperscript{20}

2.3. Issues with analysis in the original

Difficulties with interpreting the content of legacy text materials can often arise if the original author either:

1. records what they think is a distinction (phonological, morphological, syntactic) that is not actually present in the language but may exist in the author’s native language or one that they are familiar with; and/or
2. misses some crucial contrast because it is not made in the languages the author is familiar with (as a native language, languages studied).

Examples of the first type can be found in Reuther’s Diyari dictionary where he distinguishes between voiced and voiceless consonants (e.g. \(<b>\) versus \(<p>\), as in the entry for \(banbana\) in (1), however this contrast is only truly applicable for apico-domal (retroflex) stops. Similarly, Reuther writes vowels \(<e>\) and \(<o>\) although Diyari only has three contrastive vowels (\(i\), \(u\), and \(a\)). An example of the second type in Reuther’s work is his failure to record the difference between stops and nasals at various points of articulation: apico-alveolar (\(t\), \(n\) in the modern orthography), lamino-dental (\(th\), \(nh\)), and apico-domal (\(rt\), \(rn\)). Thus, we find a collapsing of distinctions in \(nganha\) ‘me’ (Reuther’s \(ngana\)) and \(nganarna\) ‘be.participle’ (Reuther’s \(nganana\)).

Examples of both types of misanalysis for grammar can also be found in missionary records of various Australian Aboriginal languages, including Diyari, as discussed in Stockigt (2016). For example, Reuther, and earlier missionaries, recorded a ‘vocative’ case for Diyari on the basis of Latin grammar models they knew, although the forms listed are actually phonological distortions used on shouted speech (see Austin 1981, 2015). On the other hand, they failed to notice that verb forms they identified as ‘Modus Conditionalis’ participate in a switch-reference system (encoding different subjects between two clauses) in paradigmatic contrast to verb forms they identified as infinitive (which encode same subject). Numerous examples of other such misanalyses can be found and the researcher needs to be careful when including legacy materials into their documentary corpus to label the original analysis and clearly distinguish it from their own, as well as cite the original source and where it can be found.

2.4. Issues arising from a lack of context

Some of the most difficult issues to deal with in legacy text materials relate to the lack of metadata (data about the data) and meta-documentation (information about the context of collection and analysis, see Austin 2013) which would help with understanding and analysing their form and content. It is often unclear, for example, where data comes from as speakers are not identified, and their geographical origins,

\textsuperscript{20} A notorious example of this kind of material actually being published and causing negative reactions is Malinowski (1989).
social positions and relationships to other contributors are not specified. We are also not told whether they learnt the language as children or adults, and what other languages they might speak that could influence their knowledge of the language in the materials. As noted above, particularly useful can be information about the collector: their background, the languages they have some knowledge of, and their education history and prior language (and linguistic) study sometimes this metadata is available in biographies or historical documents but often it is not. Darnell (1995) argues concerning anthropological fieldnotes:

A generation after the original research was carried out, the interpretative context in which the work was done a context that was easily available to contemporaries will require reconstruction. Such a task will only be possible if documents about the anthropologist, his/her research and professional milieu at the time have been preserved. Thus, records usually thought of as the history of anthropology are also crucial to the interpretation of anthropological records.

Also important is knowledge of the research training and methodology of the collector, including research methods and tools, what books and articles they were familiar with, who they studied and communicated with about the project (including mentors and colleagues), what the goals of the research were and where it fits within their own career trajectory, as well as their relationships with the language consultants and the community (see 3.1). For the latter, Good (2010) identifies ‘contact, consent, compensation, and culture’ as four important variables, that is how the community was contacted by the researcher, how consent for the work was given, what compensation was provided to participants and what were the cultural differences and expectations between the collector and the described community. For legacy materials these are often unknown or have to be reconstructed.

The issues we identify here are not unique to linguistic analysis of legacy text materials but, as Bishop (2006) points out, understanding context is a major challenge for secondary analysis of all qualitative data. Bishop (2006: 15–16) reminds us that the construction of knowledge is socially, culturally and historically embedded and needs to be (re-)studied as such:

Because secondary analysis is, in fact, re-contextualisation, as Moore (2005) suggests, this highlights how vital contextual information is to the process of reusing data. ... Both primary and secondary researchers have the responsibility to be reflexive in a manner suited to their specific projects. In the case of secondary analysis, reflexivity requires consideration of both the contemporary context and that of the original project (Fielding 2004). These challenges are significant, but not overwhelming; there may be lessons to learn from other disciplines. Historians do not lie awake nights agonising about not being able to do history because ‘they weren’t there’. They interrogate historical artefacts,
consider conditions that led to their production, while recognising that their own framework shapes what is seen (Scott 1990). Thinking deeply about context is a useful reminder that even the most knowing subject is never all-knowing. A perspective in which ‘data’ is reflexively constructed, contextualised, and re-contextualised helps us to acknowledge the inherently social character of knowledge.

3. Stakeholder issues

Documentation projects typically have many stakeholders who may have different kinds of interests in the materials collected and the analyses created. Issues of control, consultation, and decision-making are important when deciding what kind of documentary material to include in any corpus and how it can be used. This is also true of legacy sources, with the additional complication of possible mismatches between past situations and the present.

3.1. Identifying stakeholders and relationships

As O’Meara and Good (2010) note, identifying who has a stake in a given document can be complicated by the fact that the current membership of a particular contemporary ‘community’ (however that is defined) may not coincide with past membership. Indeed, people who provided legacy materials may not even now be viewed as rightful members of a given group and therefore their information may be deprecated. It is also often unclear what agreements, if any, about such things as publication and distribution existed between the original collector and the community or particular individuals at the time (and whether these agreements were documented) as well as the relationship between any such agreements and arrangements that are currently being negotiated between the contemporary researcher and other stakeholders.

3.2. Identifying rights in the materials

There may be a range of individuals and groups who could hold various types of rights with relation to legacy materials, and identifying them can be problematic. We can distinguish between:

1. intellectual property rights, which are legal rights that arise in relation to creations of the mind;
2. copyright, which relates to ownership and distribution of products with economic value. Copyright varies for different types of materials (text, sound, images, databases) and only applies to original works fixed in a tangible medium. It is a form of property law and relates to money and economic interest. As such, copyright can be inherited, given away or sold (for further details see Newman 2007);
3. **moral rights**, which concerns how representations affect reputations. As Article 6(1) of the Berne Convention identifies: ‘Independently of the author’s economic rights and even after the transfer of said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to the said work, which would be prejudicial to his honour or reputation’.

Determining who holds these for legacy materials can be particularly problematic. Thus, O’Meara and Good (2010) raise the following questions:

1. who holds what rights? Are the rights documented? How do we establish rights retroactively? What if the researcher is not sure about speaker rights?
2. how do we determine rights when there are multiple contributors and data comes from multiple media?
3. who has inherited rights between the time of the original recording and now? (e.g. descendants of the original speakers, descendants of the original researcher)
4. what happens to ‘orphan works’ where the original stakeholders can no longer be identified? (e.g. materials passed from a researcher to a later researcher)

When analysing legacy data it is important to clearly document the various contributions to the work, including those of the original author, the linguist-editor, other researchers, and current community members. The data structures and metadata set up for a project that includes legacy materials should make this clear. An example is Bowern (2003) who outlines her setup for the database (in Toolbox format) that she uses for the analysis of Laves’ Bardi data.

### 3.3. Deciding on access to materials

For language documentation projects it is essential that specification is made in relation to access to the corpus, that is spelling out permissions granted to individuals who do not have specific rights to given materials by setting out how they should be allowed to inspect or make copies of a particular resource or a representation of the content of that resource (e.g. a transcription or translation). Typically, researchers can choose between four levels of access to their data and analysis (e.g. when depositing materials in an archive, or distributing/publishing them):

1. **open access** anyone can view, use or copy the materials
2. **restricted access** access to the material is limited to a class of users, usually depending on the contributors, the type or content of materials, or the identity of the user
3. **requested access** the user must ask the researcher for permission
4. **closed access** only the researcher can view, use or copy the materials
Licence agreements, such as Creative Commons,\textsuperscript{21} may be used to formalise how the materials may be used once access has been granted, e.g. copy but not change, change with attribution.

As O’Meara and Good (2010) point out, technological changes over time can introduce complications in deciding on access and use for legacy materials. So, for example, ‘open access’ in 1980 may well have meant users who were granted access would receive an analogue copy of a cassette tape or set of fieldnotes while in 2017 ‘open access’ may mean anyone can download the materials from a website on the internet. It is unclear what ‘open access’ will mean in the future. In addition, for the categories of restricted or requested access, there is the general issue of how to verify the identity of users who want access (especially if the request comes from an internet address), and if access is restricted to a certain group (e.g. ‘community member’) how researchers (and archives) can identify such membership. To address these issues sensitive negotiations and extensive discussions with stakeholders, often over an extended time period, may be necessary.

4. Conclusions

Working with legacy text materials from a language documentation perspective involves dealing with a range of often complex issues about the form, content, context and use of the original materials and analyses arising from them. However, there are many opportunities for researchers to add substantial value to legacy text materials, especially if they are able to work with other historical sources and/or contemporary knowledge holders to elucidate them and the context surrounding their creation, analysis and current status. Maximising such opportunities will require thinking about data entities, types and relationships and being explicit about them in the project design and application (e.g. in database design or XML tagging), with a very important role for metadata and meta-documentation. By creating good meta-documentation for written materials now we can hopefully reduce legacy data problems for future researchers, compared to the issues that we face today. Careful work with legacy text materials can also be very rewarding for researchers and communities, especially in the case of unique documents on languages/varieties or areas of knowledge that are no longer available, and that can serve as important sources for language support and revitalisation.

\textsuperscript{21} See https://creativecommons.org/, accessed on 2017-04-09.
References


