Part I

Documentation

1   Language documentation and meta-documentation

Peter K. Austin

1   Introduction

The past fifteen years has seen the emergence of a new sub-field of linguistics that has been termed 'language documentation' or 'documentary linguistics' (Himmelmann 1998, 2002, 2006, Lehmann 2001, P. Austin 2010a, Grenoble 2010, Woodbury 2003, 2011a). Its major goal is the creation of lasting multi-purpose records of languages or linguistic practices through audio and video recording of speakers and signers, and annotation, translation, preservation, and distribution of the resulting materials. It is by its nature multi-disciplinary and draws on theoretical concepts and methods from linguistics, ethnography, folklore studies, psychology, information and library science, archiving and museum studies, digital humanities, media and recording arts, pedagogy, ethics, and other research areas.

The term 'language documentation' historically has been used in linguistics to refer to the creation of grammars, dictionaries, and text collections for undescribed languages (the so-called 'Boasian trilogy'; for discussion see Woodbury 2011a: 163). However, work defining language documentation as a distinct sub-field of linguistics emerged around 1995 as a response to the crisis facing the world's endangered languages, about half of which might disappear in the twenty-first century (the crisis was identified and popularized in such publications as Robins and Uhlenbeck 1991, Hale et al. 1992, Wurm 2001). Linguists drew attention to an urgent need to record and analyse language materials and speakers' linguistic knowledge while these languages (or threatened special registers and varieties within them) continued to be spoken, and to work with communities on supporting threatened languages before

opportunities to do so became reduced. The emergence of language documentation was also prompted by developments in information, media, communication, and archiving technologies which make possible the collection, analysis, preservation, and dissemination of documentary records in ways which were not feasible previously. In addition, it was facilitated by large levels of research funding support from three main sources: the DoBeS (Dokumentation Bedrohter Sprachen 'Documentation of Endangered Languages') programme sponsored by the Volkswagen Foundation in Germany (2000–13), the Endangered Languages Documentation Project (ELDP) supported by the Arcadia Trust in the United Kingdom (2002–16), and the Documenting Endangered Languages (DEL) interagency initiative of the United States National Science Foundation and the National Endowment of the Humanities (2005 onwards).

Language documentation concerns itself with principles and methods for the recording and analysis of primary language and cultural materials, and metadata about them, in ways that are transparent and accountable, and that can be archived and disseminated for current and future generations to use. Some researchers have emphasized standardization of data/metadata and analysis and 'best practices' (e.g. E-MELD, OLAC), while others have argued for a diversity of approaches which recognize the unique and particular social, cultural, and linguistic contexts within which individual languages are used (see Dobrin, Austin, and Nathan 2009), Dobrin and Berson 2011).

This chapter is concerned with the role of metadata in language documentation and argues for a broad approach to observation and documentation of the methods, processes, and outcomes of language documentation projects, which we refer to as 'meta-documentation' (or 'meta-documentary linguistics'). It argues that a theory of meta-documentation does not (yet) exist and discusses some techniques that could be adopted for developing such a theory, as well as proposing some of the components that may make it up. The need for fuller reflexivity on the part of language documenters, and linguists more generally, is particularly emphasized.

2   Language documentation (or documentary linguistics)

Language documentation (or 'documentary linguistics') is defined by Himmelmann (2006: v) 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'. A similar definition is given by Woodbury (2011a: 159) as 'the creation, annotation, preservation, and dissemination of transparent records of a language'. Woodbury (2011a: 161) also notes that the term 'language documentation' is used in another sense, as well namely the outcomes of documenting languages, but proposes to clarify the terminology: 'The sets of records, coherent or not, are often called LANGUAGE

DOCUMENTATIONS; but since that is what we are calling the activity as a whole, I will call such sets LANGUAGE DOCUMENTARY CORPORA (or just CORPORA).' The form and uses of documentary corpora have been explored somewhat by linguists (e.g. within the DoBeS programme – see DoBeS 2005) and, as Dobrin and Berson (2011: 188) argue:

It does seem clear that documentary linguists have been on relatively comfortable ground in thinking about the PRODUCTS of linguistic research: conceptually distinguishing an annotated corpus or documentation of a language from a higher order description of its patterning . . . reasserting the intellectual value of vocabulary . . . and oral discourse (as represented in texts) alongside grammar, extending the range of documentary outputs to include items like primers and orthographies that are targeted directly at non-academic audiences . . . They have also enriched the inventory of digital data models, formats, and software tools that facilitate documentary research and enable the preservation and dissemination of its results.

There has been rather less discussion about what Woodbury (2011a: 161) calls 'corpus theorization': 'I will call the ideas according to which a corpus is said to cohere or "add up" its (CORPUS) THEORIZATION. Corpus theorizations, and even principles for corpus theorization, can both offer a space for invention and become a matter of contention and debate.' Corpus theorization has been only weakly developed within documentary linguistics. Seifart (2008) attempts to address some aspects of corpus theorization, namely representativeness and sampling, and Lüpke (2010) discusses data collection methods, few scholars have been explicit about *why* and *how* they are collecting and organizing their particular corpora, other than for some vague notion of 'documenting the language' or 'saving the data'.

In addition to corpus theorization, Woodbury (2011a: 161) also mentions wider issues of what he calls 'project design':

Of special interest is the range of concerted, programmed documentary activities motivated by impending language loss and aimed at creating a final record. These activities . . . raise questions about the participants, their purposes, and the various stakeholders in the activity or program of activity or project: we may refer to this set of questions as the PROJECT DESIGN . . . of a language documentation activity.

We see corpus theorization and project design as part of meta-documentation, which we explore and elaborate in Section 3. For some speculations about a possible typology of project designs see Section 4.

3   Meta-documentation (or meta-documentary linguistics)

From the outset, those who have been working on language documentation have been clear that alongside collecting and analysing data (typically audio

6   *Peter K. Austin*

and video recordings, but also still images)[1] it is necessary to record and analyse metadata, data about the data, to ensure that its context, meaning and use can be properly determined. As Nathan (2010b) states: '[M]etadata is the additional information about data that enables the management, identification, retrieval and understanding of that data. The metadata should explain not only the provenance of the data (e.g. names and details of people recorded), but also the methods used in collecting and representing it.' Notice that metadata is required not only for archiving but also for the very management, identification, retrieval, and understanding of the data within the documentation project once processing and value-adding is to be done. The way files are named and structured in folders is itself a type of metadata (see Nathan 2010b), and as Nathan and Austin (2004) argue, any knowledge added to the recordings (including transcription, translation, annotation, summary, index etc.) should be seen as 'thick metadata' (contrasted with the 'thin' cataloguing metadata often promoted in discussions of language documentation, e.g. by the E-MELD 'School of Best Practices in Language Documentation').

Nathan (2010b) also proposes that: '[A]nother way to think of metadata is as meta-documentation, the documentation of your data itself, and the conditions (linguistic, social, physical, technical, historical, biographical) under which it was produced. Such meta-documentation should be as rich and appropriate as the documentary materials themselves.' It is my contention that alongside language documentation we need to develop a theory (and related practices) of language meta-documentation (or Meta-documentary Linguistics), the focus of which would be (to adapt the definition of Himmelmann 2006) the methods, tools, and theoretical underpinnings for setting up, carrying out, and concluding a documentary linguistics research project. It would be the documentation of the documentation research itself.

Some work on particular issues that are relevant here has been published in the language documentation literature, especially in relation to research ethics (Grinevald 2003, Dwyer 2006, Rice 2006, Thieberger and Musgrave 2006, Macri 2010), reciprocity and exchange (Yamada 2007, Czaykowska-Higgins 2009, Glenn 2009, Guerin and Lacrampe 2010, Leonard and Haynes 2010, Crippen and Robinson 2011), and researcher and community motivations (Dobrin 2008), which are part of what Dobrin and Berson (2011: 189) call 'the social processes set in motion by . . . research, from the conceptualization of fieldwork to the dissemination of its products'; however, no wider approach

---

[1] The documentary linguistics literature pays little attention to the role of still images in corpus creation. However, evidence from materials archived by documenters, e.g. at ELAR, suggests that they take large numbers of photographs and scans for a range of purposes (including documenting their field sites, recording setup, ceremonies, objects, and consultants and other people, and for copying fieldnotes and other documents, etc.).

or theorization has been undertaken. There are several reasons it would be
valuable to do so:

- to develop good ways of presenting and using language documentations
  (what Woodbury 2011b calls 'making language documentations people can
  read, use, understand and admire')
- for future preservation of the outcomes of current documentation projects
- to assist with sustainability of the field of language documentation in terms
  of ensuring continuity of projects, people, and products
- helping future researchers learn from the successes and failed experiments of
  those currently grappling with issues in language documentation (see Gawne
  2012 and comments by James Crippen)
- to document intellectual property contributions to projects, including those
  of community members, researchers, and others, along with their career
  trajectories, especially for more junior researchers (Conathan 2011a).

There are at least three possible directions that could be explored to strive
towards a theory of meta-documentation:

1. *deductive approaches*: the postulation of axioms and theorems
2. *inductive approaches*: examination of current and past documentations (so-
   called 'legacy materials') to analyse practices and identify operating princi-
   ples (as well as lacunae)
3. *comparative approaches*: examination of what other relevant and related
   fields have done in their meta-documentation to see what is applicable, and
   what not, to language documentation.

We discuss each of these in turn.

3.1   Deductive approaches

Since its establishment in the late 1990s language documentation has been
dominated by declarative deductive approaches to recommendations for creat-
ing metadata which have been primarily influenced by library concepts (e.g.
Dublin Core). Key metadata notions have been interoperability, standardiza-
tion, discovery, and access (OLAC,[2] E-MELD, Good 2002, Farrar and Lewis
2007). However, the wider goals of language documentation (including the
wider social goals relating to speaker community involvement) mean this is not
powerful enough and we need, as P. Austin (2010a: 29) argues, to 'extend the
concept of meta-documentation to include as full as possible documentation
of the documentation project itself'. It appears that meta-documentation of at
least the following aspects should be covered:

- the identity of the stakeholders and their roles in the project beyond the, so
  far, restricted concern to document people and roles such as 'speaker' and

---

[2] See the Appendix to this chapter for a listing of OLAC metadata terms and their definitions.

'recorder' (for a fuller but still incomplete listing see the OLAC roles given
in Conathan 2011b: 245). For many projects other people, organizations,
and institutions play a crucial role, e.g. funders, gate-keepers, validators of
the research etc., but they and their roles tend to be neglected in metadata
creation.

- the attitudes of language consultants, both towards their languages and
  towards the documentation project. These can of course change and develop
  over time and have a vital impact on the success or failure of a project, as
  well as the nature of the materials which can be collected and disseminated.
- the methodology of the researcher, including research methods and tools (see
  Lüpke 2010), and any theoretical assumptions encoded through abbreviations
  or glosses, as well as relationships with the consultants and the community
  (Good 2010 mentions what he called 'the 4 Cs': 'contact, consent, compen-
  sation, culture')[3]
- the biography and history of the project,[4] including the background knowl-
  edge and experience of the researcher and the main consultants[5] (e.g. how
  much fieldwork the researcher had done at the beginning of the project and
  under what conditions, what training the researcher and consultants had
  received), and how the project emerged and developed. For a funded project,
  the project biography would include the original grant application and any
  amendments, reports to the funder, e-mail communications with the funder,
  and/or any discussions with an archive, such as the reviews of sample data
  mentioned by Nathan (2010b). Both successful and unsuccessful aspects of
  the project biography should be included.
- any agreements entered into, whether formal or informal (such as a Mem-
  orandum of Understanding, payment arrangements, and any promises and
  expectations issued to stakeholders).

This kind of information is invaluable not only for the researcher and others
involved in a project but also for any other future parties wishing to make sense
of the project and its history and context. Unfortunately, linguists have typically
been poor at recording and encoding this kind of information, meaning that
work is often difficult with so-called 'legacy data', especially materials that only
become available once the researcher has died (see Subsection 3.2 and Bowern
2003, P. Austin 2010b, Innes 2010, O'Meara and Good 2010). This is an area
for further development and experimentation within language documentation
theory and practice.

---

[3] This seems to correspond to Woodbury's (2011a, b) 'corpus theorization'.
[4] Note that OLAC (see the Appendix to this chapter) allows a date specification in the metadata
for individual resources but is vague about the significance of such dates, defining it merely as
'a date associated with an event in the life cycle of the resource'.
[5] Conathan (2011b: 248) mentions biographical information about project participants but not the
historical biography of the research project itself.

3.2   *Inductive approaches*

An inductive approach to meta-documentation would involve exploring current and past practices of language documenters to see what types of metadata they collect and notate within their projects. Here we report on two examples of such an approach: (1) a review of metadata practices in the Endangered Languages Archive (ELAR) at the School of Oriental and African Studies (SOAS) carried out by Nathan (2011), and (2) the main points from P. Austin (2010b), which look at the challenges of working with Australian Aboriginal legacy materials.

Nathan (2011) is a survey of metadata practices in forty-nine deposits in ELAR. He found that about 80 per cent of the most frequently occurring categories can be mapped to OLAC labels (see the Appendix to this chapter). However, depositors added richer specifications of other kinds of metadata information, including such things as parents' and spouse's mother tongues, speaker education levels, workflow status of materials, and terms in the researched languages (such as song titles or place names), or in other locally significant languages. Across the deposits examined some of these terms appeared frequently (e.g. 1 occurred in 20 of the 49 deposits); however, there were 613 terms which were unique and only occurred once in all the deposit descriptions, giving a 'long-tail' distribution (Anderson 2006). Nathan (2011) concludes that 'for endangered languages documentation, the metadata framework is to be *discovered*, *not predefined*, and the principle of the Long Tail is the opposite of focusing on the top 10–20 keywords . . . if supported and encouraged, documenters do produce diverse and more comprehensive metadata' (my emphasis). Nathan's review is suggestive of documenter practices for one archive, but needs further elaboration if it is to serve as a counterpoint to the deductive approaches which have dominated the field so far and which have emphasized standardization and metadata templates.

A second example of induction comes from P. Austin (2010b), which looks at issues arising from working with legacy materials on the Guwamu language from southern Queensland, Australia, collected in 1955 by the late Stephen Wurm. There are practical, technical, ethical, and political issues that this legacy data raises because of a lack of meta-documentation. Exploring these gives insights into what current documenters might wish to take into account for future users.

The Guwamu materials consist of: (1) fieldnotes of language elicitation (translations from English to Guwamu) collected from Willy Willis at Goodooga and comprising forty double-sided pages of notes with phonetic transcription and glosses in Hungarian shorthand, and (2) a short tape recording. At my request, the glosses were decoded and translated into English by Wurm and recorded onto tape in 1977. I copied the fieldnotes and added the English glosses (by transcribing Wurm's tape recording), resulting in a

138-page manuscript, a copy of which was deposited with the Library of the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS).[6] The following is a sample of data from the notes:

jama inda goammu ŋalgaŋanda?	Do you speak Guwamu?
baɖariṇj ŋalla	He is sick.
balgaru ŋunan ugwɛ:ilɛja	A few days ago I camped there.
balunj ŋadju ilu iɖamanjgija juraŋu-nda	I will leave my axe here with you all.

Very little metadata was recorded with these materials, with the result that there are difficulties with them of several types:

1. Problems with the form of the original:
   a. The handwriting in the notes is sometimes difficult to decipher.
   b. Orthography – Wurm's transcription is not documented anywhere in the notes but appears to be similar to the International Phonetic Alphabet. It is quite surface phonetic but appears both to overdifferentiate (e.g. by recording gemination for consonants) and underdifferentiate (e.g. by failing to distinguish apico-alveolar and lamino-dental nasals).
   c. Word boundaries are sometimes incorrectly represented.
   d. There is sometimes cryptic glossing, or apparently wrong glossing.[7]
   e. Changing understandings over time of the language being recorded – Wurm was clearly working out the structure of Guwamu as he went along (and there are some comments in the fieldnotes which indicate his guesses about particular morphemes). so his transcription (and translation) varies from the first page to the last.[8]
2. Problems with the lack of context – we know nothing of how the material was recorded, what sessions took place, the background of the speaker and his involvement in and attitudes towards the project (on tape he sounds enthusiastic, at least when singing). No information is available about agreements entered into or any compensation or dissemination arrangements.
3. Problems of unclarity about protocol, i.e. access and usage rights to the materials in their various forms. The copy of Austin's notes at AIATSIS have the following access restrictions applied to them: 'Closed access – Principal's permission. Closed copying & quotation Principal's permission. Not for

---

[6] See www.aiatsis.gov.au/library/docs/langbibs/Guwamu_Kooma_July07.pdf.

[7] P. Austin and Crowley (1995: 60) give examples from work on legacy materials of such errors arising because the collector could not understand the consultants' accent or pronunciation, or because the semantics were misunderstood; we find instances of the latter in Wurm's notes but not the former.

[8] Bowern 2003 mentions that Gerhardt Laves began to analyse Bardi material collected in the 1930s as he was writing it down and made mistakes as a result, i.e. he did not write what he actually heard but what he thought he had heard. Also, Steele (2005: 84) comments on William Dawes' notebooks on the Sydney language: '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.'

Inter-Library Loan'. However, the notes have also been re-transcribed and typed up by Jeanie Bell and myself in Toolbox format. The status of these derivative works is unclear.

4. Stakeholder and political issues – there is a community in southern Queensland who identify themselves as Guwamu, although no one today speaks the language. What their relationship is to Willy Willis and any interest or rights they might have in the materials collected by Wurm are unclear.

Similar problems with legacy materials have been identified by Bowern (2003), Schmidt and Bennöhr (2008), Innes (2010), and O'Meara and Good (2010), and these all point to the need for as rich a meta-documentation as possible at the time material is collected and processed, especially for future researchers and users.[9]

3.3   Comparative approaches

An area for exploration for the development of a theory of meta-documentation is comparison between the place and function of metadata in language documentation and its role and application in other allied fields, including social and cultural anthropology, physical anthropology, and archaeology (the so-called 'four fields of anthropology'). The beginnings of such comparisons are made in Ember and Good (2011) and Hanks (2011). As Ember and Good (2011) point out, many of the data types used across the fields of linguistics, cultural anthropology, physical anthropology, and archaeology are similar or identical, and that 'some, but not much, metadata is now shared across the four subfields'. They suggest that linguistics could learn from cultural anthropology, which 'has devised rich, semi-standardized vocabularies to describe cultural ideas and behavior'. In addition, archaeology and physical anthropology have rich ways to deal with spatial data and taxonomy. Hanks (2011) shows that archaeology (especially that influenced by Hodder 1999) has been more reflexive about its practices in the past fifteen years than language documentation has (e.g. by regularly encoding daily field diaries, debating the role of publication of 'raw data' (field reports) versus 'cooked analyses' (academic papers), and working on ways to share comparative datasets). Its concerns have also differed from language documentation as well in that it tends to be practised almost exclusively by teams of specialists (e.g. bone specialists, ceramics experts, use-wear analysts etc.), resulting in fragmented practices among those involved in

---

[9] Schmidt and Bennöhr (2008) discuss a number of issues with recovering data and metadata from several types of digital legacy materials; one of the recommendations at the end of their paper is '[d]ocument transcription and annotation conventions as well as the design, structure, and technical realization of your corpus as early on and in as much detail as possible. Publish this documentation in a form in which it will still be accessible 50 years from now' (127).

a given project, with greater emphasis on recording and analysing the excavation itself (for linguistics this would be the documentation of the recording event(s) – see Nathan 2010a on the importance for sound recording of attention to, and documentation of, spatiality), and so 'post-excavation analysis is in fact at the periphery of the interpretive process' (Hanks 2011, see also A. Jones 2002). Perhaps language documentation could benefit from exploring this separation between meta-documentation and interpretation of observable events and processes on the one hand, and meta-documentation and interpretation of corpora and other outputs on the other.

4   A possible typology of language documentation project designs

Over the past nine years I have been involved at various times in the assessment panels for grant applications for language documentation projects submitted to the Endangered Languages Documentation Programme, the Volkswagen Foundation DoBeS project, and the Documenting Endangered Languages inter-agency programme of the National Science Foundation and the National Endowment for the Humanities. Over this period I estimate that I have assessed over 1,500 grant applications; from this I believe it is possible to obtain a sense of the kinds of project design and planned activities and outcomes that applicants consider would receive positive evaluations from the grant selection panels. Both successful and unsuccessful applications can reveal much about what the applicants wish to do (within the parameters of the application specifications of each individual funding agency).[10]

In terms of overall project design, Grinevald (2003: 58–60), drawing on Cameron et al. (1992), presents a typology of fieldwork orientations by linguists over various time periods that is a useful starting-point:

- fieldwork ON a language – this is the traditional model associated with Boasian-type linguistic description and (later) 'ethical research'
- fieldwork FOR the language community – this model emerged in the 1960s under the title 'advocacy research'
- fieldwork WITH the language speakers – this was developed in the 1980s as 'action research' or 'negotiated fieldwork'
- fieldwork BY trained language speakers – this can be labelled 'empowerment research' and, while found sporadically in the nineteenth and twentieth centuries, only became practised more generally from the 2000s onwards.

Himmelmann (2006: 15) identifies as a key feature of language documentation '*work in interdisciplinary teams*' – documentation requires input and expertise from a range of disciplines and is not restricted to mainstream ("core")

---

[10] Although in a number of cases applications *do not* conform to the application guidelines issued by the funders.

linguistics alone'. We may contrast this with 'lone-wolf' research (P. Austin 2005, 2007, Crippen and Robinson 2011), where a single individual carries out all aspects of a project (what Dwyer (2006: 54) describes as 'the go-it-alone model of research: go in, get the data, get out, publish'). My impression, gained from reviewing documentation grant applications, is that, despite Himmelmann's proposal, the majority of projects are planned as lone-wolf research ON a language or research FOR the language speakers.

Further, by analogy with the typology of human land use developed within geography (see O'Neil 2011), we can identify the following general project design types:[11]

1. *Hunter-gatherer projects*. In land-use studies, hunter-gatherers are typified by a 'primary subsistence method that involves the direct procurement of edible plants and animals from the wild, by foraging and hunting, without significant recourse to the domestication of either' (Familypedia, hunter-gatherers).[12] Within linguistics this kind of project typically involves rapid surveys relying on questionnaires (primarily lexical) for language identification and classification, and collecting basic typological data, never staying in one place for any length of time (Simpson 2007 refers to this as 'FiFo (fly-in fly-out) linguistics').

2. *Slash-and-burn swidden projects*. In land use, slash-and-burn swidden is characterized by the 'cutting and burning of forests or woodlands to create fields for agriculture . . . or for a variety of other purposes. It is sometimes part of shifting cultivation agriculture, and of transhumance livestock herding . . . and operates on a cyclic basis' (Wikipedia, slash and burn)[13]. In language documentation these are typically 3–5 year projects aimed at creation of a Boasian trilogy (grammar, texts, dictionary), after which the researcher moves on to the next language to be studied.

3. *Sedentary intensive cultivation projects*. In land use, this is a kind of fixed-location agriculture that ranges from feudal to communal sociopolitically, with employment of local serfs and artisans in temporary or specialist roles (and the necessary application of fertilizer and pesticides, or crop rotation). For language documentation, these are typically long-term projects on a single site, often with Christian missionary connections, e.g. projects carried out by SIL International (formerly the Summer Institute of Linguistics) – see Dobrin (2009) for a discussion of the scope, institutional underpinnings, and wider implications for academia of the work being carried out by missionaries.

---

[11] These are meant to be broad types for general classificatory purposes; any given project may well be a mixture of types, or change from one to another over the course of its lifespan.

[12] Available online at http://familypedia.wikia.com/wiki/Hunter-gatherer (accessed 16 April 2012).

[13] Available online at http://en.wikipedia.org/wiki/Slash-and-burn (accessed 16 April 2012).

4. *Plantation projects*. In agriculture, plantations typically involve using third-world local residents to grow consumable products according to a specified form which are then extracted to be transported to a first-world context where they are refined and expensive value-adding takes place for distribution and sale. In language documentation, such a project typically involves training native speakers to transcribe and gloss language data (in a local orthography and lingua franca) using such software tools as ELAN[14] and Toolbox.[15] The outside linguist then takes the resultant files 'home' to process them further (by 'cleaning up' transcriptions and glosses, and adding further analytical labels, or translations into an international language), and publishes academic papers based on the analysed data. Note that the skills acquired by the native-speaker artisans typically have no local application.

5. *Sustainable projects*. In land use, this requires an ecologically driven holistic approach that may include reforestation and recuperation of damaged land. It typically combines social, ecological, and economic objectives (Munasinghe 1993), and involves careful resource assessment and mixed production systems with close management and control, taking a long-term perspective. For language documentation, few exemplars of what could be sustainable projects have so far been developed. While we generally understand sustainability in a language-archiving context (Nathan 2010b), it is unclear how to develop and sustain documentation projects and their necessary relationships beyond the three- to five-year cycle that typifies research and academic life. It is also unclear what kind of research project models can best contribute to sustaining endangered languages and the communities who want to maintain and develop them.

This rough typology, based on impressions gained from reviewing research grant proposals, suggests that language documentation projects can be categorized into a set of general design and organizational types. If the aim of granting agencies, and researchers more generally, is to promote sustainable projects, there is a need to move beyond our current models, take a longer time-scale perspective, and develop (and meta-document) a range of different approaches involving more stakeholders taking different roles and contracting different relationships with each other.

5   Conclusion

The practice of language documentation over the past fifteen years suggests that we need a new development in this sub-field of linguistics, namely meta-documentation (or meta-documentary linguistics), which aims to document

---

[14] Available online at http://tla.mpi.nl/tools/tla-tools/elan/ (accessed 7 November 2012).
[15] Available online at www.sil.org/computing/toolbox/ (accessed 7 November 2012).

the goals, processes, methods, and structures of language documentation projects. We can develop this field by theorization and by investigation of current and past practices, and by exploring comparative approaches. By creating meta-documentation for projects now, we shall hopefully reduce the legacy-data problems for future researchers compared with those that we face today (because such meta-documentation was not thought about very deeply in the past).

Appendix: OLAC metadata

The following is a list of the basic metadata categories proposed by the Open Language Archives Community (OLAC):

Contributor	An entity responsible for making contributions to the content of the resource
Coverage	The extent or scope of the content of the resource
Creator	An entity primarily responsible for making the content of the resource
Date	A date associated with an event in the life cycle of the resource
Description	An account of the content on the resource
Format	The physical or digital manifestation of the resource
Format.encoding	An encoded character set used by a digital resource
Format.markup	A mark-up scheme used by a digital resource
Identifier	An unambiguous reference to the resource within a given context
Language	A language of the intellectual content of the resource
Publisher	An entity responsible for making the resource available
Relation	A reference to a related resource
Rights	Information about rights held in and over the resource
Source	A reference to a resource from which the present resource is derived
Subject	The topic of the content of the resource
Subject.language	A language which the content of the resource describes or discusses
Title	A name given to the resource
Type	The nature or genre of the content of the resource
Type.linguistic	The nature or genre of the content of the resource from a linguistic standpoint