4.1 Introduction[1]

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

p. 91

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

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

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

p. 92

This chapter is not concerned with determining how much data needs to be recorded in the course of fieldwork. Instead, we assume that researchers can decide for themselves how much is adequate for a given situation, and we focus on proper principles for managing the data that is collected. It is of no small concern to see colleagues swamped by masses of recordings, photos, and texts, confused by discussions of which standards to follow and unable to keep track of not only the items in their collection, but also the status of