4.1.1 What is data?

Linguistic data arise from a range of techniques that straddle traditional humanities and scientific methodologies, from field-based observation to corpus data to experimentation. For this chapter, 'data' is considered to be the material that results from fieldwork, which may include: primary observational notes, recordings, and transcripts; derived or secondary material such as lexical databases and annotated texts; and tertiary material, in the form of analytical writing, published collections of texts, or dictionaries. In addition, fieldwork results in elicited data: elicitation is always a part of fieldwork, and many linguists also use questionnaires and experimental field methods (see Majid's Chapter 2 above). All of these types of data and their associated metadata (see below) together make up the larger documentary corpus. Like the house with solid foundations, a documentary corpus grows and changes over time and is continually updated as we collect more recordings, refine our annotations, and publish more analyses.

4.1.2 What is metadata?

Simply stated, *metadata* is one set of data that describes another set of data. In terms of linguistic field data, your metadata is your catalogue of 'what you have'—a list, stored in a relational database or a spreadsheet, of all the important bits of information that describe each recording, transcript, photograph, and notebook
p. 94 in ↳ your corpus. Keeping accurate, up-to-date metadata is crucial, because without it, you very soon lose track of the details of the items in your collection (try to remember the contents of a recording you made last year to see how important a brief description is). You would not know, for instance, if recording *x* predates or postdates recording *y*, or if photograph *p* is of mountain *m* or mountain *n*, or if recording *z* has been transcribed yet, or if WAV file *w* was originally on cassette.

You may think that you can keep track of all of this information in your head—and indeed, you may be able to—but consider what would happen if others wanted to access your collection, perhaps when you are back home after summer fieldwork, or after your retirement or death. Without a catalogue of metadata, there would be no way for anyone else to know what your corpus contains, or the conditions under which it was collected.

Fortunately, there are two widely accepted metadata systems for linguistic material that can guide fieldworkers in which bits of information to collect as part of a catalogue. These are the standards provided by Open Language Archives Community[7] (OLAC) (see below) and the ISLE MetaData Initiative[8] (IMDI). You should not feel limited to just the fields provided by either one. Since you may want to collect types of information that neither system requires, you can build your catalogue so that it can export to a form that can be used by any archive.

4.1.3 Form and content

Critical to the goals of this chapter is the distinction between the *form* of the presentation of data and its *content* (cf. Black and Simons 2009). The content of data should be constructed, or *marked*, in such a way that it is possible to derive many different presentation forms. A simple example of this is shown in Fig. 4.1. Two different kinds of markup are shown: in (a), a form-driven markup like HTML (the encoding used on the World Wide Web), and in (b), a content-driven markup.

In this example the presentation of the data is the same for both, but the mechanism by which this presentation happens is different. In (a), the bold type would be rendered directly from the strictly presentational ⟨b⟩ (bold) tag, whereas in (b), the data would be processed through a stylesheet that assigns different styles to different tags (in this case, everything marked as ⟨sentence⟩, ⟨header⟩, and ⟨adj⟩ are bolded). Without a structural identification of a word as being, for example, an adjective, there is