those items (e.g. 'Has this video been transcribed?', 'Has this text been interlinearized?', 'Has this collection of photographs been described?'). The result is that within a short time these colleagues cannot fully access their own collections because of insufficient description of the contents, or because the data is tied to now-defunct software, or even because of permanent data loss from lack of a suitable backup procedure.

In a discussion of the general problem of data 'deluge', Borgman (2007: 6) notes: 'The amount of data produced far exceeds the capabilities of manual techniques for data management.' This has become a truism for computational tasks on the web, and it is also true of the comparatively small datasets created by documentary linguists. If we want to move beyond tedious manual techniques for manipulating our data and make use of far more efficient automatic and global operations, we need to learn how to first prepare *well-formed data* in a predictable structure that conforms to existing standards, so that we can then allow computers to do the rest of the work. After all, handling repetitive (and tedious) tasks quickly and accurately is what computers do best.

Publishing and citing the primary data on which an analysis is based is becoming more common and may soon become a requirement for major research grants (cf. Borgman 2007: 240–41) and submission to journals. Until the late 1980s it was difficult to cite primary data in a linguistic analysis (although Heath's 1984 grammar of Nunggubuyu is a valiant attempt at linking examples to textual sources). Since then, however, only a few grammatical descriptions (e.g. Morey 2004; Thieberger 2006) link the descriptions and analyses of linguistic phenomena to examples from a corpus. The methods described in this chapter build citable corpora and encourage the use of cited data in any resulting analysis, a practice that has been part of lexicographic and corpus-based linguistics for some time (cf. Liberman 2009).

We want to be clear, however, that the goal of this chapter is not to drown field linguists in the virtual ocean of stringent and ever-changing specifications of 'best practice' procedures that our discipline sometimes seems to be swimming in. In this context it is important to bear in mind Voltaire's admonition (Arouet 1877) that 'the best is the enemy of the good'. Bloomfield is quoted as saying, 'each of us should take a vow of celibacy, not teach, and dedicate the entirety of our summers to fieldwork and our winters to collating and filing our data, year after year' (Hockett 1962). We suggest it would be better for linguists to become more efficient using the methods we advocate, and to still have a life!

p. 93

This chapter aims to help find the balance between the desire to record and annotate our data to the highest possible standards, and the reality of the time constraints we all work under. In fact, if we take the time to produce well-formed, reusable data, our linguistic analyses will be more convincing, and our production of language materials will become more efficient. The initial outlay of time and effort required to understand the methods and tools for good data management may seem high, but the investment will pay off in the end.

This chapter is organized as follows. The remainder of §4.1 provides some basics about data and metadata, as well as an overview of the workflow for creating well-formed data. In §4.2 we outline the planning required before setting off on fieldwork, including preparing to use (or not use) technology in the field and making contact with an archive. The data management tasks you will face, including file naming, care for fieldnotes, metadata collection and storage, regular expressions, time-aligned transcription and interlinearization, and the use of lexical databases, are discussed in §4.3. For those who are not so technically oriented, we suggest skipping §4.3.3.1 to §4.3.4. In §4.4 we discuss the broader advantages of creating well-formed field data for linguistics as a discipline, and §4.5 contains concluding remarks.