Data management

1 The goals (EASST)

1. **Easy:** Easy to enter data and manage. Organization isn't useful if you don't use it.
2. **Accurate:** Your data should be an accurate and complete reflection of your elicitation.
3. **Safe:** Don't lose data.
4. **Searchable:** Your data isn't good if it doesn't help your job as a linguist. You may want to search for particular sound sequences, particular words, particular constructions, etc.
5. **Transparent:** Ideally, you (even years later), other linguists, or even the speaker/community should be able to later access your data and make sense of it. Consistent structure is key.

> "It is worth investing some time and thought in how you organize your fieldnotes. Don't rely on your memory; you won't remember what's in which recording in six months, or what a particular question mark meant. Did it mean that the gloss is suspicious, or that you aren't quite sure of the transcription, or that you want to check that the word is in your main database?" — Bowern 2008, p. 54

2 Data and metadata

(1) **Metadata for each session:**

- The speaker(s), date, some organizational ID (I use the date)
- The plan you used / topics you covered
- Where you stored associated recordings

(2) **The parts of a complete datum:**

- A way to refer to it (e.g. date/session ID and example number)
- (Information on the relevant context/stimuli)
- What kind of task this was
- If translation: what you asked for, the speaker's response
- If a judgement: your constructed ex, speaker's judgment, (speaker's translation)
- If volunteered: the speaker's volunteered utterance, speaker's translation
- (Comments by the speaker)

3 Storing your data

(3) **Paper:**

- Establish some general organizational conventions for your notes.
- Write large and leave extra space. During elicitation, you may want to go back to a section and correct something or add information on meanings, variations, etc.
- A notebook might be a good idea, to keep everything together.

- Scan all your notes from time to time, so you have a more permanent record.
- Disadvantages: can be disorganized, lost, damaged; hard to search...
- Advantage: a good tool during elicitation—the least distracting for your consultant

If you start with paper during elicitation, it is best to then convert those notes into a digital, searchable form later.

(4) **Text files:**

- Quickly establish a format for your files.
- Many use Word; some use Excel. Ability to format is nice but I prefer plain text.
- Plain text files (if formatted well) allow you to do all sorts of searches and manipulation across multiple files later, especially if you code or know tools like `grep`.
- Free text editors with good search (regular expressions) across multiple files:
  - *Notepad++* `https://notepad-plus-plus.org/` (Windows)
  - *TextWrangler* `http://www.barebones.com/products/textwrangler/` (Mac)
- Use a uniform orthography; if you switch later, go back and update previous files (leaving backups). Use Unicode if you can. All of this supports searching later.
- You can still print things out later to look at things on paper.

(5) **Software:**

- Popular tools include *Toolbox* `http://sil.org/computing/toolbox/` and *Fieldwork Language Explorer (FLEx)* `http://fieldworks.sil.org/flex/` from SIL.
- Native storage and searching of glossed examples, extra metadata.
- Can take some time to get used to, but potentially great benefits.
- Make sure you understand how the data is stored / how you can export the data.

(6) **Audio/video (secondary):**

- Copy to your computer immediately and make sure they play back.
- Name files informatively. Split longer files into logical chunks.
- Never edit your original files; always copy first.
- If you later transcribe from audio or compare your notes to the recording, add timecodes, which will help you find data in the recording later.
- *Praat* `http://www.fon.hum.uva.nl/praat/` is widely-used for phonetic analysis.
- *ELAN* `https://tla.mpi.nl/tools/tla-tools/elan/` is very good (but advanced) software which you can use to transcribe recordings, aligned to audio and video.