22 Managing Data in a Language Documentation Corpus

Christopher Cox

1 Introduction

This chapter presents a case study in data management in the context of documentary linguistics, a subfield of the language sciences that is concerned with issues in the development and application of records of linguistic practices and knowledge (Himmelmann 1998; Woodbury 2003, 2011; McDonnell, Berez-Kroeker, & Holton 2018). In contrast to some other areas of linguistics, attention to data management practices and procedures has been a central theme in the emergence of documentary linguistics over the past twenty years, following in part from the emphasis that definitional work in this area placed on defining linguistic data types and their relationship to one another (e.g., Himmelmann 2012) and the preservation and reuse of language records in research (e.g., as supported by language archives; cf. Henke & Berez-Kroeker 2016). Even with this attention to data management issues, what individual documentarians and language documentation teams actually *do* to develop documentation—the actual nuts and bolts of running a documentation project from start to finish—has, somewhat surprisingly, been described in the literature less often than more abstract "best practice" or "good practice" recommendations. This imbalance has begun to be addressed by published descriptions of individual documentary collections and the projects that have developed them (e.g., Schembri et al. 2013; Salffner 2015; Gawne 2018), which often include some discussion of data management issues. This chapter has a similar aim to these publications, albeit with a somewhat narrower focus on data management concerns specifically, rather than the full range of ethical and other issues that come with work in language documentation (cf. Rice 2006; Czaykowska-Higgins 2009, 2018; Holton, Leonard, & Pulsifer, chapter 4, this volume).

As important as this attention to language records and their management has been to documentary linguistics as a subfield within the language sciences, data management practices in actual language documentation projects are diverse, and what is described in this use case should not be assumed to be representative of all contexts in which documentation is being undertaken today. The sections that follow present one illustration of how common data management practices in language documentation have been applied in one particular documentation project and do not attempt to offer either exhaustive coverage of practices across the entire field or a general-purpose guide to data management in documentary linguistics, in general. Fortunately, further guidance on data management practices in language documentation can be found in sources such as Bowern (2015) and Meakins, Green, & Turpin (2018), who approach these issues from a field linguistics perspective, but whose advice applies here no less well. The discussion found in Thieberger & Berez (2012) also offers valuable insights into a wider range of data management issues in documentary linguistics. This chapter attempts to complement the treatment of data management practices found in sources such as these by illustrating the principles and practices that they describe, paying particular attention to the kinds of data that emerge in the course of a language documentation project and how they are stored, organized, and drawn on.

The overall workflow described in this chapter also shares some similarities with those described in a number of other data management use cases in this volume. Although this chapter focuses on the documentation of a spoken language that still has an appreciable number of first-language users, many of these same data management practices are also shared with work with historical documentation for language reclamation (Lukaniec, chapter 25,

this volume) and with sign languages (Crasborn, chapter 39, this volume; Palfreyman, chapter 21, this volume). As well, while the focus of the project described here is on the creation of multipurpose primary data representing language in use, rather than on particular structural or grammatical features of that language, many language documentation corpora contain records of both kinds. In these cases, there is likely to be some overlap with data management practices in descriptive linguistic fieldwork (Daniels & Daniels, chapter 26, this volume), as well, as is evident from the connection between fieldwork and documentation often made in the wider literature (e.g., Bowern 2015, among others).

2 Background, planning, and community considerations

The data management practices in language documentation that this chapter discusses are centered on Plautdietsch (ISO 639-3: pdt; Glottocode: plau1238), a diasporic West Germanic language spoken by minority communities throughout the Americas (primarily Belize, Bolivia, Brazil, Canada, Mexico, Paraguay, and the United States), Europe (Germany), and Asia (primarily Kazakhstan and the Russian Federation). Plautdietsch is spoken today primarily by Dutch-Russian Mennonites and their descendants, a pacifist Christian denomination with roots in the radical Protestant Reformation (Dyck 1993). The documentation project at the heart of this chapter was conducted from 2010 to 2015 and concentrated on Plautdietsch as spoken in central Saskatchewan, Canada, one of the larger Mennonite settlement regions in western Canada and a historically important waypoint for international Mennonite migrations during the twentieth century (Guenter et al. 1995). No prior linguistic research had been conducted in these communities before this project began, and ongoing language shift toward English had left Plautdietsch endangered in the region, with most first-language users from the region in their sixties or older.[1] Documentation activities came about as part of a larger, local response to this pattern of language shift, paralleled by a growing interest in second-language education programs for members of the local community (e.g., classes for adult learners). Outside of bilingual dictionaries developed in other Mennonite communities in Canada, there were relatively few resources available for language learners and teachers at the time that this project started. Many community members also noted considerable linguistic variation between individuals and communities in the region that needed to be taken into consideration when developing language programs and materials, but on which little linguistic or sociodemographic information was available. This was one area where (socio)linguistic research based on newly developed documentation seemed particularly promising, being well aligned with both the immediate need for serviceable information on how Plautdietsch was being spoken in the region, as well as broader questions about how linguistic variation is distributed in diasporic, religious minority communities such as these.

Preparing for documentation projects that aim to contribute both to the immediate issues identified as priorities by members of this language community and to the resources available for *language work* (which here includes efforts focused on education, revitalization, and supporting resource material development, as well as research focused on linguistic and sociolinguistic aspects of the language community's practices reflected in the final documentation) in the longer term often involves both initial and ongoing consultation with members of the language community (Czaykowska-Higgins 2009). Contacting, consulting with, and staying in touch with members of the language community present their own small data management tasks in terms of organizing contact information (e.g., contributors' phone numbers, mailing addresses, and physical addresses) and scheduling meetings related to the project. In this project, contributors' contact information was kept in a spreadsheet where each row represented a single contributor (maintained in Microsoft Excel, although in retrospect, Unicode-encoded CSV/TSV would have provided a more platform-neutral alternative), and meeting schedules were maintained in a separate word processing document (kept in Microsoft Word, although again, a simple text file or a digital calendar would have likely handled this just as well, if not better). Both of these documents were intentionally kept separate from other data in this project: because some of this information, such as phone numbers and home addresses, might be considered sensitive, an effort was made to keep these data separate from other sources of information assembled in this study.

Through the initial discussions that took place with members of the language community over the initial planning phase of this project, it quickly became clear that we would need to be mindful of the diversity of contributors'

backgrounds, varieties, and patterns of language use. While almost all of the contributors grew up in families in which Plautdietsch was the primary language of the home, most noted that their use of the language had significantly decreased over their lifetimes, with relatively few using the language on a day-to-day basis outside of interactions with their age-mates and other individuals who they knew were able to speak the language. Contributors also observed possible linguistic differences between different Mennonite settlement areas in central Saskatchewan and frequently commented on social divisions between Mennonite denominations that had historically contributed to limited interaction between groups within the Mennonite population. To help ensure that these kinds of geographical and sociolinguistic factors weren't overlooked in documentation, a basic questionnaire was developed that gave attention to individuals' places and dates of birth, levels of formal education, knowledge of languages present in the community (e.g., besides Plautdietsch, also English and Mennonite High German), and more.

These questionnaires were delivered to contributors on paper, and the completed forms were scanned (as PDF documents, although in retrospect, uncompressed TIFF at a minimum of 300 DPI would probably have been preferable from an archiving perspective) and stored alongside the meeting and contact information documents, keeping them separate from other documents in this project. The responses from these scanned forms were then entered into a contributor metadata spreadsheet (maintained as Unicode-encoded CSV), with each column representing one question and each row representing one completed questionnaire. Each contributor was assigned a unique, anonymous identifier in this spreadsheet (e.g., M01, F23), which served an important function later in this study, allowing the demographic and sociolinguistic information in this spreadsheet to be linked to transcripts of contributors' speech in project recordings (through the participant metadata in the corresponding ELAN transcripts; see section 4). Figure 22.1 shows a sample of this spreadsheet. Although the "master" copy of this document contained personally identifying details about contributors (e.g., names of contributors' parents), and was thus not immediately suitable for being shared more widely, an anonymized subset of these columns was later exported for use in analysis, allowing a large part of this information to be used in the study and shared more widely without compromising contributors' anonymity.

Another outcome of the discussions that took place at the outset of this project was general agreement about the need for more written resource materials in Plautdietsch, both to support second-language learners and adult language programs and first-language speakers who wanted to see their language in writing (both as an aid to literacy development and as a means to promoting respect for the language locally). This eventually led to the development of an illustrated *Fibel* [ˈfiːbəl], or "primer," which followed the model of books of the same name that were historically a key element of the traditional Mennonite educational system (cf. Cox 2015:56). As seen in figure 22.2, each page in this book presented a target sound, an English word whose Plautdietsch equivalent contained this sound in a consistent phonological environment, and an English sentence that presented the target word in context. By incorporating linguistically variable features reported for other Mennonite Plautdietsch speech communities in the previous literature into these example sentences, the final *Fibel* was able to serve not only as the basis for a new

	A	B	C	D	E	F	G	H
1	AnonID	Gender	POB	DOB	Age	ChurchDenomination	ChurchLocation	ParentsPOB
2	F00	F	RABBIT LAKE	1935	78	Mennonite Church Canada	SASKATOON	UkraineUkraine
3	F01	F	CARMEL	1922	91	Roman Catholic	CARMEL	UkraineUkraine
4	F02	F	Aberdeen	1938	75	Bergthaler	Warman	CanadaCanada
5	F03	F	Langham	1926	87	Mennonite Church Canada	SASKATOON	USAUSA
6	F04	F	CARMEL	1925	88	Roman Catholic	CARMEL	CanadaUkraine
7	F05	F	Kronsthal	1933	80	Old Colony	Neuhorst	CanadaCanada
8	F06	F	Hepburn	1928	85	Mennonite Church Canada	Osler	UkraineUkraine
9	F07	F	Osler	1933	80	Mennonite Church Canada	SASKATOON	UkraineUkraine
10	F08	F	Osler	1934	79	Mennonite Church Canada	SASKATOON	UkraineUkraine
11	F09	F	Osler	1939	74	Mennonite Church Canada	Osler	CanadaUkraine
12	F10	F	Hague	1940	73	Mennonite Church Canada	SASKATOON	CanadaCanada

Figure 22.1
An excerpt of the contributor metadata spreadsheet.

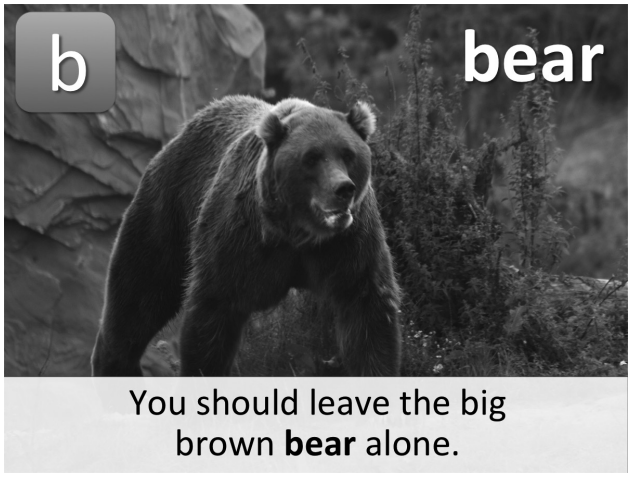


Figure 22.2
An example page from the *Fibel*.
Source: Adapted from the original photograph "Bears" by Flickr user davipt, released under a Creative Commons Attribution-NonCommercial-ShareAlike license (CC BY-NC-SA; https://creativecommons.org/licenses/by-nc-sa/2.0/).

resource for language learners, but also as a translation task that provided a baseline of information of linguistic variation among speakers of Plautdietsch in central Saskatchewan, complementing the spontaneous conversations and discussions that were also often recorded with contributors (cf. Lüpke 2009 on the use of linguistic tasks such as these in language documentation).

This document was initially developed as a rough set of text and images in Microsoft PowerPoint, which allowed for quick incorporation of contributors' feedback into the draft layouts and text of early revisions of this document. Images were drawn from public domain or Creative Commons–licensed sources, with the appropriate acknowledgments and license information kept in a separate spreadsheet (maintained as a Unicode-encoded CSV file). These slides were later exported from PowerPoint as a PDF document and as individual PNG images, which were stored alongside the other project-internal files. A copy of the *Fibel* was also printed in color, laminated, and spiral bound for easy use in the consultation sessions with contributors that followed.

3 Recording

With a printed copy of the *Fibel* in hand, consultation sessions were arranged with speakers of Plautdietsch from throughout central Saskatchewan to review and translate the *Fibel* and converse in Plautdietsch. As is common in language documentation projects, we wanted to make sure that these recordings were of the highest possible quality and that the formats and standards that we chose to follow at this stage in the project would allow for many possible future uses of the recordings that each speaker contributed (cf. Nathan 2010). This led us to use a dedicated, solid-state audio recorder (initially an Edirol R-09HR, accompanied by a stereo Sound Devices MixPre preamplifier; later, a Sound Devices 702), together with omnidirectional ear set microphones (Countryman E6i; cf. Lee 2013), for producing audio recordings of these sessions in the contributors' homes.[2] Recordings were made in uncompressed WAV format (48 kHz/24-bit samples), with each microphone recorded on a separate channel. Figure 22.3 shows one such session with Mrs. Nettie Boehr, one of the contributing speakers, as we reviewed the *Fibel* together.[3]

Immediately after each day's language meetings were finished, the audio recordings were copied from the recording devices' memory card onto a laptop, renamed and organized into directories by session, and copies then made onto multiple external hard drives for backup.[4] A standard session would result in a single folder containing files that looked something like this:

```
2011-07-23-pdt-NB-CDC-Warman/2011-07-23-pdt-NB-
CDC-Warman-Edirol-01.wav
2011-07-23-pdt-NB-CDC-Warman/2011-07-23-pdt-NB-
CDC-Warman-Edirol-01.wav.md5
2011-07-23-pdt-NB-CDC-Warman/2011-07-23-pdt-NB-
CDC-Warman-Edirol-01.wav.sha1
```

Figure 22.3
Reviewing the *Fibel* with Nettie Boehr.

While having consistent file naming and organization practices made later processing and annotation steps much more straightforward, our file naming conventions at the time were, in retrospect, unnecessarily baroque, including not only the recording date (July 23, 2011) and ISO 639-3 language code (pdt), but also the initials of the contributors (NB), of the recorder (CDC), the recording location (Warman, Saskatchewan, Canada), the recording device (Edirol R-09HR), and a track number (01). Much of this information could have safely been kept in separate, session-level metadata, whether in another text-based spreadsheet or in an XML-based format such as the Open Language Archives Community (OLAC; Bird & Simons 2003), ISLE Metadata Initiative (IMDI; Broeder et al. 2001), or Component Metadata Initiative (CMDI; Broeder et al. 2012) metadata schemas. It probably would have been better to keep these file and session names much simpler than this; in more recent documentation projects, I generally include only the language code (pdt), a unique project identifier (OS, *Onse Spröak* "Our Language"), the recording date, and the number of the session that day (01 for the first session that day, 02 for another session later that same day, etc.), with individual files being identified by track numbers (01 for the first recording made in the session, 02 for the second, etc.):

```
pdt-OS-20110723-01/pdt-OS-20110723-01-01.wav
pdt-OS-20110723-01/pdt-OS-20110723-01-01.wav.md5
pdt-OS-20110723-01/pdt-OS-20110723-01-01.wav.sha1
```

As mentioned above, once all of the day's recordings had been renamed and properly organized into sessions, we concatenated all of the tracks in each session into a single, long WAV in CD format (44.1 kHz, 16-bit stereo) and burned copies of this onto audio CDs for each of the contributors. This was done using Audacity,[5] an open source audio editing application, although this could have also easily been automated using a utility such as FFmpeg.[6] Once all of this had been done, copies of all of these files were made on several external hard drives, as well as in a private account with a cloud-based storage provider. Having multiple copies stored in multiple physical (and virtual) locations helped lessen the potential for catastrophic data loss, which is in keeping with the principle of LOCKSS (Lots of Copies Keeps Stuff Safe; cf. Austin 2006:89). Because these audio recordings were essentially static—we weren't planning on editing them directly in later work—keeping copies in multiple places was relatively easy, as we didn't need to worry about multiple versions of the same file circulating in the same project. Each of these storage options presents its own ethical and practical issues (e.g., in the case of cloud-based storage providers, the possibility of files being stored in jurisdictions where access may be granted to government agencies and other third parties without users' approval), and it is worth considering which of these options may be best suited to the particular context in which documentation is being undertaken.

4 Processing and annotating

With the recordings now properly organized and backed up safely in multiple locations, we set about annotating each recording's contents. We used ELAN, an open source software tool for annotating audiovisual materials that is the de facto standard for this kind of task in language documentation (Sloetjes 2014). Using ELAN allowed us to directly associate text (transcriptions of the original audio, as well as accompanying translations, working notes, and coding) with time-aligned segments of each recording, thereby creating a fully text-searchable speech database of all of the contributors' recordings.[7] Each recording was annotated by an ELAN transcript sharing the same file name (ending in .eaf, with a corresponding display preferences file ending in .pfsx automatically created by ELAN), for example:

```
pdt-OS-20110723-01/pdt-OS-20110723-01-01.wav
pdt-OS-20110723-01/pdt-OS-20110723-01-01.eaf
pdt-OS-20110723-01/pdt-OS-20110723-01-01.pfsx
```

Similar to the procedure that Nagy and Meyerhoff (2015) describe for sociolinguistic research, each ELAN transcript served not only as the place in which speech in the original recordings was transcribed, but also where contributors' translations of individual words and sentences in the *Fibel* were identified and linguistic variables that were represented in contributors' translations of *Fibel* prompts were coded. Thus, a translation of the *Fibel* prompt sentence 'You should leave the big, brown bear alone' would be annotated in ELAN as seen in example (1) (reproduced from Cox 2015:135, example (9), with data types added in boldface), with the third line containing JSON–like strings that identified individual linguistic variables (e.g., the form of the singular masculine

definite article in the accusative case) and their values in this utterance (e.g., here, *dän*) (see sentence display 1):

From a logistical perspective, including this kind of coding directly in the ELAN transcripts made sense: this kind of variation often required some review of the corresponding portions of the source recording to analyze, so having these codes time-aligned with the original audio made for much quicker work. This also opened the door to a much wider range of later uses of these transcripts (because we could instantly retrieve the audio associated with any text that was of interest), increasing the overall value of the investment that was made to produce time-aligned transcripts in the first place.[8]

5 Applying and sharing

Having the results of the sociolinguistic questionnaire, the original recordings, and our time-aligned annotations all organized consistently in non-proprietary, well-supported formats provided a number of benefits when it came time to apply these data to the tasks identified as priorities in this project. Managing our data in this way made it possible to develop scripts in Python and R that automatically extracted all of the tagged instances of variation directly from the corpus and fed these tokens directly into various forms of visualization and analysis (e.g., into dialectometric tools such as Gabmap [Nerbonne et al. 2011], which went a long way toward determining the extent to which differences between speakers represented in the corpus might be conditioned by features of local demographics and geography). Where these semi-automated processes drew attention to errors in the original annotations (e.g., a typo in a particular annotation), it was trivial to open the corresponding ELAN transcript, review and correct the relevant annotations, and re-run the entire analysis on the freshly corrected corpus. This kind of reproducible workflow not only made it feasible to analyze the quantity of data assembled here (contributions from nearly fifty speakers across twenty-seven hours of audio recordings), but also freed up time to delve into questions that we likely wouldn't have been able to get to otherwise if we had needed to do this kind of data wrangling by hand (including digging into apparent instances of personally patterned variation in these communities, a relatively rare finding from the perspective of sociolinguistic typology; cf. Dorian 2010; Cox 2015:244–247). Scripts that took care of data processing and analysis tasks were stored in their own "analysis" folder, separate from the primary recordings, ELAN annotations, and contributor metadata, which stayed in a consistent location that these scripts could refer to.[9]

Along with making it more straightforward to apply computational and quantitative methods in analysis, developing this collection of language materials in line with current good practice recommendations in documentary linguistics also allowed for easier citation of individual data points in later publications. In particular, using ELAN to create time-aligned transcripts of the contents of these recordings made a noticeable difference for citing excerpts from the corpus in ways that allow for quick reference back to the primary data. In one study based on these materials (Cox 2015), whenever data from the corpus were being cited, a reference was made to the corresponding session and track number, the unique identifier(s) of the contributing speaker(s), and the start and end times of this segment in the audio—all information that could be gleaned immediately from the ELAN transcripts. This is the case in example (2), where contributors M00 and F20 share their perceptions of two of the words for "girls" that are in use in the local community (reproduced from Cox 2015:178, example (13)):

(2) **M00:** *Mejalles*, [.] *Mäakjes*.
CDC: Is there a difference there, or, uh . . . ?

(1) **text:** Du su'st dän grooten, bruunen
gloss: you.SG should:2SG the.ACC big.ACC brown.ACC
Boa tochloten.
bear leave.alone:INF
coding: S07: { lxShould2S: "su'st", lxMascAccThe: "dän", lxMascAccDefBig: "grooten", lxMascAccDefBrown: "bruunen" }
free-translation: 'You should leave the big, brown bear alone.'
(*Fibel* sentence S07; F28, 2011-10-27, 4m45s890–4m48s490)

F20: *Na*, [.] "*Mejalles*" is a little more slang.

M00: Yeah, a little more crude.

F20: "*Mäakjes*" is a little more proper.

(2011-08-09 (02), 00m29s906–00m40s570)

Managing the data from this project in this way also contributed to addressing a number of the questions that motivated this work in the language community, as well. The results of the analysis of variation described herein pointed to a relatively small number of subgroups of speakers within the larger Plautdietsch-speaking community in central Saskatchewan who shared similar constellations of linguistic features. In some cases, this mirrored the intuitions of members of the language community: at least two groups of Plautdietsch varieties were widely reported within the community, although the exact nature of the linguistic differences between them wasn't always clear. Having serviceable information about a wider range of the varieties present in these communities, as well as the linguistic and sociodemographic features that typically characterized them, left us in a better position to make informed decisions about how language resource materials and language programs might be developed (e.g., in which varieties, and with which possible contributors). It also made it much simpler to see how individual *Fibel* responses, annotated in ELAN, could be transformed back into printable and/or online learning resources, merging the text and audio for individuals' responses in the ELAN transcripts with the templates that were developed for the *Fibel* to create copies that presented each contributor's translations.

It is common in many language documentation projects for teams to make sure that the assembled materials are and remain accessible to members of the contributing language community, both in the short and long terms. As mentioned above, while copies of all of the recorded sessions were returned to the contributing speakers, this in itself did not guarantee that the larger collection of language resources would be available to the larger language community in the area. Instead, arrangements were made for a complete copy of all of the materials from this project to be deposited with the archives of a local Mennonite historical society, which has a mandate to preserve and facilitate access to records such as these for community use into the future. Having this kind of support for long-term, local access was important to a documentation project like this that aimed to support both linguistic research and community language initiatives. It would also be helpful to have copies of these materials archived with an institution whose mandate also included facilitating discovery of and access to these resources for a wider, potentially non-local audience as one way of encouraging further research on Plautdietsch, although this would require further discussion with members of the language community to determine how this could best be accomplished (especially as there are presently no publicly accessible language archives in Canada, meaning that these materials would either need to be entrusted to a suitable organization outside of the country, which may be of concern to some contributors, or to one inside of Canada that does not participate in the wider community of language archives).

In this project, investing the effort required to develop language resources in ways that reflected our understanding of current recommended practices in documentary linguistics—creating our recordings using non-proprietary, uncompressed formats on equipment that would produce relatively high-quality results; favoring open standards and open source software tools wherever possible for managing our recordings, metadata, and annotations; and trying to keep materials organized consistently when it came to directory structures and file names—made many of the later uses described above not only possible, but actually practicable. While we would recommend that documentation teams consider the possible benefits that implementing similar data management practices may have in their own contexts, it is also important to recognize that not every team may necessarily find itself in a position to bring all of these recommendations to bear immediately on their own work. Although the situation is rapidly improving with step-by-step guides, learning resources, and training opportunities becoming available to documentation teams around the world, training and resources that support these kinds of data management practices are still unevenly distributed. We would likely do well to heed the observation made by Carpenter et al. (2016:4) that "unworkable standards and a dogmatic insistence on 'best practices' in digital technologies and language documentation, set by scholars and funding agencies, can have a disempowering effect on individuals and communities" and consequently take care to ensure that data management practices such as these are shared and implemented in ways that support and amplify the efforts of documentation teams that choose to adopt them.

Notes

1. The median age of contributors to this project was 79 (with a standard deviation of 8.7 years), which gives a sense of the demographic skew that had resulted from language shift in the region.

2. As one reviewer noted, it may be preferable to choose directional (e.g., cardioid or hypercardioid) microphones for capturing spoken language under circumstances such as these, both to reduce the amount of ambient noise that is recorded and to improve the degree of separation between individual speakers' voices. We agree, although it may be important to weigh these benefits against other possible trade-offs. Directional microphones sometimes exhibit proximity effects that amplify low-frequency sounds, which may affect the usability of recordings made with this kind of equipment in later acoustic analyses of phonetic features such as nasality and voice quality (cf. Plichta 2010).

3. This documentation project did not involve any video recording. While this may not have been critical to the parts of these consultation sessions that involved reviewing and translating the *Fibel*, it would almost certainly have been worthwhile for the conversations that were frequently recorded before and after this work, where being able to see each of the contributors as they spoke would no doubt expand the range of possible future uses of these materials (and likely make them more engaging on a personal level for future users, as well). If we had the chance to do this project again, I would hope that it would be possible to incorporate video to the degree that contributors thought it was appropriate.

4. Since none of these devices offered any safeguards against bit rot or other forms of data corruption over time, we also produced both MD5 and SHA-1 checksums of each of the audio files immediately after they were renamed and organized, storing them in the same folder as the corresponding recordings (the files ending in .md5 and .sha1). Although not a perfect solution, these "digital fingerprints" provided at least one way for us to confirm the integrity of the audio recordings over time by checking that each checksum algorithm returned the same value for each audio recording as it did when it was first applied shortly after the recordings were made. A copy of all of the project materials was later made on a network-attached storage system that performed automatic integrity checking and error correction and that maintained an additional off-site mirror of these files onto a remote computing cluster that offered nightly backups.

5. https://www.audacityteam.org/.

6. https://ffmpeg.org/.

7. Although this sounds attractive in theory, in practice, creating time-aligned annotations for the dozens of hours of audio recorded in this project was no small task. This "transcription bottleneck" is well known in the documentary linguistic literature (see Reiman 2010; Boerger 2011; Seifart et al. 2018:e335–e336; Himmelmann 2018; among others) and posed a serious logistical challenge to the overall success of this project. While recent work drawing on techniques from natural language processing and computational linguistics aims to lessen this burden (e.g., Cox, Boulianne, & Alam 2019), in this project, we were fortunate to have the help of two dedicated annotators, Adrienne Findlay and Chelsea Cox, who assisted in creating empty annotations around segments of speech for each of the contributors in each recording, which ultimately made the transcription and coding described here feasible.

8. At the outset of this work, we coded instances of variation by hand, relying on ELAN's multitranscript regular expression search facilities to retrieve particular variants in the transcribed text of all of our transcripts. This turned out to be quite time-consuming, because ELAN offered limited facilities for users to batch edit the results returned from these kinds of queries. In the end, we developed two small scripts that helped with this work: one that performed the same kinds of regular expression searches as ELAN did, saving the results into a Unicode-encoded CSV spreadsheet together with cross-linked audio clips that could be reviewed and edited quickly and another that reintegrated the contents of these edited spreadsheets back into the ELAN transcripts from which they had been drawn (see Cox 2015:136 for details). The final result of using these two scripts was the same as if we had done all of this annotation inside of ELAN—a fully time-aligned spoken corpus with all identified instances of variation tagged as such—but took a fraction of the time it otherwise would have to accomplish.

9. It would have been possible at the time when this project was underway (and even easier now) to integrate this analysis directly into the academic writing that presented it using tools such as Sweave (Leisch 2002) and knitr (Xie 2015), which allow snippets of "live" code to be embedded into documents. Tools like these can be used to help ensure that the output of analyses stay in sync with the contents of a documentary corpus when the latter is still being actively corrected and expanded (which is often the case in documentation projects, where annotations are continually being refined as the corpus is used and understandings of the language develop), making reproducible research practices easier to implement in this context (cf. Berez-Kroeker et al. 2018; Berez-Kroeker et al., chapter 1, this volume, for further discussion of reproducible research in linguistics).

References

Austin, Peter K. 2006. Data and language documentation. In *Essentials of Language Documentation*, ed. Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 87–112. Berlin: Mouton de Gruyter.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18. doi:10.1515/ling-2017–0032.

Bird, Steven, and Gary Simons. 2003. Extending Dublin Core metadata to support the description and discovery of language resources. *Computers and the Humanities* 37 (4): 375–388. doi:10.1023/A:1025720518994.

Boerger, Brenda H. 2011. To BOLDly go where no one has gone before. *Language Documentation and Conservation* 5:208–233. http://hdl.handle.net/10125/4499.

Bowern, Claire. 2015. *Linguistic Fieldwork: A Practical Guide*. 2nd ed. New York: Palgrave Macmillan.

Broeder, Daan, Freddy Offenga, Don Willems, and Peter Wittenburg. 2001. The IMDI metadata set, its tools and accessible linguistic databases. In *Proceedings of the IRCS Workshop on Linguistic Databases*, ed. Steven Bird, Peter Buneman, and Mark Liberman, 48–55. Philadelphia: Linguistic Data Consortium.

Broeder, Daan, Menzo Windhouwer, Dieter van Uytvanck, Thorsten Trippel, and Twan Goosen. 2012. CMDI: A component metadata infrastructure. In *Proceedings of the Workshop on Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources*, ed. Victoria Arranz, Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Monica Monachini, and Thorsten Trippel, 1–4. Istanbul: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf.

Carpenter, Jennifer, Annie Guerin, Michelle Kaczmarek, Gerry Lawson, Kim Lawson, Lisa P. Nathan, and Mark Turin. 2016. Digital access for language and culture in First Nations communities. Knowledge Synthesis Report. Vancouver: Social Sciences and Humanities Research Council of Canada. http://www.ideas-idees.ca/sites/default/files/sites/default/uploads/general/2016/2016-sshrc-ksg-turin_et_al.pdf.

Cox, Christopher. 2015. Quantitative perspectives on variation in Mennonite Plautdietsch. PhD dissertation, University of Alberta. http://hdl.handle.net/10402/era.40446.

Cox, Christopher, Gilles Boulianne, and Jahangir Alam. 2019. Taking aim at the transcription bottleneck: Integrating speech technology into language documentation and conservation. Paper presented at the 6th International Conference on Language Documentation and Conservation (ICLDC), University of Hawai'i at Mānoa, Honolulu, HI, February 28–March 3. http://hdl.handle.net/10125/44841.

Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation and Conservation* 3 (1): 15–50. http://hdl.handle.net/10125/4423.

Czaykowska-Higgins, Ewa. 2018. Reflections on ethics: Re-humanizing linguistics, building relationships across difference. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 110–121. Language Documentation and Conservation Special Publication 15. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24813.

Dorian, Nancy C. 2010. *Investigating Variation: The Effects of Social Organization and Social Setting*. Oxford: Oxford University Press.

Dyck, Cornelius J. 1993. *An Introduction to Mennonite History*. 3rd ed. Scottdale, PA: Herald Press.

Gawne, Lauren. 2018. A guide to the Syuba (Kagate) Language Documentation Corpus. *Language Documentation and Conservation* 12:204–234. http://hdl.handle.net/10125/24768.

Guenter, Jacob G., Leonard Doell, Dick Braun, Jacob L. Guenther, Henry A. Friesen, Jacob W. Loeppky, John P. Doell, et al., eds. 1995. *Hague-Osler Mennonite Reserve: 1895–1995*. Saskatoon, Canada: Hague-Osler Reserve Book Committee.

Henke, Ryan, and Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation and Conservation* 10:411–457. http://hdl.handle.net/10125/24714.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36 (1): 161–195. doi:10.1515/ling.1998.36.1.161.

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation* 6:187–207. http://hdl.handle.net/10125/4503.

Himmelmann, Nikolaus P. 2018. Meeting the transcription challenge. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. Language Documentation and Conservation Special Publication 15. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24806.

Lee, Nala Huiying. 2013. Review of Shure WH30XLR cardioid headset microphone and Countryman E6 omnidirectional earset microphone. *Language Documentation and Conservation* 7:177–184. http://hdl.handle.net/10125/4595.

Leisch, Friedrich. 2002. Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat 2002: Proceedings in Computational Statistics*, ed. Wolfgang Härdle and Bernd Rönz, 575–580. Heidelberg: Physica Verlag. https://doi.org/10.1007/978-3-642-57489-4_89.

Lüpke, Friederike. 2009. Data collection methods for field-based language documentation. *Language Documentation and Description*, vol. 6, ed. Peter K. Austin, 53–100. London: School of Oriental and African Studies. http://www.elpublishing.org/PID/071.

McDonnell, Bradley, Andrea L. Berez-Kroeker, and Gary Holton, eds. 2018. *Reflections on Language Documentation 20 Years after Himmelmann 1998*. Language Documentation and Conservation

Special Publication 15. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24800.

Meakins, Felicity, Jennifer Green, and Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. New York: Routledge.

Nagy, Naomi, and Miriam Meyerhoff. 2015. Extending ELAN into variationist sociolinguistics. *Linguistics Vanguard* 1 (1): 271–281. doi:10.1515/lingvan-2015–0012.

Nathan, David. 2010. Sound and unsound practices in documentary linguistics: Towards an epistemology for audio. In *Language Documentation and Description,* vol. 7, ed. Peter K. Austin, 262–284. London: School of Oriental and African Studies. http://www.elpublishing.org/PID/088.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap—a web application for dialectology. *Dialectologia: Revista electrònica*. Special Issue 2, 65–89.

Plichta, Bartłomiej. 2010. Microphones used in recording speech. *AKUSTYK* (website, via Wayback Machine Internet Archive). https://web.archive.org/web/20130406095648/http://bartus.org/akustyk/microphones.php. Accessed July 21, 2013.

Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation and Conservation* 4:254–268. http://hdl.handle.net/10125/4479.

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4:123–155. doi:10.1007/s10805-006-9016-2.

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation and Conservation* 9:237–267. http://hdl.handle.net/10125/24639.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the British Sign Language Corpus. *Language Documentation and Conservation* 7:136–154. http://hdl.handle.net/10125/4592.

Seifart, Frank, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language* 94 (4): e324–e345. doi:10.1353/lan.2018.0070.

Sloetjes, Han. 2014. ELAN: Multimedia annotation application. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 305–320. Oxford: Oxford University Press.

Thieberger, Nicholas, and Andrea L. Berez. 2012. Linguistic data management. In *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger, 90–118. Oxford: Oxford University Press.

Woodbury, Anthony C. 2003. Defining documentary linguistics. In *Language Documentation and Description*, vol. 1, ed. Peter K. Austin, 35–51. London: School of Oriental and African Studies. http://www.elpublishing.org/PID/006.

Woodbury, Anthony C. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

Xie, Yihui. 2015. Dynamic documents with R and knitr. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC. https://yihui.name/knitr/.

This is a section of doi:10.7551/mitpress/12200.001.0001

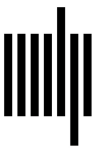
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363