CHAPTER 13

Research Methods in Recording Oral Tradition: Choosing Between the Evanescence of the Digital or the Senescence of the Analog

Nick Thieberger

In this chapter, I present methods for creating primary research records in ways that can be archived and reused in the future, with a focus on linguistic fieldwork, but with principles that apply across a range of humanities disciplines. Research is stronger if primary data can be accessed and cited so that readers can verify that the source material actually exists and so that they are able to apply their own analysis to it. This relies on the data being deposited and curated in an archive that guarantees access over time. Our research group in Australia is preserving records of intangible cultural heritage[1] in the world's small languages, by building the Pacific and Regional Archive for Digital Sources in Endangered

[1] See for example, UNESCO's page on intangible cultural heritage, https://ich.unesco.org/en/home.

N. Thieberger (✉)
The University of Melbourne and ARC Centre of Excellence for the
Dynamics of Language, Melbourne, Australia
e-mail: thien@unimelb.edu.au

l. levenberg et al. (eds.), *Research Methods for the Digital Humanities*,
https://doi.org/10.1007/978-3-319-96713-4_13

Cultures (PARADISEC.org.au). As the collection grows in size (currently 45 terabytes) and number of languages represented (currently just over 1170) we have to seriously consider the long-term viability of the collection. But the most urgent issue that faces us is to locate and digitise recordings that will not otherwise be preserved. Many of these are field recordings made by past generations of linguists, musicologists and ethnographers, from a time when emphasis was placed on the analysis of the contents of the recordings, not on preservation of the recordings themselves. As we will see, archiving is no longer an end-of-career activity, relegated as it used to be to your retirement or to your executors after you die. Archiving is now central to research methodology, creating citable records that allow research to be contextualised and verified. As Barwick notes, we are turning old ideas upside-down and this means we are training new researchers to think about the quality of the records they produce, including the content, file-naming, formats, metadata, and equipment they use.[2] There is much in all of this work that is applicable more broadly to Humanities research practices.

Our particular focus is on records of small languages for which there is otherwise little information available, and so the work of a linguist recording speakers of the language becomes all the more important. Keep in mind that there are some 7000 languages in the world and that for most of them there are few, if any, records. Thus, making recordings in the course of linguistic fieldwork becomes a critical point at which not only can scholarly work be done, but basic records of performance in another of the world's cultures can be created. The records have intrinsic value for the speakers and their descendants as well as for linguistic research. Hence there is a great responsibility, for linguists who create them, to manage these records properly.

Research and Archiving

The research questions for field linguists today focus on what each new language has to tell us about the range of possible variation in the structures and communication strategies involving language. There is a recognition that our own use of our research data will never be exhaustive, and

[2] Linda Barwick, "Turning It All Upside Down … Imagining a Distributed Digital Audiovisual Archive," *Literary and Linguistic Computing* 19, no. 3 (2004): 253–263.

that creating well-annotated and described collections will allow others to continue working with this material. I prepared the data that I created in my research on Nafsan (a language of South Efate, Vanuatu), so that it could be reused by providing a finding aid[3] to provide contextual information about the files. Other researchers are now using this material and focusing in more detail than I did on aspects of the language. They can only do this because I created and archived all the material with future reuse in mind.

Archiving is central to scholarly research. Using archives to locate primary sources is normal research practice, but it is less common for researchers to think of their own work as being an archivable resource. If your work creates new documents (be they textual or media) or touches manuscripts that no one has used or that you are providing with a new interpretation, then that interpretation and those documents need to be made available to the audience you are writing for. Creating well-formed ("archive-ready") research materials makes them reusable, and archiving ensures that you can access your own materials and that others can benefit from work you have done. In order for this virtuous work practice to be achieved, certain steps need to be put in place. These were recently summarised by an international consortium of researchers in the FAIR[4] principles, which suggest that records should be findable, accessible, interoperable, and reusable. While there is a clear case for making good records of small and maybe endangered languages, given the rarity of records for most such languages, the same methods can be applied to most humanities research. There is a logical workflow in making recordings, applying metadata, naming files systematically, transcribing files and so on, so that the files can be archived without much effort on the part of the researcher.[5] As with most Digital Humanities methods, we emphasise reusability of the research materials. For us, this means using uncompressed formats (e.g., tif rather than jpg, wav rather than mp3), providing a good description and getting licenses from the people recorded so that it is clear how the research can be used by others.

[3] http://www.nthieberger.net/sefate.html.

[4] https://www.force11.org/group/fairgroup/fairprinciples.

[5] Nicholas Thieberger and Andrea Berez, "Linguistic Data Management," in *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger (Oxford: Oxford University Press, 2012), 90–118.

The Virtue of Archiving

Fixing a performance, opportunistically captured, as an archival recording risks privileging that recording so that it can be seen as providing an authority that it was never meant to have. From an archive's point of view, these are predictable problems in the literature[6] and accepted as the necessary corollary of keeping records. Similarly, we know that archives are extremely partial, representing those with the power to make records and preserve them.[7]

If scholars work with primary materials and use them as the basis for their analysis, then the reader should be able to verify claims by going back to those primary materials. It used to be the case that the (analog) records sat in an office and were only available by visiting the researcher[8] and it is usually still the case that digital records can only be obtained in the same way. New research methods, including archiving, free the original researcher from dealing with such requests and allow her/him to specify conditions under which the materials can be used, that is, providing a licence for their use. A license can be as simple as a statement that the records can be used only for educational purposes. Rather than making up a range of different licenses, it is easiest to use an existing system, like, for example, Creative Commons.[9] These internationally recognised licenses are flexible and cover a range of options that you, as the creator of the work, can select from. You could think of questions like the following:

- Are you allowed to make copies?
- Did you get a release form from the person you recorded?
- Did the library where you copied them allow you to enrich them and make them available to others?

[6] Jacques Derrida, *Archive Fever: A Freudian Impression* (Chicago, IL: University of Chicago Press, 1996); Terry Cook, and Joan M. Schwartz, "Archives, Records and Power: From (Postmodern) Theory to (Archival) Performance," *Archival Science* 2, no. 3–4 (2002): 171–185, http://dx.doi.org/10.1007/BF02435620.

[7] Linda T. Smith, *Decolonizing Methodologies: Research and Indigenous Peoples* (Dunedin: Zed, 1999).

[8] For a humorous but nevertheless accurate portrayal of this situation, see this video produced by the TROLLING (https://dataverse.no/dataverse/trolling) archive in Norway: https://www.youtube.com/watch?v=uEf0c0NT9.

[9] https://creativecommons.org/.

- Was the original you made of sufficient quality to allow other users or is it too compressed and poorly captured to be used again?
- Did you make enough backups and have you archived the files?
- Did you give the files you created unique names that allow you to find them again?
- Do you have a description of what is in the files so you do not need to open each file every time you want to look at or listen to it again?

Each of these questions suggests simple issues of data management. Fortunately, there are well-established standards for file formats and data management[10] that linguists can adopt. But even if these files do make it into a curated repository, how long is digital data going to last? There is no proven storage medium, so the key to preservation of data is migration to the next storage system. Given that all research data will need to be properly curated over time, we are confident that national storage infrastructure will soon become the norm and that our collection will become part of that larger effort. Academic institutions will increasingly provide archival storage for primary research data, but you may need to curate your collections in some more temporary solution until a suitable archive is available.

In our experience of working with many legacy collections (those made in the past by retired or deceased researchers), we have learned what it takes to get from an undifferentiated box of tapes to a well-organised collection. If a researcher is still able to provide information about their tapes, archiving them is usually the first time that they compile their notes in an organised way. This creates a catalog that lists enough information for someone else to make sense of the item that they have described. It is very rare for collections of recordings to have even the most minimal metadata or cataloging information. In some cases, in our work on archiving language records, we have preserved tapes for which there is no metadata, but we assume they are worth digitising because of the value of the collection they are part of. At some point, we intend to put this kind of unannotated material online for crowdsourced annotation (see e.g., Zooniverse[11] as a platform for online annotation). An example of a collection of primary records we have worked with was

[10] Louise Corti et al., *Managing and Sharing Research Data: A Guide to Good Practice* (London: Sage, 2014), 56.

[11] https://www.zooniverse.org/.

created by Arthur Capell, who was a professor of linguistics and a prolific collector of information in many languages. When he died in 1986, he left many boxes of papers and audio recordings. We put 15,000 pages of Arthur Capell's notes online[12] (but untranscribed) with sufficient metadata to make them locatable and archived the same collection with the same metadata to ensure longevity of access.[13] Another current project is a set of 24,000 pages of manuscripts and typescripts created by the ethnographer Daisy Bates since 1904, which represents information about many Western Australian Aboriginal languages. We are now working to put these page images online[14] with textual versions of some 4000 pages. This is more complex than the Capell project in that it requires typing the manuscript text (and encoding it in XML), but it has the benefit of allowing the text to be searched, always with the ability to see the original page image, so in each of these cases a user can cite the original document which is a necessary step for scholarly practice.

Metadata for Findability

Metadata is the cataloging information about an item that typically includes basic information about what it is, when and where it was made, by whom and with whom. Within the PARADISEC collection, there are hundreds of recordings that have only the information written on their tape covers as metadata. In some cases, this is a single and cryptic word that would clearly have made sense at the moment it was written but is not particularly informative about the contents of the recording for anyone else now searching the catalog. We have improved our methods as an academic discipline since these recordings were made and are in a better position to create materials that we can access ourselves in future. So, if you record someone telling a story or being interviewed, then noting down all these details will help you remember what is on the recording later. For example, at the beginning of a recording, I will say who is being recorded, on what date and where. If I know what they will be talking about I will summarise it here too. This means the recording has some internal identification in case it is separated from other metadata.

[12] http://paradisec.org.au/fieldnotes/AC2.htm.

[13] http://catalog.paradisec.org.au/collections/AC2.

[14] http://bates.org.au.

I also take note on paper of what the recording device has named the recording and also note the contents and this will later be typed into the metadata notes I keep for all my files. The archive you will deposit your files with can advise on the metadata they need, so that you can start getting your catalog into the right shape to meet the archive's requirements. For example, each archive has its own notion of what an "item" is and how much information should be provided for each item. Some standard metadata terms (typically modeled on Dublin Core,[15] the librarian's standard) include: "Title" of the item; "Role" of the people involved; "Date" recorded (in a specific format); "Coverage", or what geographic area does it cover; "Content language", or the language used in the item; "Subject language", or what language the item is about (these last two use the ISO-639-3 code for language identification); and, "Roles" of people involved (speaker, singer, writer etc.).

Any files you create in the course of your research should be uncompressed and high resolution to allow them to be used in various ways later on. If you are going to make copies of a manuscript that only exists on paper or microfilm in one library then consider making good quality copies and taking careful notes about the context of the images (the library identifier or the pages and book title), so that you can relate them to the originals later on.

How to Name Files

File-naming is one of the simplest first steps for managing research data and a lack of proper file-naming can result in a loss of data. If you use a device like a camera or recorder then it will typically name files sequentially, often restarting the same names every month (like STE-0023, STE-0024, and so on). You need to change these to a unique name (like 201702-0023 for example) and then never change them again! All of your metadata will then relate to that filename and, when the file is archived, it should ideally retain that name so that you can cite it in the course of your own research. And, the case of the name is critical. If you use uppercase in the filename then the metadata listing of that filename has to be identical. This is because when you copy your metadata to an archive's catalog, the name will be the lynchpin between the

[15] http://dublincore.org.

metadata and the file in the archive. A mismatch in case will mean the file is not related to the metadata. The Bates project mentioned earlier includes over 24,000-page images that each needed to be named according to a principle that relates a page back to its original document in the National Library of Australia. In my (2006) PhD dissertation, I was able to cite each example sentence in my grammar back to its source in archival digital audio files and to provide them on a DVD with the book.[16] The archival references will continue to work after the DVD is no longer playable, but only because the same filenames are used in the book as are used in the archive.

All of these methods rely on training new researchers with better methods for recording, annotating, transcribing, and curating their research data. As part of our work, we run regular training sessions in linguistic data management built into normal fieldwork practice. The ability to develop this research infrastructure further depends on the degree to which academic institutions recognise that holding such repositories enhances their prestige and hence requires their support. In our case, it helps that PARADISEC has been listed by the UNESCO Register of Intangible Cultural Heritage for the Memory of the World and has been awarded the European Data Seal of Approval. Our main problem is how to make sure this digital material can be shepherded through the next period and survive the evanescence that otherwise awaits.

As research produces more and more digital material, often of considerable heritage importance, the onus is on the research community to provide long-term repositories for this material. In fact, there is no other choice for preservation of analog recordings but to digitise them. In this chapter, I have addressed key factors in mitigating digital evanescence, so that our precious research materials will survive into the future. When it comes to media recordings, the evanescence of the digital is preferable to the senescence of the analog.[17]

[16] Nicholas Thieberger, *A Grammar of South Efate: An Oceanic Language of Vanuatu* (Oceanic Linguistics Special Publication, No. 33. Honolulu: University of Hawai'i Press, 2006).

[17] Also see the UCLA Library's guide to audio archiving: http://guides.library.ucla.edu/ethno/ethnomuc200.

References

Barwick, Linda. “Turning It All Upside Down … Imagining a Distributed Digital Audiovisual Archive.” *Literary and Linguistic Computing* 19, no. 3 (2004): 253–263.

Cook, Terry, and Joan M. Schwartz. “Archives, Records and Power: From (Postmodern) Theory to (Archival) Performance.” *Archival Science* 2, no. 3–4 (2002): 171–185. http://dx.doi.org/10.1007/BF02435620.

Corti, Louise, Veerle van den Eynden, Libby Bishop, and Matthew Woollard. *Managing and Sharing Research Data: A Guide to Good Practice*. London: Sage, 2014.

Derrida, Jacques. *Archive Fever: A Freudian Impression*. Chicago, IL: University of Chicago Press, 1996.

Smith, Linda T. *Decolonizing Methodologies: Research and Indigenous Peoples.* Dunedin: Zed, 1999.

Thieberger, Nicholas. *A Grammar of South Efate: An Oceanic Language of Vanuatu*. Oceanic Linguistics Special Publication, No. 33. Honolulu: University of Hawai‘i Press, 2006.

Thieberger, Nicholas, and Andrea Berez. “Linguistic Data Management.” In *The Oxford Handbook of Linguistic Fieldwork*, edited by Nicholas Thieberger, 90–118. Oxford: Oxford University Press, 2012.