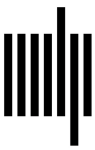
This is a section of doi:10.7551/mitpress/12200.001.0001

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 MIT Press

27 Managing Historical Data in the Chirila Database

Claire Bowern

1 Introduction

For as long as linguists have been researching language variation, typology, and change, they have been organizing their data. Historical linguistics is both one of the oldest areas of linguistics and one of the most cutting-edge. It is the focus of interdisciplinary work together with genetics, anthropology, and archaeology, a crucial source of information about the human past. Contemporary historical linguistics includes reconstruction and philology, but it also includes work involving computational approaches to historical linguistics, such as phylogenetics (see Bowern 2018; List, Greenhill, & Gray 2017; Greenhill, Blust, & Gray 2008).

This use case is based on the Chirila database, a database of the contemporary and historical lexical data for Australian languages. The project was begun in 2007 and is described in some detail in Bowern (2016).[1] This use case concentrates on the use of data in historical linguistic and language reclamation research. I briefly describe the database, before providing some discussion of the structure of the database in terms of data types in historical linguistics. Because the data structures for Chirila were already described in detail in Bowern (2016), I concentrate here on the consequences of decisions made early in the Chirila project. Many of these decisions have consequences for how the database can be used later on. This database is both repository and research tool and serves a varied audience. Here I use the metaphor of a "choose your own adventure" novel to describe the process of decision making for a complex research database; with twelve years of work on Chirila, we can now look back at where the "choices" went right and where they could (or should) have been different. In the next section, I discuss some of the issues that arise in historical linguistics research based on the points made in the contributions to part I of this volume.

2 General considerations

Historical databases can either contain single languages at different points in time, or be comparative, with multiple languages, possibly also from different periods. Chirila is a comparative database, containing both multiple languages and records that span two centuries.

2.1 Overview of the database

First, following Good (chapter 3, this volume) I describe the "scope" of a database. Chirila is a set of relational databases, currently stored in FileMaker Pro, along with limited web viewing. The rationale for using FileMaker Pro was discussed in Bowern (2016). However, so much of the data pipeline is now done outside of FileMaker, I am investigating the feasibility of moving the database entirely to a plain text file + Python and R. See also sections 2.4 and 4 for further discussion. A summary of the database structure is given in figure 27.1.

The database covers primarily lexical data; however, there is also a collection of grammatical features and a secondary database of geocoded languages that includes information about other linguistic ontologies (Glottolog, ISO-639, and the AIATSIS[2] language codes), georeferencing, and classification. Its uses include comparative, typological, and historical research involving both traditional comparative and phylogenetic methods.

A lexical database was chosen because so many of the languages are recorded and exhibited only in the form of word lists. If the aim is continent-wide consistent comparison, it needs to be word lists, because that is the only data available for much of the country. Contexts of use were never recorded for these languages. However, that does not prevent us from using ancillary data to check different hypotheses. For example, we can generate ideas about phonological structure from studying the sound systems of a few languages.

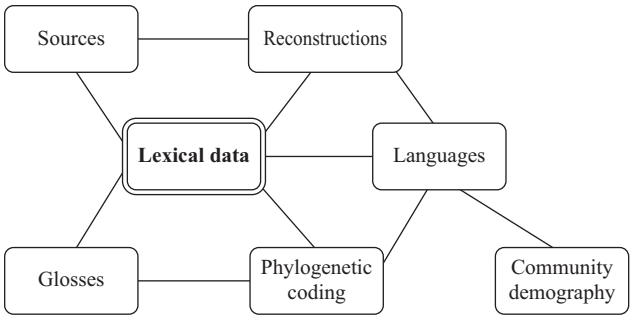


Figure 27.1
Core data structures for the Australian lexical database (from Bowern 2016).

Chirila includes both public and restricted data. The restrictions arise mostly from archives. Several archives place restrictions on reuse of data; one Aboriginal community preferred their language material be not shared, and several linguists placed limits on data use pending community consultation and permission. These restrictions have been respected.

A few sample entries from the linking databases are given in table 27.1. It shows the kinds of data that are collected from original sources, the processing of the original data, and the linking to other material within the databases.[3] The fields are further explained in Bowern (2016).

Chirila also contains grammatical data. Ninety languages were coded for 250 structural features in morphology, semantics, and syntax. These features were based on a combination of the Constenla Umaña (1991) typological survey, WALS features, and the Sahul project "pioneers of Island Melanesia."

Holdings of Chirila (as of September 2019), include 1,140 sources, 474 standard languages with more than 50 items, and 2,988 "variety" names (not necessarily distinct varieties, but records of language/variety names across sources). There are 874,165 items in the lexical database.

The database has been used over the last ten years for research involving structural reconstructions of phonology, morphology, and semantics on the one hand, and language change and population spreads using lexical data on the other.[4] Its metadata have been used to make public language maps. It has also been used as an informal research tool for sourcing early word attestations.

Crosscutting the interlocking structures illustrated in figure 27.1 and table 27.1 are the research strata in the database. These strata do not concern the type of data directly (in the sense of lexical data or morphology data types). Rather, they concern the relationship between the data items and their place in the research process, more akin to the distinction between primary and secondary data or the documentation materials versus analytical materials described by Himmelmann (1998) and subsequent work. These strata are described next.

First, there is the part of the database that is concerned with the keeping of the actual *primary* data as represented in the original source materials. This part of the database aims to reproduce the original language items as faithfully as possible, including the glosses and orthographic representations from the original source. For example, if a word for the bird of the genus *Dacelo* is glossed as "laughing jackass"—rather than the much more common and contemporary word "kookaburra"—the original gloss would be recorded. This is illustrated in table 27.1 by material in the columns labeled "original." This component of the database is primarily archival.

Another set of fields comprise the *regularization* of the primary data. This allows systematic comparison of disparate sources through the regularization of forms. Producing standardized forms from language-specific orthographies is automatic in some cases, but not for the majority of nineteenth- and early twentieth–century sources.[5] Even quasi-automatic regularization and conversion requires specialist knowledge of the language and documentation conventions of different regions (see, for example, Thieberger 1995). At the very least, it involves conversion of orthographies. Glosses are also problematic (see Bowern 2016 and Everson et al., in prep., for examples specific to Australia). In my previous example, "laughing jackass" is standardized to "kookaburra," because the latter is the term that is more widely used and much more likely to be searched for. An example of orthography standardization is given in the Ngiyambaa example in table 27.1.

Han (chapter 6, this volume) discusses data transformation, which as can be seen is a crucial aspect of making data in Chirila comparable. Note also that evaluating the accuracy of data import and transformation often also requires specialist knowledge; automatic computational validation of entries alone will not catch errors. Automated validation lets us see whether a field is empty when it should be filled, IDs are not unique, or a variety is assigned to multiple standard languages. But it won't tell us if a form has been standardized incorrectly.

Table 27.1
Schematic sample entries for different tables in the Chirila database

Data

Variety	[a]Std language	Original gloss	[b]Std gloss	Original form	Std form	Part of speech	[c]Abbrev source
Bardi	Bardi	boat	boat	yandilybar(a)	yantilypara	N	akl99
Ngarluma	Ngarluma	boat, ship	boat	yandilybara	yantilypara	N	ASEDA0037
Biri	Biri	boat	boat	gumbara	kumparra	N	Terrell
The Lower Macleay River	Ngiyambaa	boat	boat	olpino	wulpinu [unc]	N	Curr185
DDY	Gumatj	boat, ship	boat	mitjiyaŋ	mityiyaŋ	N	Zorc
DDY	Gumatj	large boat, steamship	ship	gapala(')	kapala(ʔ)	N	Zorc

[a]Language

Standard language name	ISO639	Glottolog	AIATSIS code	Family	Latitude	Longitude	Number of entries (all varieties)
Bardi	bcj	bard1255	K15	Nyulnyulan	-16.5	122.93	9801
Ngiyambaa	wyb	ngiy1239	D22, D18	Pama-Nyungan	-31.28	146.05	2589
Gumatj	guf	guma1253	N141	Pama-Nyungan	-12.25	136.8	4949

[b]Gloss

Standard gloss	Entry count	Part of speech	Wordnet form	Concepticon name	Semantic field(s)	Sutton/ Walsh (1980) category	Cognate code
boat	201	N	boat.n.01	BOAT	water, material culture	D	—
ship	94	N	ship.n.01	SHIP	water, material culture	D	—

[c]Source

Abbreviated source name	Author/ Compiler	Title	Access	Published	Reliability	Year	Number of items
akl99	Aklif, Gedda	Ardiyooloon Bardi Ngaanka: One Arm Point Bardi Dictionary	open	yes	1	1999	1547
Curr185	Spencer, Charles		open	yes	3	1886	215

Superscript letters show links between database tables.

Third, historical data, just like other data, require metadata. Metadata are the data about the data, the data about the languages in the sample, the source materials, the linguists who collected them, and the like (Good 2003). Metadata are part of the archival component of the database but may require additional research of themselves and may also be the subject of research.[6] Bowern (2014, 2015) describes some of the Chirila outputs based on the metadata research. Metadata for a historical database should also include documentation of the regularization conventions that were required for each data set. This was not done explicitly in the Chirila database, but there is a three-point scale for reliability of the transcription at the level of the source. A score of 1 is given to sources that are phonemically represented and likely to be accurate representations of the lexical items; 2 is given to sources that are regular, but which are idiosyncratic in some way (e.g., under-differentiated phonemes); and 3 is given to materials that are not consistently represented and whose phonological representation cannot be automatically inferred.

Finally, there is the comparison part of the database: this is the identification of cognate forms, the identification of loans, and the tracking of the aspects of language history that are crucial to doing research involving diachrony. This also requires expert knowledge of the data. Only at that point, once we have primary data, metadata, and cognate links, do we have the database forms that are the output for the historical research itself: that is, inferring trees, loan patterns, and other research. We might call this the iceberg principle of databases: that underneath the visible outputs of research questions are all the invisible decisions and codings that nonetheless impact how data can be obtained and used.

Ideally, the curatorial aspects of a database are kept distinct from research questions involving historical analysis of the data. I did not separate these parts of the database in earlier work, and they are now intertwined, making some aspects of data use unwieldy. For example, in 2013 I experimented with mapping the outputs of cognate sets by writing a script that produced kml-formatted entries for display in Google Earth. This required several columns that I put in the main lexical database, along with match field queries in the database structure. Although this mapping is now superseded (by scripts, in R), the match queries are now used in other database functions, and the mapping columns, though not actually used, cannot easily be removed.

2.2 Reproducibility and recoverability of hidden decisions

The contents of the Chirila data are primary data in that it is an original research contribution and feeds into the secondary analysis of language change. Such work does not clearly fit into a description versus documentation dichotomy (cf. Himmelmann 1998), as it is itself a combination of reporting and analysis, of compilation and synthesis in a loop that becomes more accurate as more knowledge is obtained. The intertwined nature of the data and analysis for historical databases has, in my opinion, been underappreciated in work on synchronic language documentation based on field methods, where authors have focused greater attention on the field to write-up "pipeline." One may also compare the Chirila database with the TerraLing/SSWL database (Koopman & Guardiano, chapter 55, this volume), which includes statements about grammatical features in languages but does not explicitly track how those decisions are made.

A further way in which the primary/secondary data distinction fails here is that diachronic databases contain material which is all, to a greater or lesser extent, *secondary*, in that the database records are abstracted representations of speech signals, presented in isolation.[7] These discussions have practical relevance when considering the analytical paper trail: how much manipulation of data is done, to what end, and how is that tracked and recovered if necessary (see also Han, chapter 6, this volume).

While data processing may be a pipeline, the decisions about *how* to process the data perhaps resemble a "choose your own adventure" novel. More technically, it resembles a process where prior choices shape the possibilities for subsequent work. Regularizing data, for example, allows comparisons to be made more easily and systematically. But irrecoverably regularizing means that it is then impossible to examine the ways in which words are variably represented in earlier sources. Another example of consequence is how to treat synonyms and multiple glosses or senses of a lexical item. Should the entry be split, so that each entry contains a single gloss or sense, or should all senses be in a single entry? The former makes it easier to tag disparate senses, to make it clear which words are recorded in some sentences but

not others, and to track semantic change. However, this comes at the expense of proliferation of entries, which becomes unwieldy in cognate coding. The alternative is to add another layer of structure (a "lexeme" layer) where cognates are marked, and which has senses and variant spellings as daughters.

Furthermore, it is vital to have some way of documenting decisions about such choices in analysis and considering what implications of those choices might be. For example, in Chirila, doculects are assigned to *reference languages* (similar to a standard language, allowing for typological comparison across varieties of roughly equivalent genetic distance).[8] In many cases, assigning a doculect to a reference language is straightforward; in others, it requires extensive additional research and may rely on implicit assumptions about the accuracy of underlying sources. If later work reveals that the doculect was mis-assigned, it can affect later analyses. This is particularly problematic where all the sources are sparse. For example, there are three vocabularies that are all tentatively assigned to the standard language Bigambal. Because they do not contain a lot of overlapping vocabulary, the cognate list for Bigambal includes words attested from only one of each of the three word lists. Where the word lists do overlap, they do not always have the same word. If later research shows that more than one language is represented in these vocabulary lists, this affects not just the mapping between varieties and languages, but cognate coding as well.

Gawne and Styles (chapter 2, this volume) and earlier Berez-Kroeker et al. (2018) describe a framework by which linguistic research, even if not fully replicable, should be reproducible at key points. That is, even if we cannot do all the steps to redo an analysis from scratch, it should be possible to trace all the steps and understand where all the data came from. This is also key for Chirila, where original sources are fully sourced, and major points in analysis are documented. Data sets like these are easier to use to replicate analyses; for example, as in the case of the discussion in PNAS between Haynie and Bowern (2016) and Nash (2017), and the reply at Bowern and Haynie (2017).

2.3 Summary: The "data science" of historical linguistics

This description of the database is somewhat similar to taking a data science view of the work of the database; that is, seeing the data as part of a research pipeline from the original research questions, to data questions and identification, to the actual analysis and its write-up. While I have treated the historical database as a single entity, in practice identical data sets may be used in very different ways depending on the particular research question. For example, Swadesh word lists have been used for Australian language historical research in studying the distribution of sound correspondences and sound change (Babinski & Bowern 2018); studying the phylogenetic signal in phonological inventories and phonotactics (Macklin-Cordes & Round 2015; Macklin-Cordes, Bowern, & Round, 2021; building a tree of languages (Bouckaert, Bowern, & Atkinson 2018; Bowern & Atkinson 2012); studying sound symbolism (Haynie, Bowern, & LaPalombara 2014); examining the distribution of loans and retentions across a large number of languages (Haynie et al. 2014); and public outreach.

3 Use case problems

The most challenging issues for this database have been the software components, an adequate pipeline for the many different types of data, and the separation of the research and simple data processing aspects of the database. Each of these issues is discussed in turn.

3.1 Software

Interfacing with pipelines for analysis requires other software. For example, the advantage of structured data of this type is that they can eventually feed in, possibly automatically, to tools for language analysis such as LingPy (List, Greenhill, & Forkel 2017; List, Greenhill, & Gray 2017). Statistical packages such as those in the R platform (R Core Team 2017) can use data from text formats and manipulate strings, but they are not very graphically user-friendly. FileMaker Pro, on the other hand, is user friendly for some types of operations. Complex searching, for example, is straightforward; building, changing, and manipulating joins in relational databases to construct complex queries is far easier in FileMaker than in SQL. Its spreadsheet visualization functions make it straightforward to use for manual data comparison, loan investigation, and lexical reconstruction using the principles of the comparative method (Weiss 2015). But these same advantages also bring disadvantages. For example,

it is difficult to undo operations in a FileMaker database. It is possible, with three mouse clicks, to irrecoverably delete all the data in a table (unless a backup has been made). It does not easily fit into a data pipeline, such as can be constructed in R or Python, where the original underlying data set is untouched and subsequently manipulated.

A reviewer of this chapter brought up the point of advocating for proprietary software and the dangers of using such software, which may cease to be maintained or may render data inaccessible when the software is upgraded. This is a serious concern. However, on these dimensions, free open source is not necessarily better. Having used open source software for many years, I see two issues that cause particular problems for language projects: the lack of stable development and problems with backward compatibility. For example, R packages that have become integral to a pipeline cease to be maintained and become incompatible with platform updates. While such issues do not affect the underlying data, they can render a project unusable or results undisplayable.

My choices for databases at the time were to use FileMaker (with its imperfections), learn SQL and web interfaces and create an architecture from scratch, or pay someone to create a database. A major advantage of FileMaker was that I could experiment with database structures while still doing language research; I did not have to wait for the database to be ready and be beholden to someone else's schedule for development (something I have subsequently found to be a problem for databases such as this). At the time, Chirila was unique—the models of other comparative databases either weren't exactly equivalent (such as the Austronesian Basic Vocabulary Database; cf. Greenhill & Gray 2008) or were not available to be viewed. I wanted to create something that would be an archive, a research tool, and a data and analysis collation tool together. Using FileMaker allowed both historical-linguistic and database structure research to take place organically. Given the options available, FileMaker was the right choice.

If I were starting again from scratch now, however, I would probably use plain text files, stored in .json format or an equivalent, with field manipulation through R or Python, and data viewing in something like a Jekyll site (that serves content from static pages). Alternatively, I would keep the master files as text or .json files and use FileMaker for the research coding. That is, I would use FileMaker as a data visualization end point, not as the main repository. However, I still believe that for most of what I do interactively with the database, there is no better solution than FileMaker.

In summary, whatever software is used for storage, analysis, and display, the principles of data sustainability apply. Data should be backed up, code should be backed up, and both in ways that retain accessibility as much as possible. Chirila holdings are frequently exported as Unicode plain text tables, and waypoints in the FileMaker databases are also saved before any major changes (and periodically at other times; see section 3.4).

3.2 Pipelines, bottlenecks, and "life cycles" of data

Three metaphors are in common usage in talking about data and its processing: *data pipelines* refer to the process of working with original data, processing and cleaning it and analyzing it for a specific purpose. *Bottlenecks* are points in that pipeline where the processing takes more time, leading to less data being available. For example, in the Chirila database, one bottleneck is the assignment of standardized English glosses to dictionary entries, which has to be done manually. Another is the digitization of handwritten notes. The *data life cycle* is a third metaphor and will be discussed further in section 3.4.

Data pipelines are structured by protocols. These are procedures for processing data in a particular way, in a predefined order, to make sure that all necessary steps are followed correctly. Protocols ensure that crucial steps are followed in the appropriate order and that data are treated consistently. However, it is difficult to get student researchers to keep to detailed protocols when many temporary student workers are involved. The training may end up taking more time than they spend on the data itself. This has meant for Chirila that some languages are better represented than others, simply because the original data were more accessible to student researchers, and that some protocols were not followed, meaning subsequent emendations were required. For example, one student did not follow instructions about not representing underlined characters with Excel underline markup; when the data were imported into the main database, all underlines (and the distinction between retroflex and apical consonants) were lost.

3.3 Research and data

For long-term and intensive database work, it is crucial to prepare for an *evolving* research project; that is difficult when the nature of the research means that future possibilities cannot be known. For example, when Chirila was first constructed, none of the software tools that I now most commonly use in my research were available. Both historical linguistics and linguistics more generally were considerably less quantitative. Being transparent and flexible in the data structures is crucial.

3.4 Ethics

Ethical considerations (Holton, Leonard, & Pulsifer, chapter 4, this volume) are everywhere. The ethical issues are complex, because individuals have different views of what should happen to their language data. Three particular issues that arise with Chirila are (1) consultation about data use, (2) negotiation with archives, and (3) sharing.

Language work in Aboriginal Australia typically proceeds with the assumption that many members of a language community will be consulted before work begins. This was simply not feasible with Chirila, given the number of languages involved and the number of languages without clear community presence. When compiling Chirila, I consulted with representative bodies, such as local language centers, and individuals at the 2007 inaugural Aboriginal Languages Conference in Adelaide. My aim was to do things ethically, but also with the concern that when negotiating access and publication where different people have different opinions, it's impossible to please everyone, and both restricting and releasing data have positive and negative ethical consequences. Too often, I think, we focus on the positives of restricting and the negatives of releasing, without considering the positives of releasing data.[9] Indeed, one of the primary community user groups of Chirila has been members of the Stolen Generations using the resource to find out more about their languages. Some have done this in part by looking up words (e.g., words for grandparents) that they remember from early childhood. By definition they would not have been able to do this if the availability of resources had been dependent on giving permission in advance.

In discussions about language data, Aboriginal people have stressed that consultation is important. The communities and individuals I have consulted with have been almost unanimously in favor of some type of data release. Linguists have usually also been in favor of data release. The exceptions have been mostly where linguists have not wished to give an answer on behalf of the community or communities they worked with and have needed time to consult them.

Data backup and curation are ethical issues, not just logistic ones. This database is a record of thousands of hours of work between Aboriginal communities and linguists and represents a major repository of cultural patrimony. Taking care of that material should not be taken lightly. Chirila is itself an archive of a sort. The solution for Chirila has been a version of LOCKSS (Lots of Copies Keeps Stuff Safe; Bird & Simons 2003). In addition to multiple copies of the FileMaker databases (backed up in Dropbox and Yale's institutional backup service and the FileMaker server), the material is privately backed up with Zenodo (zenodo.org), the data archive run by CERN.

Zenodo allows the publication and citation of data sets with DOIs, both increasing version control and making citations more visible and trackable (cf. Conzett & De Smedt, chapter 11, this volume). Chirila's citation issues are complicated, because someone working with Chirila data needs to be able to cite both the original source(s) and the Chirila database itself.

Finally, some work has recently discussed the data life cycle (e.g., Mattern, chapter 5, this volume). I'm not in favor of the metaphor of the life cycle for linguistic data of this type, given what it implies about end of cycle and how the very existence of this database shows that the metaphor is misplaced.[10] I also do not think that ending up in an inaccessible archive is good "end of life" care for data, particularly given the social backdrop against which so many Australian languages have been documented.

3.5 Recommendations

In summary, here are the recommendations of decisions taken for Chirila that turned out to be good ones, even if the choice was not obvious at the time. I also include some decisions that in retrospect were not good ones. This summary builds on the discussion of key points in this chapter.

Including both standardized and original data was a good decision. Some other databases have only presented standardized materials (or only original materials). Having both data types allows for new projects, such as studying the ways in which nineteenth-century researchers represented the sounds they were hearing. It

also made it possible to catch errors that would otherwise have gone undetected.

Where possible, separate the basic data from particular views or additions to the database that are used for particular projects. For example, I created some mapping fields which proved useful for a particular representation of some lexical data for a subset of languages, but these are in columns in the main lexical database. A better decision would be to pull the underlying database material into a new frame, using virtual links.

Geocoding both standard languages and varieties, where possible (Bowern 2017), was also very helpful in unforeseen ways. This has made it possible to do much more geospatial analysis of cognacy than was dreamed of when the database was first constructed. It has also made it possible to construct outreach materials (e.g., maps based on words for items in particular languages).

A problem that is still ongoing is how to represent recursive subgrouping in a hierarchical database. When I started compiling Chirila, we had family data, major subgroup data (within Pama-Nyungan), but no fully articulated tree (that came with Bowern & Atkinson 2012 and Bouckaert, Bowern, & Atkinson 2018). Hierarchical subgroups with different layers are not easy to represent in a table-based database such as this. Currently, I include family, subgroup, and major grouping as per Bowern and Atkinson (2012).

FileMaker does not make it straightforward to incorporate changelogs, but we do have some basic data tracking (such as the date and time last edited, and the username of the person who last modified the entry). This has proven invaluable on more than one occasion. Issue tracking is highly recommended. Document decisions and back up before and after major data changes. I once accidentally replaced all the cognate codes with language codes, destroying hundreds of hours of work. However, I was able to restore a prior copy of the database (from ten minutes before).

Consider the consequences of decisions not just for immediate projects but for all conceivable future projects. It is helpful to ask variations of the question "If I organize it this way, will I be able to do X, Y, Z?" For example, if I do this, how will I get data back out? How will I be able to analyze the forms with another software program? If I split all glosses, will I be able to recover the words that were glossed together in the original sources?

Considering carefully what properties are properly attributes of which pieces of data is crucial. For example, words can be geocoded so they can be represented on a map, but words are part of doculects, so the geocoded entity is technically the doculect, and words inherit their geocoding by which doculect they represent.

Most of the biggest problems have stemmed from areas of data processing where we lack good data management practices. One issue is that just as Chirila is a work in progress, so too are some of the data sources that underlie it. Language centers such as Wangka Maya and the Goldfields continue to work on languages and release updates, corrections, and new works. Some of those works are based in part on earlier editions. We do not want to include errors from previous sources, nor can we simply replace one source with its update.

Another area where problems have arisen is in language names. It has been difficult to standardize language names across all projects, leading to confusion and time lost in having to match languages to codes. For example, the first geocoding of polygons for language names (Bowern 2017) was done in Google Earth before some language names were standardized. Thus, the FileMaker database and the Google Earth files ended up with alternative spellings for some languages. Matching these needs to be done manually. Furthermore, there are three major standard codes for Australian languages (the AIATSIS codes, Glottolog, and the ISO-639 three-letter codes). These are not directly comparable, as they have slightly different coverage and classification.

4 Conclusions

Computational and statistical approaches that combine with historical linguistics require complex data management. The Chirila database has evolved over its twelve years of development as the field has changed; this has created both a test of flexibility in data structures and an illustration of the need to be careful about data curation decisions.

Notes

1. Although the database is still under active use and data are still being added and processed, the underlying structure of the database has been stable since 2013. Development of the database has been supported by NSF BCS-0844550 and BCS-1423711.

2. See https://collection.aiatsis.gov.au/austlang/search/.

3. While all links in the Chirila database are done with unique and persistent ID numbers, here only the labels are given. Fields are omitted for clarity; only the most common fields are given.

4. References include (Bowern 2011; Bowern 2012; Haynie & Bowern 2016; Bowern et al. 2014; Epps et al. 2012; Bouckaert, Bowern, & Atkinson 2018), among others.

5. There were 379 sources published before 1920, comprising 84,635 words.

6. For example, the information about how many standard languages are represented in the database is the main data for a chapter in the forthcoming *Oxford Handbook of Australian Languages* on how many languages were used in Australia at European settlement.

7. For my purposes, I distinguish original field documentation, published analyses (which might be based on preprocessed original documentation, such as Wafer, Lissarrague, & Harkins 2008; Ash, Giacon, & Lissarrague 2003; and others), and work that I have subsequently done that regularizes materials for the purpose of comparison. That is, we could distinguish "interpretation" from "regularization"; the latter is a substitution algorithm, whereas the former requires special linguistic knowledge of the particular language. Others may make different distinctions.

8. The term *doculect* is due to Good and Cysouw (2013).

9. To take one example, I have an informal data-sharing arrangement with the Resource Network for Linguistic Diversity (rnld.org), who run workshops on language reclamation for individuals who are members of the Stolen Generations and their descendants. Keeping language data highly restricted would further deprive those communities of their cultural rights. Is it truly ethical to say that members of the Stolen Generation cannot further discover details of their cultural heritage (which colonial institutions deprived them of) because of rules restricting access that are created by archives (another type of colonial institution)? Compare further projects such as http://www.decolonisingthearchive.com.

10. For further discussion of metaphors of death in documentary linguistics, see Perley (2012), Davis (2017), and Hill (2002).

References

Ash, Anna, John Giacon, and Amanda Lissarrague. 2003. *Gamilaraay, Yuwaalaraay and Yuwaalayaay Dictionary*. Alice Springs, Australia: IAD Press.

Babinski, Sarah, and Claire Bowern. 2018. Mergers in Bardi: Contextual probability and predictors of sound change. *Linguistics Vanguard* 4 (s2): 20170024. https://doi.org/10.1515/lingvan-2017-0024.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, and Stanley Dubinsky. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79 (3): 557–582.

Bouckaert, Remco R., Claire Bowern, and Quentin D. Atkinson. 2018. The origin and expansion of Pama–Nyungan languages across Australia. *Nature Ecology and Evolution* 2:741–749.

Bowern, Claire. 2011. Loans in the basic vocabulary of Pama-Nyungan languages. LSA Conference Presentation presented at the 85th Annual Meeting of the Linguistic Society of America, Pittsburgh, Pennsylvania, January 6–9.

Bowern, Claire. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences 279 (1747): 4590–4595*. https://doi.org/10.1098/rspb.2012.1842.

Bowern, Claire. 2014. Data "big" and "small"—Examples from the Australian lexical database. *Linguistics Vanguard* 1 (1): 295–303. https://doi.org/10.1515/lingvan-2014-1009.

Bowern, Claire. 2015. Pama-Nyungan phylogenetics and beyond. Presented as the Plenary Address to Leiden Lorentz Center Workshop on Phylogenetic Methods in Linguistics, Leiden, The Netherlands, October 26–30. https://zenodo.org/record/3032846.

Bowern, Claire. 2016. Chirila: Contemporary and historical resources for the Indigenous languages of Australia. *Language Documentation and Conservation* 10:1–44. http://scholarspace.manoa.hawaii.edu/handle/10125/24685.

Bowern, Claire. 2017. Files for Australian language locations. Zenodo. doi:10.5281/zenodo.848646. https://zenodo.org/record/848646. Accessed May 19, 2019.

Bowern, Claire. 2018. Computational phylogenetics. *Annual Review of Linguistics* 4 (1): 281–296. http://www.annualreviews.org/doi/full/10.1146/annurev-linguistics-011516-034142.

Bowern, Claire, and Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88 (4): 817–845.

Bowern, Claire, and Hannah J. Haynie. 2017. Reply to Nash: Color terms are lost, despite missing data. *Proceedings of the National Academy of Sciences* 114 (39): E8132–E8133. doi:10.1073/pnas.1714258114.

Bowern, Claire, Hannah Haynie, Catherine Sheard, Barry Alpher, Patience Epps, Jane Hill, and Patrick McConvell. 2014. Loan and inheritance patterns in hunter-gatherer

ethnobiological systems. *Journal of Ethnobiology* 34 (2): 195–227. doi:10.2993/0278-0771-34.2.195.

Constenla Umaña, Adolfo. 1991. *Las lenguas del área intermedia: Introducción a su estudio areal*. San José: Editorial de la Universidad de Costa Rica.

Davis, Jenny L. 2017. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. In *Language Documentation and Description*, vol. 14, ed. Wesley Y. Leonard and Haley De Korne, 37–58. London: EL Publishing.

Epps, Patience, Claire Bowern, Cynthia Hansen, Jane Hill, and Jason Zentz. 2012. On numeral complexity in hunter-gatherer languages. *Linguistic Typology*16 (1): 41–109.

Everson, Rebecca, R Tom McCoy, and Claire Bowern. In preparation. Standardizing glossing from dictionary sources: A case study from Australia. Unpublished manuscript, Yale University.

Good, Jeff. 2003. *A Gentle Introduction to Metadata*. UBIR. http://ubir.buffalo.edu/xmlui/handle/10477/38681. Accessed September 30, 2019.

Good, Jeff, and Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion "language." *Language Documentation and Conservation* 7:331–359. http://scholarspace.manoa.hawaii.edu/handle/10125/4606.

Greenhill, S. J., R. Blust, and R. D. Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics Online* 4:271–283.

Haynie, Hannah J., and Claire Bowern. 2016. Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences* 113 (48): 13666–13671.

Haynie, Hannah, Claire Bowern, Patience Epps, Jane Hill, and Patrick McConvell. 2014. Wanderwörter in languages of the Americas and Australia. *Ampersand* 1:1–18.

Haynie, Hannah, Claire Bowern, and Hannah LaPalombara. 2014. Sound symbolism in the languages of Australia. *PLoS One* 9 (4): e92852.

Hill, Jane H. 2002. "Expert rhetorics" in advocacy for endangered languages: Who is listening, and what do they hear? *Journal of Linguistic Anthropology* 12 (2): 119–133.

Himmelmann, Niklaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36 (1): 161–196. doi:10.1515/ling.1998.36.1.161.

List, Johann-Mattis, Simon Greenhill, and Robert Forkel. 2017. *LingPy: A Python Library for Quantitative Tasks in Historical Linguistics*. Jena, Germany: Max Planck Institute for the Science of Human History.

List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLoS One* 12 (1): e0170046. doi:10.1371/journal.pone.0170046.

Macklin-Cordes, Jayden L., and Erich R. Round. 2015. *High-definition Phonotactics Reflect Linguistic Pasts*. Tübingen, Germany: Universitätsbibliothek Tübingen.

Macklin-Cordes, Jayden L., Claire Bowern, and Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38. https://doi.org/10.1075/dia.20004.mac.

Nash, David. 2017. Loss of color terms not demonstrated. *Proceedings of the National Academy of Sciences* 114 (39): E8131. doi:10.1073/pnas.1714007114.

Perley, Bernard C. 2012. Zombie linguistics: Experts, endangered languages and the curse of undead voices. *Anthropological Forum* 22 (2): 133–149. doi:10.1080/00664677.2012.694170.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

Thieberger, N., ed. 1995. *Paper and Chalk: Manual or Reconstituting Materials in Australian Indigenous Languages from Historical Sources*. Canberra, Australia: Aboriginal Studies Press.

Wafer, Jim, Amanda Lissarrague, and Jean Harkins. 2008. *A Handbook of Aboriginal Languages of New South Wales and the Australian Capital Territory*. Nambucca Heads, Australia: Muurrbay Aboriginal Language and Culture Co-operative.

Weiss, Michael. 2015. The comparative method. In *The Routledge Handbook of Historical Linguistics*, ed. Claire Bowern and Bethwyn Evans. Routledge Handbooks Online. Abingdon, UK: Routledge. June 27, 2014. doi:10.4324/9781315794013–16. Accessed November 19, 2018.

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