Author/s:
Thieberger, N

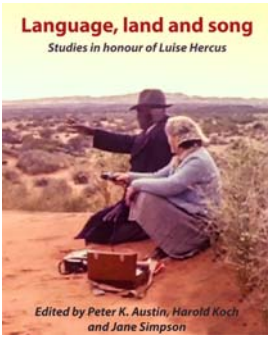
Title:
Daisy Bates in the digital world

Date:
2016

Citation:
Thieberger, N. (2016). Daisy Bates in the digital world. Austin, PK (Ed.). Koch, H (Ed.). Simpson, J (Ed.). Language, land and song: Studies in honour of Luise Hercus, (1), pp.102-114. EL Publishing.

Persistent Link:
https://hdl.handle.net/11343/282440

elpublishing



This item is Chapter 8 of

Language, land & song:
Studies in honour of Luise Hercus

Editors: Peter K. Austin, Harold Koch & Jane Simpson

ISBN 978-0-728-60406-3

http://www.elpublishing.org/book/language-land-and-song

Daisy Bates in the digital world

Nick Thieberger

Cite this item:

Nick Thieberger (2016). Daisy Bates in the digital world. In *Language, land & song: Studies in honour of Luise Hercus*, edited by Peter K. Austin, Harold Koch & Jane Simpson. London: EL Publishing. pp. 102-114

Link to this item:

http://www.elpublishing.org/PID/2008

This electronic version first published: March 2017

EL Publishing

Open access, peer-reviewed electronic and print journals, multimedia, and monographs on documentation and support of endangered languages, including theory and practice of language documentation, language description, sociolinguistics, language policy, and language revitalisation.

For more EL Publishing items, see http://www.elpublishing.org

8

Daisy Bates in the digital world

Nick Thieberger
School of Languages and Linguistics, University of Melbourne

1. Introduction[1]

I am pleased to offer this paper in tribute to Luise Hercus who has always been quick to adopt new approaches to working with older sources on Australia's Indigenous languages (see also Nathan, this volume). In that spirit, I offer an example of using a novel method of working with a large set of material created by Daisy Bates (1859-1951) in the early 1900s. The masses of papers she produced over her lifetime have been an ongoing source of information for Aboriginal people and for researchers (e.g. White 1985; McGregor 2012; Bindon & Chadwick 1992). The collection at the National Library of Australia (NLA) takes up 51 boxes and 8.16m of shelf space and contains a range of material, but here I will focus on the vocabularies of Australian languages. Bates sent out a questionnaire in 1904 that was filled in by various people by hand, creating a set of manuscript pages. She then supervised the typing of these manuscripts. Over the past two years I have been working with the NLA to make digital images of some 23,000 pages of these vocabulary manuscripts, and to create digital text versions of the 4,368 typescripts, which can then be linked back to the page images of both the typescripts and handwritten questionnaire manuscripts.

This project has now digitised, transcribed, and encoded this valuable collection of wordlists of Australian languages to make it accessible both for language research and for community access.[2] The dataset has been constructed

[1] This paper reports on work done together with Conal Tuohy who developed the TEI representation of the text. Thanks to Claire Bowern for providing the digital form of the Sutton & Walsh questionnaire. Thanks also to two anonymous reviewers for helpful comments. This work was initially funded by a research grant from the Faculty of Arts at the University of Melbourne and supported in part by ARC grants DP0984419 & FT140100214.

[2] The website (in construction in 2016/2017) described in this chapter can be viewed at: http://bates.org.au/

Thieberger, Nick. 2016. Daisy Bates in the digital world. In Peter K. Austin, Harold Koch and Jane Simpson (eds.) *Language, Land and Song*, 102-114. London: EL Publishing.

according to the Text Encoding Initiative (TEI) Guidelines[3], to embody both a (partial) facsimile of the original set of manuscripts and a structured dataset for examining complex research questions. The dataset will be open to reuse, in particular providing access for Indigenous people in remote areas to vocabularies of their ancestral languages. The model will also be an exemplar of how a text and document-based project, typical of humanities research, can benefit from new methods of encoding for subsequent research and reuse.

This paper presents the method used to deal with this set of data, and argues that large scale manuscript collections such as this one are suited to an XML encoding (similar to the work discussed by Henderson 2008) that permits reorganising the text in new ways while maintaining its original context. Visualisation of the text can be created as facsimiles of the original, or via geographic maps, or alphabetic sorting, among others. This questionnaire is compared to other questionnaires for Australian languages and the results obtained by Bates are quantified for the first time. This and other kinds of visualisation of the material in the collection are only possible due to their digital and structured nature. Earlier work could not take advantage of these methods, e.g. in my work in the Noongar Native Title case (Thieberger 2004), I extracted lexical information from the Bates papers and used it to help show linguistic continuity over time. My work in selecting, sorting, and comparing the wordlists with other sources resulted in a spreadsheet of information that was disconnected from the originals and was essentially unusable by anyone else. In retrospect, a better method would have been to create an encoded version of the vocabularies that allowed parts to be selected for export, corrected, enriched with annotations, and stored with all of that explicit encoding for others to use again.

2. The Bates vocabulary collection

The Bates vocabularies are extraordinarily valuable as little else was recorded in the same time period and nothing of the same scale has been attempted before or since. However, despite their value, the wordlists, often including grammatical information in the form of example sentences, remain relatively inaccessible due to being held in paper form only in the Battye Library in Perth, the Barr Smith Library in Adelaide (who hold copyright in this collection) and the National Library in Canberra. By processing the wordlists and making them accessible online, we will prepare material that will be of use to Indigenous Australians today, as well as embodying an open research dataset which may be linked to other data and from which we can determine what the languages are that are represented in the Bates collection. This computational analysis has not been possible while the dataset has remained in its original analog format. By building accessible content for Indigenous Australians this project takes historical records out of the archive and into the community, potentially supporting current language initiatives.

---

[3] TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 1.7.0. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (accessed 2015-03-24).

The Bates papers have a great deal of information on many aspects of Aboriginal life ranging from kinship to local affiliation and 'waters' with which individuals are associated. But, as noted by McGregor (2012:81):

> although her life, anthropological research, and her contribution to Aboriginal policy have been the subject of a number of publications [...] her linguistics has received virtually no scholarly attention.

The focus of the present project is section XII of the collection which is titled *Language, grammar and vocabularies*. In 1904 Bates organised a printed questionnaire booklet that was filled in by hand by various people, including policemen and property owners, mainly around Western Australia, so creating a set of manuscript pages. The Commonwealth provided a typist in Adelaide, while Bates was camped at Pyap (some 160km north of Adelaide) (White 1993: 57). Bates then supervised the typing of these manuscripts, with subsequent corrections and additions to the typescript, and omission of some of the information on the manuscript pages. In 1938 Bates moved to Adelaide but had already sent the papers to the Parliamentary Library (the forerunner of the National Library of Australia) in February 1934.[4] At some later stage, copies were made for the Barr Smith Library in Adelaide and the Battye Library in Perth.

As noted by Reece (2007:46-47),

> At the outset, Daisy's official responsibility had been to prepare a blank vocabulary 'which contained a sufficient number of words expressing the ideas essential to a language in the form of substantives, adjectives, verbs etc., and a few simple sentences, which would enable the philologist to ascertain the structure of the grammar and vocables [sic]'[5]. Five hundred copies of this were sent out to postmasters, police and station owners requesting their assistance. It soon became clear, however, that she was by no means limiting herself to linguistic information-gathering: a number of questions concerning customs were also added to elicit information about social classification.

The present work began by analysing the collection and establishing that the questionnaire provided a standard format that, in theory, is the basis of each typescript. Responses to the questionnaire were filled out in booklets which had been microfilmed by the NLA in the past. TIFF images of the pages were downsampled to web-deliverable size, but even so, they total 4.52 gigabytes in size. The image files of the typescripts were then sent to an agency for keyboarding. A major task was renaming the files to allow automated access to their contents. Thus a typescript page numbered by the NLA as 16 in folio 59 was uniquely named 59-016T (where 'T' indicates it is a typescript see Table 1). This step allows citation to the level of a page image so that all words can link to the context in which they occur.

---

4 Memorandum, Department of the Interior 35/1066, National Archives of Australia, box A659

5 Reece notes that this comes from a letter from Bates to Under-Secretary James Donnelly, dated 6th February 1909. SROWA Acc. 1023.

The Barr Smith Library in Adelaide has been producing PDF/text versions of some of these manuscripts that increase their availability. Each PDF document represents a complete questionnaire typescript. What the present work offers in addition is a searchable text version and a unique identifier for each page of the typescript and the manuscript, allowing text to resolve to the source images of both the typescript and manuscript.

3. Encoding digital text of Australian language material

The National Lexicography Project (Simpson & Nash 1989) of the late 1980s and its successor, the Aboriginal Dictionaries Project (1990-1994), both based at the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS), supported the creation of over 50 dictionaries and provided a repository to store digital source files for those dictionaries and others. The Aboriginal Studies Electronic Data Archive[6] (Thieberger 1994, 1995), also at AIATSIS, was an early repository for curating and storing the underlying data files for these dictionaries and making them available for re-use, often in Aboriginal language centres.[7] It was noted that this set of material could become the basis for a pan-Australian dictionary (Goddard & Thieberger 1997:192) – a project yet to eventuate but still a possibility, and one that will benefit from the kind of structured lexical dataset described here. The methods used for dealing with manuscript pages of lexical information in the past can be broadly categorised into the following distinct types:

1. words are transcribed, and perhaps standardised, into a word processor document (not linked to an image or text of the original)
2. the content is extracted, and perhaps standardised, into a spreadsheet or database management system (DBMS) (not linked to an image or text of the original)
3. images are placed online for access with some descriptive metadata, but with no textual transcript
4. images are placed online with a textual facsimile, allowing the original forms (both typed and as images) to be linked and available to the reader

The problem with methods A, B & C is that there is no way to trace back from a form given in a database to its source, especially if it has been standardised, based on decisions made by the linguist. Method C does allow access to manuscript images, but they are not searchable. An example of encoded texts of Australian language manuscripts using method D is Henderson's rendering of the legacy material produced by the linguist Gerhardt Laves in the 1930s (Henderson 2008),

6 A list of some of the contents of ASEDA can be found at: https://web.archive.org/web/20130108143102/http://www.aiatsis.gov.au/aseda/docs/index.html

7 There are a number of indigenous language centres around Australia, listed at: http://www.ourlanguages.net.au/language-centres.html [accessed 2015-04-07]

which involved creating an XML facsimile of some 3,000 pages of handwritten notes consisting of a mix of dictionary and texts of Nyungar (Southern Western Australia). Ideally the Bates work described here will, in future, be able to work together with the Laves papers to provide a richer view of historical material related to the south-west of Western Australia.

4. Encoding the Bates collection

Analysis of the texts in this collection identified three types of document: (1) the original handwritten questionnaires, containing some 2,000 prompt words and sentences and each typically made up of around 100 pages; (2) typescript versions of those questionnaires (each made up of varying numbers of pages), and; (3) manuscripts that could be either questionnaires or additional material. Filenames reflect these three types as shown in Table 1.

Table 1: Types of files in the Bates collection

Type of information and filename structure	**Example filename**	**Number of items**
Typescripts Folio – page number & T	51-007T.tif	142 (with varying numbers of pages per item, totaling 4,368 pages)
Questionnaires Folio – booklet & Q & page number	39-241Q99.tif	167 (with some 100 pages per item)
Other Manuscripts Folio – page number for other manuscripts (M)	39-308M.tif	84 (with varying numbers of pages per item)

Various Bates vocabularies have been keyboarded in the past and her work was used in Native Title claims, often in comparative spreadsheets, which did not have a life beyond that use (e.g. Thieberger 2004). On reflection, this lack of re-use is due to extracting just the terms needed for a particular purpose, typically comparing a given set of terms to show continuity over time for a land claim. Other uses of historical sources for linguistic analysis have adopted the same method, with the result that it is often difficult, if not impossible, to trace a word back to its original manuscript context. If, instead, the Bates files had been created as digital facsimiles of the original files so that extracts used in comparison were cited back to the original documents then the primary material would have been of more use, both to Aboriginal people and to subsequent researchers. This is the method adopted in this project. The digital images of pages of the manuscripts were each renamed to reflect their NLA identifier and the typescript images keyboarded into a simple tabular format, with page numbers marked in the text to allow links to be maintained between the text and the image. Subsequently the text encoding was automated using XSLT (for

more details see Thieberger & Tuohy forthcoming). This has allowed the text to be presented on maps as seen on pages 109 and 110, and will, in future, allow the text to be extracted for comparative purposes. Links between the typescript and manuscript questionnaire are inferred because we know the page on which each prompt word occurs in the questionnaire and so can resolve a link to that page. The direction of links is from the text to the image of the typescript and then from there to the image of the manuscript, but not vice-versa.

5. Nature of the Bates questionnaire

Of the 500 copies of the questionnaire (Reece 2007:46) that were printed and distributed to magistrates, pastoralists and the police force[8] eventually some 120 were completed and returned. When the responses to the questionnaire came in they were 'of mixed quality and often unreliable as to the original location of the Aboriginal people questioned' (Reece 2007:48). Some were filled in by Bates from earlier sources, for example item 39-216 is 'Compiled from vocabularies by Richard Helms (Fraser Range, Esperance, Hampton Plains, Knutsford); A. Wells (Fraser Range); W. Williams (Eucla and Eyre's Sandpatch); Campbell Taylor (Doubtful Bay, Israelite Bay)'. Another list (43-095) was filled out from sources by 'Police Constable Hackett, Bishop Salvado, H. J. Monger, E. K. Parker, G. Whitfield'. It would be useful to be able to identify which words came from which source in these cases, and, where multiple questionnaires were combined into one typescript, that will be possible in the next stage of our project. McGregor (2012: 88-92) provides a summary of the questionnaire that need not be repeated here, but it is worth noting that it was clearly devised before Bates had had much experience of Aboriginal languages. Terms like 'aunt', 'brother/sister', 'grandfather/mother', and 'nephew/niece' have multiple possible referents. For example, aunt may refer to 'father's sister' but perhaps not to 'mother's sister' (which often is the same term as 'mother'). Sibling terms may relate to seniority and not have gender characteristics.

The front-matter of the questionnaire included two pages of instructions about the requested responses, noting that Aboriginal languages 'have distinct names for every minute portion of the human frame and other natural objects'. The prompts were listed under the following headings: Vocabulary (Man, his Relationships, etc.; Parts and functions of the body; Animals; Birds; Fishes; Reptiles; Insects; The Elements, etc.; General Vocabulary); Short Sentences; Questions. It seems there was no suggestion made about the kind of spelling system to be used, but the Royal Geographical Society had, in 1891, issued guidelines for an orthography for native names of places which were then published by the Intelligence Division of the War Office in 1892, with the following heading:

---

8 Under the authority of the Registrar General of Western Australia, Malcolm A. C. Fraser, see 'A History of Natives.' (1904, August 25). *The Daily News* p.1. Available at http://nla.gov.au/nla.news-article82473284 [accessed 2015-03-30]

> The following system of orthography for native names of places adopted by the Council of the Royal Geographical Society, the Foreign and Colonial Offices, Admiralty, and War Office is to be adhered to in all Intelligence Division Publications[9].

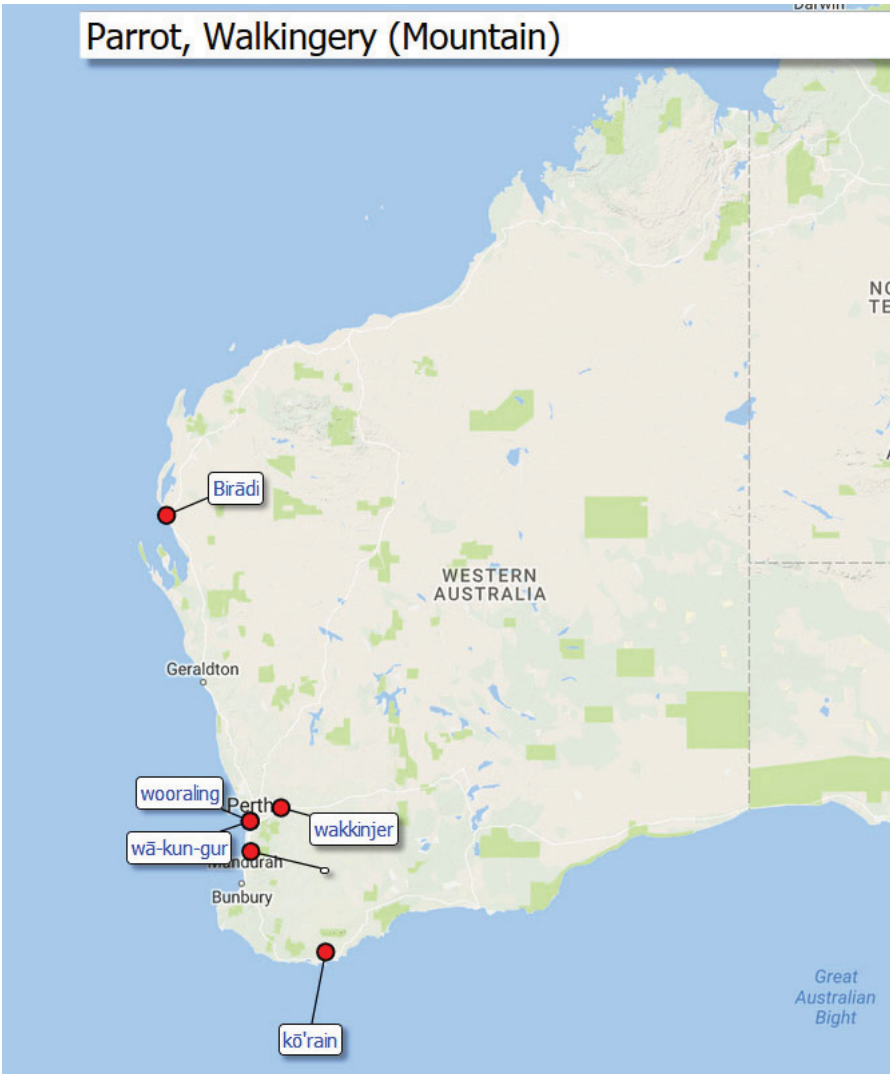
This document provided a list of letters to be used and noted that 'vowels are pronounced as in Italian and consonants as in English', and that 'every letter is pronounced, and no redundant letters are introduced'. It is quite possible that this document was known to colonial officials and used in their transcription of Aboriginal words. In her biography of Bates, Salter (1971: 216) suggests that Bates 'had kept to the two thousand word structures recommended by the Royal Geographical Society.'

The current work reveals that, of the 1,829 lexical prompt terms in the questionnaire, only 13 have more than 100 responses (these are: eye; ear; nose; teeth; hand; mother; father; tongue; beard; head; crow; emu; moon). The most commonly provided term is the equivalent for 'eye' which has 109 responses (see map 2). 1,309 prompt words have between 10 and 99 responses and 431 have between 1 and 9. For 74 words there is not a single form in the responses (examples include: albatross – light wandering; anger; another; anxious; autumn; avoid, to; blackboy; ceremony; come, I, from; contest, a; country (desert); cuckoo, lesser bronze; donkey; volcano; will). These figures suggest that there was no field-test of the wordlist before it was printed to identify the more productive terms to include.

A further problem for our encoding is that the prompts provided in the questionnaire were modified by some respondents so that, instead of having 1,829 prompt forms and responses in the resulting material, there were in the order of 4,400 English forms and local equivalents. The distance between the original and the variant can be due to simple transcription differences, sometimes transcription errors that need to be corrected, as in these examples of variation in the form of the prompt written into the questionnaire response (where the bold term is the original prompt):

> **wife** is that your / wife is that yours;
>
> **white man's house where is the /** white man's house where / white man's house where is;
>
> **white man** or woman / white man or woman spirit ghost / white man spirit of dead native ghost / white man spirits evil spirit / white man white fellow / white man woman;
>
> **stone** / stone black / stone hills eye / stone into which the spirit of a dead woman has entered / stone millstone;
>
> **pigeon bronzewing** / pigeon bronze wind / pigeon bronze wing / pigeon bronzewing brush bronzewing;
>
> **displease** I / displease to.

9 http://gallica.bnf.fr/ark:/12148/btv1b84410133



Map 1: Illustration of the distribution of four terms for 'Parrot, Walkingery (Mountain)' Map: Nick Thieberger.

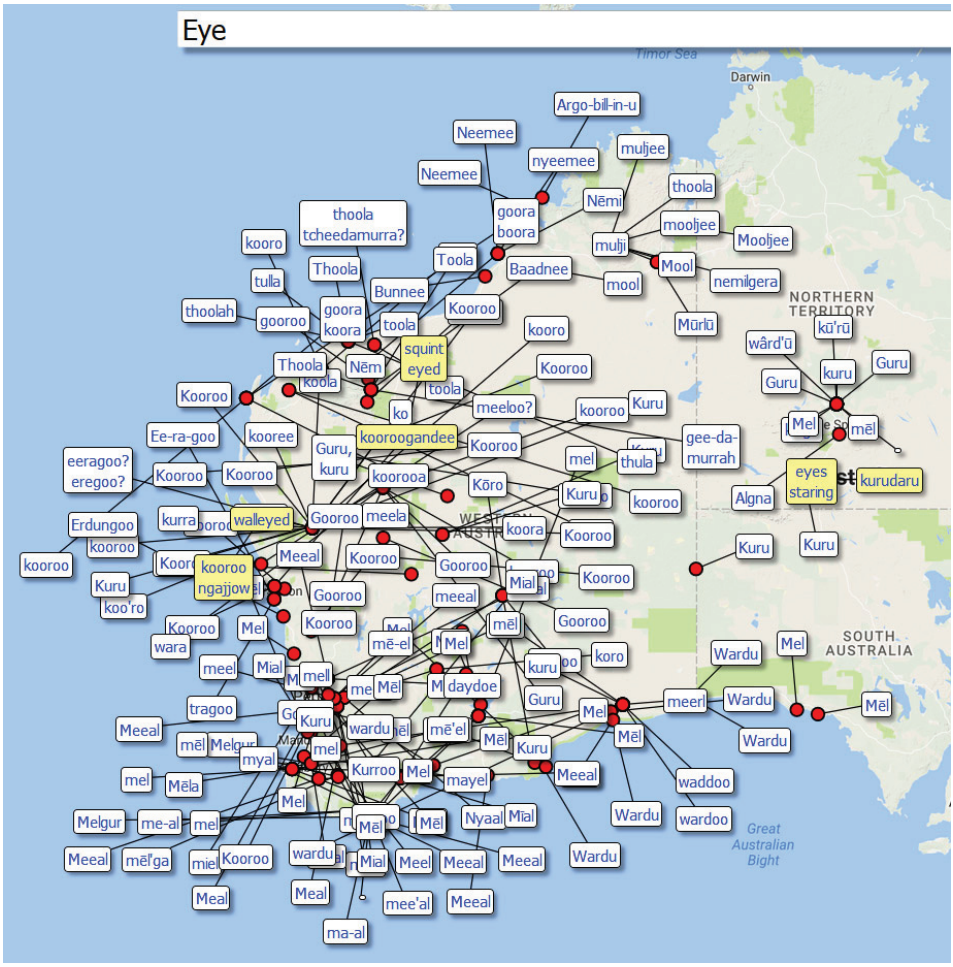
Some responses have no corresponding questionnaire form, e.g. 'large long sacred board' / 'large sacred board long'. We are currently marking such exceptions and allowing the standard forms to drive the representation in the first stage of the work, and providing a mechanism for a separate view of all responses, including non-standard and otherwise extra information.

As might be expected, the spelling of terms is occasionally internally inconsistent, so, for example, these three terms are given in one response: *thoonoo, tuninga, thoonong* ('below' 50-063). Is the initial segment likely to be alveolar or dental? And *tchoo-gi-go, joo-je-koo* ('boots' 50-064) look like they could be two ways of writing what could be rendered as *jujiku*. *Beeja, baia, baja, batcha* ('anger' 50-062) suggest there could be variation between *pija*, *paya*, and *paja*. When there are two forms like *to-nee, thornee* is the medial nasal retroflex, and what

is the nature of the initial consonant? If we assume that specifying /th/ and /rn/ is more likely to reflect the sounds than the underspecified *to-nee* then *thurni* is the likely form. The present work makes no attempt to standardise the forms, but if in future a standard form is hypothesised (for example, in a pan-Australian lexicon, as discussed earlier), a strength of the present method is that all forms are linked to their context, allowing any reader to also see the originals.

As each list is provided with a location, it is possible to visualise the geographic range of the lists, and this allows other anomalies to appear, e.g. the word 'alligator' has two correspondences in the south-west of WA (*bung-arra, bong-yar?* 45-001) where crocodiles are not found and so the most likely referent is 'goanna'.

There are also terms that are not currently known, for example: parrot, walkingery (mountain) (see Map 1). Searching with Google does not identify this parrot, nor is there anything in the National Library's Trove. But looking at the responses (four in the south-west for 'parrot, walkingery', and one north of Carnarvon for 'parrot, walkingery (mountain)', it is clear that the English term includes the (presumably) Nyungar term 'wā-kun-gur' that may have been current in the early 1900s but not since.



Map 2: Illustration of the distribution of correspondences for the term 'eye' Map: Nick Thieberger.

6. Comparison with other questionnaires

The questionnaire format has been used in Australia by Curr (1886), Capell (1963), and Sutton & Walsh (1979) [S&W], among others, as a means of getting a basic wordlist in Aboriginal and Torres Strait Islander languages, and of checking on vocabulary elicitation for comparative purposes. I compared the Bates list with these in order to determine if there were enough words in common to allow for a combined list to be constructed from them. This work is in part a contribution to the study of questionnaires used in linguistic fieldwork and typological studies (e.g. Plank 2003; Lahaussois 2014).[10] Table 2 should be read as follows: Bates has 1,829 items of which just 6% correspond to Curr, while 85% of Curr's list (which has only 136 terms) correspond to Bates. A comparison of Bates with the Swadesh-equivalent for Australian languages, as represented in Alpher & Nash's (1999) adaptation of O'Grady's 100-word list [GNOG-AN] (O'Grady & Klokeid 1969: 303-307) and then also in the Sourcebook for Central Australian languages [SCAL] (Menning & Nash 1981), shows that, of the 165 possible terms in SCAL, 76% are found in Bates, with comparable similarity with GNOG-AN.

Table 2: Comparison of the English terms in Bates and five other wordlists.

	Bates (1829)	Curr (136)	Capell (558)	S&W (2098)	SCAL (165)	GNOG-AN (151)
Bates (1829)		6%	22%	35%	7%	6%
Curr (136)	85%		72%	90%	46%	45%
Capell (558)	72%	18%		69%	20%	20%
S&W (2098)	30%	6%	18%		8%	7%
SCAL (165)	76%	38%	68%	96%		64%
GNOG-AN (151)	78%	40%	74%	95%	70%	

As may be expected, some terms, especially sensitive body-part terms, are found in the more recent lists but are not in the earlier ones, for example, 'testicles' is in Capell and S&W but not in Bates or Curr, but 'coition' is in Bates and 'penis' occurs in all but Curr. 'Generative Organ (female)' is in Bates corresponding to 'vagina' in S&W; 'clitoris' is only found in S&W. S&W has 'faeces' corresponding to 'excrement' in Curr, but neither term appears in the other lists.

The following are the only items that are common to all six lists: blood; bad; big; bite, to; bone; by-and-by; cloud; ear; egg; eye; fat; woman; fire; fish (generic); fly (generic); foot; good; hair; hand; head; hungry; leaf; moon; mouth; near; nose; now; one; rain; see, to; sit down, to; skin; snake (generic); spear (generic term); star; sun; thigh; tomorrow; tongue; two; where.

Comparison of the lists is necessarily imprecise as there are similar terms but not always with an exact match between them. For example, 'spearthrower',

10 The spreadsheet used to arrive at these results is available on request.

'throwing stick' and 'woomera' occur as separate entries in S&W and correspond to 'board, for throwing spear' and 'throwing board' in Bates (counting as a correspondence for each term for present purposes). More problematic are relationship terms which, as noted above, are quite underspecified in Bates' list, so, for example, Bates has 'aunt' corresponding to the following terms in S&W: FZ-, FZ+, MZ-, MZ+, FFZD (boosting the correspondence figures). FF and MF in S&W are correlated with 'grandfather' in Bates. Bates has no correspondences for the following in S&W: 'FZD', 'FZDC (man speaking)', 'FZDC (woman speaking)', 'FZS', 'FZSC (man speaking)', 'FZSC (woman speaking)'.

In some cases, Bates is more specific, so S&W have 'foot' while Bates has 'foot (left)' and 'foot (right)'. S&W have 'wind' while Bates has terms for winds from all cardinal directions. Other terms in Bates are more detailed than in the other lists, so 'heron' occurs in Capell, while Bates has 'heron, black with white neck' and 'heron, egret (white)'; S&W have 'blue heron', 'reef heron', and 'herons'. In summary, the results in Table 2 should be taken as showing that the correspondences of these four wordlists do not readily allow for complete comparability.

7. Novel ways of viewing these records

Simply digitising the text of this collection has made it searchable. Previously, the size of the paper collection was such that it was not feasible to look through several thousand pages to locate particular items, but that can now be done by a simple text search. For example, David Nash was researching the origin of the cossid moth species *Xyleutes* (now *Endoxyla*) *biarpiti*, named by Norman Tindale (1953: 60-62) who wrote of specimens collected with Aboriginal children at Ooldea in April 1951: 'The larvae from the roots of the Zygophyllum bushes or biarpiti are called by them mako biarpiti, i.e. mako or grubs of the biarpiti'. No source for this term was apparent in any record, but it was possible to search the current set of Bates vocabularies. Looking for 'biarpiti' failed, as did subsequent searches for 'biahbidi, biarpiti, biarbidi, piarpiti'. Finally, searching for 'biabirdi' was successful (in the document *XII2G8a Central Districts Folio 60/185*), also pointing to the need for a 'soundex' search to be available for this data set.

Wherever possible, a questionnaire has a geographic location encoded (as a point), allowing a map presentation, such as those seen in Map 1 and Map 2 above. Named entities in the questionnaires, like person names, places, tribes, or languages, are all explicitly marked in the TEI transcriptions, as follows:

```
<orgName type='tribename'>Yaggangui</orgName> of
<placeName>Ooleroo</placeName>
<lang>Wija Baggani</lang> or <lang>Wijaŋga</lang>
North of <lang>Maŋgunga</lang> <lang>Gunbai-ija woŋga</lang> and
<lang>Kularda</lang>
```

These terms will then feed an index of personal names, languages, tribes and locations. Where the language can be identified, and where it has a standard code (ISO-639-3) then it is included as the default language code for that list, with any exceptions being marked within the list.

8. Conclusion

To date, this project has identified the names of over 150 speakers in section XII of the Bates papers, and in the order of 40 languages. One of the outcomes of the project will be to determine the relationships between these wordlists and later sources, and to identify more securely what languages are represented. As the underlying material is in a re-usable format it will be possible to publish electronic editions of the manuscripts (in various formats, RTF, PDF or DOCX), output sets of wordlists for various uses, and to do comparative analysis of the vocabularies to infer locations (as in Nash 2002; Embleton et al. 2013). Earlier, I worked with a collection of over 15,000 pages from Arthur Capell's estate, putting images online[11] with minimal metadata. The effort of writing summaries of items took Capell's executor, Peter Newton, several weeks of work. Taking images of the papers, naming them, and loading them into the viewer also took a few weeks. The resulting webpages have made the papers openly available to anyone on the web since 2005, but with no possibility to search the text as it is stored only in image files. Clearly it takes more work to create a textual facsimile of a paper-based collection, but this effort then allows greater access to the original source via the text itself.

The method described here encodes the information in the vocabularies and allows various forms of navigation and visualisation based on that encoding. What distinguishes this method primarily from earlier treatment of work like Bates' is that it reproduces the whole manuscript as text, and thus allows each item within the manuscript to be cited in its original context, and with reference to the manuscript image.

References

Alpher, Barry & David Nash. 1999. Lexical replacement and cognate equilibrium in Australia. *Australian Journal of Linguistics* 19(1), 5-56.

Bindon, Peter & Chadwick, Ross. 1992. *A Nyoongar wordlist from the south west of Western Australia*. Perth: Western Australian Museum.

Capell, Arthur. 1963. *Linguistic survey of Australia*. Canberra: AIAS.

Curr, Edward M. 1886. *The Australian race, Vol.1*. Melbourne: Government Printer.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2013. Defining dialect regions with interpretations: Advancing the multidimensional scaling approach. *Literary & Linguistic Computing* 28(1), 13-22.

Goddard, Cliff & Nick Thieberger. 1997. Lexicographic research on Australian Aboriginal languages 1968-1993. In Darrell Tryon & Michael Walsh (eds.) *Boundary rider: Essays in Honour of Geoffrey O'Grady*, 175-208. Canberra: Pacific Linguistics.

Henderson, John. 2008. Capturing chaos: Rendering handwritten language documents. *Language Documentation & Conservation* 2(2), 212-243. http://hdl.handle.net/10125/4347

11 http://paradisec.org.au/fieldnotes/AC2.htm

Lahaussois, Aimée. 2014. *Les questionnaires: recensement, analyse, valorisation et réflexion épistémologique* (ms) http://www.typologie.cnrs.fr/spip.php?rubrique105

McGregor, William. 2012. Daisy Bates' documentations of Kimberley languages. *Language and History* 55(2), 79-101.

Menning, Kathy, & David Nash. 1981. *Sourcebook for Central Australian Languages*. Alice Springs: Institute for Aboriginal Development.

Nash, David. 2002. Historical linguistic geography of south-east Western Australia, In John Henderson & David Nash (eds.) *Language in Native Title*, 205-230. Canberra: AIATSIS Native Title Research Unit, Aboriginal Studies Press.

O'Grady, Geoffrey N., & Terry J. Klokeid. 1969. Australian linguistic classification: A plea for co-ordination of effort. *Oceania* 39(4), 298-311.

Plank, Frans. 2003. Unanswered questions, wasted answers, loose leaves lost (ms).http://ling.uni-konstanz.de/pages/home/plank/for_download/unpublished/08_FP_UnansweredQuestions_2003.pdf [accessed 2016-11-01]

Reece, Bob. 2007. *Daisy Bates: Grand dame of the desert*. Canberra: National Library of Australia.

Salter, Elizabeth. 1971. *Daisy Bates*. Sydney: Angus & Robinson.

Simpson, Jane, & David Nash. 1989. AIAS archive of machine-readable files of Australian languages: the National Lexicography Project. Australian Aboriginal Studies 1/1989, 57-59.

Sutton, Peter, & Michael Walsh. 1979. *Revised Linguistic Fieldwork Manual for Australia*. AIAS New Series no.8. Canberra: Australian Institute of Aboriginal Studies.

Thieberger, Nick. 1994. Report on the AIATSIS Visiting Research Fellowship, Aboriginal Studies Electronic Data Archive. (ms)

Thieberger, Nick. 1995. The Aboriginal Studies Electronic Data Archive, *International Journal on the Sociology of Language* 113, 147-150.

Thieberger, Nick. 2004. Linguistic report on the Single Noongar Native Title Claim. (ms) http://repository.unimelb.edu.au/10187/6935

Thieberger, Nick & Conal Tuohy. 2017. From Small to Big Data: paper manuscripts to RDF triples of Australian Indigenous Vocabularies. *In Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages*. http://altlab.artsrn.ualberta.ca/wp-content/uploads/2017/03/CEL2_5.pdf

Tindale, Norman B. 1953. On some Australian Cossidae including the moth of the witjuti (witchety) grub. *Transactions of the Royal Society of South Australia* 76, 56-65

White, Isobel. (ed.) 1985. *The native tribes of Western Australia.* Canberra: National Library of Australia.

White, Isobel. 1993. Daisy Bates: Legend and reality. In Julie Marcus (ed.) *First in their field: Women and Australian anthropology,* 47-65. Carlton: Melbourne University Press.