THE DIGITISED HOLLE LIST PROJECT: BUILDING A DATABASE FROM LEGACY MATERIALS FOR CONSERVING INDIGENOUS INDONESIAN LANGUAGES

Gede Primahadi Wijaya Rajeg
Computer-assisted Lexicology and Lexicography (CompLexico) Research Group
Centre for Interdisciplinary Research on the Humanities and Social Sciences (CIRHSS)
Bachelor of English Literature, Faculty of Humanities, Udayana University

Abstract

Advances in cloud computing, as well as computational tools for extracting text from images, offer an opportunity to scale up the development of digital databases for Indigenous languages. This paper reports on the application of these advances to the digitalisation of old, paper-based lexical items of over a hundred Indigenous languages in Indonesia; these items are part of the so-called Holle List (HL). After introducing the (structure of the) HL, the paper underlines the motivation for the HL digitalisation project. It then provides an overview of *Google Colab* as a free cloud-computing platform for executing a series of optical character recognition (OCR) operations on hundreds of scanned pages of the HL, utilising *pytesseract*, a Python interface for *Google's Tesseract-OCR engine*. Advantages (e.g., computational searchability and manipulability), as well as issues (especially typos and unrecognised characters) in the plain-text OCR outputs, are discussed. In conclusion, the paper highlights the importance of digital technology in conserving Indigenous languages via digital platforms, despite some unavoidable challenges that require human intervention.

Keywords: *Holle List; Indigenous Indonesian languages; Digital Humanities; Lexical databases*

1. Introduction

This paper[31] reports on a Digital Humanities (Drucker, 2021) project of digitalising and curating large volumes of word lists, the so-called *Holle List vocabulary*. The Holle List project was initiated in the late 19th century by Karel Frederik Holle, a Dutch colonial administrator. His aim was to gain knowledge about the linguistic situation of the Dutch East Indies, corresponding to the present-day state of Indonesia. In the first edition of the Holle List (Holle, 1894), K. F. Holle set up a list of elicitation concepts (i.e., 905 concepts to be exact) given in Dutch (Holle, 1894, pp. 8–38). This list was distributed throughout the Indonesian archipelago. The goal was to collect the corresponding expressions/words of these elicited concepts (from

[31] Some of the works reported here (especially the digitalisation of the main, reference Holle List, the New Basic List, the Enggano Holle List, and some languages of the Barrier Islands, Sumatra), have been initiated when the author was a postdoctoral researcher at the University of Oxford, UK (2023-2025), working on developing lexical resources for Enggano, funded by the *Arts and Humanities Research Council* (AHRC), UK (AH/W007290/1).

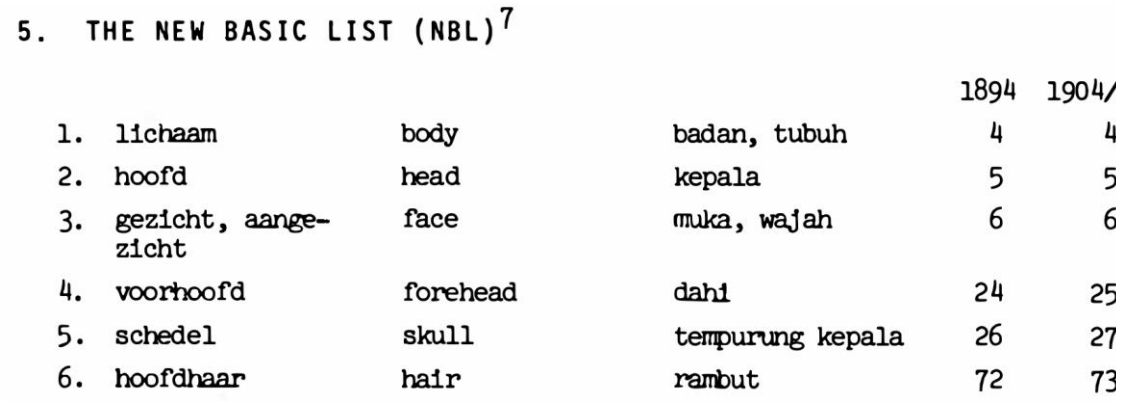
various semantic domains) across more than two hundred indigenous regional language varieties in Indonesia.

Between 1980 and 1987, W. A. L. Stokhof and colleagues (viz. Lia Saleh-Bronckhorst and Alma E. Almanar) edited, collated, and published (i) the different versions of the reference/master, elicitation Holle List as well as (ii) the corresponding expressions/words in the regional language-varieties into an eleven-volume publication series[32]. These publications are available as open access under the Creative Commons License (see Figure 2 in § 2.1) and consist of two main parts. The first one is the volume containing just the reference (or master) Holle List (Stokhof, 1980), comprising elicitation concepts given in Dutch, English, and Indonesian/Malay together with their index numbers (see Figure 1 (a)); this is called *The New Basic List* (hereafter NBL) in Stokhof (1980) because Stokhof and colleagues collated three different versions of the Holle List (namely those published in 1894, 1904/1911, and 1931; see Figure 1 (a)). The second part of the Holle List publications is the separate volumes containing the expressions/words of the regional language-varieties and their index numbers (see Figure 1 (b) for an example from the Enggano language); these index numbers for words in the regional language-varieties correspond to the index numbers of the concepts in the reference Holle List/NBL. It is important to note that there is only one volume of the NBL; the content of the NBL is not repeated in the remaining volumes for the expressions/words in the regional language-varieties, but only the index numbers.

In the above two-part publication setup, linguists, who are interested in the Dutch, English, and Indonesian/Malay translations of *a* given word in *a* given regional language, must manually match the index number of that regional word with the corresponding index number in the reference Holle List. Let us use the data snippet in Figure 1 as an example.

[32] See these Holle List publication series on this page.

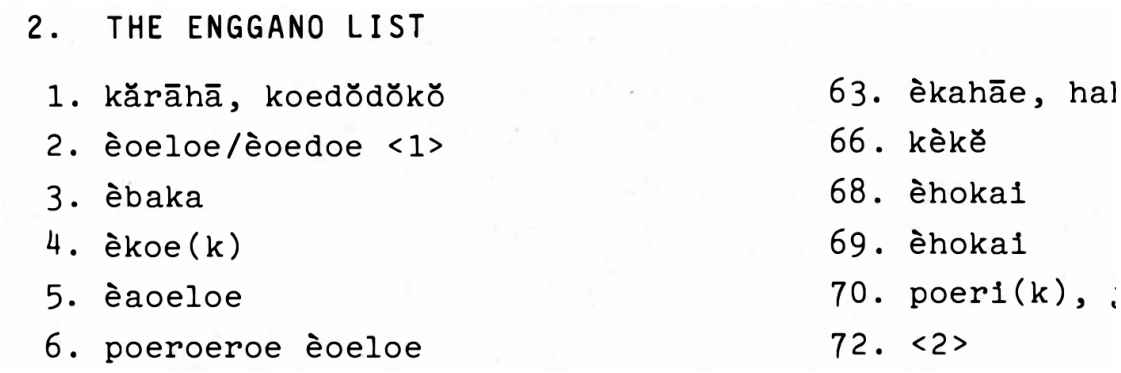
(a) the reference Holle List/NBL (Stokhof, 1980)



5. THE NEW BASIC LIST (NBL)[7]

				1894	1904/
1.	lichaam	body	badan, tubuh	4	4
2.	hoofd	head	kepala	5	5
3.	gezicht, aange-zicht	face	muka, wajah	6	6
4.	voorhoofd	forehead	dahi	24	25
5.	schedel	skull	tempurung kepala	26	27
6.	hoofdhaar	hair	rambut	72	73

(b) Enggano regional list (Stokhof & Almanar, 1987)



2. THE ENGGANO LIST

1. kărāhā, koedŏdŏkŏ
2. èoeloe/èoedoe <1>
3. èbaka
4. èkoe(k)
5. èaoeloe
6. poeroeroe èoeloe

63. èkahāe, hal
66. kèkĕ
68. èhokai
69. èhokai
70. poeri(k),
72. <2>

Figure 1: Correspondence between index numbers of the regional list of Enggano and of the reference Holle List (or the New Basic List [NBL]).

Consider the Enggano word *èbaka* (ID number 3 in Figure 1 (b)). To understand what the word refers to in Dutch, English, and Indonesian/Malay, one must look up the ID number 3 in the separate NBL publication (Figure 1 (a)). In this case, *èbaka* in Enggano refers to '*gezicht, aangezicht*' in Dutch, 'face' in English, and '*muka, wajah*' in Indonesian/Malay. Alternatively, from the perspective of the NBL, linguists could have asked how a given concept is lexicalised in a given language. For example, the concept of '*lichaam*' or 'body' in English (and '*badan, tubuh*' in Indonesian) (ID number 1 in the NBL) can be lexicalised by two forms in Enggano, as shown by those given for the ID number 1 in the Enggano list, viz. *kărāhā* and *koedŏdŏkŏ* (cf. Table 3 and Table 4).

With such paper-based, separate arrangement between words in the regional language-varieties and their translations, one could imagine the amount of manual back-and-forth procedure needed to link the words and their translations. The development of modern data science (Donoho, 2017) allows us to navigate such a challenge predominantly in a computational manner. The Holle List setup can be conceived as disjointed relational data with common keys; these shared keys are the index numbers present in both datasets (the NBL and the given regional list). Then, they can be computationally joined at scale once they are both in computer-readable format (see Wickham et al., 2023, Ch. 19, for the description of table-joining and its computational implementation in the R programming language).

In order to tackle the issue of manual matching, with a desideratum for computational matching between the NBL and the regional lists, the PDF file containing the NBL table (Stokhof, 1980) has been digitalised. The NBL is now available as a computer-readable, searchable, and manipulable database (Rajeg, 2023b); this is also available online as a webpage at https://engganolang.github.io/digitised-holle-list/. The joining of the translations in the digitalised NBL into the regional list data is via the matching keys, viz. the index numbers. This digitalised NBL (in a tab-separated plain-text file) has first been implemented in joining (i) the (also digitalised) regional word list for Enggano with (ii) the corresponding Dutch, English, and Indonesian glosses in the master Holle List (Rajeg, 2023a; Rajeg et al., 2025).

Building on the Enggano research, the current project envisages computational matching between the NBL and all words from the remaining regional languages in the Holle List. To achieve this, the first step is to digitalise the other regional languages from the PDF files into plain texts. Since there are more than one hundred regional lists (comprising ten volumes) in the Holle List, we need to scale-up the digitalisation process.

This paper leverages a cloud computing platform, that is "Google Colaboratory" (https://colab.google/) (Google, 2026), to handle the computational resources (such as the Central Processing Unit and Memory) to run a large-scale digitalisation process of many PDF files into plain-text file. The software that performs the remediation from PDF to plain text is the *Tesseract* O(ptical) C(haracter) R(ecognition) engine (Smith, 2007) (see § 2.2 for further details). Once the digitalisation output of the regional lists has been checked, edited for errors (cf. § 4 esp. in Table 2), and tagged to separate each language in the source PDF file (Figure 11), it is possible to computationally collate what was two-parts paper-based publications into a digital cross-linguistic lexical database in which the words in the regional

languages are matched with their corresponding Dutch, English, and Indonesian/Malay glosses (cf. Table 3).

The future potentials of these large lexical data are diverse. It will open new possibilities for systematic computational historical linguistic analysis in finding relationships between languages (Lai & List, 2023). The database can also be used to study diachronic changes of the same language, combining older datasets with present-day datasets (where available) (Krauße et al., 2024; Rajeg et al., 2024). In the area of lexical semantics, the database could be used to investigate collexification patterns (François, 2008; Rzymski et al., 2020) (Table 3, Table 4). Last but not least, the database contributes to the preservation and empowerment of Indonesian regional languages, especially the older varieties, in the digital realm, corresponding to UNESCO's *Digital Initiatives for Indigenous Languages* (Llanes-Ortiz, 2023).

2. Methodology

This section covers two main points. First, accessing the scanned Holle List publications in PDFs (§ 2.1). Second, the computational processing (i.e., the coding components) on Google Colab to remediate these PDFs into text files (§ 2.2).

Data source acquisition

The PDF files for all eleven volumes of the Holle List series are available as open access on the Open Research Repository of the Australian National University (ANU) library, under the "ANU Asia-Pacific Linguistics/Pacific Linguistics Press" collection. The Holle List publications can be looked up using "Holle Lists" as the search term (see Figure 2).

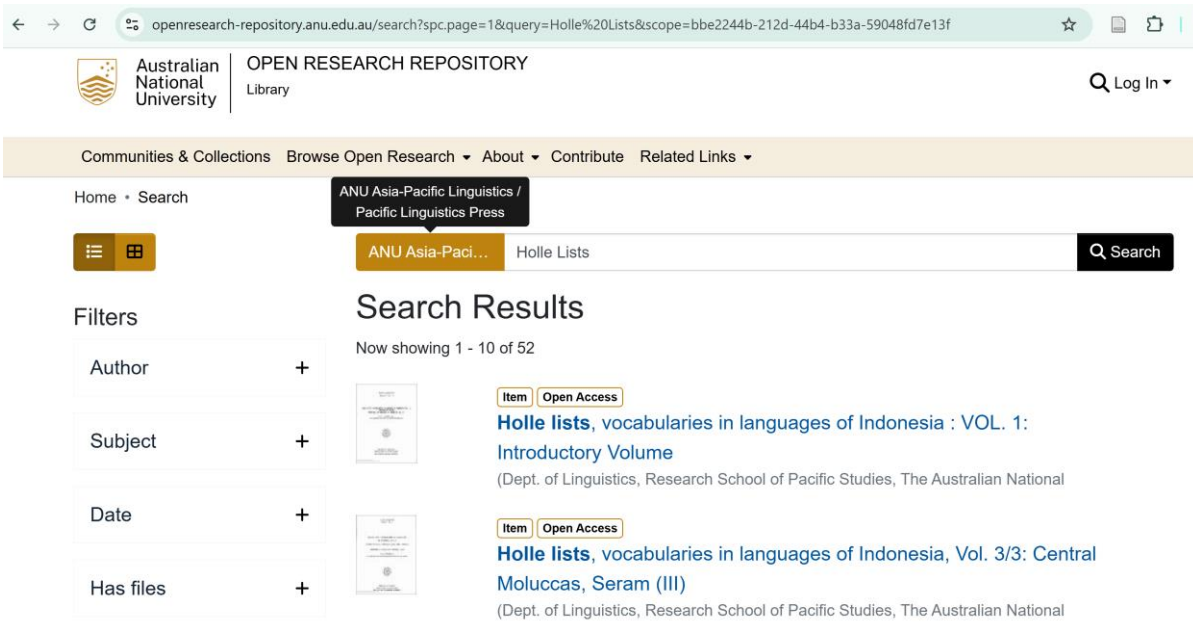


Figure 2: A snippet of the search results for the Holle List publications on the ANU Open Research Repository

The PDF file for every volume was downloaded and uploaded onto the Google Drive of the Holle List project so that it can be accessed during processing in the Google Colab coding environment. This is explained next.

Data processing

To use Google Colab, one only needs to sign up for a Google Account (if they do not already have one). After signing-in to one's Google Account, go to https://colab.google/ and choose the New Notebook option. A computational Jupyter Notebook will be created and stored in the Google Drive folder. This Notebook runs the Python programming language (see Figure 3). All computations for the digitalisation happened on this online, cloud computer on Google Colab.

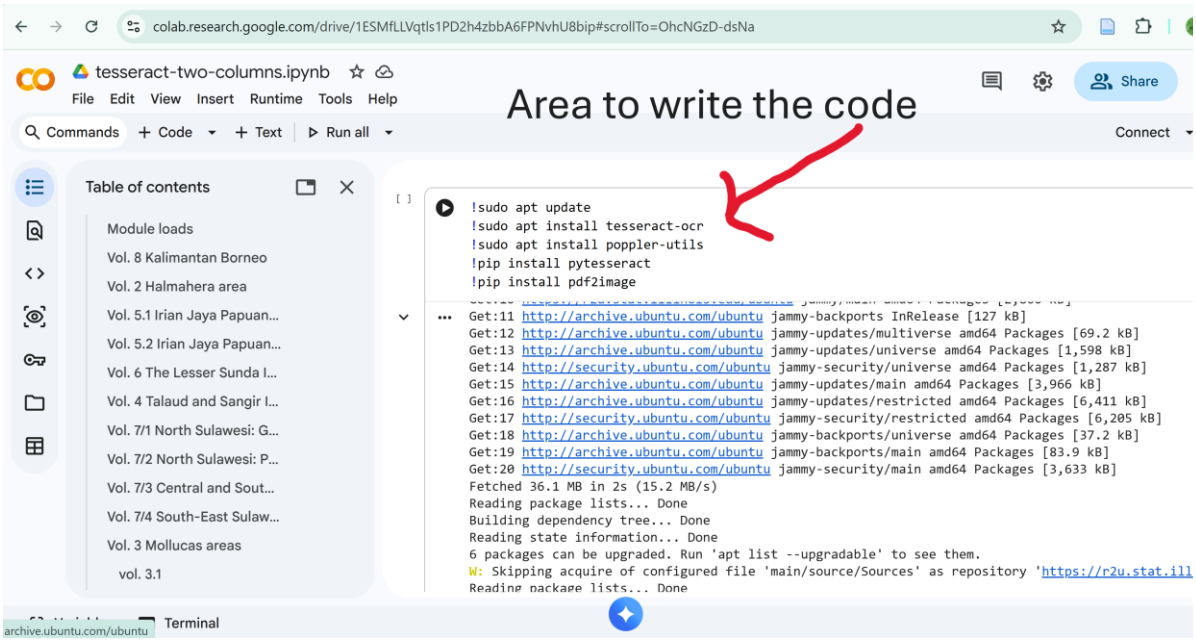


Figure 3: A snippet of an interface of the Jupyter Notebook in Google Colab

The codes shown in Figure 3 are for installing relevant software in the remediation process from PDF into plain text. The *Tesseract OCR* engine (see the code !sudo apt install tesseract-ocr) as well as the Python package/module *pytesseract* (Hoffstaetter, 2024) were installed to allow access to *Tesseract* via Python. Before converting the PDF into plain text with the *pytesseract*, the PDF files must first be converted into images using the pdf2image module (Belval, 2024).

The next step is to load the necessary functionality from the installed modules for PDF-to-text conversion, including a function that allows access to the downloaded PDFs stored on Google Drive (cf. Figure 2). This is shown in Figure 4.

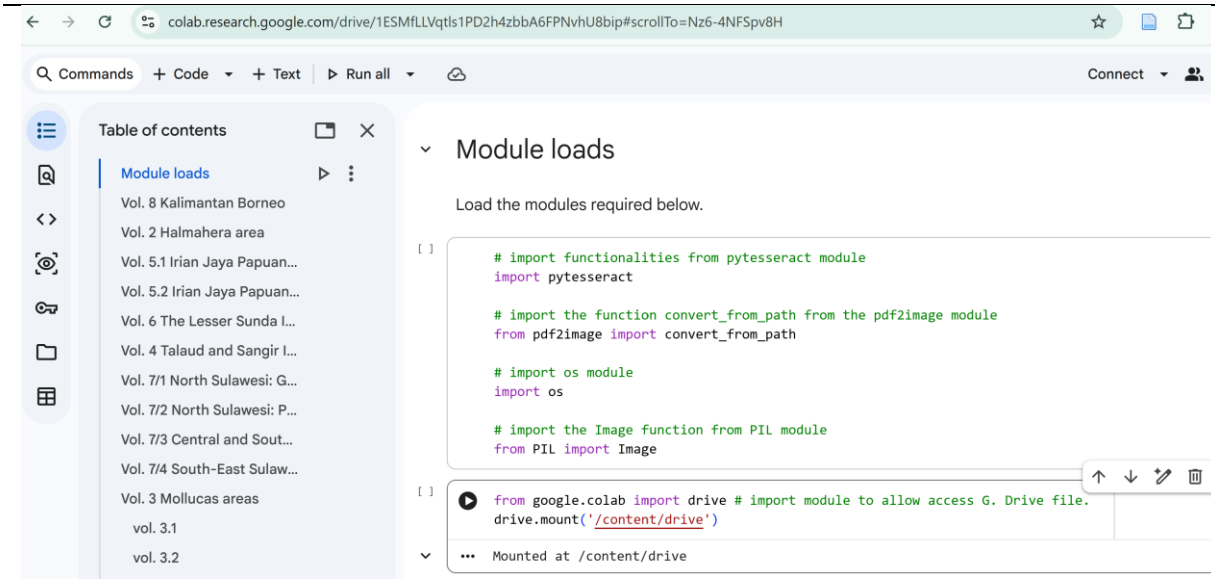


Figure 4: Loading the relevant functions from the installed modules

After loading the relevant functions by executing codes in Figure 4, we need to write custom Python codes for the digitalisation processes. An example is shown in Figure 5 for the processing of the regional Holle List vol. 5/1 for the Papuan and Austronesian languages in the Digul Area, Irian Jaya/West Papua, Indonesia (Stokhof et al., 1982).

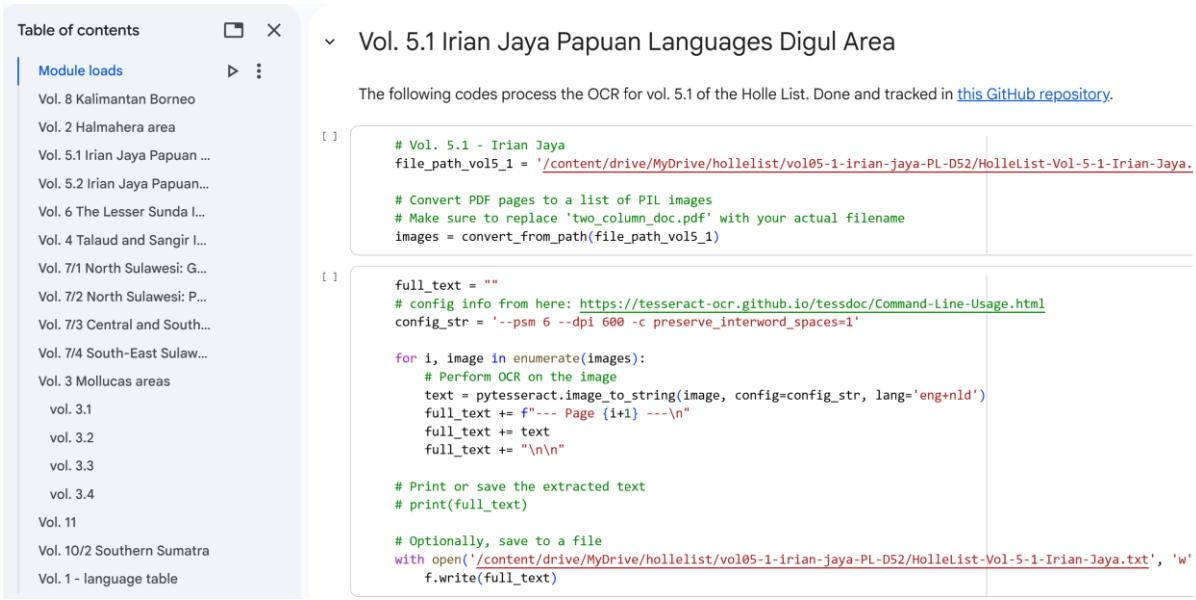


Figure 5: Custom Python codes for the digitalisation of the PDF into plain text

Codes in the upper code block/box in Figure 5 deal with converting the PDF into images by providing the Google Drive path of the PDF.

After that come the codes in the lower code block/box in Figure 5. They cover three aspects. First, setting up the parameters or configuration for the output format of the plain text. Details can be found at https://tesseract-ocr.github.io/tessdoc/Command-Line-Usage.html. Second, creating the main processing iteration to convert the images into a single text file. This is shown from the code line containing for i, image in enumerate(images): up until the line stating full_text += "\n\n". Finally, saving the output into a plain text file; this file is stored on a Google Drive folder specified by the user. The code line shown in Figure 5 indicates that the output is saved with the file name HolleList-Vol-5-1-Irian-Jaya.txt under a folder for the Vol. 5/1, which is in turn inside the hollelist sub-folder in my main Google Drive folder (MyDrive) (see Figure 6).

Figure 6: A Google Drive folder containing the plain-text (.txt) output of the OCR operation for the PDF of the Holle List vol. 5/1

Note that writing the codes as presented in Figure 5 does not necessarily mean that the codes are automatically executed/instructed to produce the output. To run the codes inside a given code block, hover the cursor in the top-left area of the code block until a white rightward arrow (with black background) appears (see the point of the blue arrow in Figure 7), then click on that white arrow.

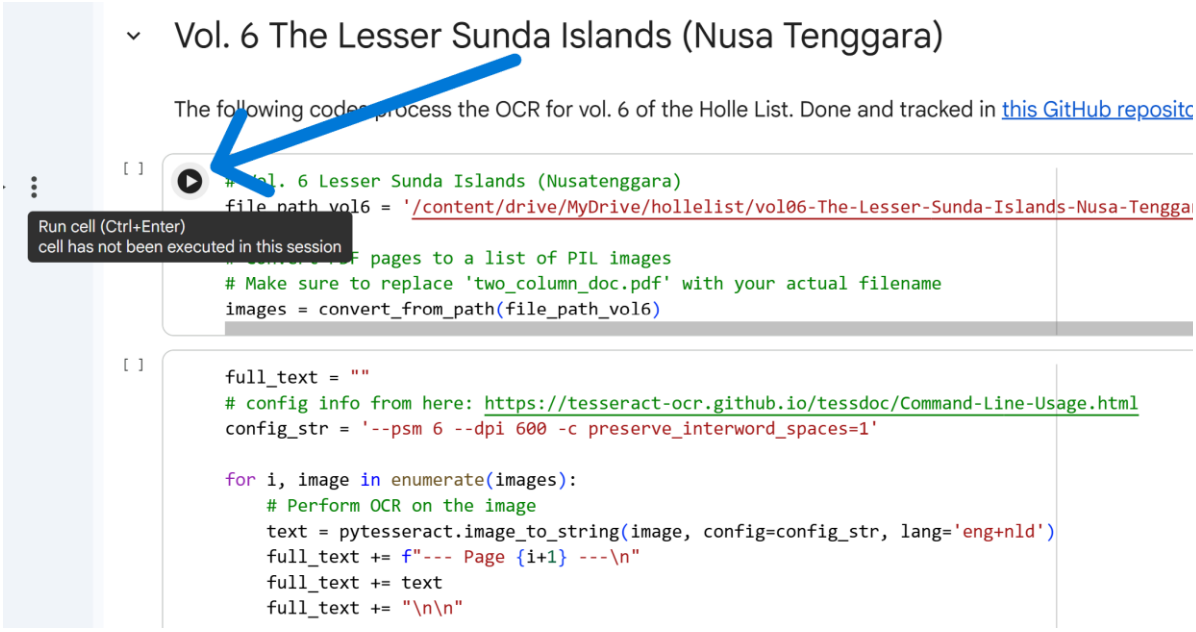


Figure 7: The execution of the codes using the graphical user interface button or keyboard shortcut

Alternatively, ensure the cursor is in the relevant code block and then use the keyboard shortcut Ctrl+Enter to execute the codes in that code block.

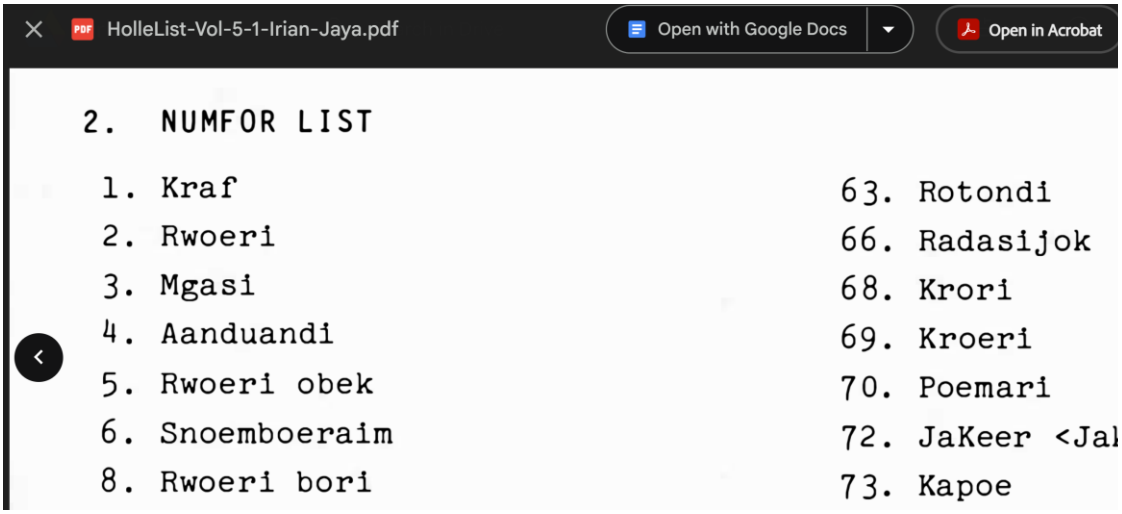
As mentioned in § 1, the execution of these codes is not performed locally on our laptop but online on the cloud, using the Central Processing Unit (CPU) and memory provided by Google Colab. It is also important to note that the conversion processes from image files into plain texts for a given volume could take more than one hour. In this study, the processing times for digitalising each volume were not recorded because the primary goal is not to assess the efficiency and processing times. This paper was written in Quarto and the source code is available at https://github.com/complexico/snbi2026-holle-list.

3. Results

This section presents two sets of results from a single Holle List volume, namely vol. 5/1 (Stokhof et al., 1982) for Austronesian and Papuan languages in the Digul area, Irian Jaya (West Papua), Indonesia. They were chosen to illustrate differences in the quality of the conversion output, depending on the nature of the input characters in the source PDF file. These two sets represent two different languages in that volume, namely Numfor (Stokhof et al., 1982, p. 17) and Digul Mappi (spoken in the area between the Digul River and the Mappi River) (Stokhof et al., 1982, p. 133).

Figure 8 (a) captures several words in Numfor in the PDF source file while Figure 8 (b) shows their corresponding OCR conversion output in plain-text.

(a) the PDF source (Stokhof et al., 1982, p. 17)



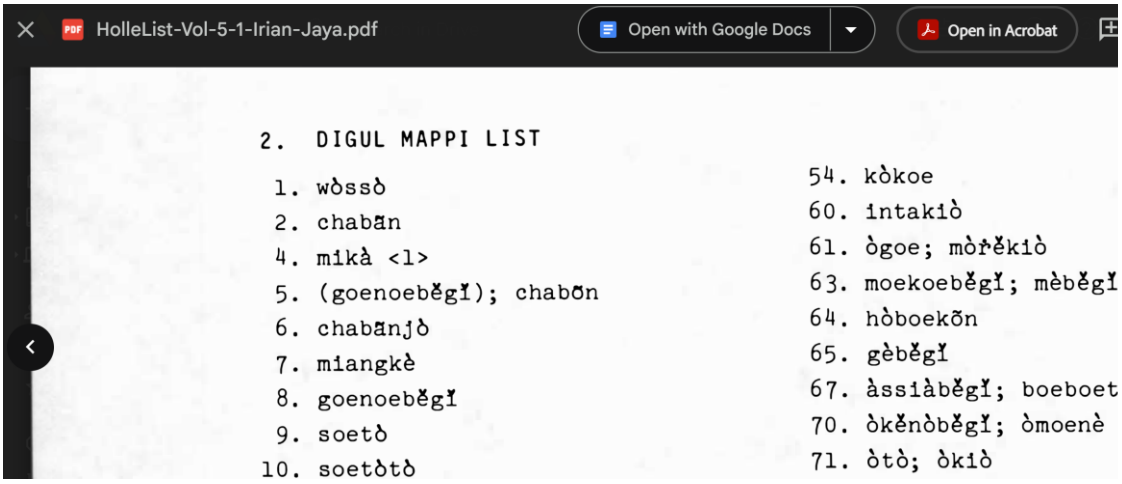
(b) the plain-text output

```
18                                                    NUMFOR

2.  NUMFOR LIST

1. Kraf                                               63. Rotondi

2. Rwoeri                                             66. Radasijok

3. Mgasi                                              68. Krori

4. Aanduandi                                          69. Kroeri

5. Rwoeri obek                                        70. Poemari

6. Snoemboeraim                                       72. JaKeer <Jakee
8. Rwoeri bori                                        73. Kapoe
```

Figure 8: Snippets of the PDF source for the Numfor list and its corresponding plain-text format after the image-to-text conversion

The pair in Figure 8 can now be contrasted with that for the Digul Mappi list in Figure 9.

(a) the PDF source (Stokhof et al., 1982, p. 133)



(b) the plain-text output

```
134                                DIGUL MAPPI
2. DIGUL MAPPI LIST
1. wdssd                                    54. kdkoe
2. chaban                                   60. intakid
|                    4, mika <1>                          61. dgoe;
5. (goenoebégt); chabon                     63. moekoebégi; mébégiki
6. chabanjd                                 64. hdboekdn
7. miangké                                  65. gebégt
8. goenoebégt                               67. Assidbégi; boeboeté
9. soetd                                    70. dkéndbégt; dmoené
10. soetdtd                                  71. dtd; dkid
```

Figure 9: Snippets of the PDF source for the Digul Mappi list and its corresponding plain-text format after the image-to-text conversion

In § 4, these results will be discussed, tailored with a recent data publication of the Holle List vol. 10/3 (Stokhof & Almanar, 1987) when the master NBL has been joined with the regional words into a digital database.

4. Discussion

Careful inspection of the result snippets in Figure 8 and Figure 9 reveal differences in output quality. The words shown for Numfor consist of standard alphabetic characters without diacritics. The OCR output for these words appears correct and closely resembles the source PDF without the need to edit them. In addition, the index numbers are rendered correctly.

Let us now contrast that result with that for Digul Mappi in Figure 9. The Digul Mappi words contain accented characters (i.e., those with diacritics). For example, the word for 'body' (ID no. 1) is *wòssò* in Digul Mappi. The OCR engine did not render or convert the character ò correctly. Hence, from *wòssò* (original) into *wdssd* in the OCR output (Figure 9 (b)). A quick look at some other words containing *ò* in the list in Figure 9 (a) suggests that these combined characters (o + ◌̀) are rendered into *d*:

Table 1: OCR rendering output of some words with *ò* in Figure 9

Holle ID	Gloss	Source PDF	OCR output
6	'hair'	*chabãnj**ò***	*chabanj**d***
9	'ear'	*soet**ò***	*soet**d***
10	'earwax'	*soet**ò**t**ò***	*soet**d**td***
54	'belly'	*k**ò**koe*	*k**d**koe*
60	'side'	*intaki**ò***	*intaki**d***
61	'navel'	***ò**goe*	***d**goe*

With such a pattern from the snippet of the data in Table 1, we might assume that all other occurrences of *ò* will be rendered as *d*. With that assumption, and to expedite correction, we might also be tempted to perform a global find-and-replace procedure: replacing all occurrences of *d* with *ò*. However, the assumption is not fully supported and a one-time find-and-replace procedure is probably not ideal.

Consider another word in Figure 9 (a) referring to 'navel' (ID no. 61), namely *mòřěkiò*. The OCR output of this word shows that the two *ò*-s inside it are rendered differently: as *o* in the first syllable and as *d* in the penultimate syllable. What is more, *d* is not only the OCR rendering for *ò* in the original PDF, but also for the second *à* in the word *àssi**à**bĕgĭ* (ID no. 67) 'buttocks'; this word is rendered as *Assi**d**bégi* in the OCR output (Figure 9 (b)).

With all these issues, the OCR output from image to text requires further *manual* verification and correction. Nevertheless, computational remediation from image to text on the

cloud with *Tesseract* lets us obtain computationally searchable text data relatively quickly, rather than typing manually (cf. below). As we have seen from Figure 8, typically the conversion works well for simple characters (i.e., without diacritics or without other formatting, like underscore). This means that we reduce the time and effort to re-type (or manually type) these well-recognised characters, with effort focused primarily on verifying accented characters[33].

Even if we decided that we would perform the digitalisation task through manual typing, the result of the typing still needs further manual checking and editing for *human error* and/or inconsistencies. This is what was done for an on-going digitalisation work of the Holle List for the remaining Barrier Islands Languages[34] (other than Enggano), off the west coast of Sumatra (Rajeg & Arka, 2024/2025; Stokhof & Almanar, 1987). The results from students' group project[35] (Rajeg & Arka, 2025) to digitalise these lists as an introduction to *WeSay* were first processed, combined, and then manually checked on Google Spreadsheet for tracking changes (cf. Fomin & Toner, 2006, p. 84).

Table 2 shows how tracking changes are organised as a table. The regional lexical items that were manually typed by students (lx_all column) are in a separate column from that containing the corrected forms (called lx_all_correct). The ID column refers to the Holle List index number and a separate column for its correction is also provided but not shown here.

Table 2: A sample of manually corrected lexical items from the languages of Barrier Islands in the Holle List

lang_name	ID	lx_all	lx_all_correct
SalangSigule	1259	*nifeu-eu*	*nifieu-eu*
Semalur	1060	*mamboeïh*	*mamboeĭh*
SiguleSalang	1429	*adénaĕn*	*adénaĕn*
Nias1905	824	*onorafati*	*onarafati*

[33] At the moment, all OCR outputs for the Holle List are still in private repository of the Holle List GitHub (https://github.com/Holle-List) because the results need manual checking and editing.
[34] This work (i) continues a previous digitalisation sub-project of the Enggano Holle List (funded by the Arts and Humanities Research Council [Grant ID: AH/W007290/1] led by the University of Oxford, UK) and (ii) is now part of the Australian Research Council (ARC) research (Grant ID: DP230102019) on Languages of Barrier Islands in Sumatra, Indonesia (led by the Australian National University in Canberra, Australia).
[35] The list of the students' names contributing to this work and the languages they work on is available at https://github.com/complexico/lexico-holle-list-barrier-islands?tab=readme-ov-file#student-contributors. They are also listed as the co-authors for the data publication for each language; access this information at https://github.com/complexico/holle-list-barrier-islands?tab=readme-ov-file#updates-from-students-contributions

Nias1911	912	*dofi*	*dôfi*
Mentawai (n.d.)	536	*goêlai*	*go͡elai*

With that correction set-up, we can (i) compare the original and the edited version, as well as, (ii) for each language, quantify the proportion of manually entered items requiring correction versus those that are already correct. Such a quantification is visualised in Figure 10.

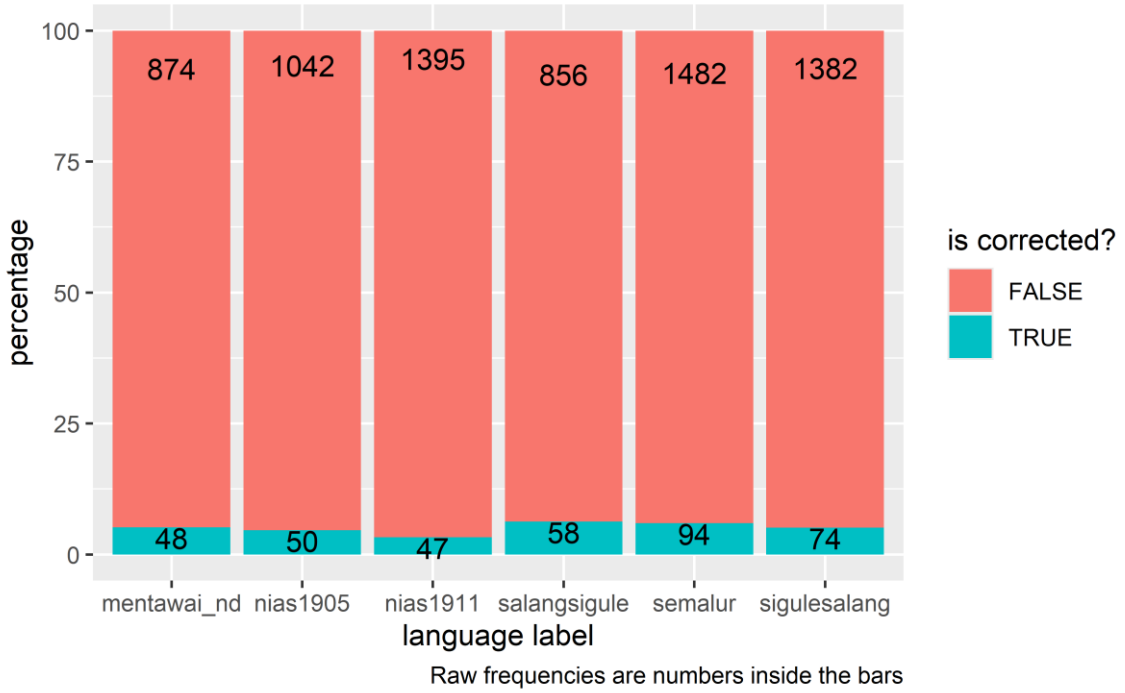


Figure 10: Proportion of lexical items corrected after the manual digitalisation for the Holle List of the Barrier Islands Languages (Stokhof & Almanar, 1987).

Once the word list has been in digital form and corrected, they can be joined with their translations from the reference Holle List (or the *New Basic List*) (cf. § 1). Then, we can explore and computationally search across a set of languages (e.g., from a given volume of the Holle List) how a certain concept is expressed in these languages. An illustration will be given for languages of the Barrier Islands in Sumatra, Indonesia, whose manual digitalisation work has nearly been completed (pending further work for orthography standardisation and phonemic transliteration). Another example is given for languages of Kalimantan where the *Tesseract* OCR output can be processed computationally which is not feasible in PDF format.

Table 3: Forms referring to 'land' as a physical landscape opposing the sea (ID 942), as a state (ID 943), and as a country (ID 326) in the Holle List of the Barrier Islands Languages (Stokhof & Almanar, 1987)

lang_name	Index	Forms	English	Indonesian
Lekon	326	*banò*	Country	negara
Lekon	942/943	*angkal*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
Tapah	326	*banò*	Country	negara
Tapah	942/943	*angkal*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
Simalur	326	*banò*	Country	negara
Simalur	942/943	*angkal*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
Seumalur1912	326	*làntja*	Country	negara
Seumalur1912	942	*banoh*	Land	darat
Seumalur1912	943	*negri*	land (state)	negara
SiguleSalang1912	326	*banoèwa'*	Country	negara
SiguleSalang1912	942/943	*banoewà'*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
SalangSigule1920	326	*banoea*	Country	negara
SalangSigule1920	942/943	*taneu*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
Mentawai (n.d.)	942	*kapi*	Land	darat
Mentawai1933	942	*boeggei*	Land	darat
Nias1905	326	*banoea*	Country	negara
Nias1905	942	*tan̲o*	Land	darat
Nias1905	943	*banoea*	land (state)	negara
Nias1911	326	*banoea*	Country	negara

Nias1911	326	*ori*	Country	negara
Nias1911	326	*tano*	Country	negara
Nias1911	942	*reli danô*	land	darat
Nias1911	942	*tanô*	land	darat
Nias1911	943	*banoea*	land (state)	negara
Nias1911	943	*ôri*	land (state)	negara
Nias1911	943	*tanô*	land (state)	negara
Enggano	326	*èloffo*	country	negara
Enggano	942	*ijokie*	land	darat
Enggano	943	*èlŏpŏ*	land (state)	negara
Enggano	326	*èloppo*	country	negara

Table 3 illustrates the joint database from the Barrier Islands Languages in the Holle List, filtered specifically for word-forms expressing the concept of LAND literally (opposing the sea) (ID 942) and metaphorically as a state and a country respectively (IDs 943 and 326)[36]. It can be seen, for example, that Lekon, Tapah, and Simalur colexify (François, 2008, p. 170; 2022, p. 95) the concept of LAND as a physical landscape (in contrast to sea) (ID 942) and as a state (ID 943) with a single form, namely *angkal*. Yet, interestingly, these three languages have a different form (i.e., *banò*) to lexicalise 'country' (ID 326) (in Dutch elicitation given as *land (staat)*). *Banò* in Lekon, Tapah, and Simalur appears to be cognate with Seumalur (1912) *banoh* (but for ID 942 for land as a landscape), Sigule and Salang (1912) *banoewà'*, Salang and Sigule (1920) *banoea*, and Nias (1905, 1911) *banoea*.

Then, the two Sigule and Salang lists from two different periods of collection and publications also colexify these two concepts (IDs 942 and 943) using two different forms in these two periods. Other languages, such as Nias in 1905, distinguish between land as a physical landscape (*tan̲o*) and land as a state (ID 943) as well as a country (ID 326) (i.e., *banoea* respectively). An interesting observation can be made about Nias (1911). While in Nias 1905 data the form *tan̲o* only refers to land (as an opposition to sea) (ID 942), the same form in 1911 (*tanô*) can also refer to land as a state (ID 943) as well as a country (ID 326); this might suggest

[36] The Dutch elicitation concepts for IDs 326 and 943 are the same, namely *land (staat)* (glossed as 'country' and 'land (state)' respectively in English and as 'negara' in Indonesian). Meanwhile, the Dutch elicitation concept for ID 942 is given as *land (in tegenstelling van zee)* (glossed as 'land' and 'darat' respectively in English and Indonesian).

a semantic extension of cognates for land as a physical landscape in diachronic varieties of Nias. This assumption needs further verification. Moreover, there is a new form lexicalising country or land (state) in Nias (1911), namely *ori*; this form did not exist in Nias (1905), suggesting a lexical expansion to express country or land/state in Nias (1911).

The OCR output in plain-text format from running the *Tesseract* engine (§ 2.2) can also be computationally (via programmatic coding) processed and searched for certain word forms. Before doing this, the output needs to be manually tagged for the language boundary in the text (see the yellow-highlighted line 120 in Figure 11). That is, which part of the output belongs to which language in the original PDF so that computationally we can write code to detect which form belongs to which language.

```
119     <floatingText subtype="PART XV: BORNEO: SOUTH BORNEO">
120     <group xml:lang="OT DANUM DAYAK" xml:id="151">
121
122     <text>
123     <front>
124     --- Page 7 ---
125     OT DANUM DAYAK
126     1. GENERAL INFORMATION
127     1.1. BASIC DATA
128     Language/dialect      : Ot Danum Dayak
129     Number of the list     : 151
130     Mentioned in         : NBG.1917
131     Year of investigation   : 1916-1917
132     Place of investigation : South Borneo (Katingan River)
133     Name of investigator   : H.P. Loing
134     1.2 OTHER DETAILS
135     1.2.1 Collected during the South Borneo Scientific Expedition
136     1916-1917.
```

Figure 11: Snippet of grouping per-language word list using XML tag in the OCR output of a volume

Once all languages in a volume have been tagged as in Figure 11, a programmatic script in R or Python can be designed to process and access the text file. Table 4 shows the tabular output of extracting (with programmatic edit of the OCR output) word forms referring to land as a landscape (ID 942) and/or as a state (ID 943) or country (ID 326) in the Holle List for languages of Borneo/Kalimantan (Stokhof & Almanar, 1986).

Table 4: Forms referring to 'land' as a physical landscape (ID 942) and as a state/country (ID 943/ID 326) in the Holle List of the languages of Kalimantan (Borneo) vol. 8 (Stokhof & Almanar, 1986)

lang_name	Index	Forms	English	Indonesian
Ot Danum Dayak	942	*tana*	land	darat
Banjar	942	*darata n*	land	darat
Ngaju Dayak	942	*petak*	land	darat
Katingan Dayak	942/943	*pètak*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
Maanyan	326	*tanae*	country	negara
Maanyan	942	*tanae*	land	darat
Maanyan	943	*tanae*	land (state)	negara
Ulu Malay	942	*darata n*	land	darat
Kenyah Dayak	326	*tănà*	country	negara
Kenyah Dayak	942	*tănà'*	land	darat
Kenyah Dayak	943	*lĕpo̊*	land (state)	negara
Penihing Dayak	942/943	*tanah*	land [ID_942]; land (state) [ID_943]	darat [ID_942]; negara [ID_943]
Dialect Spoken in The West Kutei	942	*darat*	land	darat

It is important to mention that not all languages/varieties represented in the Holle List vol. 8 for Kalimantan contain forms referring to the concept of LAND in IDs 326, 942, and 943 (i.e., our programmatic search returned null results for these varieties with respect to these two IDs). These varieties are Martapura, Sekajang Dayak, Language Spoken In Matan, and an unidentified variety (Semitau?). The reasons why they are not attested need further investigation.

5. Conclusion

This paper reports on a project to digitalise the scanned PDF files of eleven volumes of the Holle List vocabulary (Stokhof, 1980) into plain texts that are both computationally searchable and manipulatable. The computational tools (the *Tesseract* OCR engine and its Python

implementation) and resources (the Google Colab) used to achieve that have been discussed (§ 2). The aim here is to demonstrate how a paper-based, disjointed set of information relevant to the Humanities, especially Indigenous language preservation, can be re-conceptualised and re-mediated digitally for further uses. The issues arising from the automatic OCR output (§ 4) as well as from the manual typing of the list (Table 2) have also been described. We hope to have provided simple illustrations of the potential that the cross-linguistic database could offer once it is fully prepared in computer-searchable, digital format. One linguistic example that is presented is how the semantic concept of LAND as a physical landscape and as a state is lexicalised in languages of the Barrier Islands, Sumatra (vol. 10/3 in Stokhof & Almanar, 1987) (Table 3) and of the Kalimantan (Borneo) region (vol. 8 in Stokhof & Almanar, 1986) (Table 4). One caveat is that the regional language data in the Holle List represent the state of the language investigated circa the late 19th century and/or early 20th century. Other intricacies include variation of orthography/spelling between investigators of different languages and the need to standardise the Dutch orthography (e.g., the use of *oe* to refer to */u/*), which is the future desideratum of this project.

6. References

Belval, E. (2024). *Pdf2image: A wrapper around the pdftoppm and pdftocairo command line tools to convert PDF to a PIL Image list.* (Version 1.17.0) [Computer software]. https://badge.fury.io/py/pdf2image

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Drucker, J. (2021). *The Digital Humanities coursebook: An introduction to digital methods for research and scholarship*. Routledge. https://doi.org/10.4324/9781003106531

Fomin, M., & Toner, G. (2006). Digitizing a dictionary of Medieval Irish: The eDIL Project. *Literary and Linguistic Computing*, *21*(1), 83–90. https://doi.org/10.1093/llc/fqh050

François, A. (2008). *Semantic maps and the typology of colexification: Intertwining polysemous networks across languages* (M. Vanhove, Ed.; pp. 163–215). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.106.09fra

François, A. (2022). Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift Für Sprachwissenschaft*, *41*(1), 89–123. https://doi.org/10.1515/zfs-2021-2041

Google. (2026). *Google Colaboratory*. https://colab.google/

Hoffstaetter, S. (2024). *pytesseract: A python wrapper for Google's Tesseract-OCR* (Version 0.3.13). https://pypi.org/project/pytesseract/

Holle, K. F. (1894). *Blanco woordenlijst*. Landsdrukkerij. https://hdl.handle.net/2027/coo.31924023363215

Krauße, D., Rajeg, G. P. W., Pramartha, C. R. A., Zobel, E., Nothofer, B., Hemmings, C., Ogilvie, S., Arka, I. W., & Dalrymple, M. (2024). *EnoLEX: A diachronic lexical database for the Enggano language*. https://doi.org/10.25446/oxford.28282169.v1

Lai, Y., & List, J.-M. (2023). Lexical data for the historical comparison of Rgyalrongic languages. *Open Research Europe*, *3*, 99. https://doi.org/10.12688/openreseurope.16017.2

Llanes-Ortiz, G. (2023). *Digital initiatives for indigenous languages*. United Nations Educational, Scientific; Cultural Organization (UNESCO) & STICHTING GLOBAL VOICES. https://unesdoc.unesco.org/ark:/48223/pf0000387186

Rajeg, G. P. W. (2023a). *CLDF dataset of the Enggano word list from 1895 in Stokhof and Almanar's (1987) Holle List*. https://doi.org/10.25446/oxford.23515788

Rajeg, G. P. W. (2023b). *Digitised, searchable Holle List in Stokhof (1980)*. https://doi.org/10.25446/oxford.23205173

Rajeg, G. P. W., & Arka, I. W. (2025). *Group Work in the Lexicography Class for the Holle List of the Barrier Islands Languages of Indonesia* (Version 0.0.1) [Dataset]. https://doi.org/10.17605/OSF.IO/7TQG6

Rajeg, G. P. W., & Arka, I. W. (2025). *The digitised and annotated Holle List of the Barrier Islands languages, off the west coast of Sumatra, Indonesia* [Dataset]. Open Science Framework (OSF). https://doi.org/10.17605/OSF.IO/P8A3R (Original work published 2024)

Rajeg, G. P. W., Arka, I. W., Pramartha, C. R. A., & Sangian, E. Z. (2025). *The data science behind the curation of the Holle List: A case study from the Enggano Holle List and its neighbouring Barrier Islands Languages* [Presentation]. Oceanic and Southeast Asian Navigators (OCSEAN) Conference, Faculty of Humanities, Udayana University. University of Oxford. https://doi.org/10.25446/oxford.29625407.v1

Rajeg, G. P. W., Krauße, D., & Pramartha, C. (2024). EnoLEX: A diachronic lexical database for the Enggano language. In A. Inoue, N. Kawamoto, & M. Sumiyoshi (Eds.), *AsiaLex 2024 Proceedings: Asian Lexicography - Merging cutting-edge and established approaches* (pp. 123–132). https://doi.org/10.25446/oxford.27013864

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., Chang, S., Lai, Y., Morozova, N., Arjava, H., Hübler, N., Koile, E., Pepper, S., Proos, M., Van Epps, B., … List, J.-M. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, *7*(1, 1), 13. https://doi.org/10.1038/s41597-019-0341-x

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *2*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Stokhof, W. A. L. (Ed.). (1980). *Holle lists, vocabularies in languages of indonesia, vol. 1: Introductory volume: Vols. Materials in Languages of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University. https://doi.org/10.15144/PL-D17

Stokhof, W. A. L., & Almanar, A. E. (Eds.). (1986). *Holle lists, vocabularies in languages of Indonesia, Vol. 8: Kalimantan (Borneo)*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University. https://doi.org/10.15144/PL-D69

Stokhof, W. A. L., & Almanar, A. E. (Eds.). (1987). *Holle lists: Vocabularies in languages of indonesia, vol. 10/3: Islands off the west coast of sumatra: Vols. Materials in Languages of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University. http://hdl.handle.net/1885/144589

Stokhof, W. A. L., Saleh-Bronckhorst, L., & Almanar, A. E. (Eds.). (1982). *Holle lists, vocabularies in languages of Indonesia, Vol. 5/1: Irian Jaya: Austronesian languages; Papuan languages, Digul area: Vols. Materials in Languages of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, The Australian National University. http://hdl.handle.net/1885/144577

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data* (Second edition). O'Reilly. https://r4ds.hadley.nz/