Electronically Cataloging Butterflies

Part IB Group Project - Group D

Tutorial

James Alner, Yulong Huang, Francesca Iovu, Jack Parkinson, Suzie Welby and Abigail Wilkinson
2020

Prerequisites

Scanning Instructions:

1) We assume that the user will scan the logbooks with a phone app since they are fragile documents.

2) The user can use a desk phone holder clip to hold the phone in the same position above the pages and avoid unnecessary shadows.

3) The user should try their best to avoid glares in the scans.

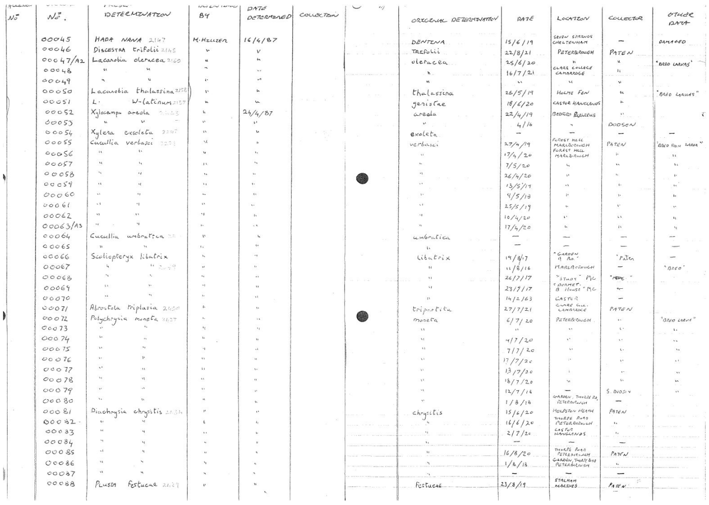
4) They need to make sure that the file is of type PDF of JPEG, and has no cover pages or pages other than simple spreadsheets. Our program only works with consistently formatted spreadsheets.

5) The user should make sure the PDF or JPEG is oriented correctly, with the writing horizontal. The program does not care if the file is in landscape or portrait mode.

6) The user should make sure to scan one page at a time, even if a logbook page is made out of two physical pages. This way a higher resolution image will be produced and word recognition will be more accurate.

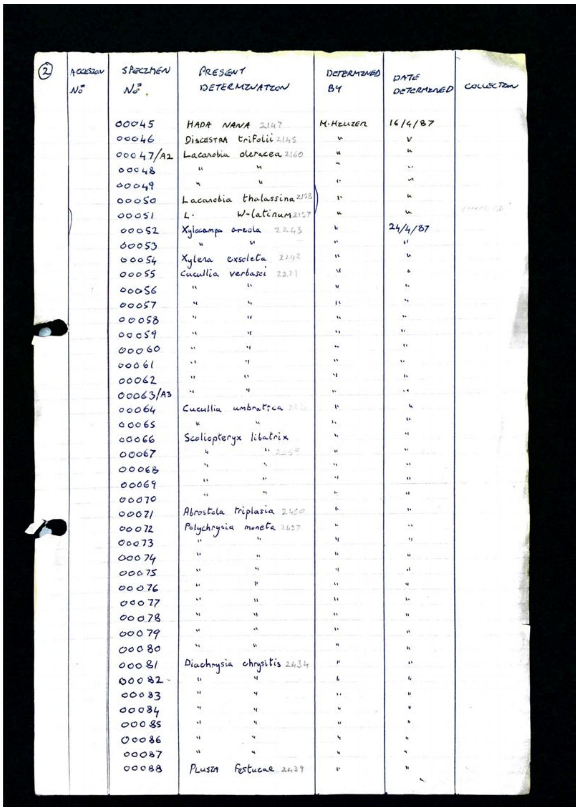
For example:

The scan below comes from a logbook that has 2 physical pages per logbook page (you can see the black dots partially on the first page and fully on the second one) This gives a bad image quality and makes the words harder to get out. Plus, the corners are hard to find on a white background.

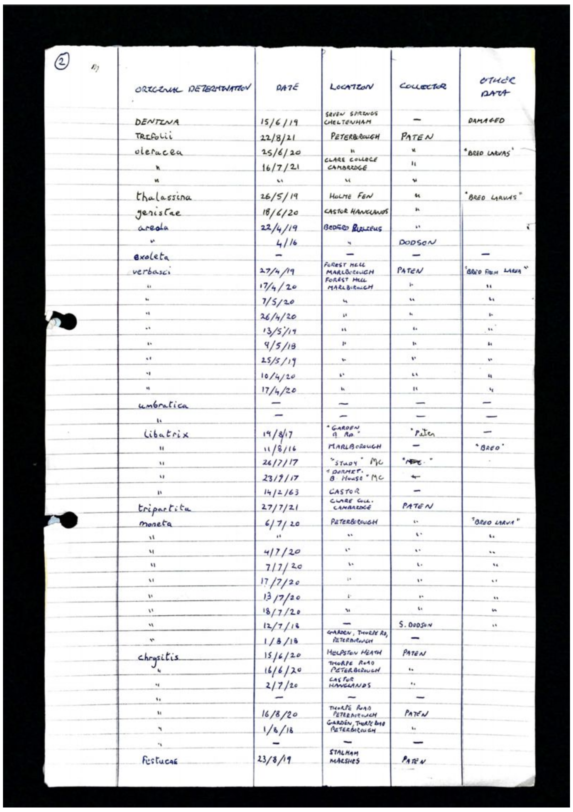


No	No.	DETERMINATION	BY	DATE DETERMINED	COLLECTION	ORIGINAL DETERMINATION	DATE	LOCATION	COLLECTOR	OTHER DATA
	00045	HADA NANA 2147	M.HILLIER	16/4/87		DENTINA	15/6/19	SEVEN SPRINGS CHELTENHAM	—	DAMAGED
	00046	DISCESTRA trifolii 2145	"	"		TRIFOLII	22/8/21	PETERBOROUGH	PATEN	
	00047/A2	Lacanobia oleracea 2160	"	"		oleracea	25/6/20	"	"	"BRED LARVAS"
	00048	" "	"	"		"	16/7/21	CLARE COLLEGE CAMBRIDGE	"	
	00049	" "	"	"		"	"	"	"	
	00050	Lacanobia thalassina 2158	"	"		thalassina	26/5/19	HOLME FEN	"	"BRED LARVAS"
	00051	L. w-latinum 2157	"	"		genistae	18/6/20	CASTOR HANGLANDS	"	
	00052	Xylocampa areola	"	24/4/87		areola	22/4/19	BEDFORD PURLIEUS	"	
	00053	" "	"	"		"	4/16	"	DODSON	
	00054	Xylena exsoleta	"	"		exoleta	—	—	—	—
	00055	Cucullia verbasci	"	"		verbasci	27/4/19	FOREST HILL MARLBOROUGH	PATEN	"BRED FROM LARVA"
	00056	" "	"	"		"	17/4/20	FOREST HILL MARLBOROUGH	"	"
	00057	" "	"	"		"	7/5/20	"	"	"
	00058	" "	"	"		"	26/4/20	"	"	"
	00059	" "	"	"		"	13/5/19	"	"	"
	00060	" "	"	"		"	4/5/18	"	"	"
	00061	" "	"	"		"	25/5/19	"	"	"
	00062	" "	"	"		"	10/4/20	"	"	"
	00063/A3	" "	"	"		"	17/4/20	"	"	"
	00064	Cucullia umbratica	"	"		umbratica	—	—	—	—
	00065	" "	"	"		"	—	—	—	—
	00066	Scoliopteryx libatrix	"	"		libatrix	19/8/17	"GARDEN 9 Rd."	"Paten	—
	00067	" "	"	"		"	11/8/16	MARLBOROUGH	—	"BRED"
	00068	" "	"	"		"	26/7/17	"STUDY" MC	" "	
	00069	" "	"	"		"	23/9/17	"DORMIT. B House" MC	—	
	00070	" "	"	"		"	14/2/63	CASTOR	—	
	00071	Abrostola triplasia 2450	"	"		tripartita	27/7/21	CLARE COLL. CAMBRIDGE	PATEN	
	00072	Polychrysia moneta 2437	"	"		moneta	6/7/20	PETERBOROUGH	"	"BRED LARVA"
	00073	" "	"	"		"	"	"	"	"
	00074	" "	"	"		"	4/7/20	"	"	"
	00075	" "	"	"		"	7/7/20	"	"	"
	00076	" "	"	"		"	17/7/20	"	"	"
	00077	" "	"	"		"	13/7/20	"	"	"
	00078	" "	"	"		"	18/7/20	"	"	"
	00079	" "	"	"		"	12/7/18	—	S. DODSON	"
	00080	" "	"	"		"	1/8/18	GARDEN, THORPE RD, PETERBOROUGH	—	
	00081	Diachrysia chrysitis 2434	"	"		chrysitis	15/6/20	HELPSTON HEATH	PATEN	
	00082	" "	"	"		"	16/6/20	THORPE ROAD PETERBOROUGH	"	
	00083	" "	"	"		"	2/7/20	CASTOR HANGLANDS	"	
	00084	" "	"	"		"	—	—	—	
	00085	" "	"	"		"	16/8/20	THORPE ROAD PETERBOROUGH	PATEN	
	00086	" "	"	"		"	1/8/18	GARDEN, THORPE ROAD PETERBOROUGH	"	
	00087	" "	"	"		"	—	—	—	
	00088	PLUSIA festucae 2439	"	"		festucae	23/8/19	STALHAM MARSHES	PATEN	

Instead, you can scan each page separately and on a black page background so we can identify the corners easier. Also, make sure that the corners are flat (can use blu tack for that). This is how the logbook page above looks when properly scanned:

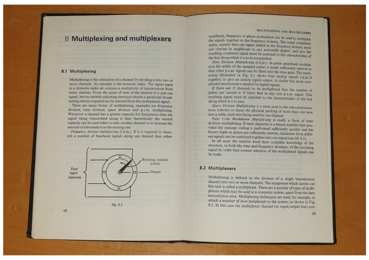


②	ACCESSION No	SPECIMEN No.	PRESENT DETERMINATION	DETERMINED BY	DATE DETERMINED	COLLECTION
		00045	HADA NANA 2147	M. HELLZER	16/4/87	
		00046	DISCESTRA trifolii 2145	"	"	
		00047/A2	Lacanobia oleracea 2160	"	"	
		00048	" "	"	"	
		00049	" "	"	"	
		00050	Lacanobia thalassina 2158	"	"	
		00051	L. W-latinum 2157	"	"	
		00052	Xylocampa areola 2243	"	24/4/87	
		00053	" "	"	"	
		00054	Xylena exsoleta 2242	"	"	
		00055	Cucullia verbasci 2221	"	"	
		00056	" "	"	"	
		00057	" "	"	"	
		00058	" "	"	"	
		00059	" "	"	"	
		00060	" "	"	"	
		00061	" "	"	"	
		00062	" "	"	"	
		00063/A3	" "	"	"	
		00064	Cucullia umbratica [illegible]	"	"	
		00065	" "	"	"	
		00066	Scoliopteryx libatrix	"	"	
		00067	" " 2469	"	"	
		00068	" "	"	"	
		00069	" "	"	"	
		00070	" "	"	"	
		00071	Abrostola triplasia 2450	"	"	
		00072	Polychrysia moneta 2437	"	"	
		00073	" "	"	"	
		00074	" "	"	"	
		00075	" "	"	"	
		00076	" "	"	"	
		00077	" "	"	"	
		00078	" "	"	"	
		00079	" "	"	"	
		00080	" "	"	"	
		00081	Diachrysia chrysitis 2434	"	"	
		00082	" "	"	"	
		00083	" "	"	"	
		00084	" "	"	"	
		00085	" "	"	"	
		00086	" "	"	"	
		00087	" "	"	"	
		00088	PLUSIA festucae 2439	"	"	



②	ORIGINAL DETERMINATION	DATE	LOCATION	COLLECTOR	OTHER DATA
	DENTINA	15/6/19	SEVEN SPRINGS CHELTENHAM	—	DAMAGED
	TRIFOLII	22/8/21	PETERBOROUGH	PATEN	
	oleracea	25/6/20	"	"	"BREO LARVAS"
	"	16/7/21	CLARE COLLEGE CAMBRIDGE	"	
	"	"	"	"	
	thalassina	26/5/19	HOLME FEN	"	"BREO LARVAS"
	genistae	18/6/20	CASTOR HANGLANDS	"	
	areola	22/4/19	BEDFORD PURLIEUS	"	
	"	4/16	"	DODSON	
	exoleta	—	—	—	—
	verbasci	27/4/19	FOREST HILL MARLBOROUGH	PATEN	"BREO FROM LARVA"
	"	17/4/20	FOREST HILL MARLBOROUGH	"	"
	"	7/5/20	"	"	"
	"	26/4/20	"	"	"
	"	13/5/19	"	"	"
	"	9/5/18	"	"	"
	"	25/5/19	"	"	"
	"	10/4/20	"	"	"
	"	17/4/20	"	"	"
	umbratica	—	—	—	—
	"	—	—	—	—
	libatrix	19/8/17	"GARDEN 9 RD"	"Paten	—
	"	11/8/16	MARLBOROUGH	—	"BREO"
	"	26/7/17	"STUDY" MC	"~~MC~~"	
	"	23/9/17	"DORMIT. B House" MC	—	
	"	14/2/63	CASTOR	—	
	tripartita	27/7/21	CLARE COLL. CAMBRIDGE	PATEN	
	moneta	6/7/20	PETERBOROUGH	"	"BREO LARVA"
	"	"	"	"	"
	"	4/7/20	"	"	"
	"	7/7/20	"	"	"
	"	17/7/20	"	"	"
	"	13/7/20	"	"	"
	"	18/7/20	"	"	"
	"	12/7/18	—	S. DODSON	"
	"	1/8/18	GARDEN, THORPE RD, PETERBOROUGH	—	
	chrysitis	15/6/20	HELPSTON HEATH	PATEN	
	"	16/6/20	THORPE ROAD PETERBOROUGH	"	
	"	2/7/20	CASTOR HANGLANDS	"	
	"	—	—	—	
	"	16/8/20	THORPE ROAD PETERBOROUGH	PATEN	
	"	1/8/18	GARDEN, THORPE ROAD PETERBOROUGH	"	
	"	—	—	—	
	Festucae	23/8/19	STALHAM MARSHES	PATEN	

For an actual book, in order to make sure one page is scanned at a time,
a black page (larger than the book) can be put in between the top page currently being scanned and the previous one. Another black page can be used to cover up the second page on top. For example, for a book looking like this:



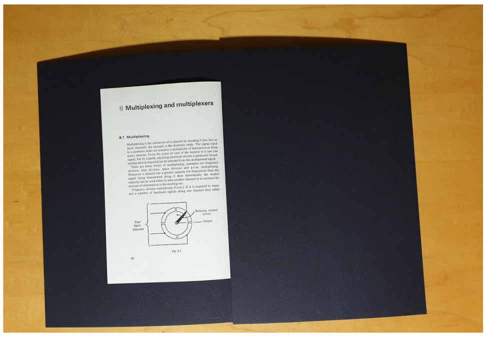
8 Multiplexing and multiplexers

8.1 Multiplexing

8.2 Multiplexers

Fig. 8.1

Place black paper like so to scan the 1st page:

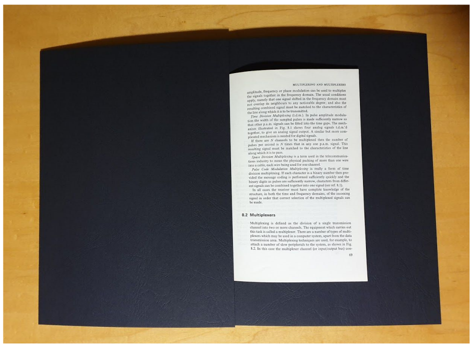


8 Multiplexing and multiplexers

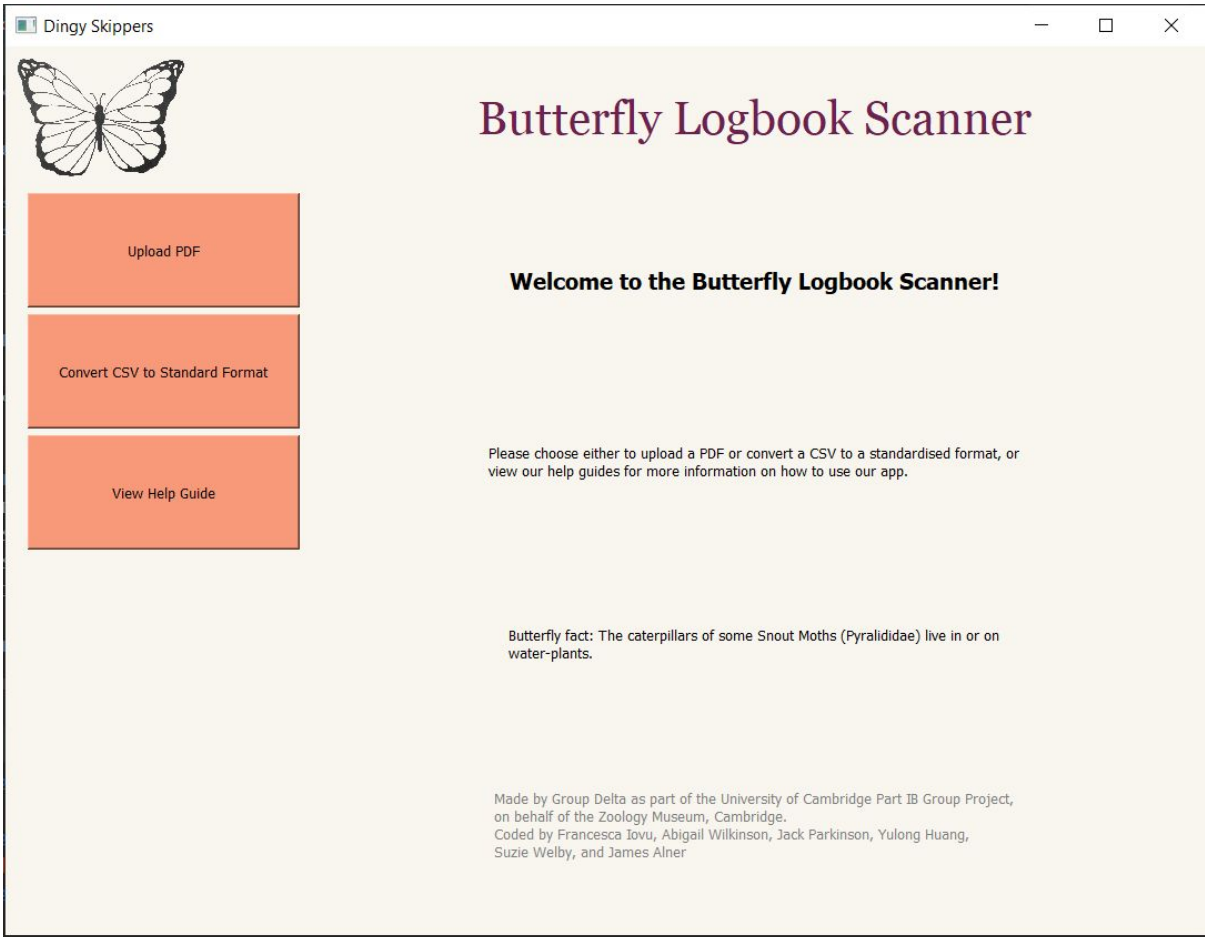
8.1 Multiplexing

Fig. 8.1

And this to scan the 2nd page;



After opening the app, you will see this:



Butterfly Logbook Scanner

Welcome to the Butterfly Logbook Scanner!

Please choose either to upload a PDF or convert a CSV to a standardised format, or view our help guides for more information on how to use our app.

Butterfly fact: The caterpillars of some Snout Moths (Pyralididae) live in or on water-plants.

Made by Group Delta as part of the University of Cambridge Part IB Group Project, on behalf of the Zoology Museum, Cambridge.
Coded by Francesca Iovu, Abigail Wilkinson, Jack Parkinson, Yulong Huang, Suzie Welby, and James Alner

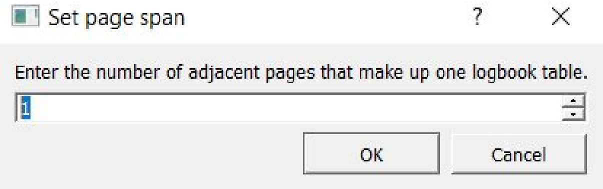
Using the interface

There are three buttons available from the homepage:

- Upload PDF
- Convert CSV to Standard Format
- View Help Guide

Upload pdf

1. Click file icon to upload
2. Select pdf file
3. Click open
4. Click Read Page
5. You will be asked the following:

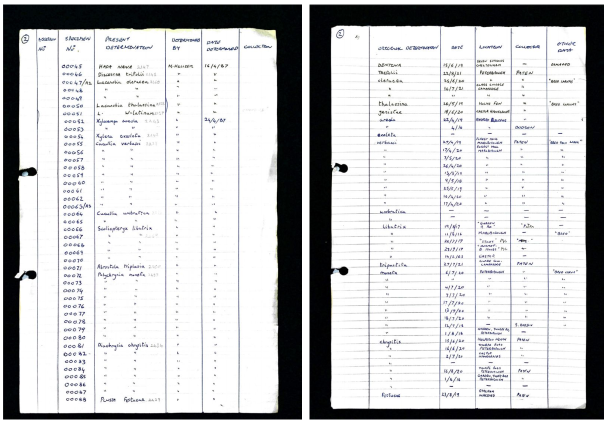


Set page span

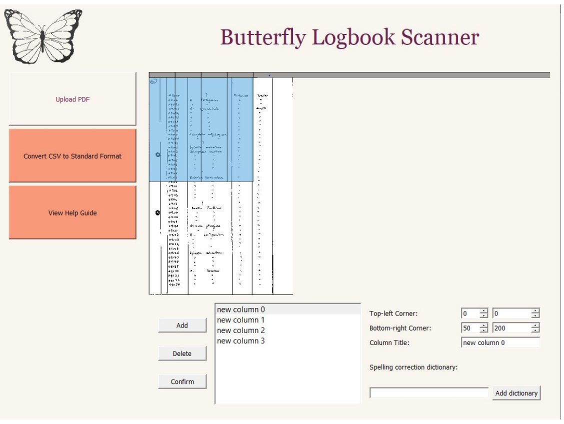
Enter the number of adjacent pages that make up one logbook table.

1

6. Input the number of scanned pages that make up a logbook page:
   For example, for the logbook below, the number is 2:



7. The PDF will be displayed like so:

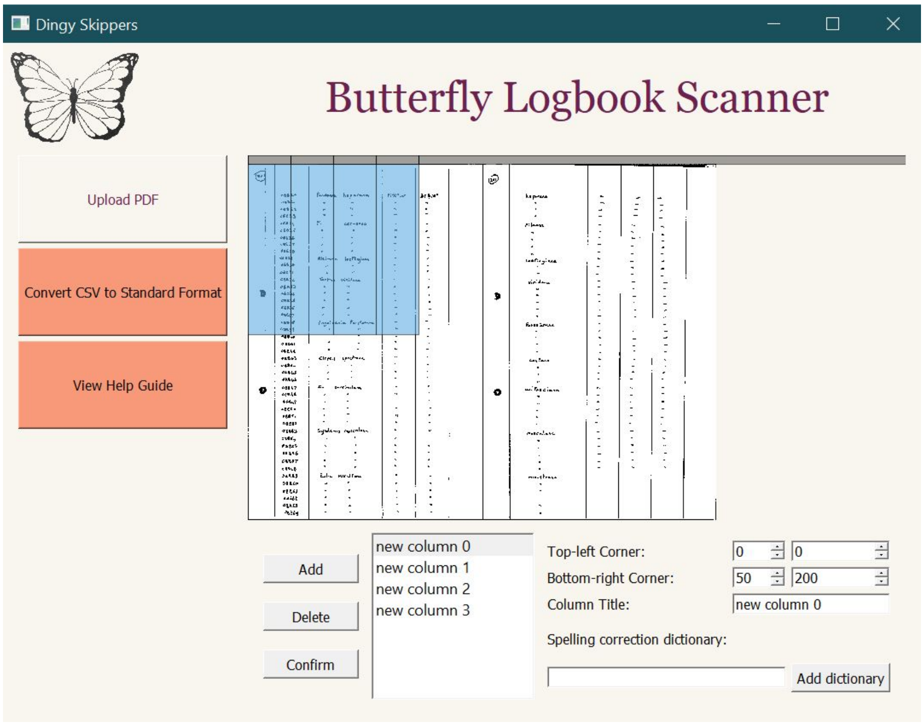


Butterfly Logbook Scanner

new column 0
new column 1
new column 2
new column 3

Top-left Corner: 0 0
Bottom-right Corner: 50 200
Column Title: new column 0
Spelling correction dictionary:

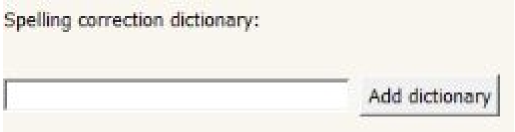
This is an example for one scanned page per logbook page, for more than one, they would be merged together. Like this:



new column 0
new column 1
new column 2
new column 3

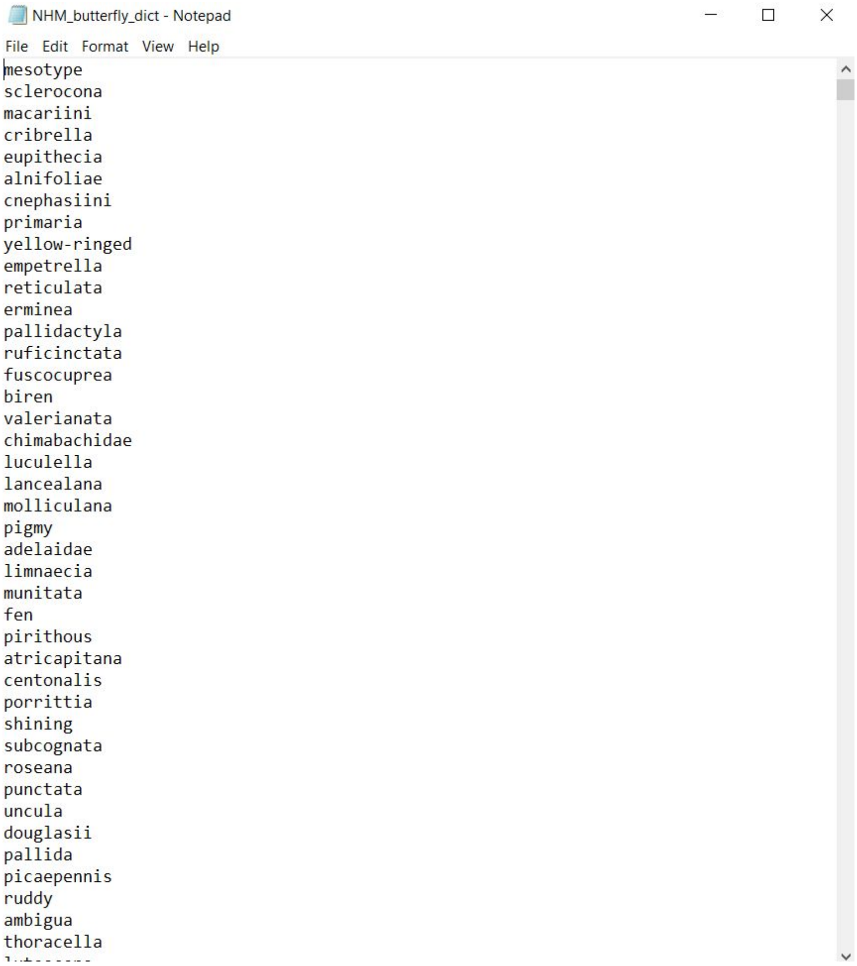
Top-left Corner: 0 0
Bottom-right Corner: 50 200
Column Title: new column 0
Spelling correction dictionary:

8. Now you can add or delete columns and move them around. Make sure the columns line up with the actual ones. Please also remove the header and the footer. You can also get rid of margin columns that are just white and have no header name.
9. Rename all the columns to have the correct headers, these are the headers that will appear in the finished CSV.
10. Before confirming, you can add a txt dictionary that will be used for spelling correction



Spelling correction dictionary:

The dictionary should look like this:



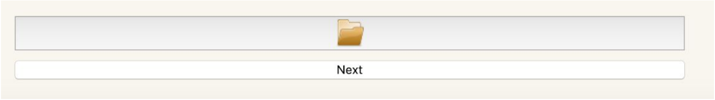
```
mesotype
sclerocona
macariini
cribrella
eupithecia
alnifoliae
cnephasiini
primaria
yellow-ringed
empetrella
reticulata
erminea
pallidactyla
ruficinctata
fuscocuprea
biren
valerianata
chimabachidae
luculella
lancealana
molliculana
pigmy
adelaidae
limnaecia
munitata
fen
pirithous
atricapitana
centonalis
porrittia
shining
subcognata
roseana
punctata
uncula
douglasii
pallida
picaepennis
ruddy
ambigua
thoracella
```

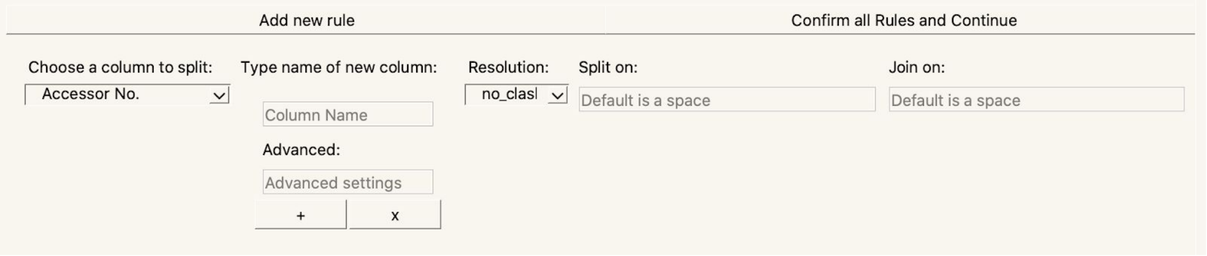
11. After you've made sure the columns are correctly placed and the header and footer removed, you can click confirm.
12. The program will work its magic now!

Convert CSV to Standard format

This feature can be used to convert an existing CSV to a standard format (i.e. to be uploaded to an existing database in a specific format)

1. First select a CSV using the file browser by clicking the folder icon and click next to continue. The only accepted format is a non-empty .CSV file.





Choose a column to split: Accessor No.

Type name of new column:

Resolution: no_clasl

Split on: Default is a space

Join on: Default is a space

2\.

Once on the Rules Page, choose a column to split using the drop-down selection. Type the name of a new column and include the advanced parameters. This parameter can take multiple forms:

- A single index (i.e. “0” to select the first word in each cell), with indices starting from 0.
- A set of indices (i.e. “[0, 1, 3]” selects the 1st, 2nd and 4th words and joins them in that order with the joiner)
- A range of indices (i.e. “1:5” selects the 2nd to 6th words and joins them with the joiner).
- A wildcard “*” which will take the words which are not assigned to any other column for each cell.

Note that negative indices can be used to index from the end (i.e. “-2” refers to the 2nd to last word).

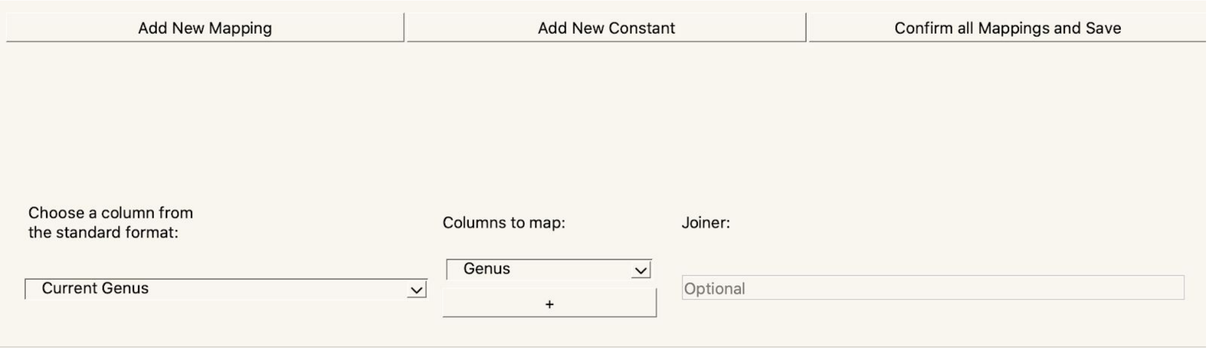
Multiple new columns can be added using the + button, and can be removed using the x button. To remove the current rule press the x button when there is only one new column.

Select the resolution from the drop down list, the default is that there is no clash.

You can also provide which characters to split or join words on, the default is a space.

Once all rules have been filled in and checked, click the Confirm Rules and Continue button to move to the Mappings page.

3. Having created new columns using the rules page, you can then specify what to put in each column of the standard format.
   You can either define a mapping - takes the values from 1 or more columns and combines them - or a constant value for every record in that column.
   Once done with these mappings, you can save the CSV by clicking ‘Confirm all Mappings and Save’



Choose a column from the standard format: Current Genus

Columns to map: Genus

Joiner: Optional

View Help Guide:

1. Will open this tutorial pdf in the operating system’s default application.