用OpenRefine清理資料

TaiBIF 內容經理 劉璟儀

劉璟儀

TaiBIF 內容經理/ GBIF 亞洲區諮詢顧問

melissaliu0520@gmail.com

大學唸公共衛生，碩士轉領域到海洋科學
主要研究鯨豚的重金屬和碳氮同位素

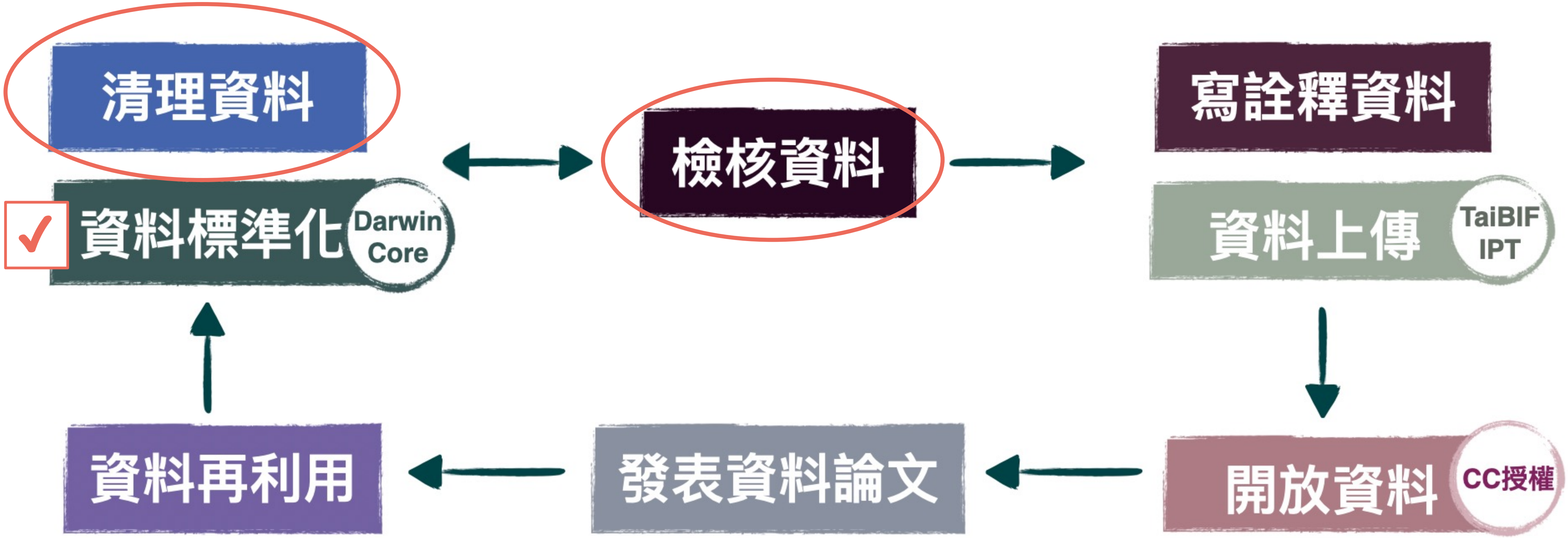
TaiBIF

- 推動及宣傳國內生物多樣性資料標準、資料開放
- 國內生物多樣性資料庫相關單位合作
- TaiBIF 資料庫管理諮詢

GBIF

- 亞洲計畫團隊開放資料諮詢
- 推動亞洲國家節點合作
- 促進亞洲區域開放資料

上傳資料前...
你應該準備好這些事



資料清理小工具

OpenRefine

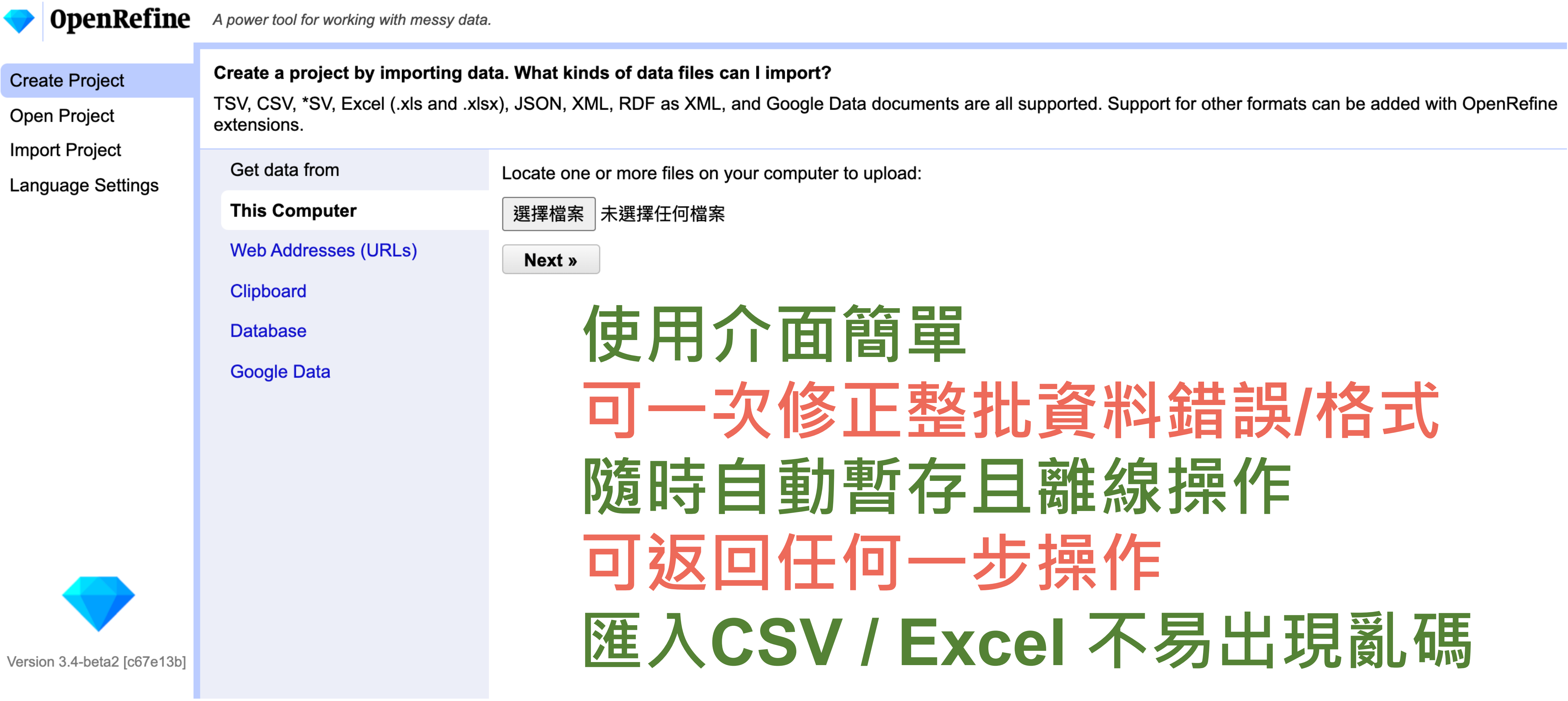
不是資料庫

(無法儲存資料)

與 Excel 的使用方式不同

（只能清理資料）

用 OpenRefine 清資料



OpenRefine *A power tool for working with messy data.*

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

使用介面簡單

可一次修正整批資料錯誤/格式

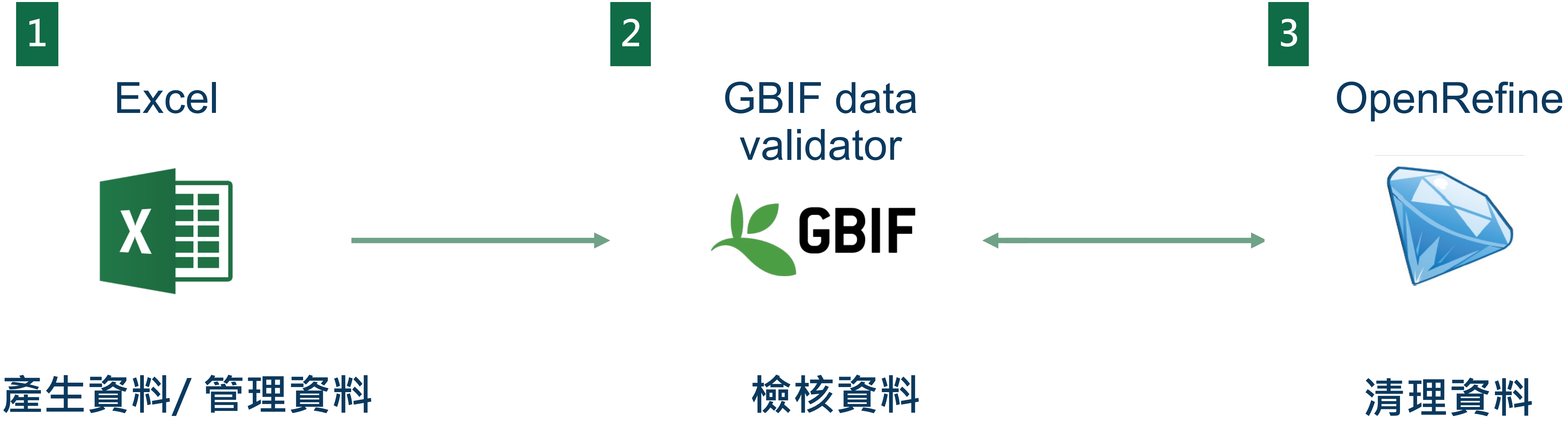
隨時自動暫存且離線操作

可返回任何一步操作

匯入CSV / Excel 不易出現亂碼

資料清理小工具

OpenRefine



1

Excel

產生資料/ 管理資料

2

GBIF data validator

檢核資料

3

OpenRefine

清理資料

清理資料流程

1 先產生並彙整資料

2 驗證資料
GBIF Data Validator

3 查看資料問題
Validation Issues

4 清理資料
OpenRefine

5 上傳資料
TaiBIF IPT

6 再次確認資料問題
GBIF dataset 的 Issues & flags

實作練習

1. 下載 **Data-cleaning-open-refine v20220927**
2. 使用 **GBIF Data Validator** 找出資料錯誤
3. 試著用 **OpenRefine** 找出個別錯誤並修正
4. 利用 **NomenMatch** 比對有效學名/ **Canadensys Coordinate conversion** 轉換座標格式
5. 進階題：使用 GBIF backbone API 新增分類階層欄位

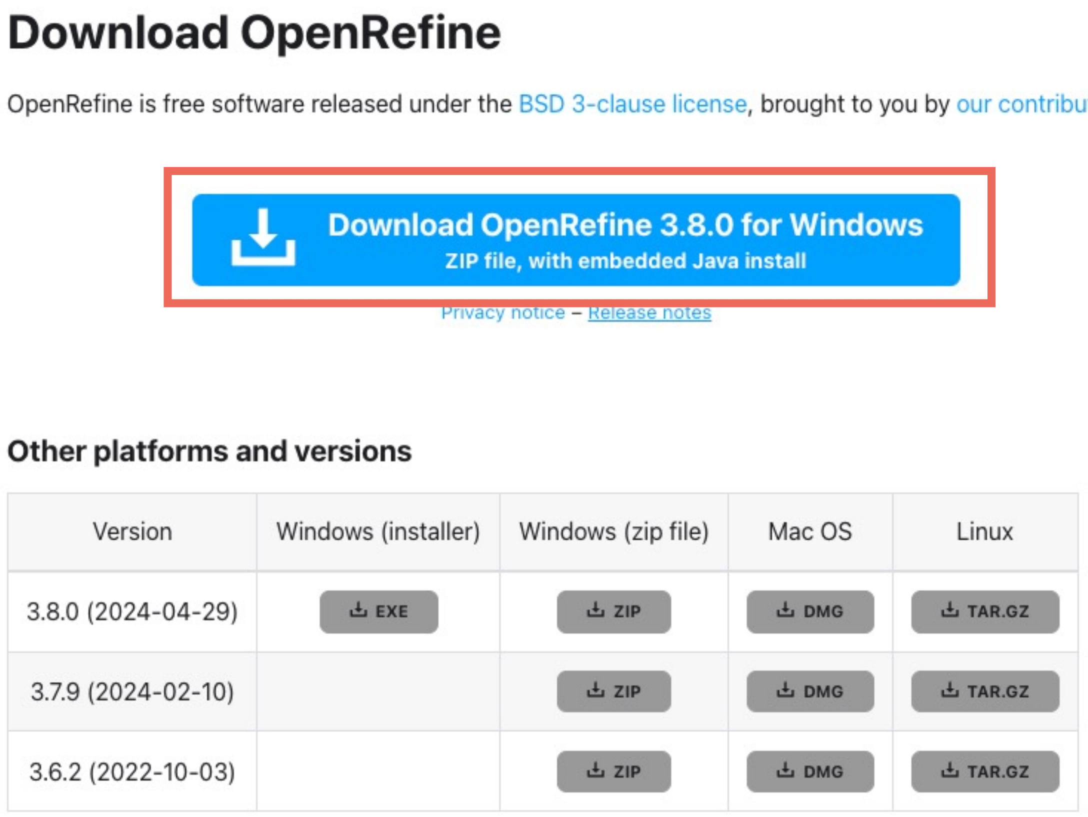
用OpenRefine清理資料

等等會需要用到的連結

練習檔案下載
GBIF 資料驗證工具
NomenMatch 學名比對
Canadensys 座標/日期轉換工具

用 OpenRefine 清資料

下載並安裝在電腦 https://openrefine.org/download.html



Download OpenRefine

OpenRefine is free software released under the BSD 3-clause license, brought to you by our contributors.

Download OpenRefine 3.8.0 for Windows
ZIP file, with embedded Java install

Privacy notice – Release notes

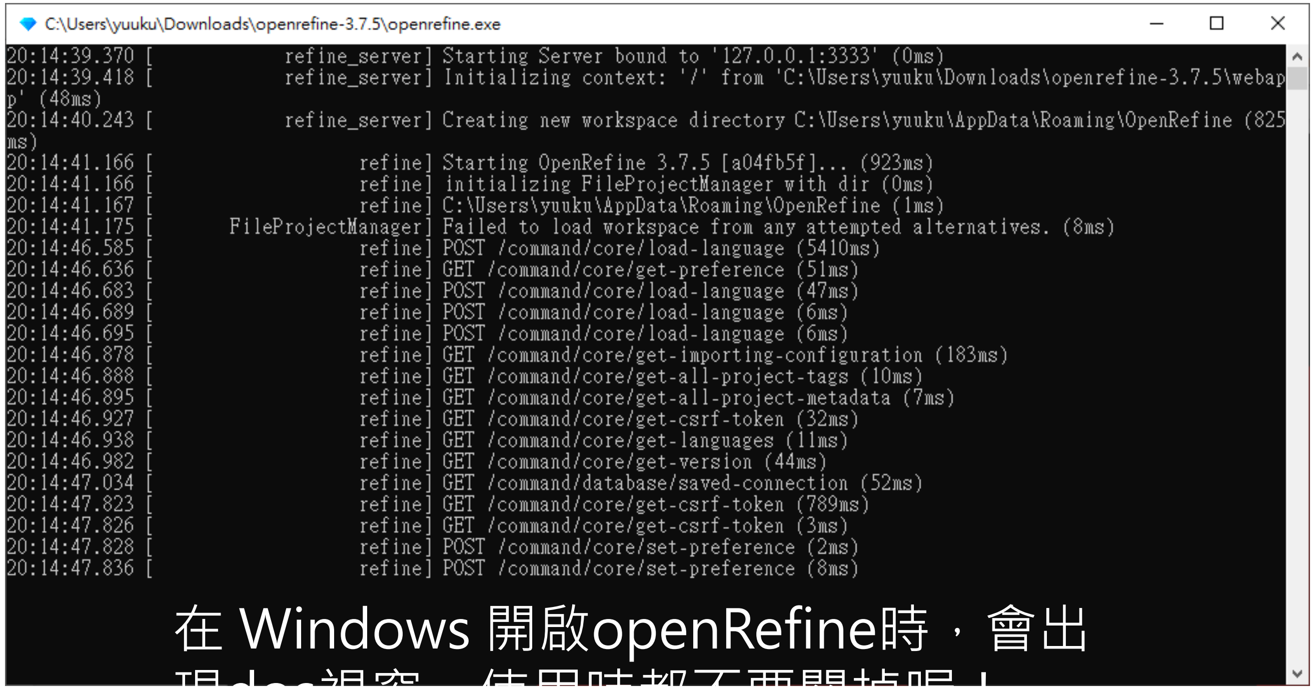
Other platforms and versions

Version	Windows (installer)	Windows (zip file)	Mac OS	Linux
3.8.0 (2024-04-29)	EXE	ZIP	DMG	TAR.GZ
3.7.9 (2024-02-10)		ZIP	DMG	TAR.GZ
3.6.2 (2022-10-03)		ZIP	DMG	TAR.GZ

用 OpenRefine 清資料

下載到電腦打開exe檔

https://openrefine.org/download.html



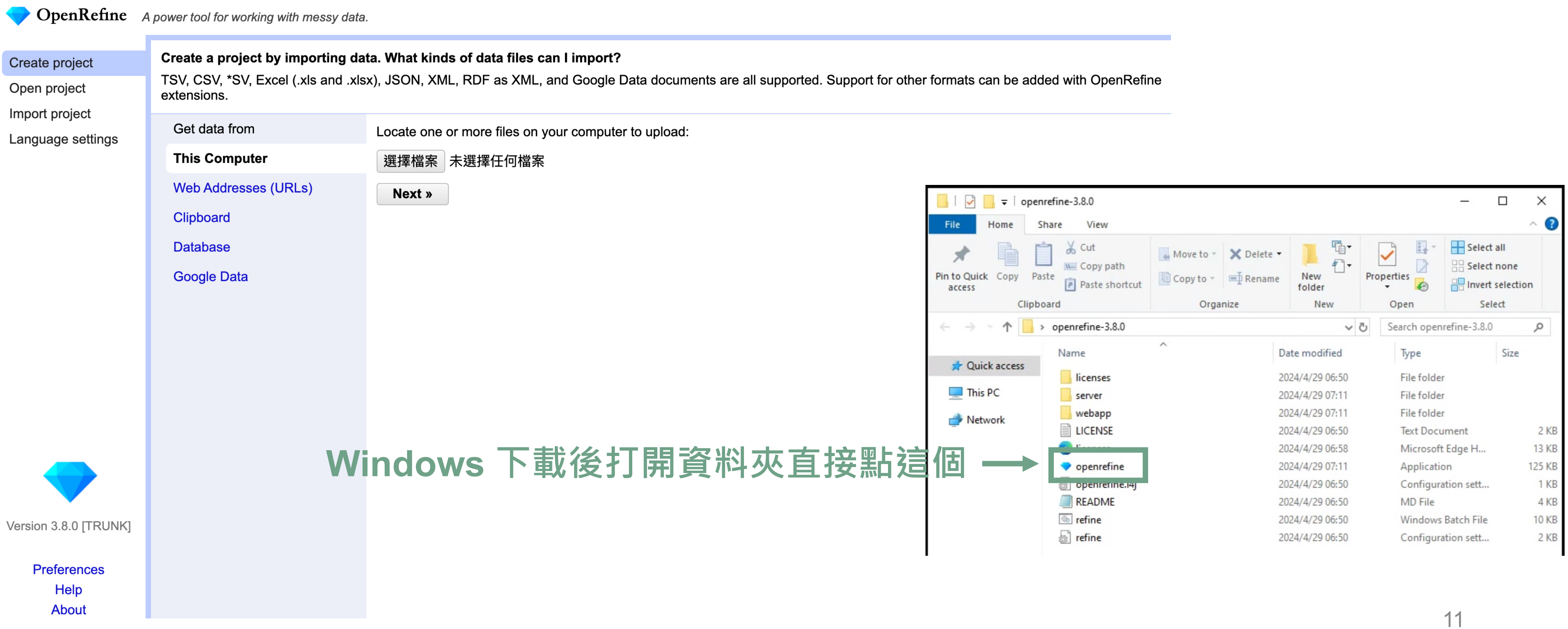
C:\Users\yuuku\Downloads\openrefine-3.7.5\openrefine.exe

```
20:14:39.370 [            refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
20:14:39.418 [            refine_server] Initializing context: '/' from 'C:\Users\yuuku\Downloads\openrefine-3.7.5\webap
p' (48ms)
20:14:40.243 [            refine_server] Creating new workspace directory C:\Users\yuuku\AppData\Roaming\OpenRefine (825
ms)
20:14:41.166 [                   refine] Starting OpenRefine 3.7.5 [a04fb5f]... (923ms)
20:14:41.166 [                   refine] initializing FileProjectManager with dir (0ms)
20:14:41.167 [                   refine] C:\Users\yuuku\AppData\Roaming\OpenRefine (1ms)
20:14:41.175 [       FileProjectManager] Failed to load workspace from any attempted alternatives. (8ms)
20:14:46.585 [                   refine] POST /command/core/load-language (5410ms)
20:14:46.636 [                   refine] GET /command/core/get-preference (51ms)
20:14:46.683 [                   refine] POST /command/core/load-language (47ms)
20:14:46.689 [                   refine] POST /command/core/load-language (6ms)
20:14:46.695 [                   refine] POST /command/core/load-language (6ms)
20:14:46.878 [                   refine] GET /command/core/get-importing-configuration (183ms)
20:14:46.888 [                   refine] GET /command/core/get-all-project-tags (10ms)
20:14:46.895 [                   refine] GET /command/core/get-all-project-metadata (7ms)
20:14:46.927 [                   refine] GET /command/core/get-csrf-token (32ms)
20:14:46.938 [                   refine] GET /command/core/get-languages (11ms)
20:14:46.982 [                   refine] GET /command/core/get-version (44ms)
20:14:47.034 [                   refine] GET /command/database/saved-connection (52ms)
20:14:47.823 [                   refine] GET /command/core/get-csrf-token (789ms)
20:14:47.826 [                   refine] GET /command/core/get-csrf-token (3ms)
20:14:47.828 [                   refine] POST /command/core/set-preference (2ms)
20:14:47.836 [                   refine] POST /command/core/set-preference (8ms)
```

在 Windows 開啟openRefine時，會出現dos視窗，使用時都不要關掉喔！

用 OpenRefine 清資料

要看到這個畫面出現在瀏覽器上才是對的



OpenRefine *A power tool for working with messy data.*

- Create project
- Open project
- Import project
- Language settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

- **This Computer**
- Web Addresses (URLs)
- Clipboard
- Database
- Google Data

Locate one or more files on your computer to upload:

選擇檔案 未選擇任何檔案

Next »

Version 3.8.0 [TRUNK]

Preferences
Help
About

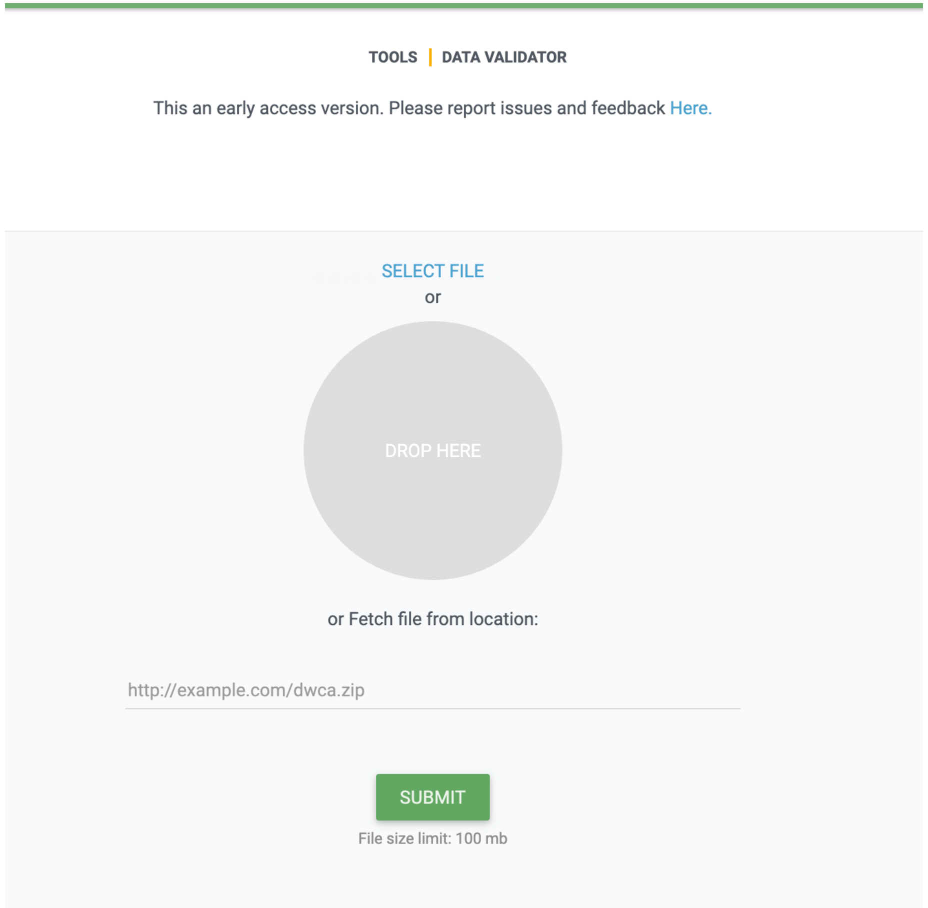
openrefine-3.8.0

Name	Date modified	Type	Size
licenses	2024/4/29 06:50	File folder	
server	2024/4/29 07:11	File folder	
webapp	2024/4/29 07:11	File folder	
LICENSE	2024/4/29 06:50	Text Document	2 KB
licenses	2024/4/29 06:58	Microsoft Edge H...	13 KB
openrefine	2024/4/29 07:11	Application	125 KB
openrefine.l4j	2024/4/29 06:50	Configuration sett...	1 KB
README	2024/4/29 06:50	MD File	4 KB
refine	2024/4/29 06:50	Windows Batch File	10 KB
refine	2024/4/29 06:50	Configuration sett...	2 KB

Windows 下載後打開資料夾直接點這個 →

檢核資料—先找出可能的資料錯誤

- **GBIF data validator** https://www.gbif.org/tools/data-validator



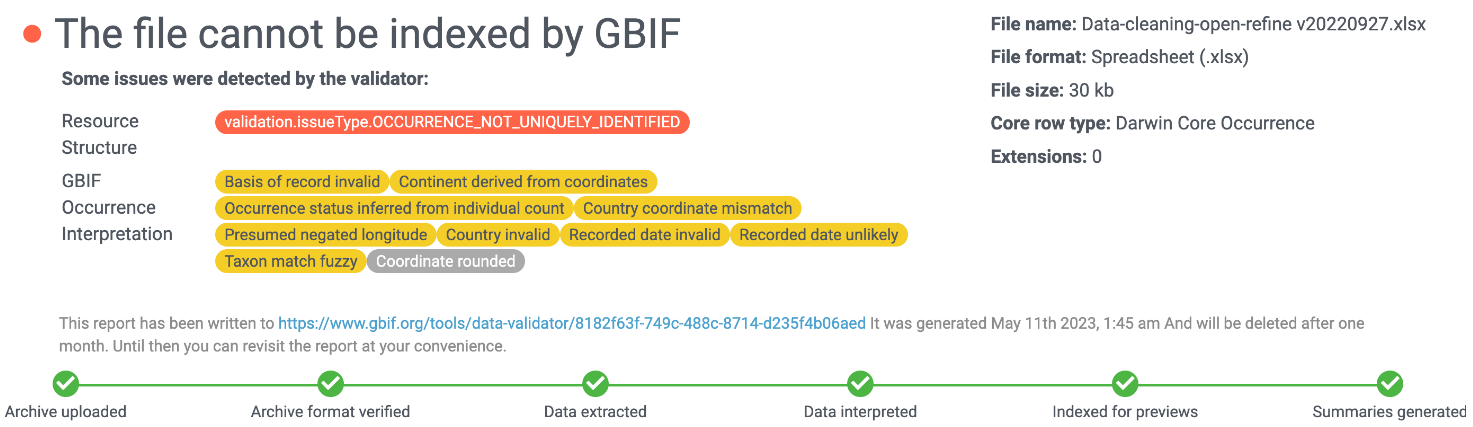
TOOLS | DATA VALIDATOR

This an early access version. Please report issues and feedback Here.

or Fetch file from location:

http://example.com/dwca.zip

File size limit: 100 mb



● The file cannot be indexed by GBIF

Some issues were detected by the validator:

Resource Structure	validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED
GBIF Occurrence Interpretation	Basis of record invalid Continent derived from coordinates Occurrence status inferred from individual count Country coordinate mismatch Presumed negated longitude Country invalid Recorded date invalid Recorded date unlikely Taxon match fuzzy Coordinate rounded

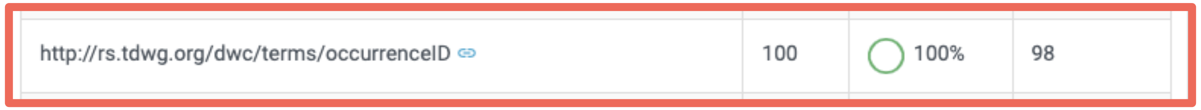
File name: Data-cleaning-open-refine v20220927.xlsx
File format: Spreadsheet (.xlsx)
File size: 30 kb
Core row type: Darwin Core Occurrence
Extensions: 0

This report has been written to https://www.gbif.org/tools/data-validator/8182f63f-749c-488c-8714-d235f4b06aed It was generated May 11th 2023, 1:45 am And will be deleted after one month. Until then you can revisit the report at your convenience.

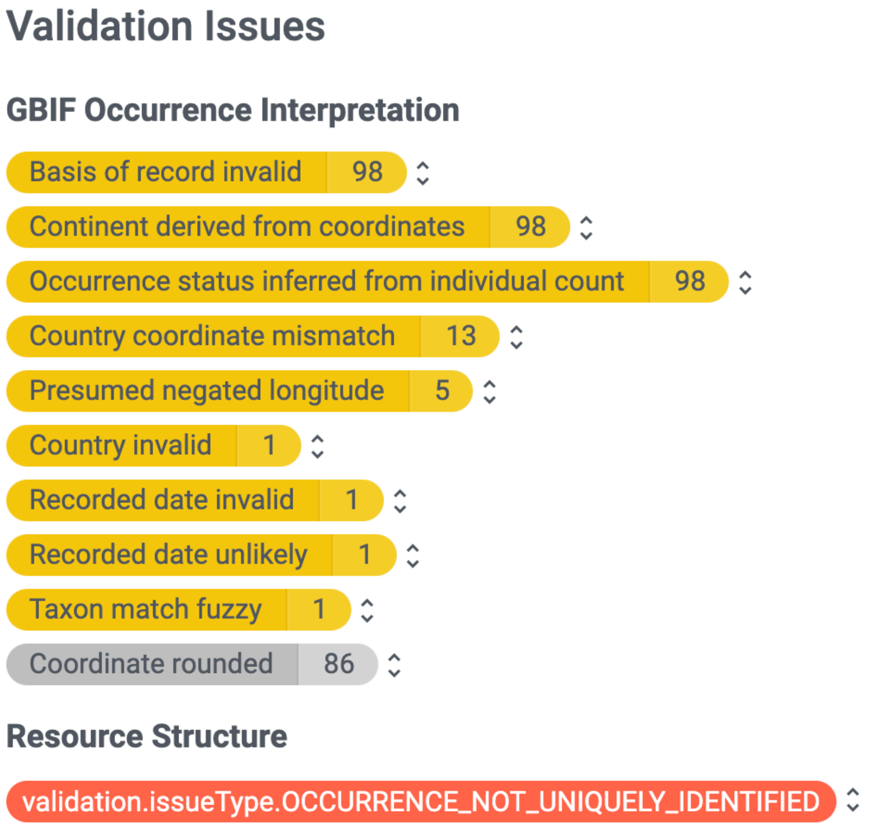
案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符
  decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 County



http://rs.tdwg.org/dwc/terms/occurrenceID	100	○ 100%	98



Validation Issues

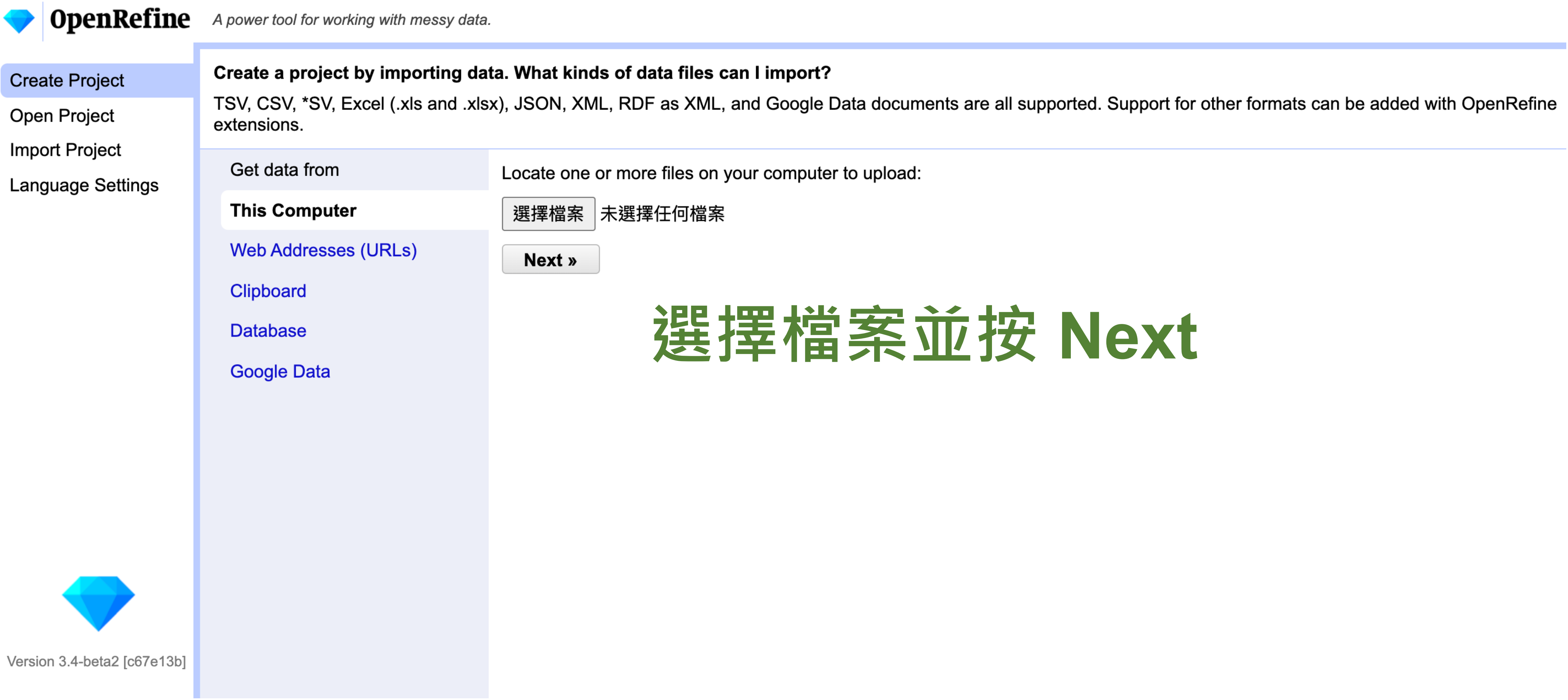
GBIF Occurrence Interpretation

Basis of record invalid	98
Continent derived from coordinates	98
Occurrence status inferred from individual count	98
Country coordinate mismatch	13
Presumed negated longitude	5
Country invalid	1
Recorded date invalid	1
Recorded date unlikely	1
Taxon match fuzzy	1
Coordinate rounded	86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料



OpenRefine *A power tool for working with messy data.*

Create Project
Open Project
Import Project
Language Settings

Version 3.4-beta2 [c67e13b]

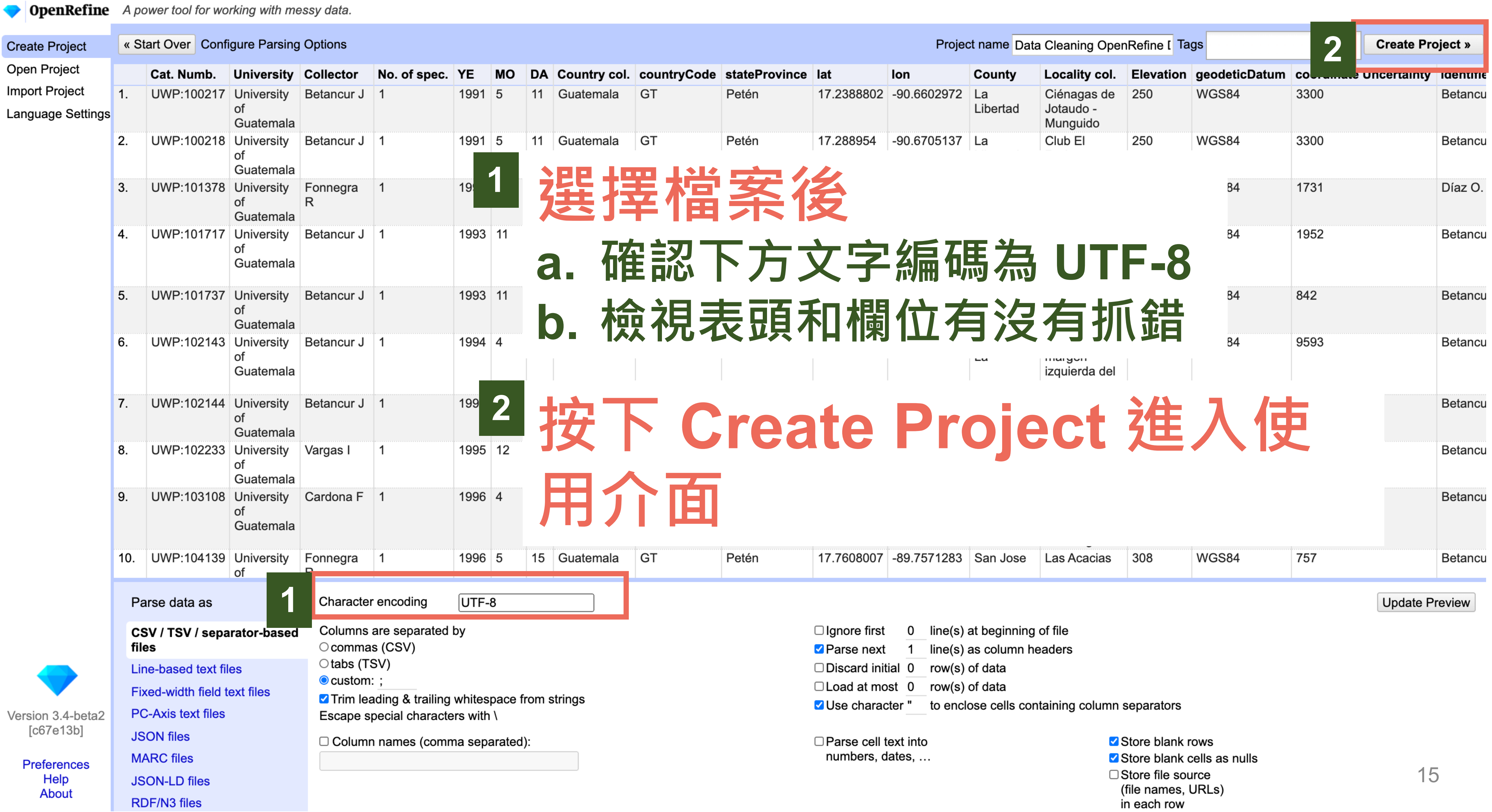
Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

選擇檔案並按 Next

用 OpenRefine 清資料



OpenRefine *A power tool for working with messy data.*

Create Project
Open Project
Import Project
Language Settings

« Start Over Configure Parsing Options

Project name Data Cleaning OpenRefine [Tags Create Project »

	Cat. Numb.	University	Collector	No. of spec.	YE	MO	DA	Country col.	countryCode	stateProvince	lat	lon	County	Locality col.	Elevation	geodeticDatum	coordinate Uncertainty	Identifie
1.	UWP:100217	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT	Petén	17.2388802	-90.6602972	La Libertad	Ciénagas de Jotaudo - Munguido	250	WGS84	3300	Betancu
2.	UWP:100218	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT	Petén	17.288954	-90.6705137	La	Club El	250	WGS84	3300	Betancu
3.	UWP:101378	University of Guatemala	Fonnegra R	1	19											84	1731	Díaz O.
4.	UWP:101717	University of Guatemala	Betancur J	1	1993	11										84	1952	Betancu
5.	UWP:101737	University of Guatemala	Betancur J	1	1993	11										84	842	Betancu
6.	UWP:102143	University of Guatemala	Betancur J	1	1994	4							La	margen izquierda del		84	9593	Betancu
7.	UWP:102144	University of Guatemala	Betancur J	1	199													Betancu
8.	UWP:102233	University of Guatemala	Vargas I	1	1995	12												Betancu
9.	UWP:103108	University of Guatemala	Cardona F	1	1996	4												Betancu
10.	UWP:104139	University of	Fonnegra R	1	1996	5	15	Guatemala	GT	Petén	17.7608007	-89.7571283	San Jose	Las Acacias	308	WGS84	757	Betancu

1 選擇檔案後
a. 確認下方文字編碼為 UTF-8
b. 檢視表頭和欄位有沒有抓錯

2 按下 Create Project 進入使用介面

Parse data as

CSV / TSV / separator-based files
Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
MARC files
JSON-LD files
RDF/N3 files

Character encoding UTF-8

Update Preview

Columns are separated by
○ commas (CSV)
○ tabs (TSV)
● custom: ;
☑ Trim leading & trailing whitespace from strings
Escape special characters with \

□ Column names (comma separated):

□ Ignore first 0 line(s) at beginning of file
☑ Parse next 1 line(s) as column headers
□ Discard initial 0 row(s) of data
□ Load at most 0 row(s) of data
☑ Use character " to enclose cells containing column separators

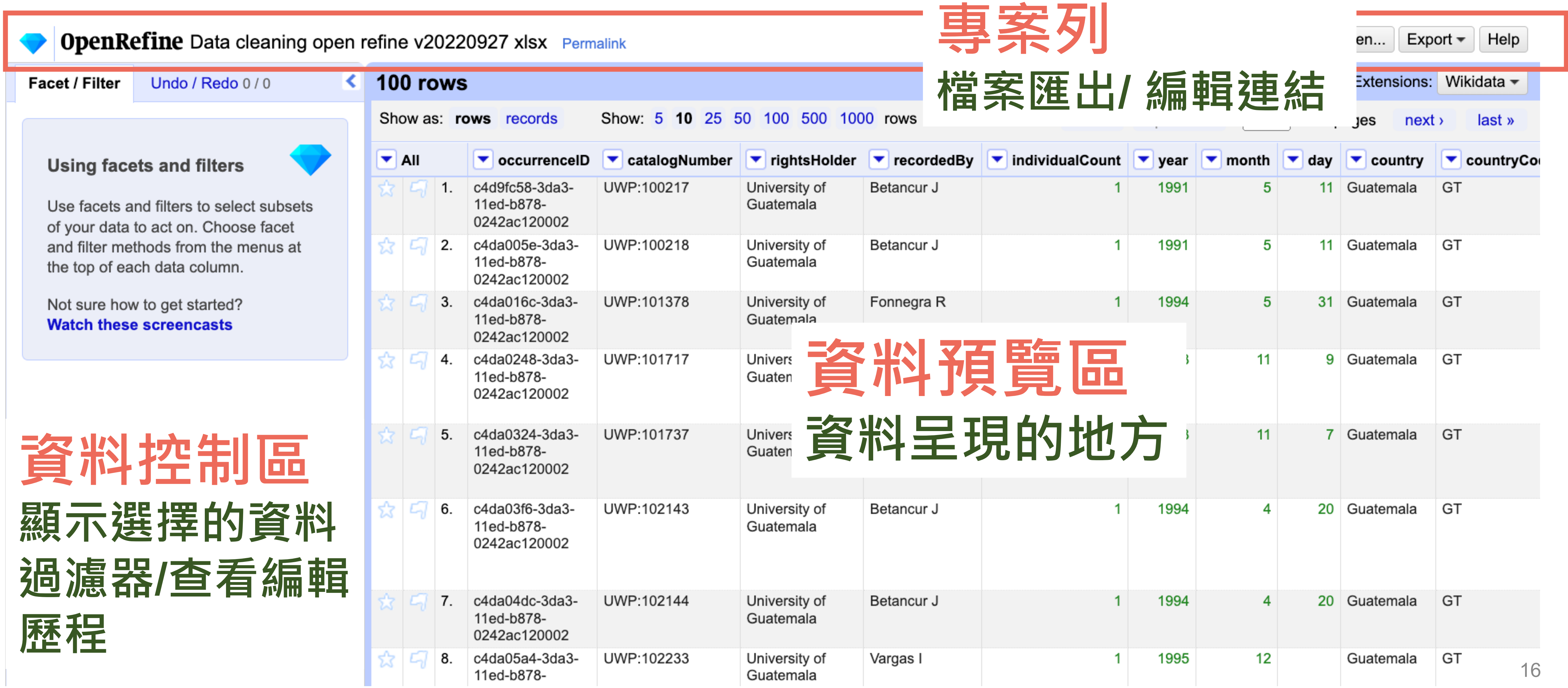
□ Parse cell text into numbers, dates, ...

☑ Store blank rows
☑ Store blank cells as nulls
□ Store file source (file names, URLs) in each row

Version 3.4-beta2
[c67e13b]

Preferences
Help
About

用 OpenRefine 清資料



OpenRefine Data cleaning open refine v20220927 xlsx Permalink

en... Export ▾ Help

專案列

檔案匯出/ 編輯連結

Facet / Filter | Undo / Redo 0 / 0

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

100 rows

Extensions: Wikidata ▾

Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows

ges next › last »

All			occurrenceID	catalogNumber	rightsHolder	recordedBy	individualCount	year	month	day	country	countryCo
☆		1.	c4d9fc58-3da3-11ed-b878-0242ac120002	UWP:100217	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT
☆		2.	c4da005e-3da3-11ed-b878-0242ac120002	UWP:100218	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT
☆		3.	c4da016c-3da3-11ed-b878-0242ac120002	UWP:101378	University of Guatemala	Fonnegra R	1	1994	5	31	Guatemala	GT
☆		4.	c4da0248-3da3-11ed-b878-0242ac120002	UWP:101717	Univers Guaten				11	9	Guatemala	GT
☆		5.	c4da0324-3da3-11ed-b878-0242ac120002	UWP:101737	Univers Guaten				11	7	Guatemala	GT
☆		6.	c4da03f6-3da3-11ed-b878-0242ac120002	UWP:102143	University of Guatemala	Betancur J	1	1994	4	20	Guatemala	GT
☆		7.	c4da04dc-3da3-11ed-b878-0242ac120002	UWP:102144	University of Guatemala	Betancur J	1	1994	4	20	Guatemala	GT
☆		8.	c4da05a4-3da3-11ed-b878-	UWP:102233	University of Guatemala	Vargas I	1	1995	12		Guatemala	GT

資料預覽區

資料呈現的地方

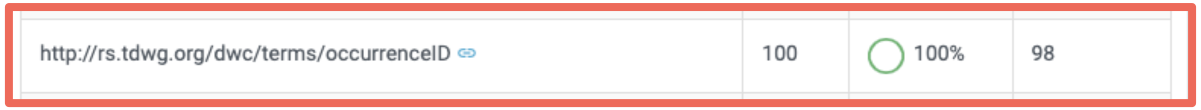
資料控制區

顯示選擇的資料過濾器/查看編輯歷程

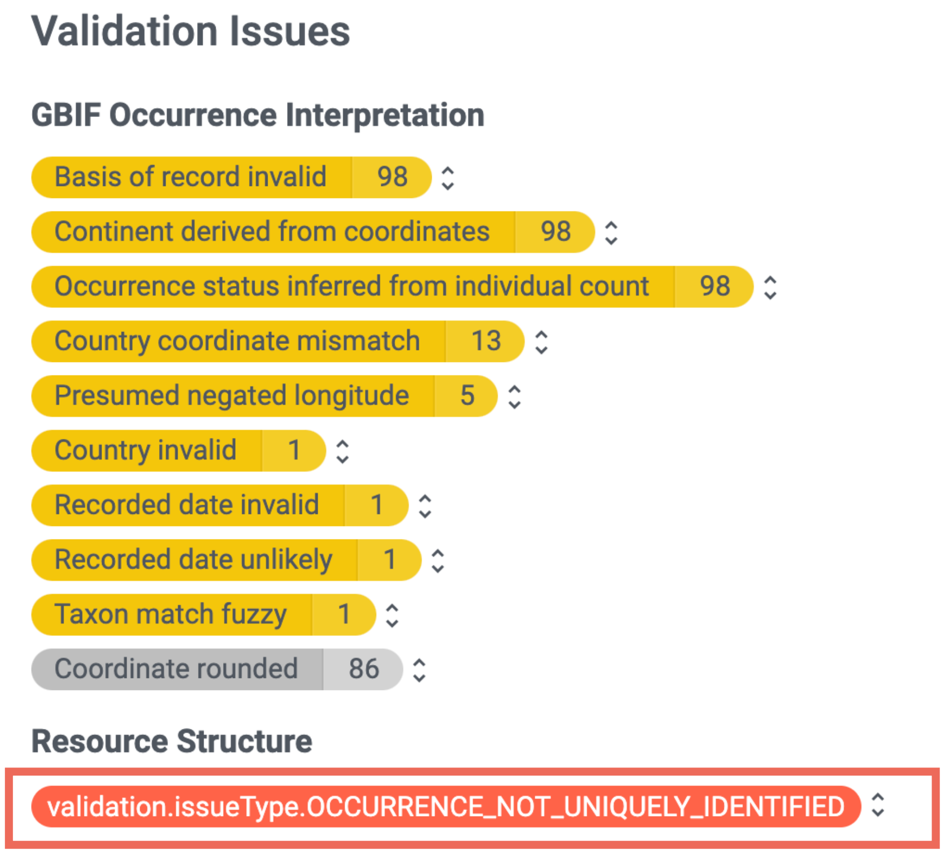
案例練習- 進階作業

資料問題

- **找出重複 ID** occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符 decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 County



http://rs.tdwg.org/dwc/terms/occurrenceID	100	○ 100%	98



Validation Issues

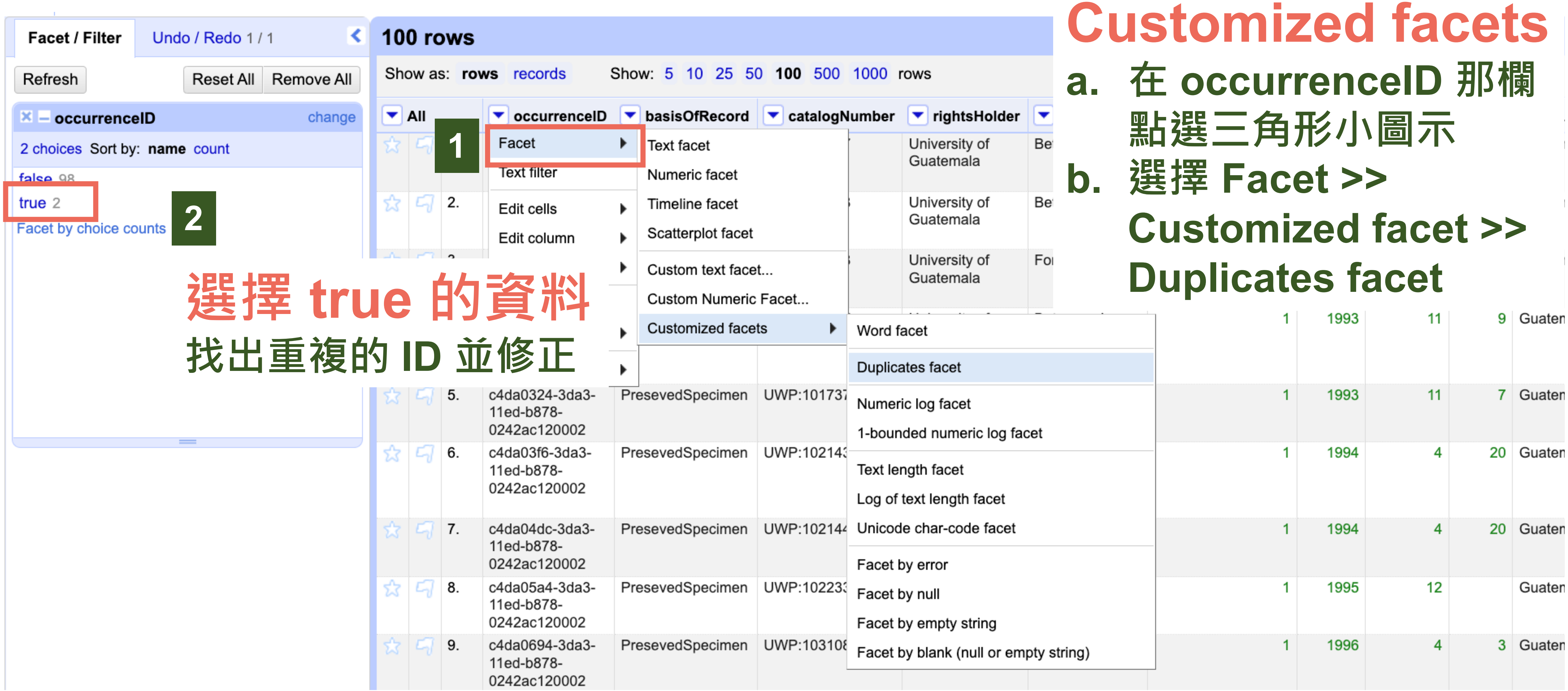
GBIF Occurrence Interpretation

Basis of record invalid	98
Continent derived from coordinates	98
Occurrence status inferred from individual count	98
Country coordinate mismatch	13
Presumed negated longitude	5
Country invalid	1
Recorded date invalid	1
Recorded date unlikely	1
Taxon match fuzzy	1
Coordinate rounded	86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料- 找出重複 ID



	occurrenceID	basisOfRecord	catalogNumber	rightsHolder						
				University of Guatemala	Be					
2.				University of Guatemala	Be					
3.				University of Guatemala	Fo					
						1	1993	11	9	Guaten
5.	c4da0324-3da3-11ed-b878-0242ac120002	PresevedSpecimen	UWP:101737			1	1993	11	7	Guaten
6.	c4da03f6-3da3-11ed-b878-0242ac120002	PresevedSpecimen	UWP:10214			1	1994	4	20	Guaten
7.	c4da04dc-3da3-11ed-b878-0242ac120002	PresevedSpecimen	UWP:10214			1	1994	4	20	Guaten
8.	c4da05a4-3da3-11ed-b878-0242ac120002	PresevedSpecimen	UWP:10223			1	1995	12		Guaten
9.	c4da0694-3da3-11ed-b878-0242ac120002	PresevedSpecimen	UWP:10310			1	1996	4	3	Guaten

Customized facets

a. 在 occurrenceID 那欄點選三角形小圖示

b. 選擇 Facet >> Customized facet >> Duplicates facet

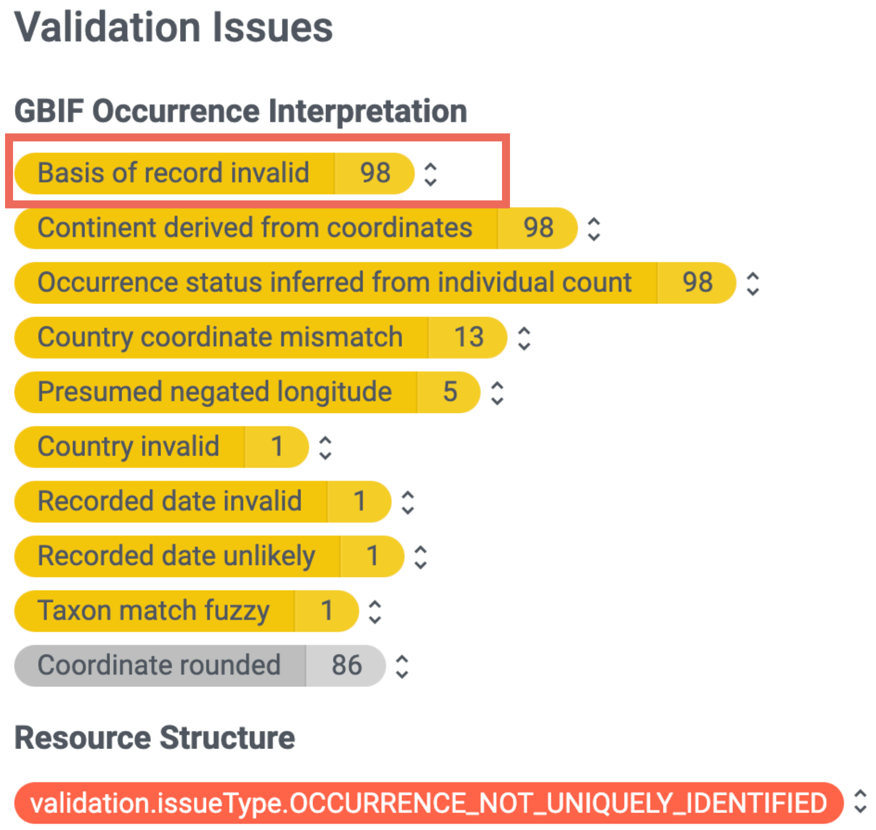
選擇 true 的資料

找出重複的 ID 並修正

案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符 decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 County



Validation Issues

GBIF Occurrence Interpretation

Basis of record invalid 98

Continent derived from coordinates 98

Occurrence status inferred from individual count 98

Country coordinate mismatch 13

Presumed negated longitude 5

Country invalid 1

Recorded date invalid 1

Recorded date unlikely 1

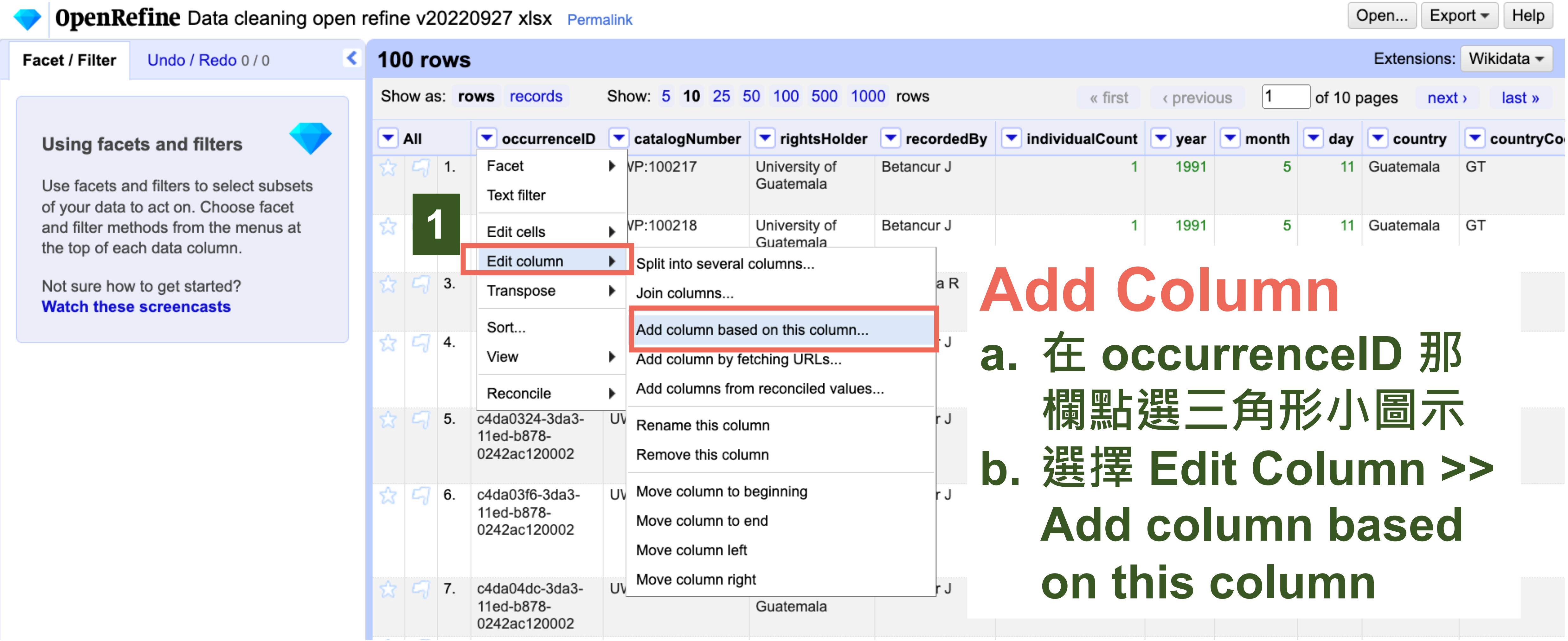
Taxon match fuzzy 1

Coordinate rounded 86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料- 新增欄位



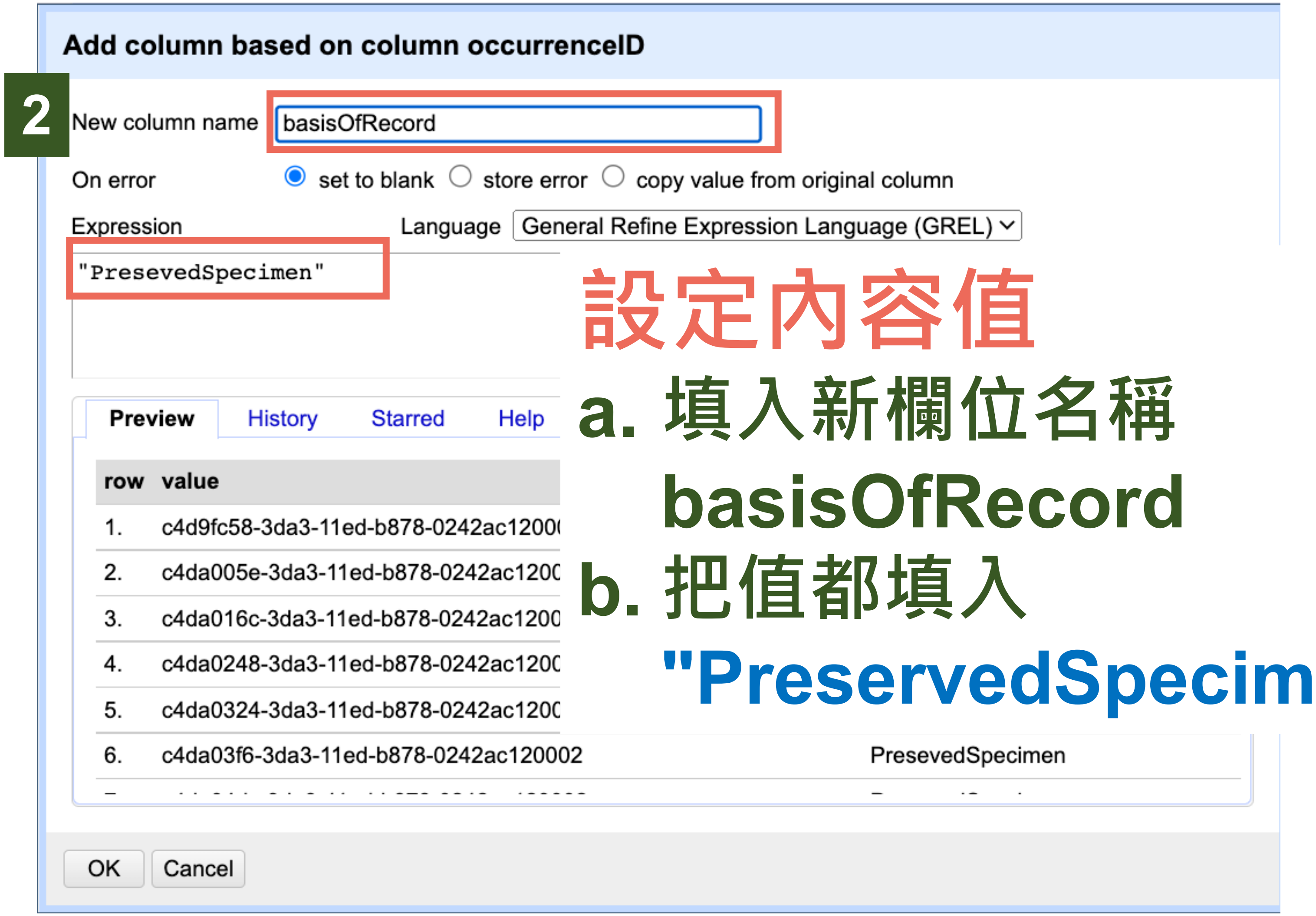
	occurrenceID	catalogNumber	rightsHolder	recordedBy	individualCount	year	month	day	country	countryCo
1.		VP:100217	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT
2.		VP:100218	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT
3.				a R						
4.				r J						
5.	c4da0324-3da3-11ed-b878-0242ac120002	UV		r J						
6.	c4da03f6-3da3-11ed-b878-0242ac120002	UV		r J						
7.	c4da04dc-3da3-11ed-b878-0242ac120002	UV	Guatemala	r J						

Add Column

a. 在 occurrenceID 那欄點選三角形小圖示

b. 選擇 Edit Column >> Add column based on this column

用 OpenRefine 清資料- 新增欄位



Add column based on column occurrenceID

2

New column name: basisOfRecord

On error: ● set to blank ○ store error ○ copy value from original column

Expression — Language: General Refine Expression Language (GREL)

```
"PresevedSpecimen"
```

row	value	
1.	c4d9fc58-3da3-11ed-b878-0242ac1200	
2.	c4da005e-3da3-11ed-b878-0242ac1200	
3.	c4da016c-3da3-11ed-b878-0242ac1200	
4.	c4da0248-3da3-11ed-b878-0242ac1200	
5.	c4da0324-3da3-11ed-b878-0242ac1200	
6.	c4da03f6-3da3-11ed-b878-0242ac120002	PresevedSpecimen

設定內容值

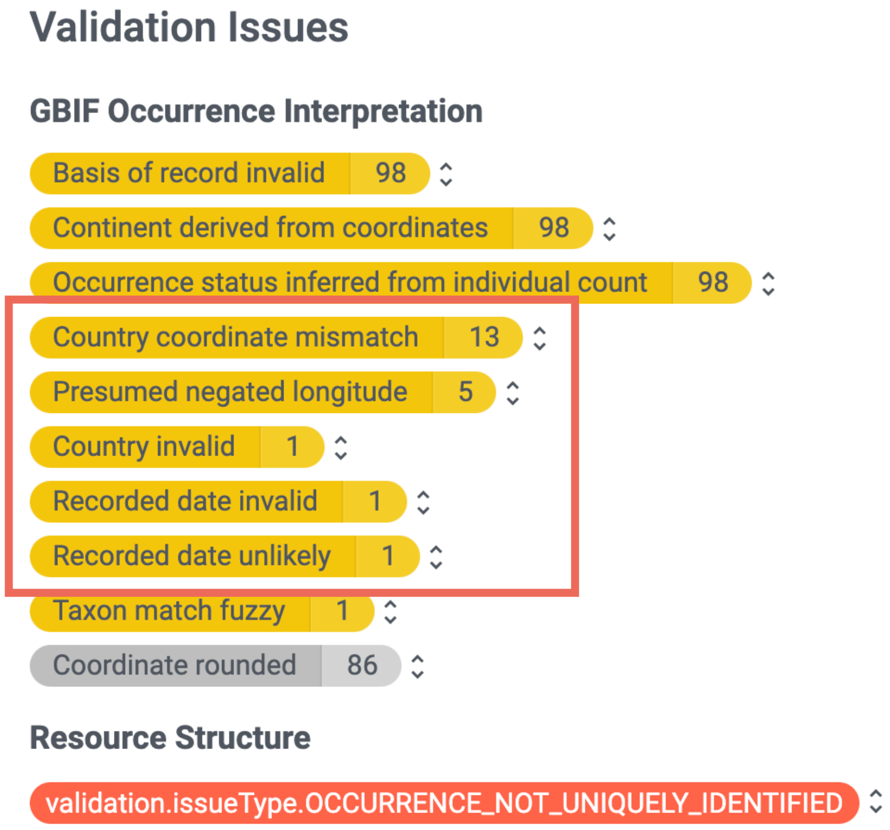
a. 填入新欄位名稱 basisOfRecord

b. 把值都填入 "PreservedSpecimen"

案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- **內容錯誤或與欄位不符** decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 County



Validation Issues

GBIF Occurrence Interpretation

Basis of record invalid 98

Continent derived from coordinates 98

Occurrence status inferred from individual count 98

Country coordinate mismatch 13

Presumed negated longitude 5

Country invalid 1

Recorded date invalid 1

Recorded date unlikely 1

Taxon match fuzzy 1

Coordinate rounded 86

Resource Structure

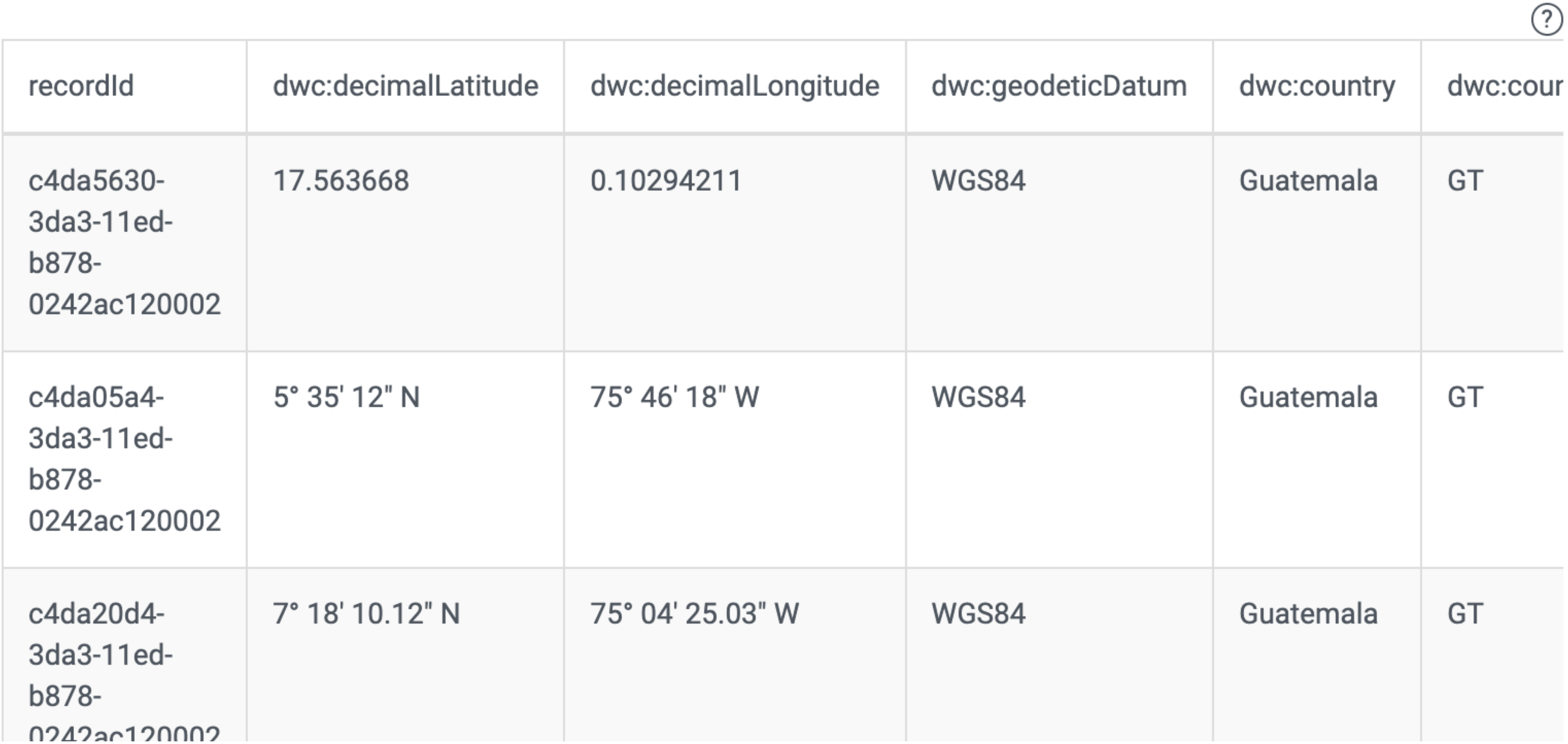
validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

案例練習- 進階作業

資料問題

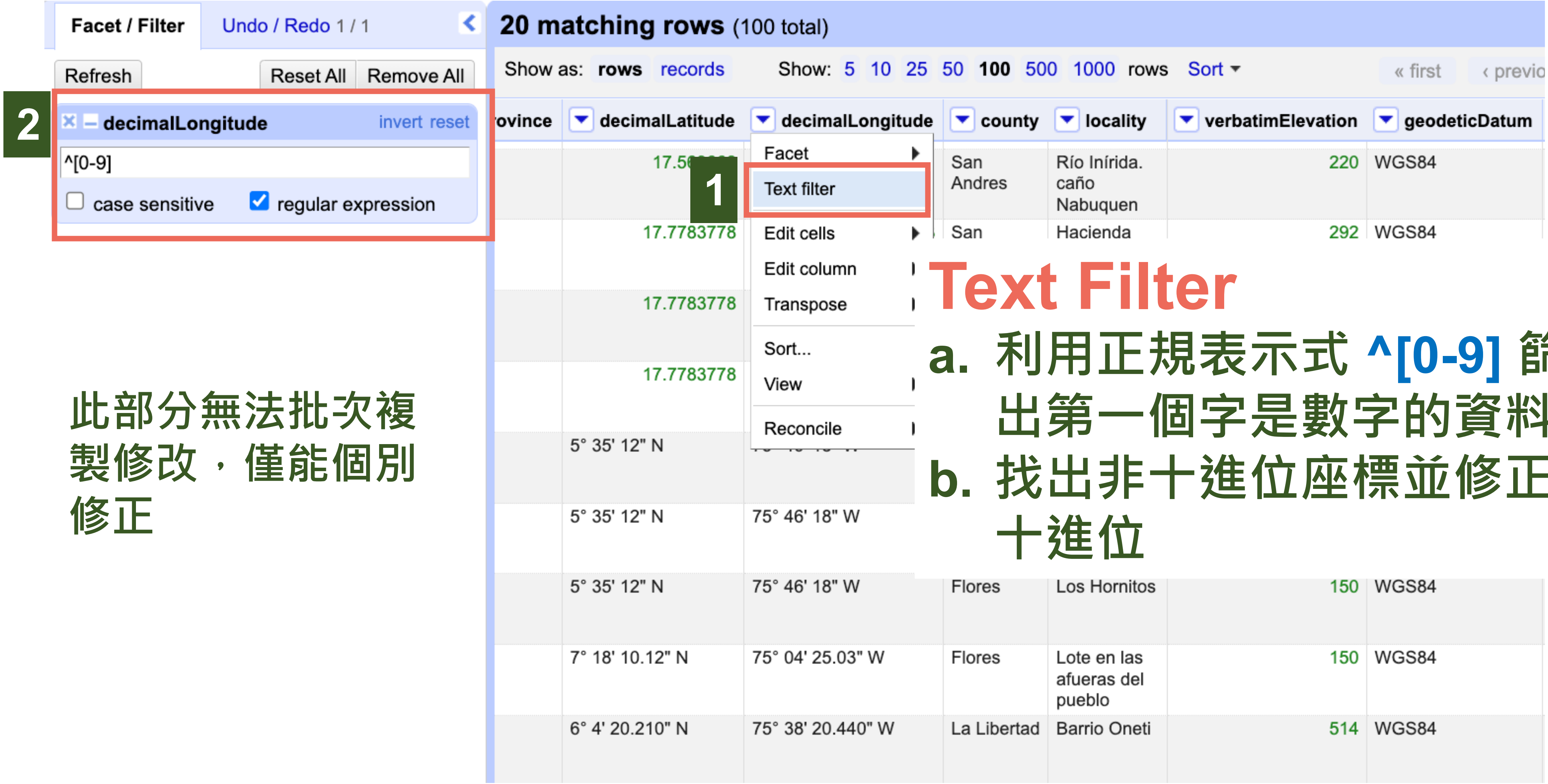
Country coordinate mismatch 13

座標和國家不符



recordId	dwc:decimalLatitude	dwc:decimalLongitude	dwc:geodeticDatum	dwc:country	dwc:cour
c4da5630-3da3-11ed-b878-0242ac120002	17.563668	0.10294211	WGS84	Guatemala	GT
c4da05a4-3da3-11ed-b878-0242ac120002	5° 35' 12" N	75° 46' 18" W	WGS84	Guatemala	GT
c4da20d4-3da3-11ed-b878-0242ac120002	7° 18' 10.12" N	75° 04' 25.03" W	WGS84	Guatemala	GT

用 OpenRefine 清資料- 內容錯誤(座標)



20 matching rows (100 total)

decimalLongitude: ^[0-9]

□ case sensitive ☑ regular expression

ovince	decimalLatitude	decimalLongitude	county	locality	verbatimElevation	geodeticDatum
	17.56		San Andres	Río Inírida. caño Nabuquen	220	WGS84
	17.7783778		San	Hacienda	292	WGS84
	17.7783778					
	17.7783778					
	5° 35' 12" N					
	5° 35' 12" N	75° 46' 18" W				
	5° 35' 12" N	75° 46' 18" W	Flores	Los Hornitos	150	WGS84
	7° 18' 10.12" N	75° 04' 25.03" W	Flores	Lote en las afueras del pueblo	150	WGS84
	6° 4' 20.210" N	75° 38' 20.440" W	La Libertad	Barrio Oneti	514	WGS84

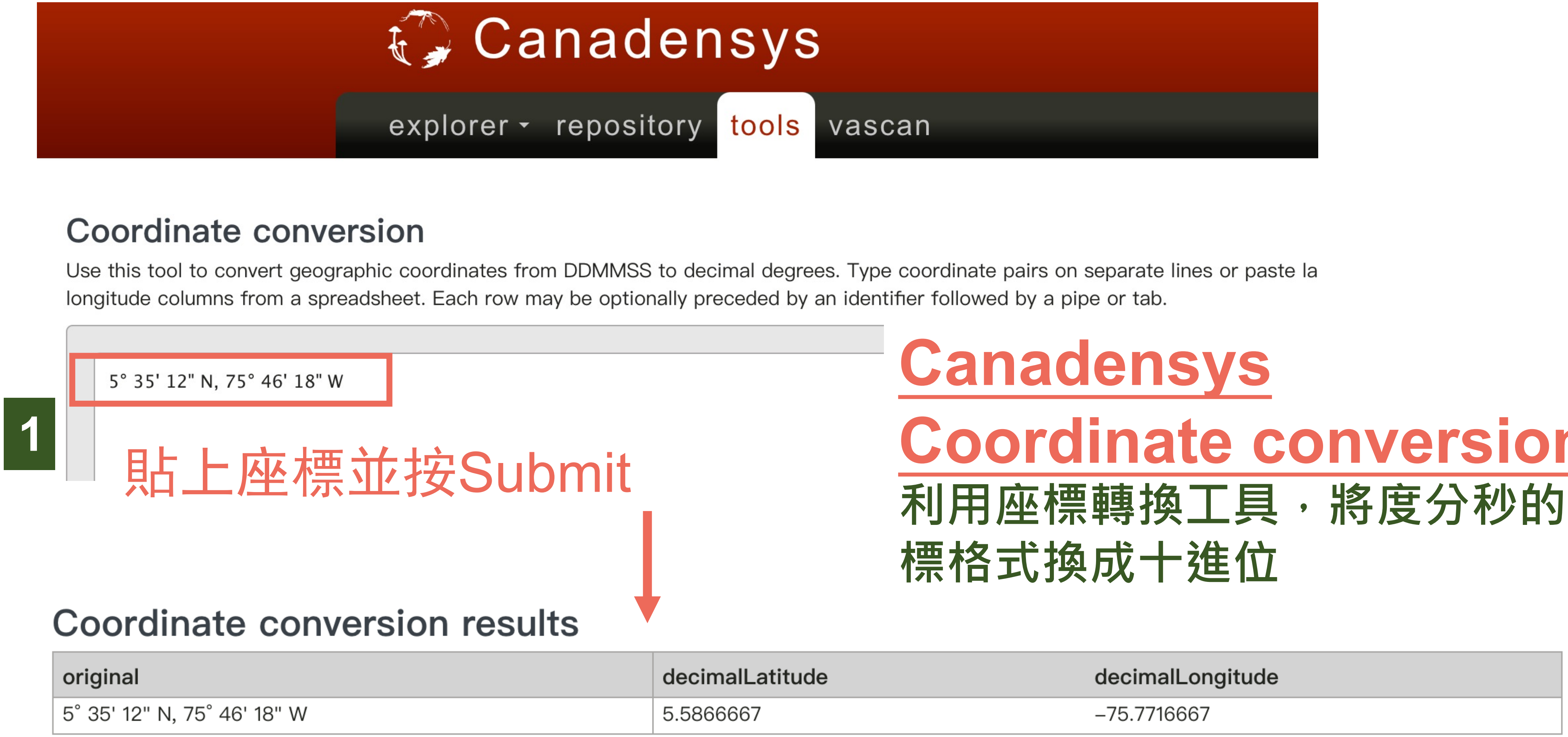
此部分無法批次複製修改，僅能個別修正

Text Filter

a. 利用正規表示式 ^[0-9] 篩選出第一個字是數字的資料

b. 找出非十進位座標並修正成十進位

用 OpenRefine 清資料- 內容錯誤(座標)



Coordinate conversion

Use this tool to convert geographic coordinates from DDMMSS to decimal degrees. Type coordinate pairs on separate lines or paste la
longitude columns from a spreadsheet. Each row may be optionally preceded by an identifier followed by a pipe or tab.

5° 35' 12" N, 75° 46' 18" W

1

貼上座標並按Submit

Coordinate conversion results

original	decimalLatitude	decimalLongitude
5° 35' 12" N, 75° 46' 18" W	5.5866667	–75.7716667

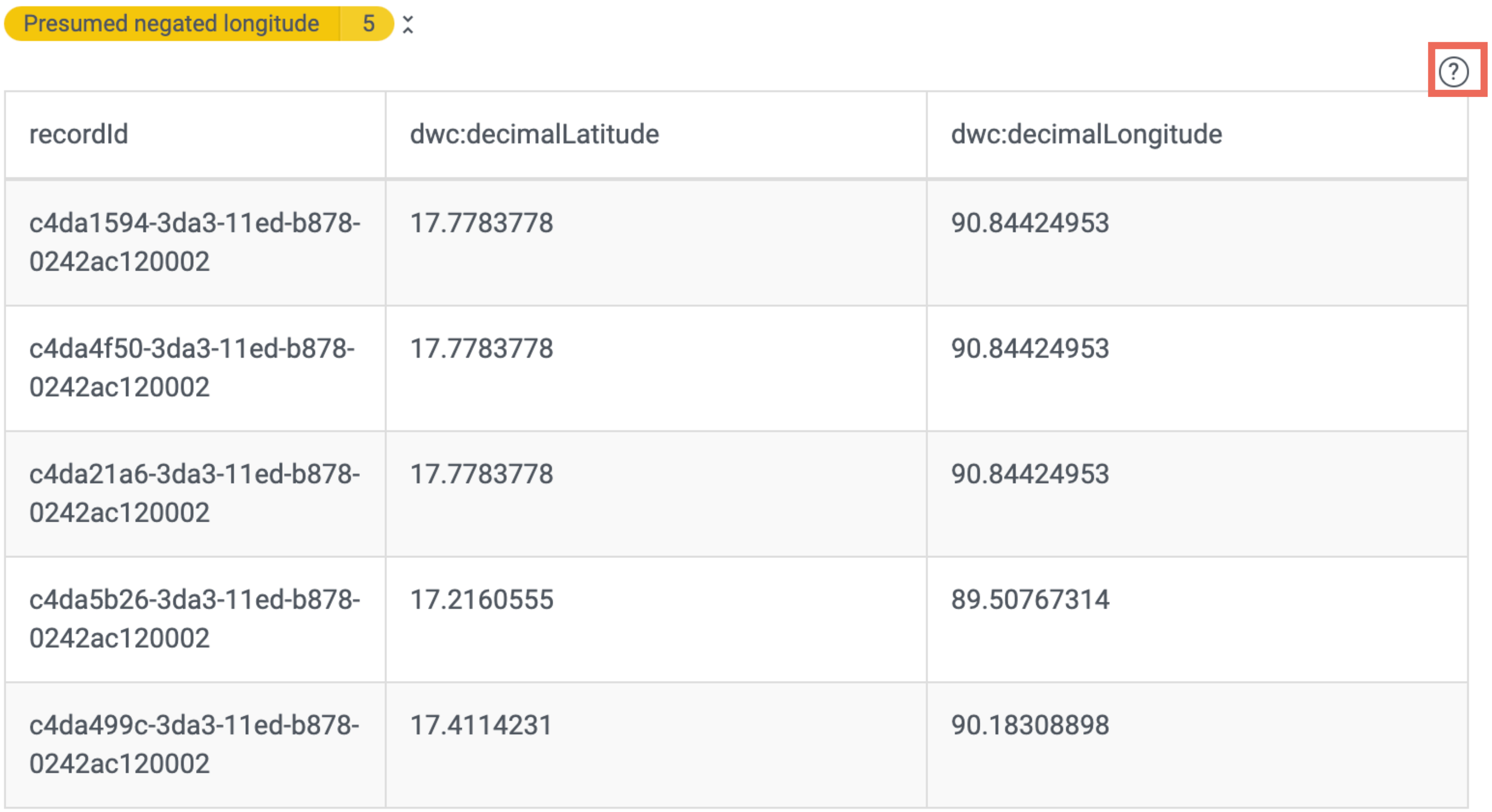
Canadensys Coordinate conversion

利用座標轉換工具，將度分秒的座標格式換成十進位

案例練習- 進階作業

資料問題

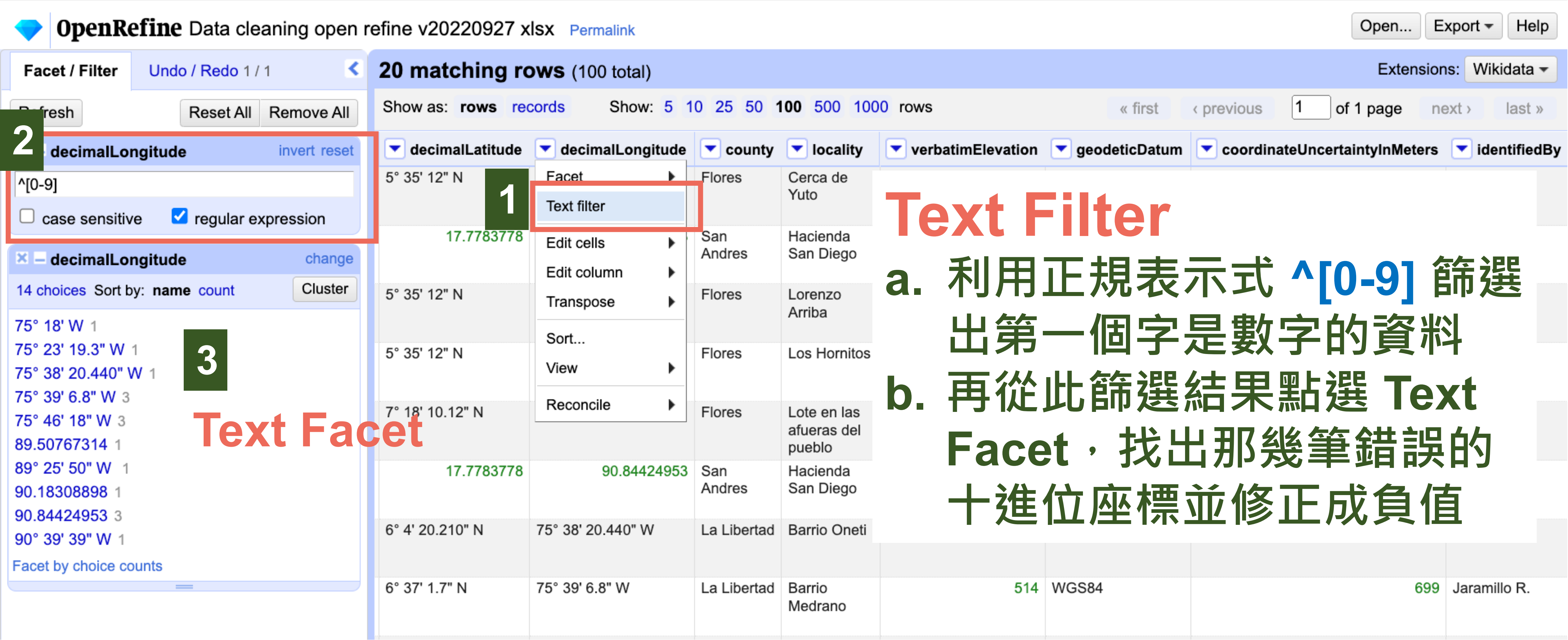
推定經度應為負值



Presumed negated longitude 5

recordId	dwc:decimalLatitude	dwc:decimalLongitude
c4da1594-3da3-11ed-b878-0242ac120002	17.7783778	90.84424953
c4da4f50-3da3-11ed-b878-0242ac120002	17.7783778	90.84424953
c4da21a6-3da3-11ed-b878-0242ac120002	17.7783778	90.84424953
c4da5b26-3da3-11ed-b878-0242ac120002	17.2160555	89.50767314
c4da499c-3da3-11ed-b878-0242ac120002	17.4114231	90.18308898

用 OpenRefine 清資料- 內容錯誤(座標)



OpenRefine Data cleaning open refine v20220927 xlsx Permalink

decimalLongitude

^[0-9]

□ case sensitive ☑ regular expression

decimalLongitude

14 choices Sort by: **name** count

75° 18' W	1
75° 23' 19.3" W	1
75° 38' 20.440" W	1
75° 39' 6.8" W	3
75° 46' 18" W	3
89.50767314	1
89° 25' 50" W	1
90.18308898	1
90.84424953	3
90° 39' 39" W	1

20 matching rows (100 total)

Show as: **rows** records Show: 5 10 25 50 **100** 500 1000 rows

decimalLatitude	decimalLongitude	county	locality	verbatimElevation	geodeticDatum	coordinateUncertaintyInMeters	identifiedBy
5° 35' 12" N		Flores	Cerca de Yuto				
17.7783778		San Andres	Hacienda San Diego				
5° 35' 12" N		Flores	Lorenzo Arriba				
5° 35' 12" N		Flores	Los Hornitos				
7° 18' 10.12" N		Flores	Lote en las afueras del pueblo				
17.7783778	90.84424953	San Andres	Hacienda San Diego				
6° 4' 20.210" N	75° 38' 20.440" W	La Libertad	Barrio Oneti				
6° 37' 1.7" N	75° 39' 6.8" W	La Libertad	Barrio Medrano	514	WGS84	699	Jaramillo R.

Text Facet

Text Filter

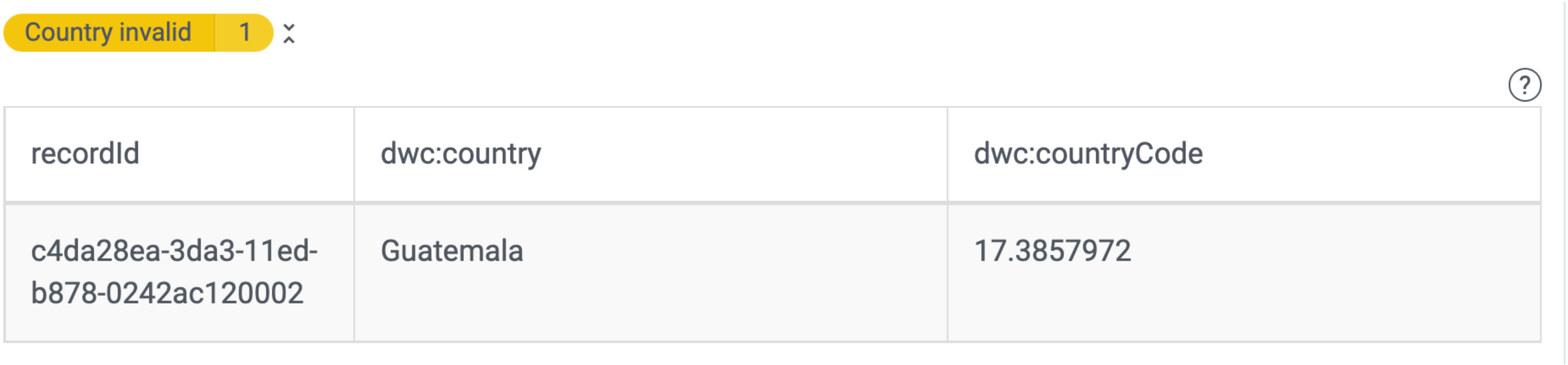
a. 利用正規表示式 ^[0-9] 篩選出第一個字是數字的資料

b. 再從此篩選結果點選 Text Facet，找出那幾筆錯誤的十進位座標並修正成負值

案例練習- 進階作業

資料問題

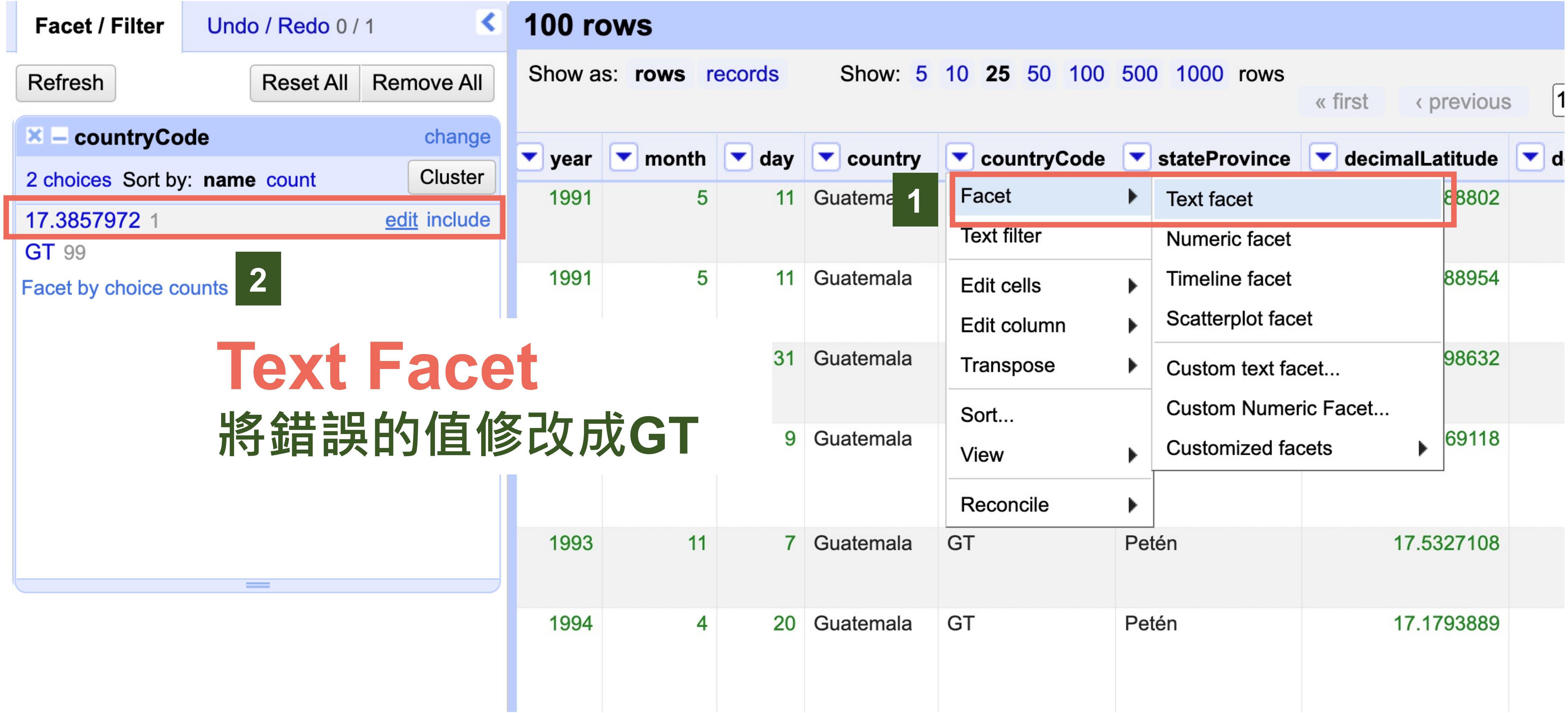
國家代碼無效



Country invalid 1

recordId	dwc:country	dwc:countryCode
c4da28ea-3da3-11ed-b878-0242ac120002	Guatemala	17.3857972

用 OpenRefine 清資料- 內容錯誤(countryCode)



year	month	day	country	countryCode	stateProvince	decimalLatitude
1991	5	11	Guatema			88802
1991	5	11	Guatemala			88954
		31	Guatemala			98632
		9	Guatemala			69118
1993	11	7	Guatemala	GT	Petén	17.5327108
1994	4	20	Guatemala	GT	Petén	17.1793889

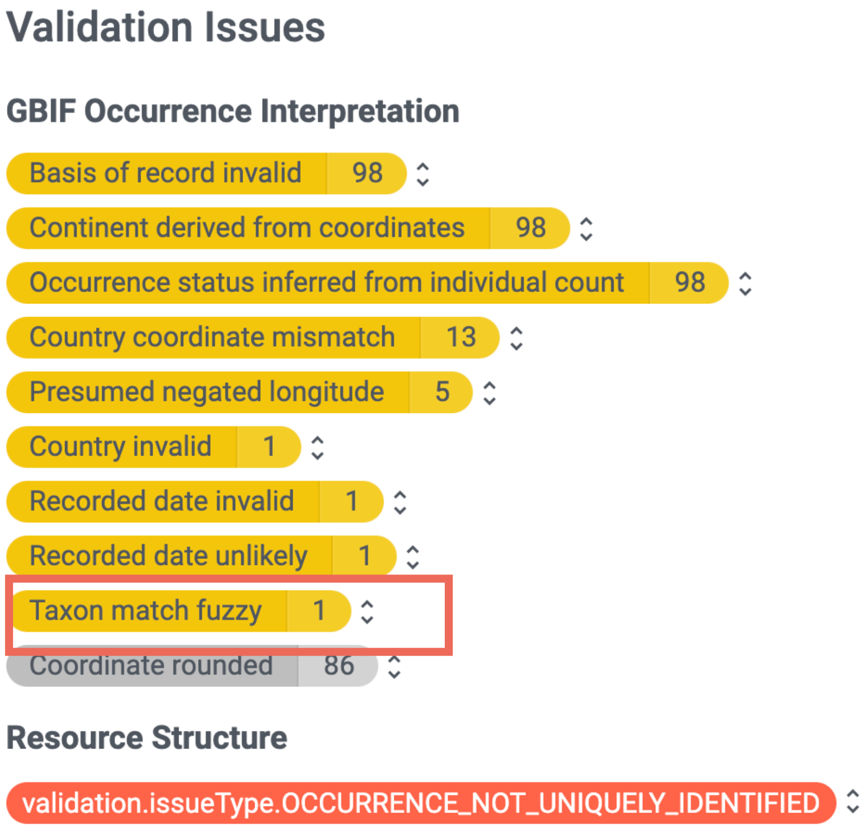
Text Facet

將錯誤的值修改成GT

案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符 decimalLatitude, decimalLongitude, countryCode, country, day, year
- **學名比對&清理** scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 County



Validation Issues

GBIF Occurrence Interpretation

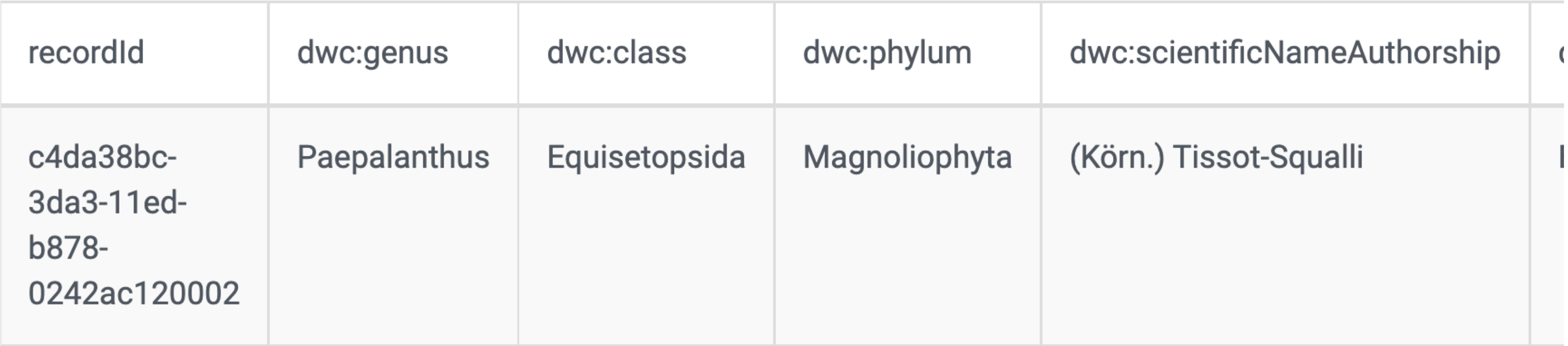
Basis of record invalid	98
Continent derived from coordinates	98
Occurrence status inferred from individual count	98
Country coordinate mismatch	13
Presumed negated longitude	5
Country invalid	1
Recorded date invalid	1
Recorded date unlikely	1
Taxon match fuzzy	1
Coordinate rounded	86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

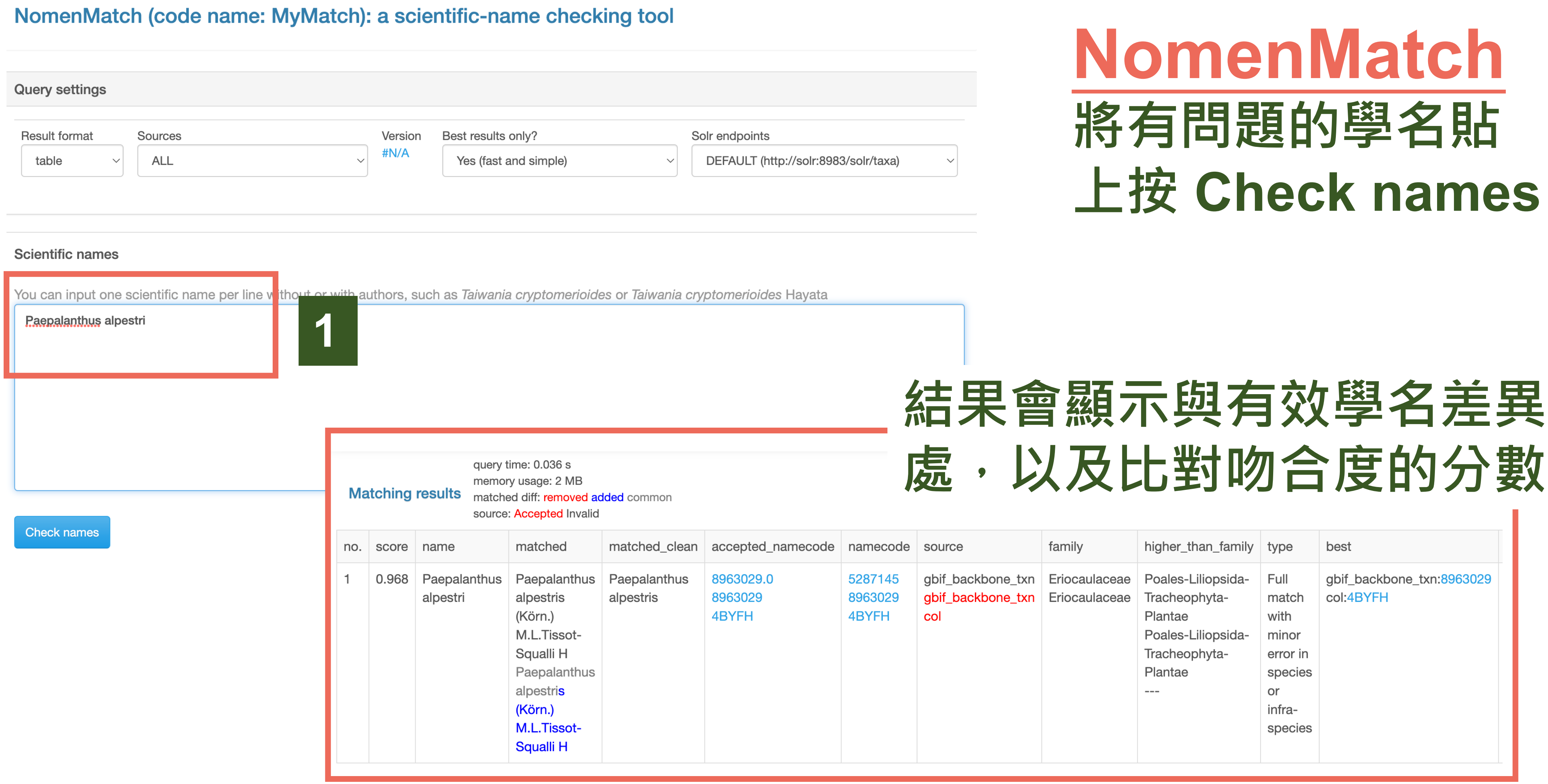
用 OpenRefine 清資料- 學名比對

分類末對應 GBIF backbone



recordId	dwc:genus	dwc:class	dwc:phylum	dwc:scientificNameAuthorship
c4da38bc-3da3-11ed-b878-0242ac120002	Paepalanthus	Equisetopsida	Magnoliophyta	(Körn.) Tissot-Squalli

用 OpenRefine 清資料- 學名比對



NomenMatch (code name: MyMatch): a scientific-name checking tool

Query settings

Result format: table

Sources: ALL

Version: #N/A

Best results only?: Yes (fast and simple)

Solr endpoints: DEFAULT (http://solr:8983/solr/taxa)

Scientific names

You can input one scientific name per line without or with authors, such as *Taiwania cryptomerioides* or *Taiwania cryptomerioides* Hayata

Paepalanthus alpestri

Matching results

query time: 0.036 s
memory usage: 2 MB
matched diff: removed added common
source: Accepted Invalid

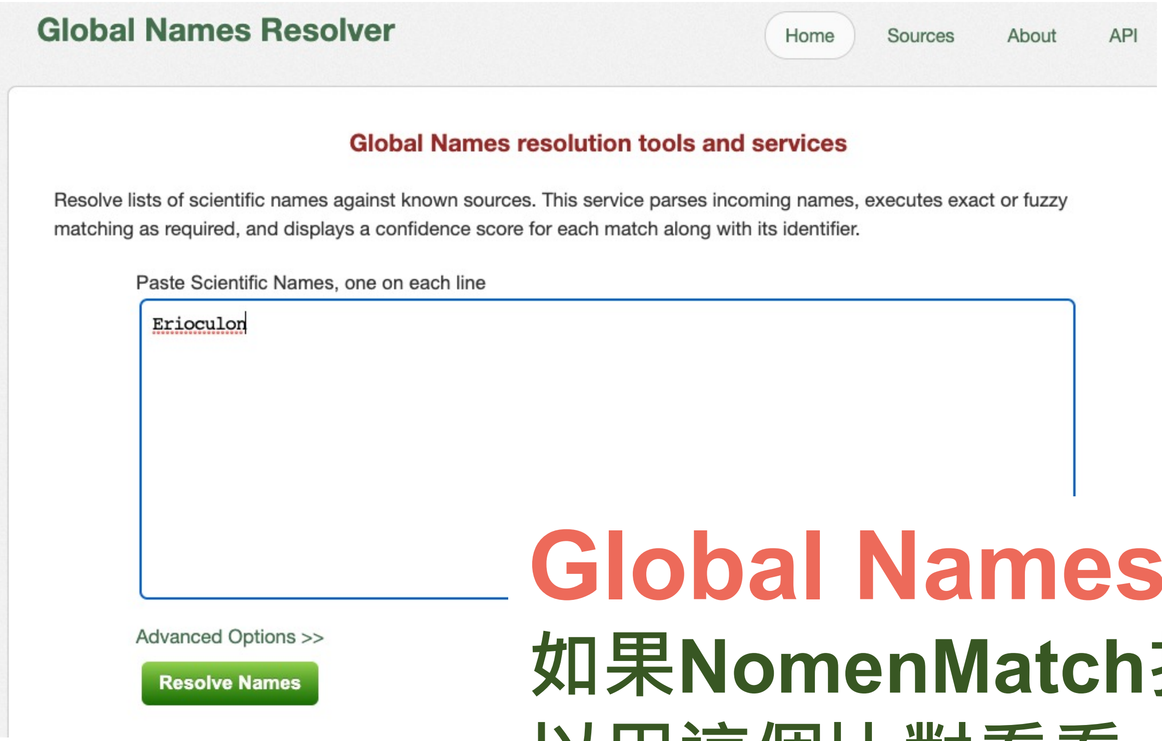
no.	score	name	matched	matched_clean	accepted_namecode	namecode	source	family	higher_than_family	type	best
1	0.968	Paepalanthus alpestri	Paepalanthus alpestris (Körn.) M.L.Tissot-Squalli H Paepalanthus alpestris (Körn.) M.L.Tissot-Squalli H	Paepalanthus alpestris	8963029.0 8963029 4BYFH	5287145 8963029 4BYFH	gbif_backbone_txn gbif_backbone_txn col	Eriocaulaceae Eriocaulaceae	Poales-Liliopsida-Tracheophyta-Plantae Poales-Liliopsida-Tracheophyta-Plantae ---	Full match with minor error in species or infra-species	gbif_backbone_txn:8963029 col:4BYFH

NomenMatch

將有問題的學名貼上按 Check names

結果會顯示與有效學名差異之處，以及比對吻合度的分數

用 OpenRefine 清資料- 學名比對

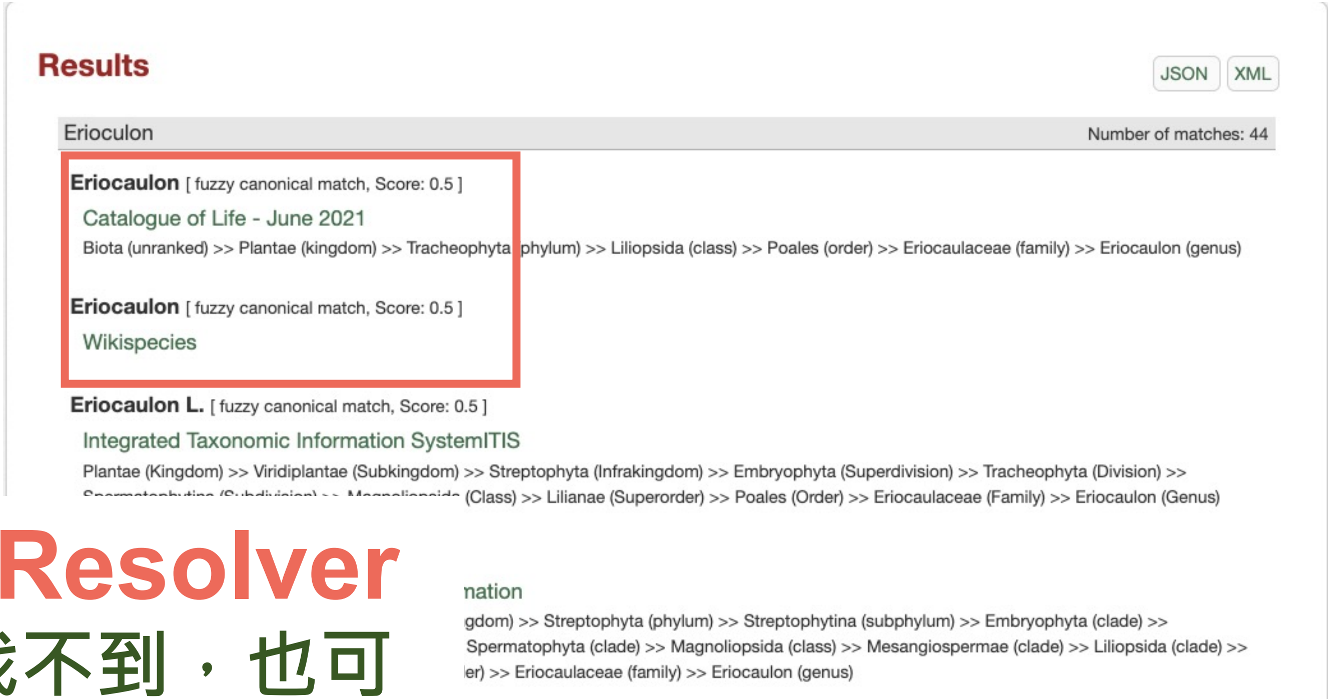

Global Names Resolver

Global Names resolution tools and services

Resolve lists of scientific names against known sources. This service parses incoming names, executes exact or fuzzy matching as required, and displays a confidence score for each match along with its identifier.

Paste Scientific Names, one on each line

Erioculon


Results

Erioculon Number of matches: 44

Eriocaulon [fuzzy canonical match, Score: 0.5]
Catalogue of Life - June 2021
Biota (unranked) >> Plantae (kingdom) >> Tracheophyta phylum) >> Liliopsida (class) >> Poales (order) >> Eriocaulaceae (family) >> Eriocaulon (genus)

Eriocaulon [fuzzy canonical match, Score: 0.5]
Wikispecies

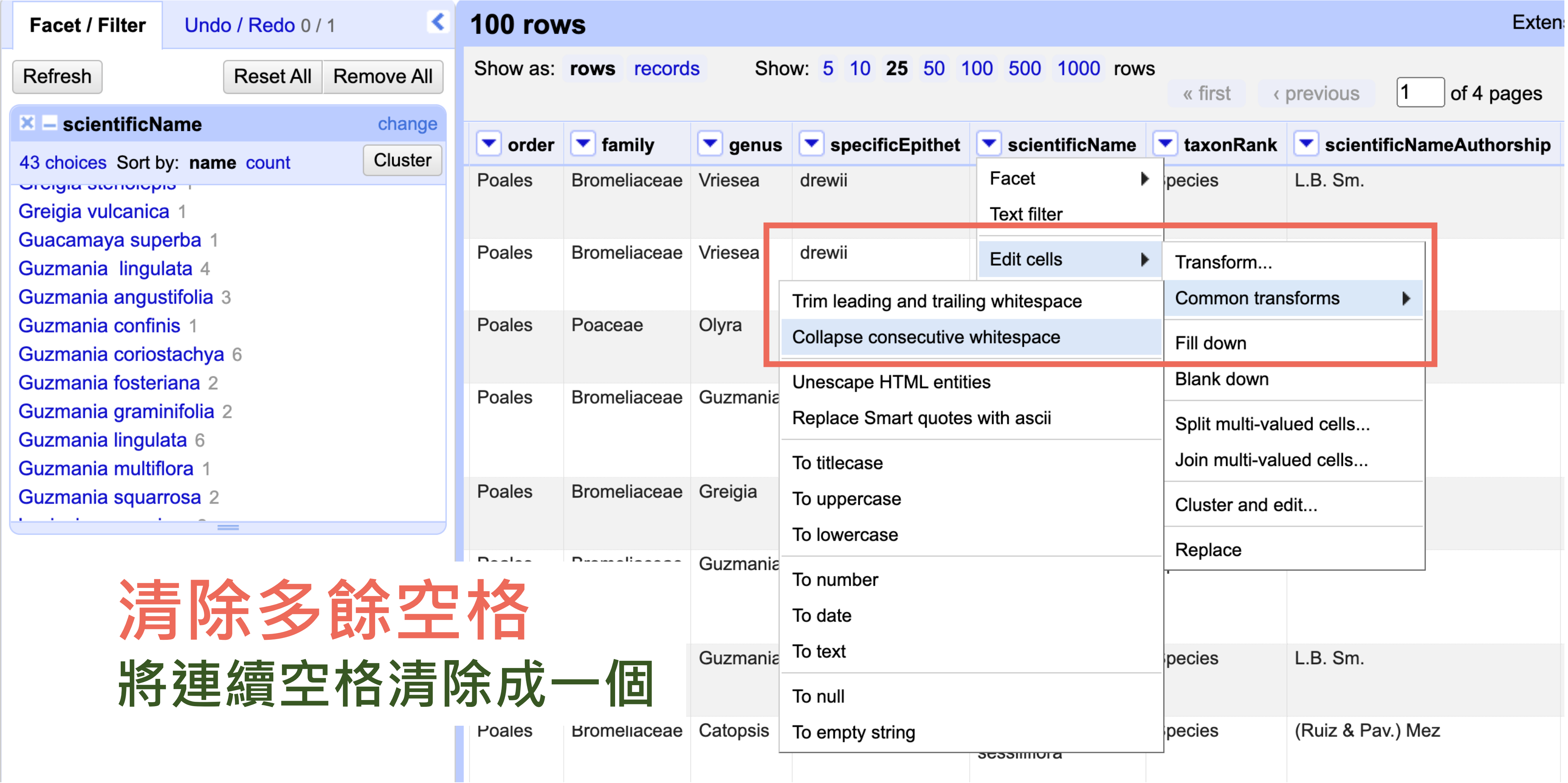
Eriocaulon L. [fuzzy canonical match, Score: 0.5]
Integrated Taxonomic Information SystemITIS
Plantae (Kingdom) >> Viridiplantae (Subkingdom) >> Streptophyta (Infrakingdom) >> Embryophyta (Superdivision) >> Tracheophyta (Division) >> Spermatophytina (Subdivision) >> Magnoliopsida (Class) >> Lilianae (Superorder) >> Poales (Order) >> Eriocaulaceae (Family) >> Eriocaulon (Genus)

nation
gdom) >> Streptophyta (phylum) >> Streptophytina (subphylum) >> Embryophyta (clade) >>
Spermatophyta (clade) >> Magnoliopsida (class) >> Mesangiospermae (clade) >> Liliopsida (clade) >>
er) >> Eriocaulaceae (family) >> Eriocaulon (genus)

Global Names Resolver

如果NomenMatch找不到，也可以用這個比對看看

用 OpenRefine 清資料- 學名清理



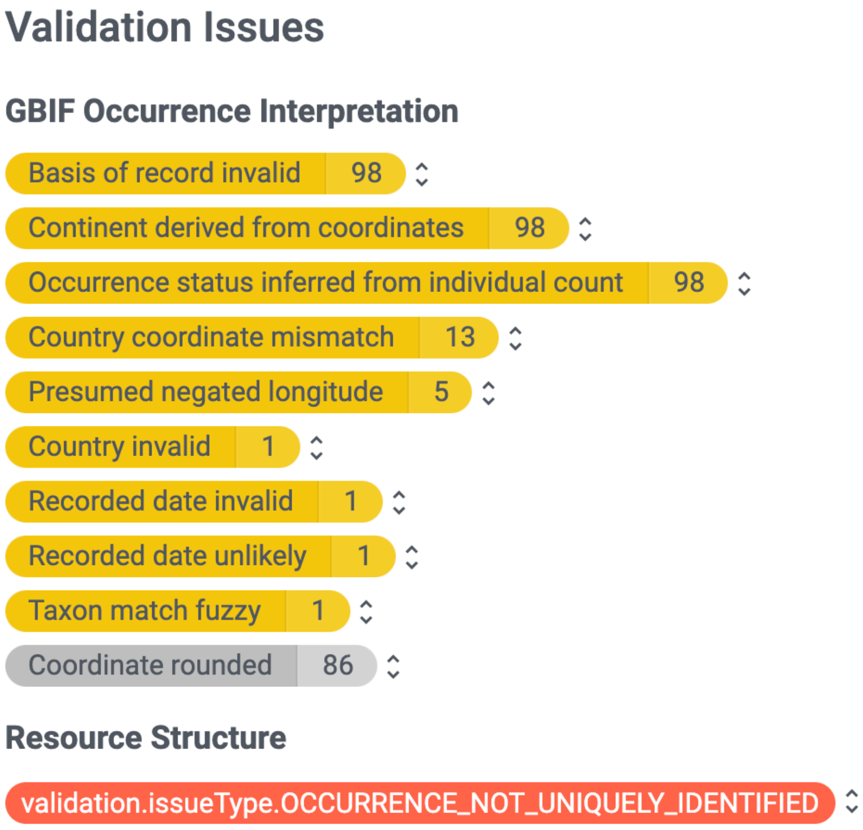
清除多餘空格

將連續空格清除成一個

案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符 decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- **修正學名格式 ^[A-Z].*\s[A-Z]**
- 清除多餘空格 country
- 找出相似文字並合併 County



Validation Issues

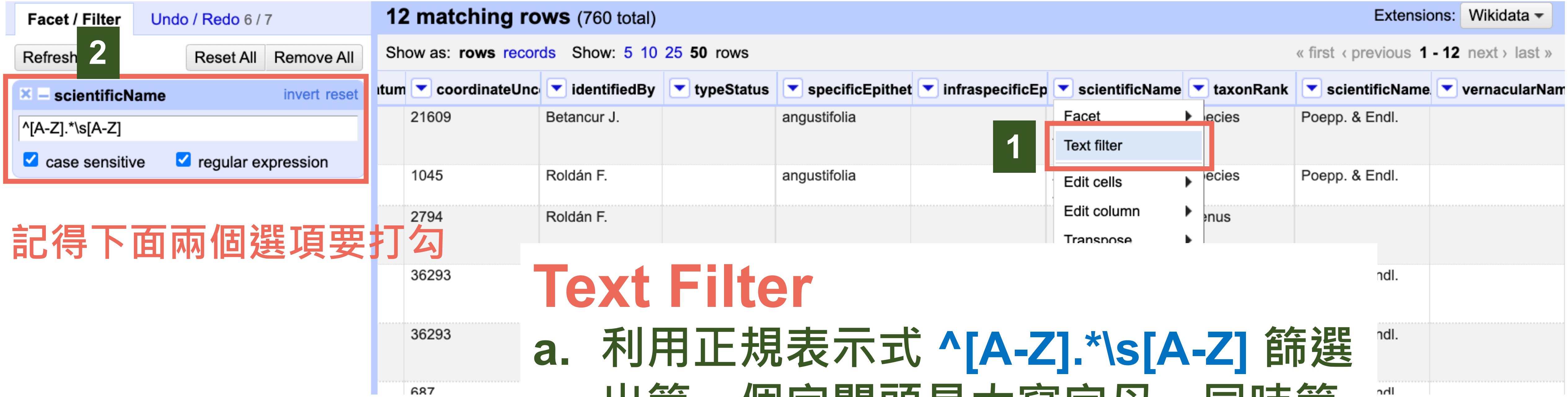
GBIF Occurrence Interpretation

Basis of record invalid	98
Continent derived from coordinates	98
Occurrence status inferred from individual count	98
Country coordinate mismatch	13
Presumed negated longitude	5
Country invalid	1
Recorded date invalid	1
Recorded date unlikely	1
Taxon match fuzzy	1
Coordinate rounded	86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料- 修正學名格式



scientificName invert reset

^[A-Z].*\s[A-Z]

☑ case sensitive ☑ regular expression

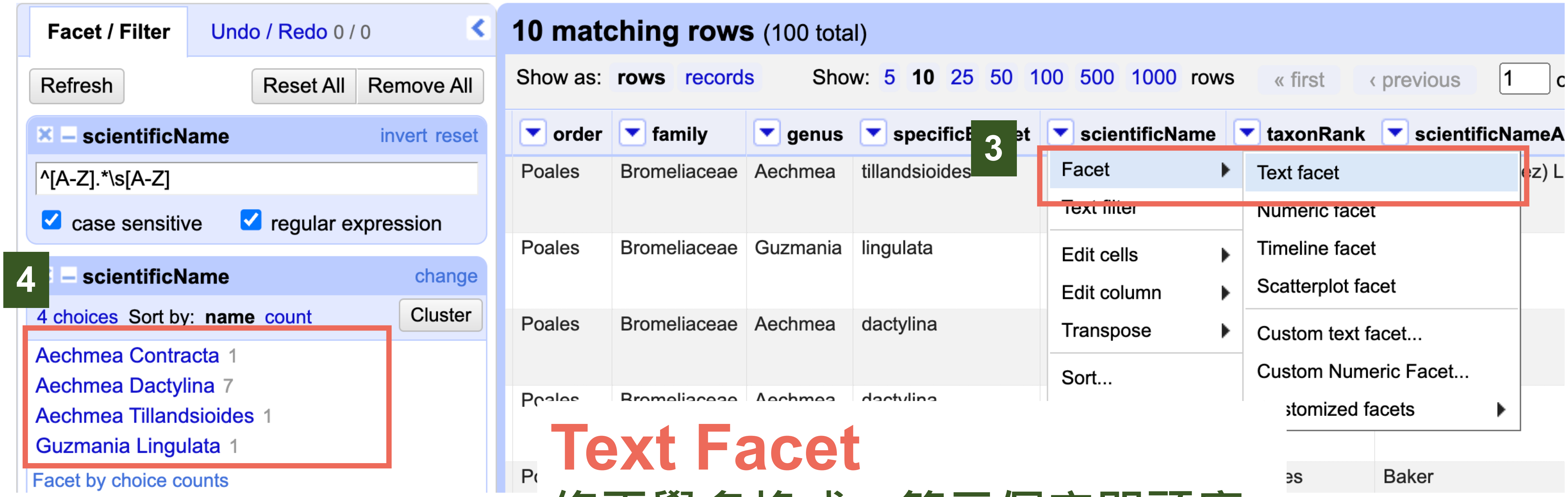
tum	coordinateUnc	identifiedBy	typeStatus	specificEpithet	infraspecificEp	scientificName	taxonRank	scientificName	vernacularNam
	21609	Betancur J.		angustifolia			ecies	Poepp. & Endl.	
	1045	Roldán F.		angustifolia			ecies	Poepp. & Endl.	
	2794	Roldán F.					nus		
	36293							ndl.	
	36293							ndl.	
	687							ndl.	

記得下面兩個選項要打勾

Text Filter

a. 利用正規表示式 ^[A-Z].*\s[A-Z] 篩選出第一個字開頭是大寫字母，同時第二個字開頭也是大寫字母的資料

用 OpenRefine 清資料- 修正學名格式



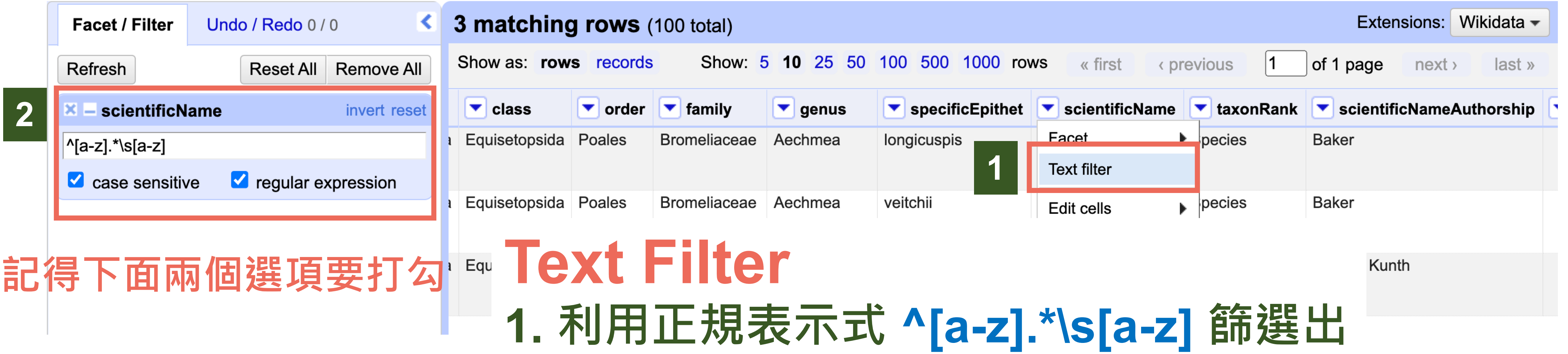
order	family	genus	specificEpithet	scientificName	taxonRank	scientificNameA
Poales	Bromeliaceae	Aechmea	tillandsioides			ez) L
Poales	Bromeliaceae	Guzmania	lingulata			
Poales	Bromeliaceae	Aechmea	dactylina			
Poales	Bromeliaceae	Aechmea	dactylina			
Po					es	Baker

Text Facet

修正學名格式，第二個字開頭應為小寫字母

可以批次修改

用 OpenRefine 清資料- 修正學名格式



^[a-z].*\s[a-z]

☑ case sensitive ☑ regular expression

3 matching rows (100 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first ‹ previous 1 of 1 page next › last »

class	order	family	genus	specificEpithet	scientificName	taxonRank	scientificNameAuthorship
Equisetopsida	Poales	Bromeliaceae	Aechmea	longicuspis		pecies	Baker
Equisetopsida	Poales	Bromeliaceae	Aechmea	veitchii		pecies	Baker
Equ							Kunth

記得下面兩個選項要打勾

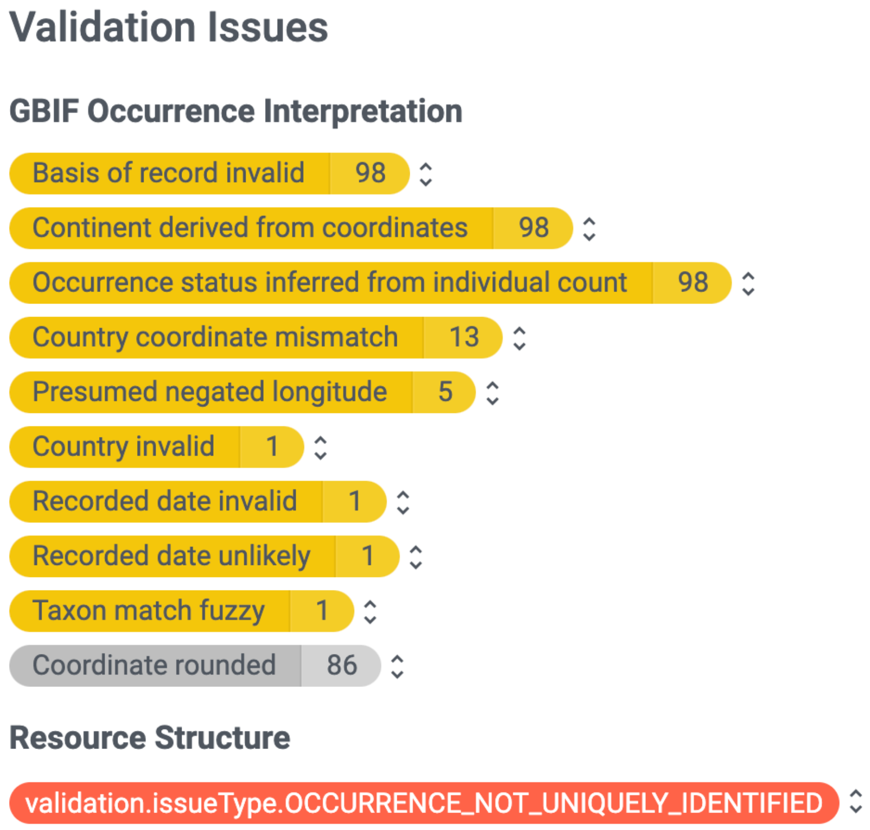
Text Filter

1. 利用正規表示式 ^[a-z].*\s[a-z] 篩選出第一個字開頭是小寫字母，同時第二個字開頭也是小寫字母的資料
2. 將第一個字開頭修正為大寫

案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符 decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 County



Validation Issues

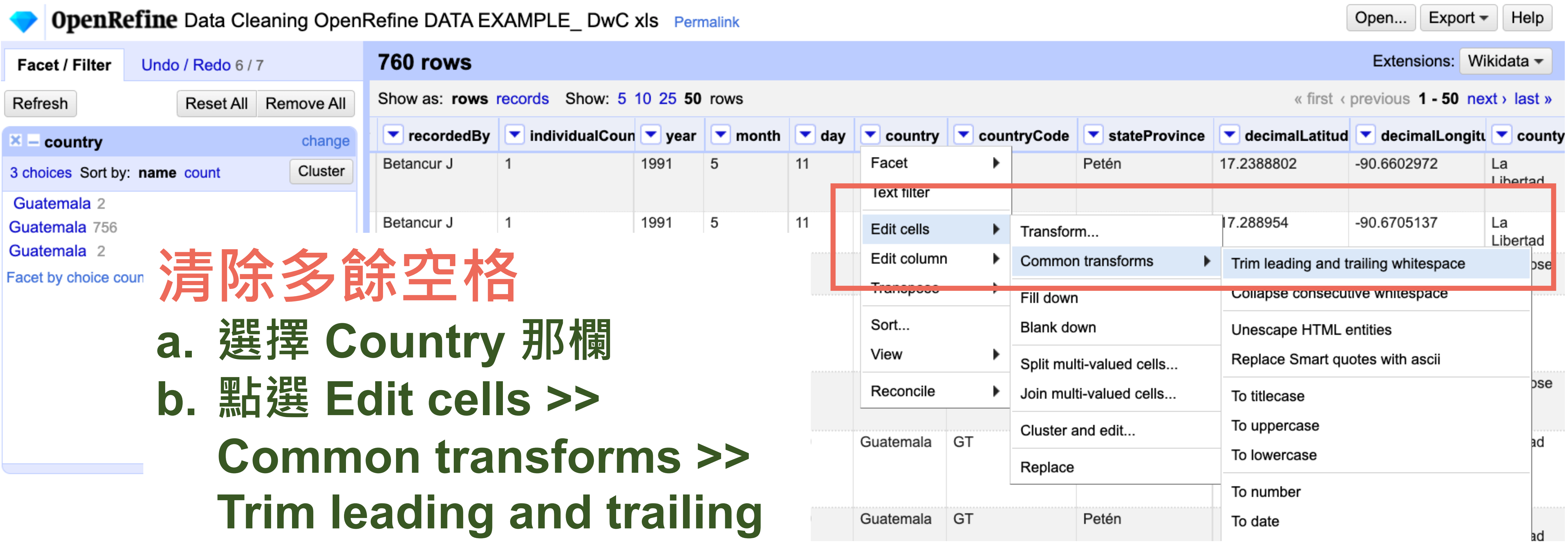
GBIF Occurrence Interpretation

Basis of record invalid	98
Continent derived from coordinates	98
Occurrence status inferred from individual count	98
Country coordinate mismatch	13
Presumed negated longitude	5
Country invalid	1
Recorded date invalid	1
Recorded date unlikely	1
Taxon match fuzzy	1
Coordinate rounded	86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料- 清除多餘空格2



清除多餘空格

a. 選擇 Country 那欄

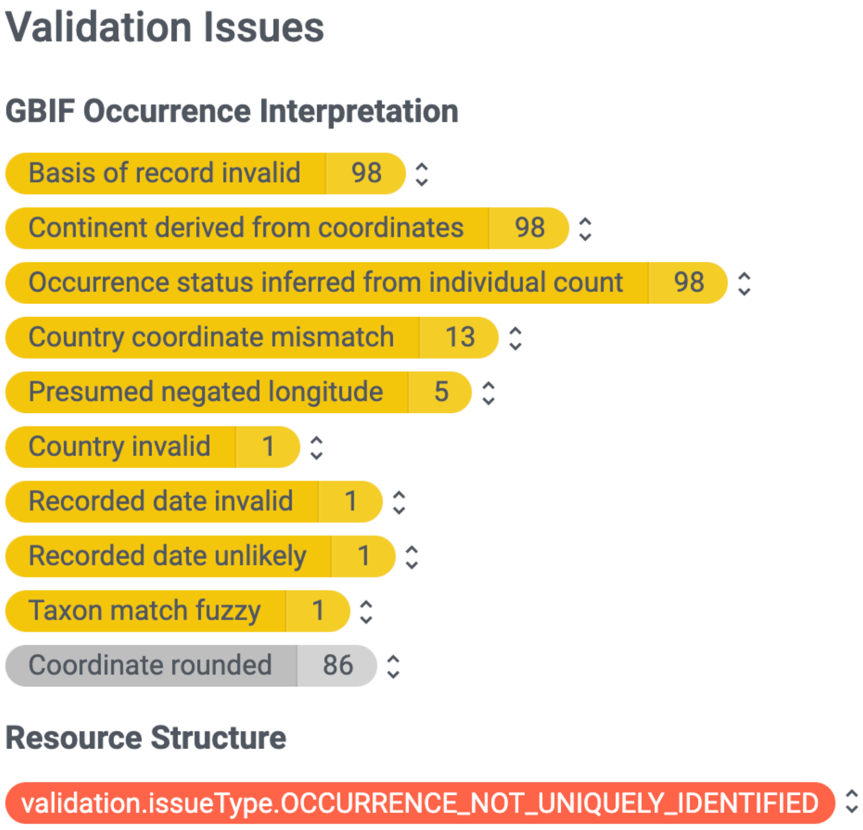
b. 點選 Edit cells >> Common transforms >> Trim leading and trailing whitespace

c. 將文字前後的多餘空格去除

案例練習- 進階作業

資料問題

- 找出重複 ID occurrenceID
- 新增欄位 basisOfRecord
- 內容錯誤或與欄位不符 decimalLatitude, decimalLongitude, countryCode, country, day, year
- 學名比對&清理 scientificName
- 修正學名格式 ^[A-Z].*\s[A-Z]
- 清除多餘空格 country
- 找出相似文字並合併 county



Validation Issues

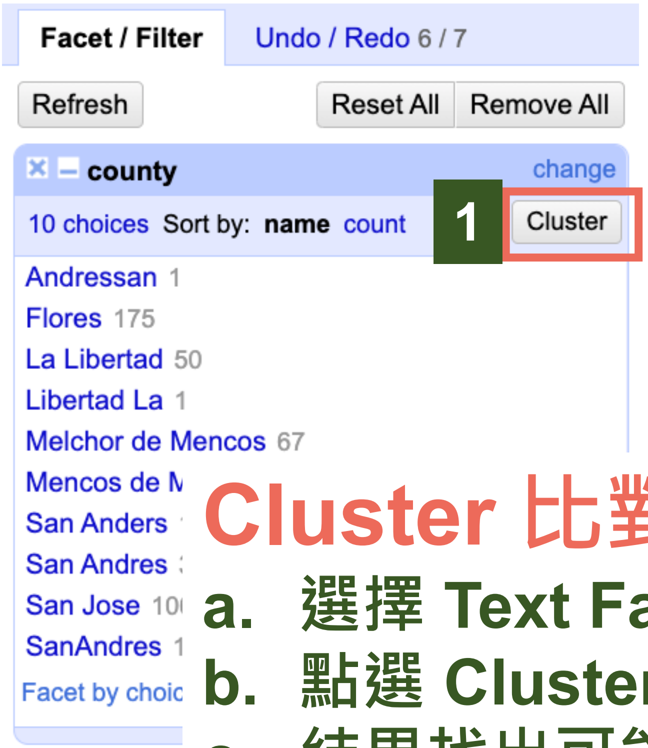
GBIF Occurrence Interpretation

Basis of record invalid	98
Continent derived from coordinates	98
Occurrence status inferred from individual count	98
Country coordinate mismatch	13
Presumed negated longitude	5
Country invalid	1
Recorded date invalid	1
Recorded date unlikely	1
Taxon match fuzzy	1
Coordinate rounded	86

Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料- 統一資料格式



Facet / Filter | Undo / Redo 6 / 7

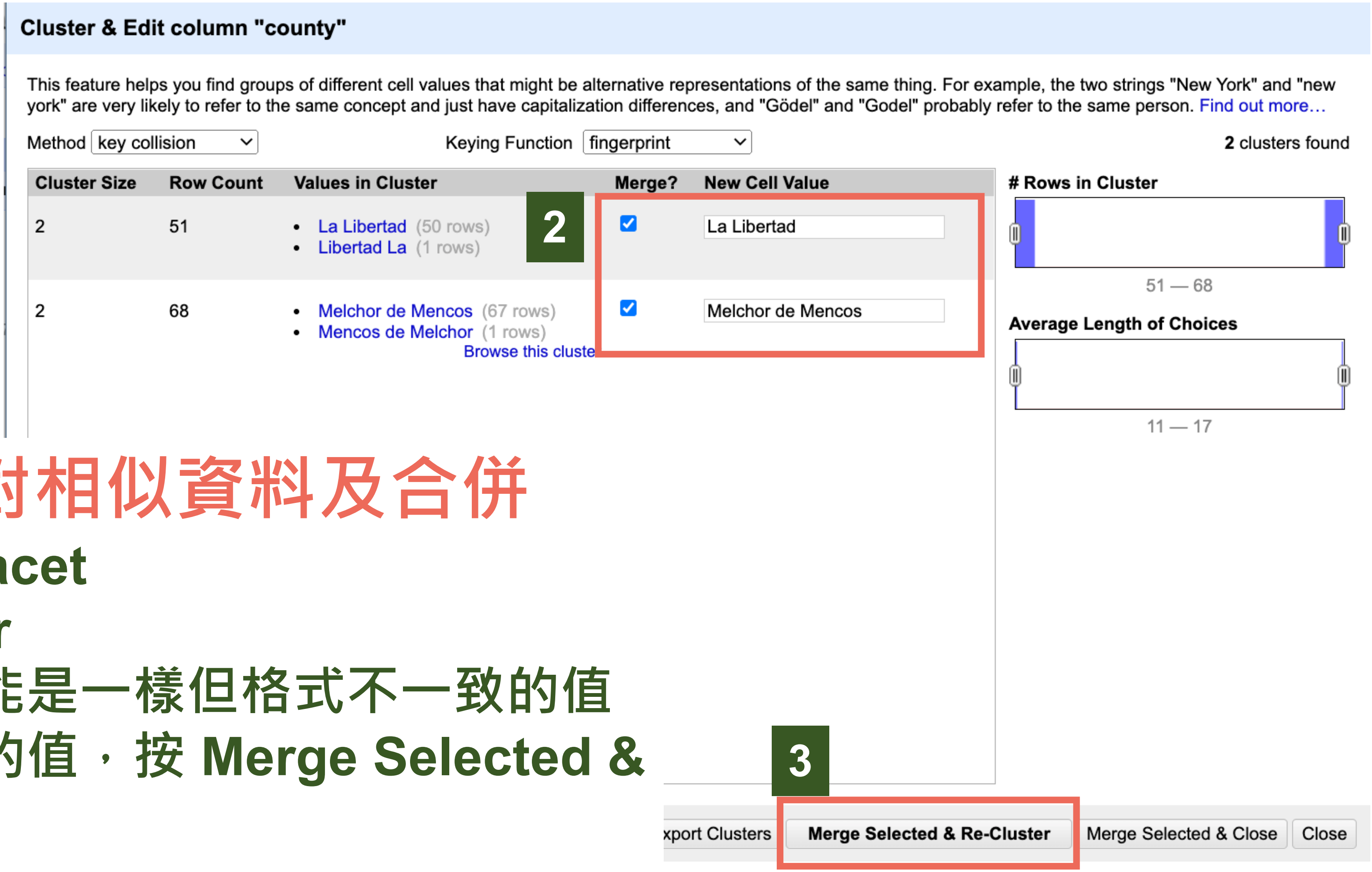
Refresh | Reset All | Remove All

county — change

10 choices Sort by: name count

Andressan	1
Flores	175
La Libertad	50
Libertad La	1
Melchor de Mencos	67
Mencos de M	
San Anders	
San Andres	
San Jose	10
SanAndres	1

Facet by choic



Cluster & Edit column "county"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: key collision — Keying Function: fingerprint — 2 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	51	La Libertad (50 rows) Libertad La (1 rows)	☑	La Libertad
2	68	Melchor de Mencos (67 rows) Mencos de Melchor (1 rows) Browse this cluste	☑	Melchor de Mencos

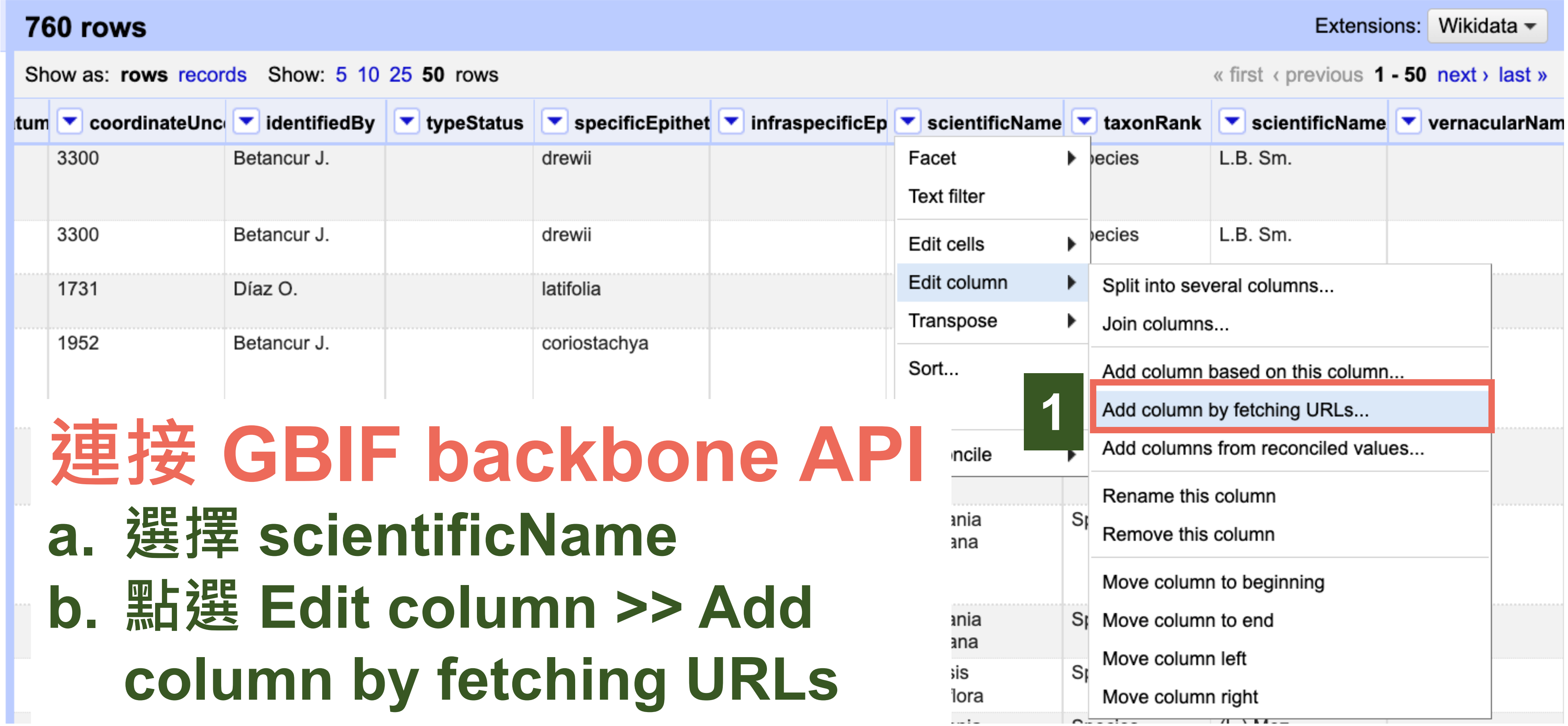
Rows in Cluster: 51 — 68

Average Length of Choices: 11 — 17

Cluster 比對相似資料及合併

a. 選擇 Text Facet
b. 點選 Cluster
c. 結果找出可能是一樣但格式不一致的值
d. 勾選要合併的值，按 Merge Selected & Re-cluster

進階題-自動匯入高階層分類欄位



760 rows

Show as: rows records Show: 5 10 25 50 rows

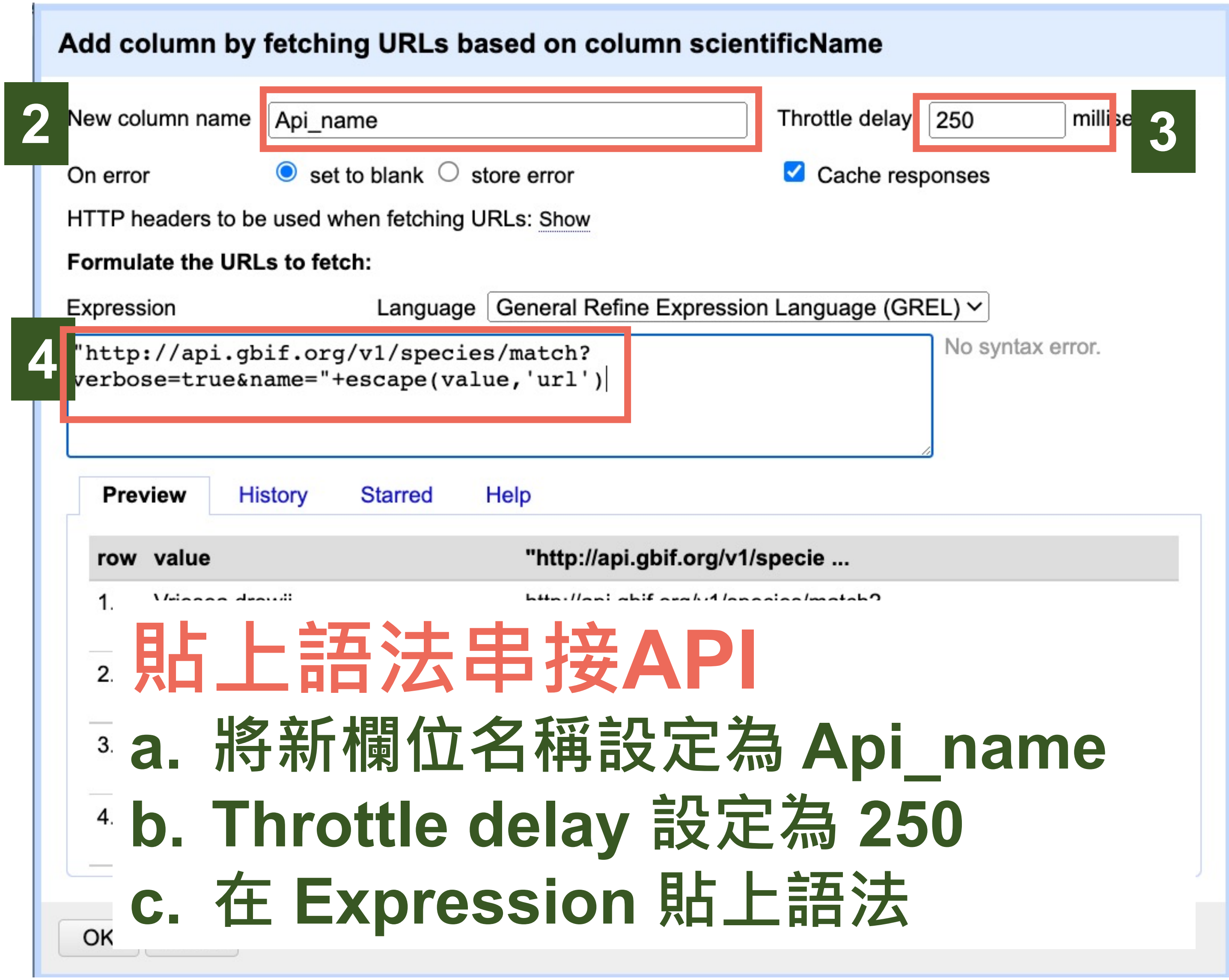
tum	coordinateUnc	identifiedBy	typeStatus	specificEpithet	infraspecificEp	scientificName	taxonRank	scientificName	vernacularNam
	3300	Betancur J.		drewii			pecies	L.B. Sm.	
	3300	Betancur J.		drewii			pecies	L.B. Sm.	
	1731	Díaz O.		latifolia					
	1952	Betancur J.		coriostachya					

連接 GBIF backbone API

a. 選擇 scientificName

b. 點選 Edit column >> Add column by fetching URLs

進階題-自動匯入高階層分類欄位



Add column by fetching URLs based on column scientificName

New column name: Api_name

Throttle delay: 250 millise

On error: ● set to blank ○ store error

☑ Cache responses

HTTP headers to be used when fetching URLs: Show

Formulate the URLs to fetch:

Expression

Language: General Refine Expression Language (GREL)

```
"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value,'url')
```

No syntax error.

row	value	"http://api.gbif.org/v1/specie ...
1.	Vriesea drewii	http://api.gbif.org/v1/species/match?
2.		
3.		
4.		

貼上語法串接API

a. 將新欄位名稱設定為 Api_name

b. Throttle delay 設定為 250

c. 在 Expression 貼上語法

語法在下一頁，
請整串複製貼上

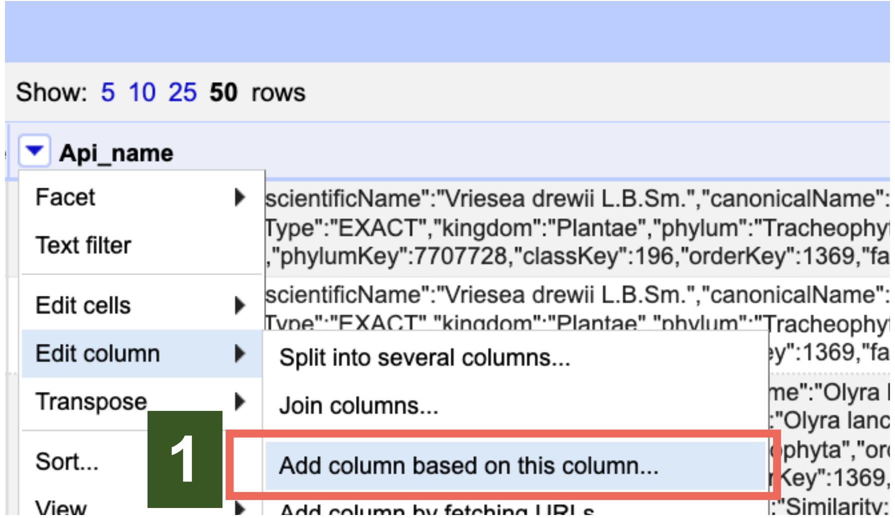
進階題-自動匯入高階層分類欄位

語法在此，請整串複製貼上

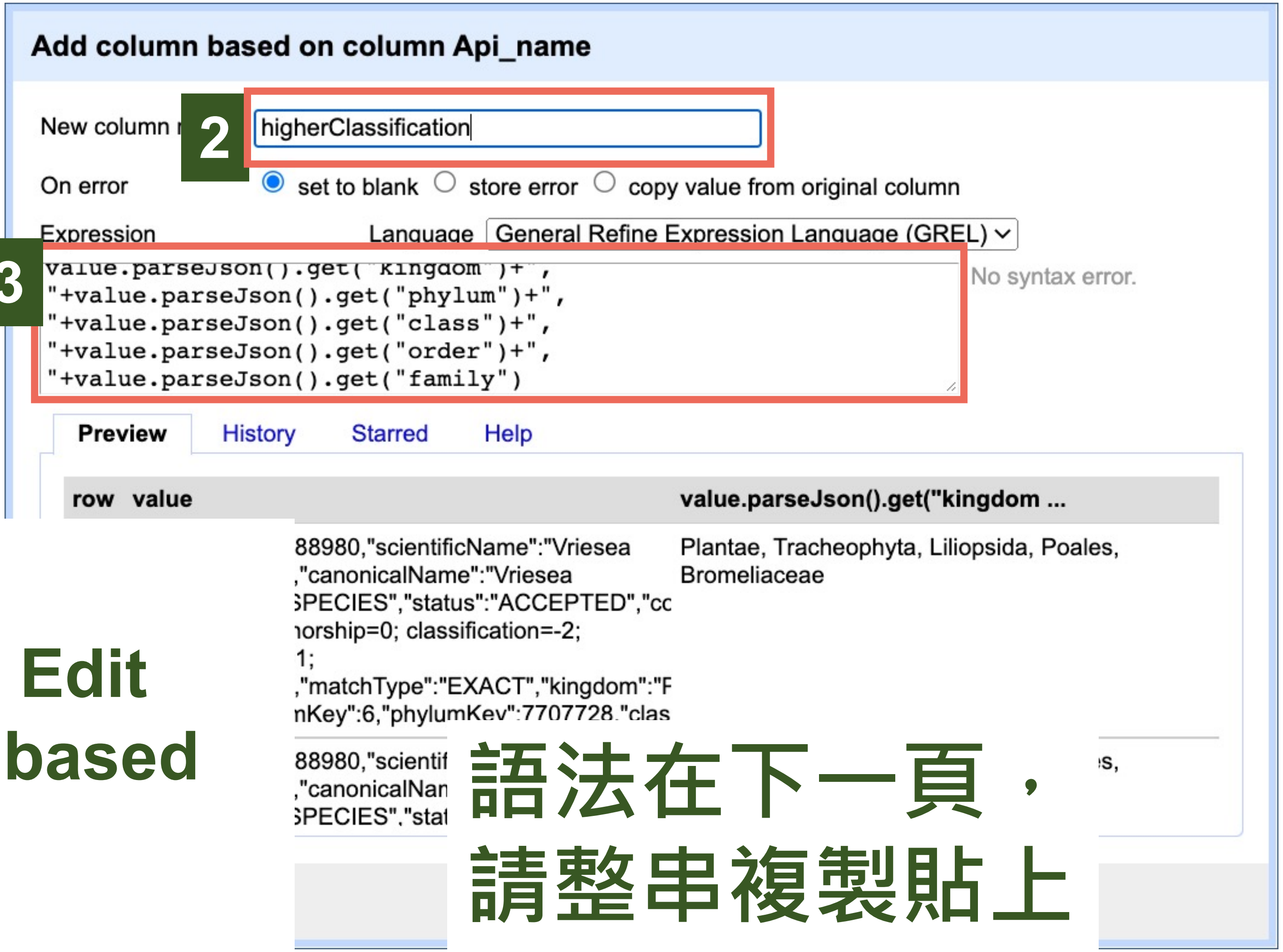
⬇

```
"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value,'url')
```

進階題-自動匯入高階層分類欄位



scientificName":"Vriesea drewii L.B.Sm.","canonicalName":
Type":"EXACT","kingdom":"Plantae","phylum":"Tracheophy
,"phylumKey":7707728,"classKey":196,"orderKey":1369,"fa
scientificName":"Vriesea drewii L.B.Sm.","canonicalName":
Type":"EXACT","kingdom":"Plantae","phylum":"Tracheophy
y":1369,"fa
me":"Olyra l
:"Olyra lanc
phyta","or
Key":1369,
:"Similarity:



Add column based on column Api_name

New column name: higherClassification

On error ● set to blank ○ store error ○ copy value from original column

Expression Language General Refine Expression Language (GREL)

```
value.parseJson().get("kingdom")+",
"+value.parseJson().get("phylum")+",
"+value.parseJson().get("class")+",
"+value.parseJson().get("order")+",
"+value.parseJson().get("family")
```

No syntax error.

row	value	value.parseJson().get("kingdom ...
	88980,"scientificName":"Vriesea ,"canonicalName":"Vriesea SPECIES","status":"ACCEPTED","cc 1orship=0; classification=-2; 1; ,"matchType":"EXACT","kingdom":"F nKey":6,"phylumKey":7707728,"clas	Plantae, Tracheophyta, Liliopsida, Poales, Bromeliaceae
	88980,"scientif ,"canonicalNan SPECIES","stat	s,

語法在下一頁，請整串複製貼上

呼叫各分類階層的值

a. 到 Api_name 欄位並選擇 Edit column >> Add column based on this column
b. 將新欄位名稱寫為 higherClassification
c. 貼上語法按 OK

進階題-自動匯入高階層分類欄位

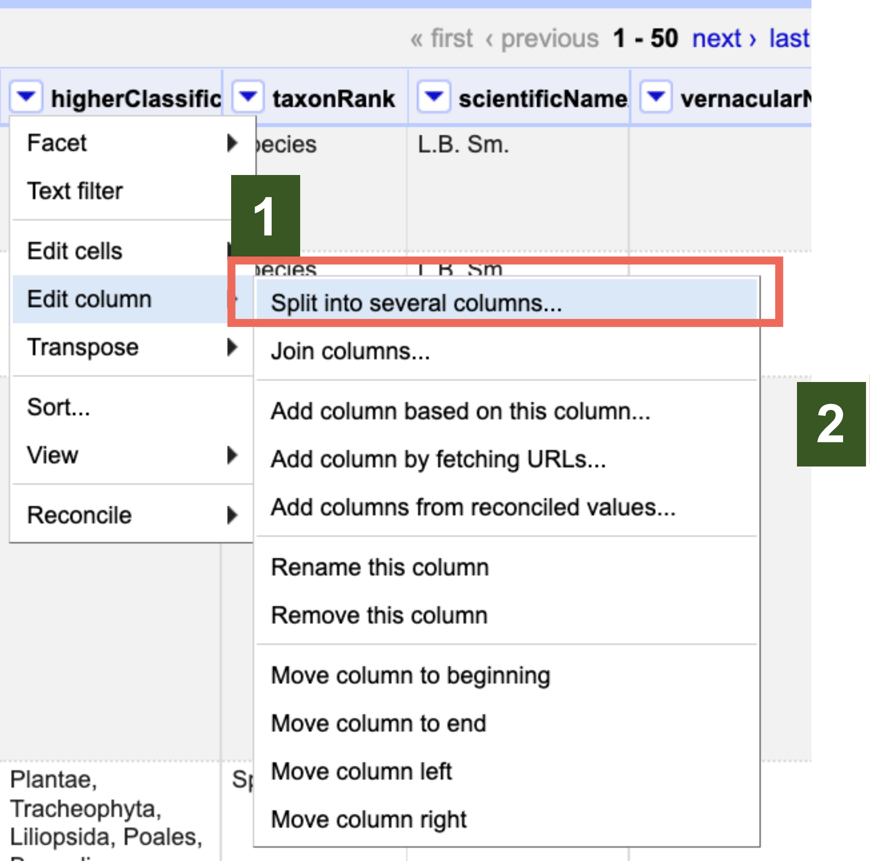
語法在此，請整串複製貼上

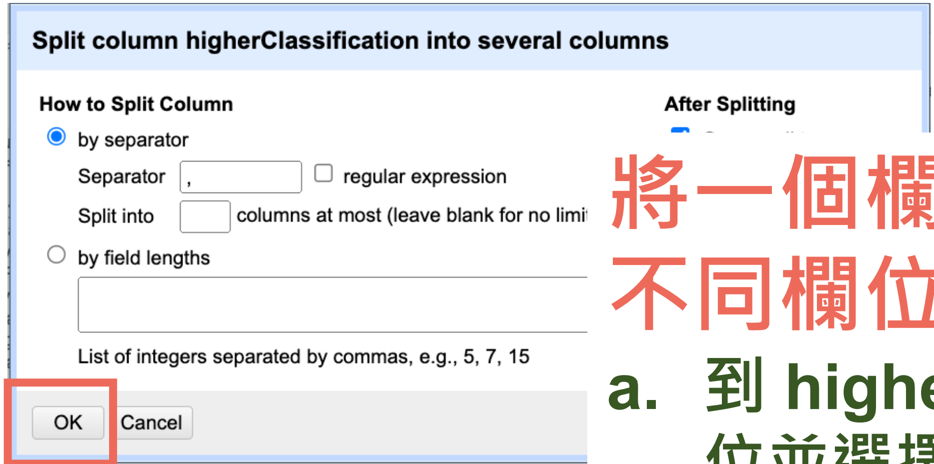
⬇

```
value.parseJson().get("kingdom")+", "+value.parseJson().get("phylum")+",
"+value.parseJson().get("class")+", "+value.parseJson().get("order")+", "+value.parseJson().get("family")
```

複製貼上請注意語法是否有空格和空行，請刪除

進階題-自動匯入高階層分類欄位




Split column higherClassification into several columns

How to Split Column

● by separator

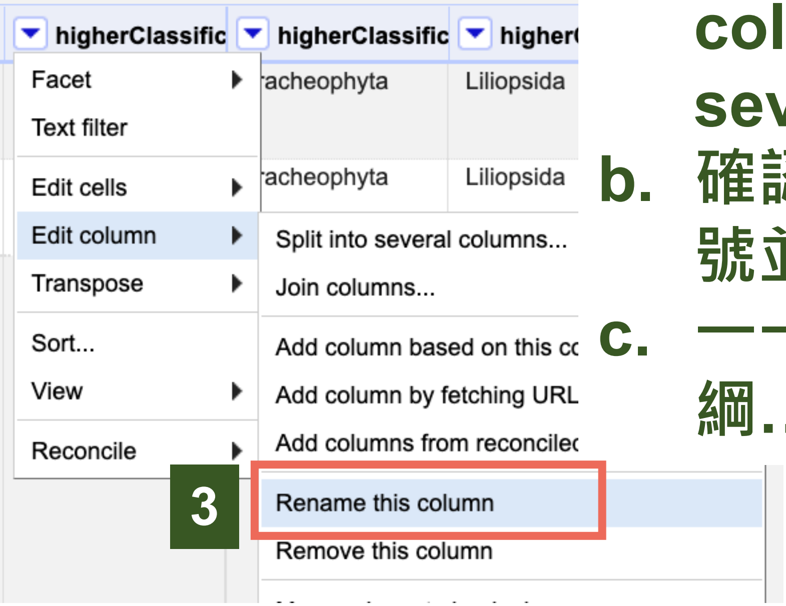
Separator , □ regular expression

Split into columns at most (leave blank for no limi

○ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting



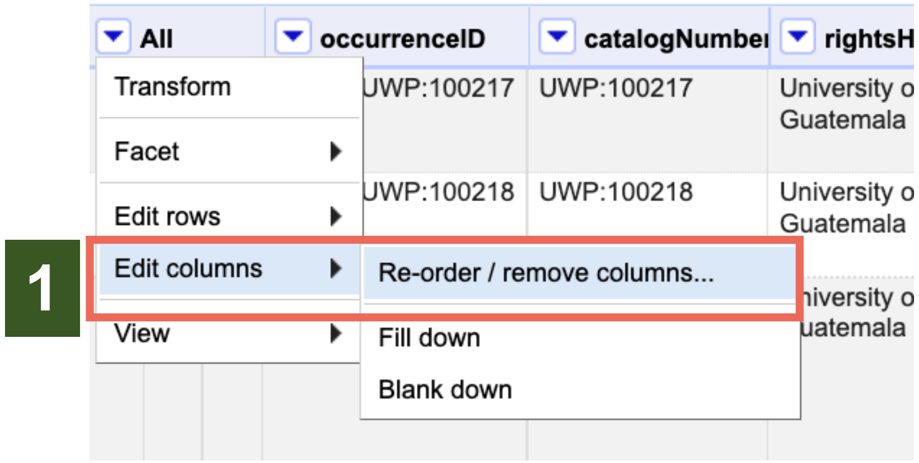
將一個欄位中的值分成不同欄位

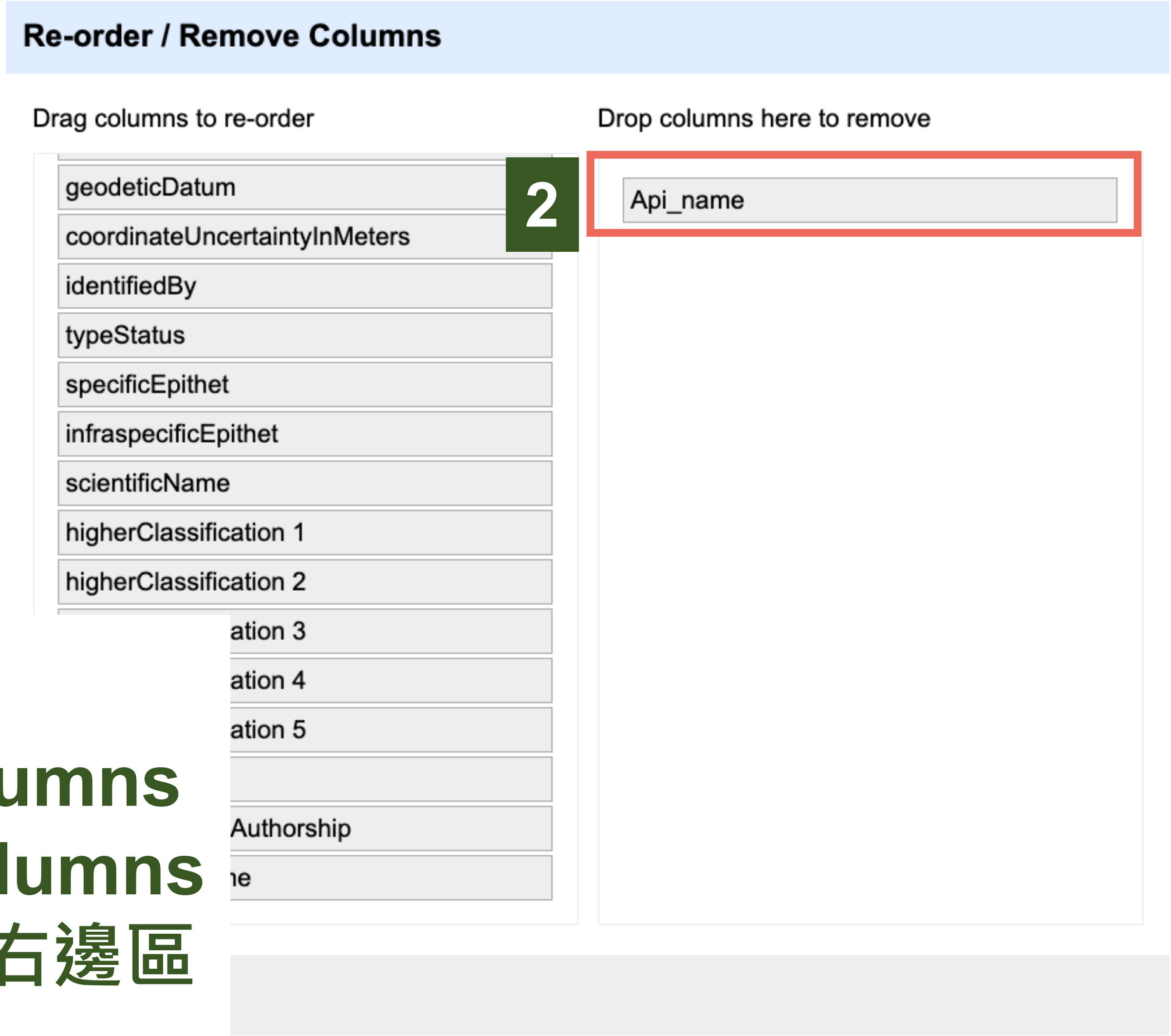
a. 到 higherClassification 欄位並選擇 Edit column >> Split into several columns

b. 確認該欄位的分隔符號是逗號並按 OK

c. 一一將欄位名稱改為界、門、綱…

進階題-自動匯入高階層分類欄位





Re-order / Remove Columns

Drag columns to re-order

geodeticDatum
coordinateUncertaintyInMeters
identifiedBy
typeStatus
specificEpithet
infraspecificEpithet
scientificName
higherClassification 1
higherClassification 2
ation 3
ation 4
ation 5
Authorship
ne

Drop columns here to remove

Api_name

將不要的欄位刪除

a. 到 All欄位並選擇Edit columns >> Re-order/ remove columns
b. 拖曳左邊不想要的欄位到右邊區域並按 OK

TaiBIF Open Data Toolkit
開放資料整合工具

TaiBIF 內容經理 劉璟儀

開發緣由與背景

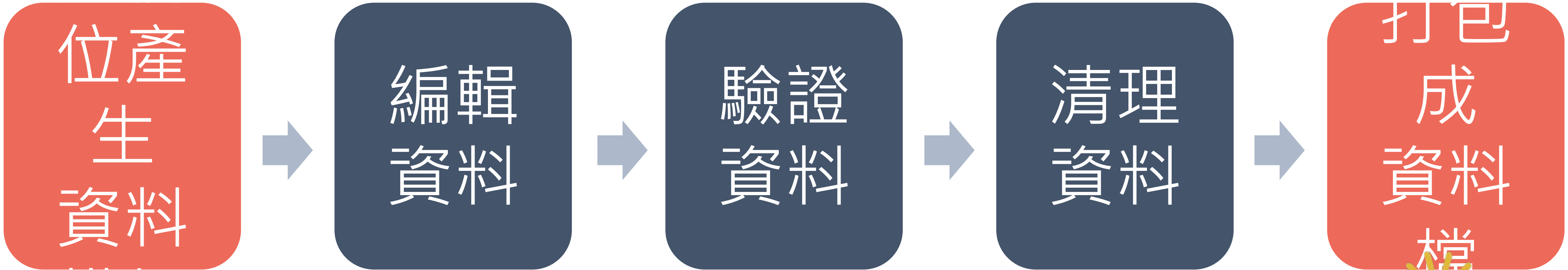
大家開放資料挫敗感太重

- DwC 那麼多要怎麼選
- 資料太多錯誤很難一次找到修正

很多工具可以用但尚未整合

- 如果有一個可以從資料欄位產生到編輯到清理都可以在同一個地方做完的工具就好了…

TaiBIF Open Data Toolkit
開放資料整合工具出爐！



位產生資料

編輯資料

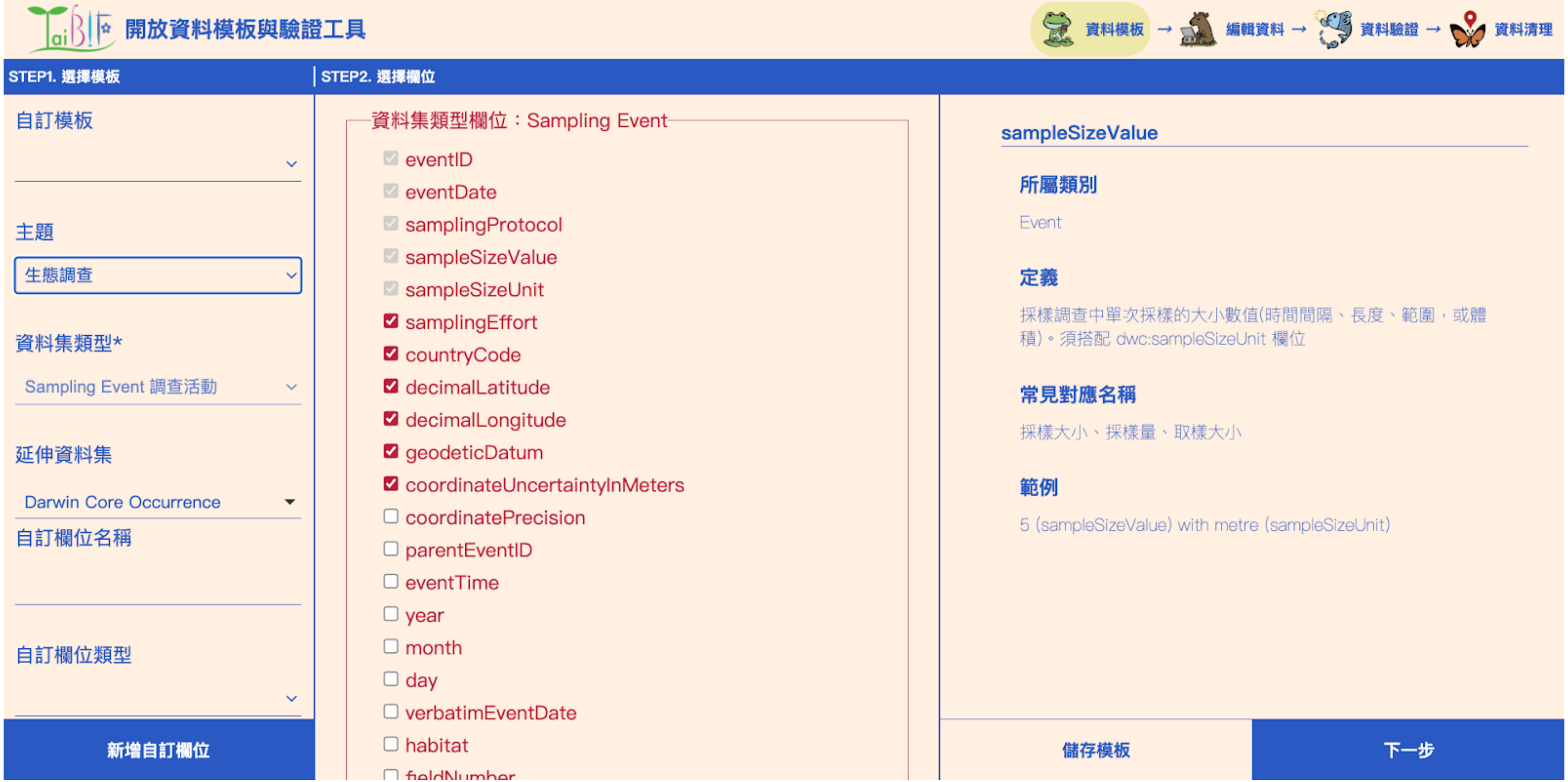
驗證資料

清理資料

打包成資料檔

直接可以上傳IPT～

第一步：建立資料模板



STEP1. 選擇模板

自訂模板

主題: 生態調查

資料集類型*: Sampling Event 調查活動

延伸資料集: Darwin Core Occurrence

自訂欄位名稱

自訂欄位類型

新增自訂欄位

STEP2. 選擇欄位

資料集類型欄位：Sampling Event

☑ eventID
☑ eventDate
☑ samplingProtocol
☑ sampleSizeValue
☑ sampleSizeUnit
☑ samplingEffort
☑ countryCode
☑ decimalLatitude
☑ decimalLongitude
☑ geodeticDatum
☑ coordinateUncertaintyInMeters
□ coordinatePrecision
□ parentEventID
□ eventTime
□ year
□ month
□ day
□ verbatimEventDate
□ habitat
□ fieldNumber

sampleSizeValue

所屬類別

Event

定義

採樣調查中單次採樣的大小數值(時間間隔、長度、範圍，或體積)。須搭配 dwc:sampleSizeUnit 欄位

常見對應名稱

採樣大小、採樣量、取樣大小

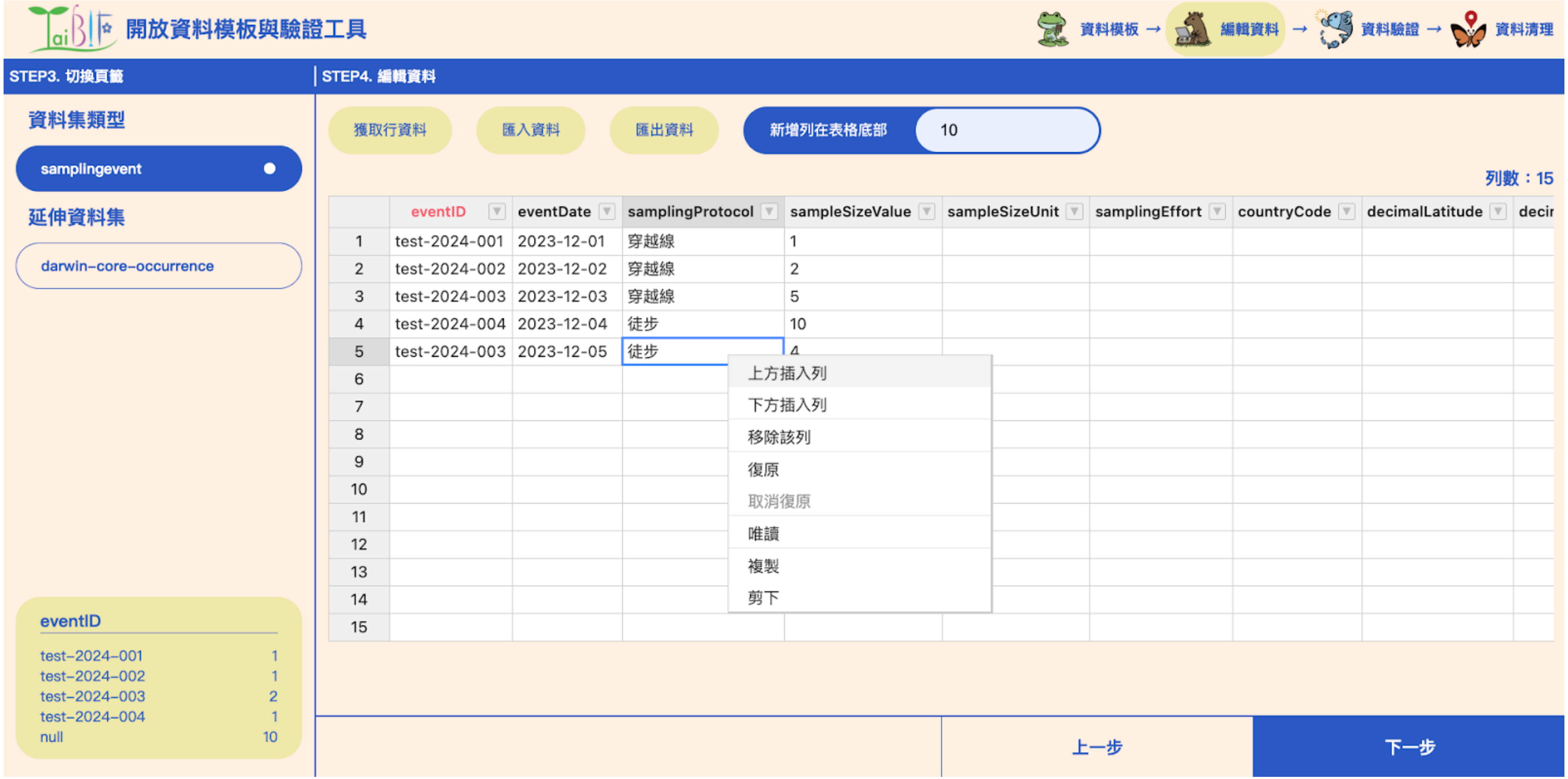
範例

5 (sampleSizeValue) with metre (sampleSizeUnit)

儲存模板

下一步

第二步：編輯資料



STEP3. 切換頁籤

STEP4. 編輯資料

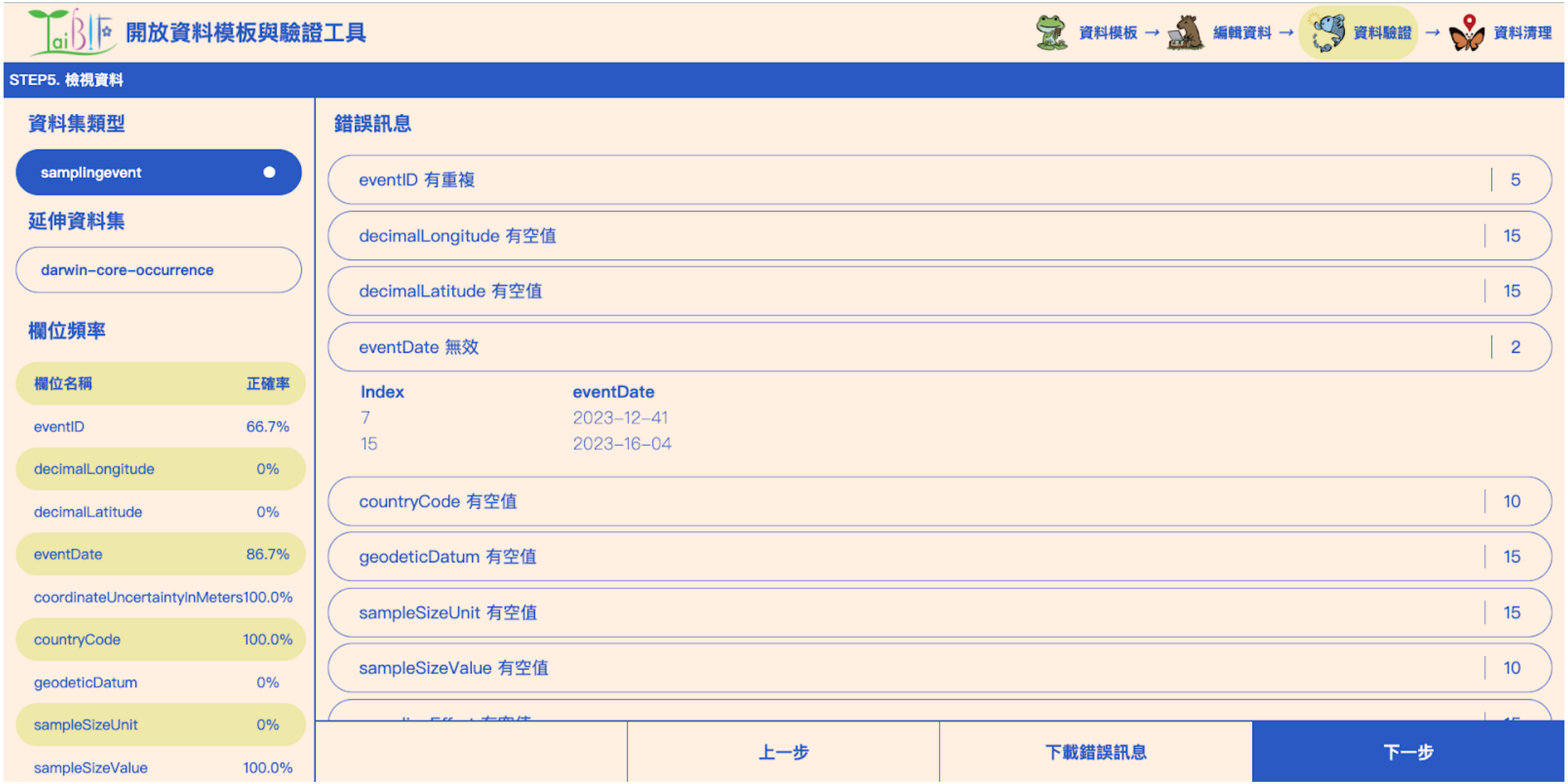
10

列數：15

	eventID	eventDate	samplingProtocol	sampleSizeValue	sampleSizeUnit	samplingEffort	countryCode	decimalLatitude	decir
1	test-2024-001	2023-12-01	穿越線	1					
2	test-2024-002	2023-12-02	穿越線	2					
3	test-2024-003	2023-12-03	穿越線	5					
4	test-2024-004	2023-12-04	徒步	10					
5	test-2024-003	2023-12-05	徒步	4					
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

eventID	
test-2024-001	1
test-2024-002	1
test-2024-003	2
test-2024-004	1
null	10

第三步：驗證資料



STEP5. 檢視資料

資料集類型

● samplingevent

延伸資料集

darwin-core-occurrence

欄位頻率

欄位名稱	正確率
eventID	66.7%
decimalLongitude	0%
decimalLatitude	0%
eventDate	86.7%
coordinateUncertaintyInMeters	100.0%
countryCode	100.0%
geodeticDatum	0%
sampleSizeUnit	0%
sampleSizeValue	100.0%

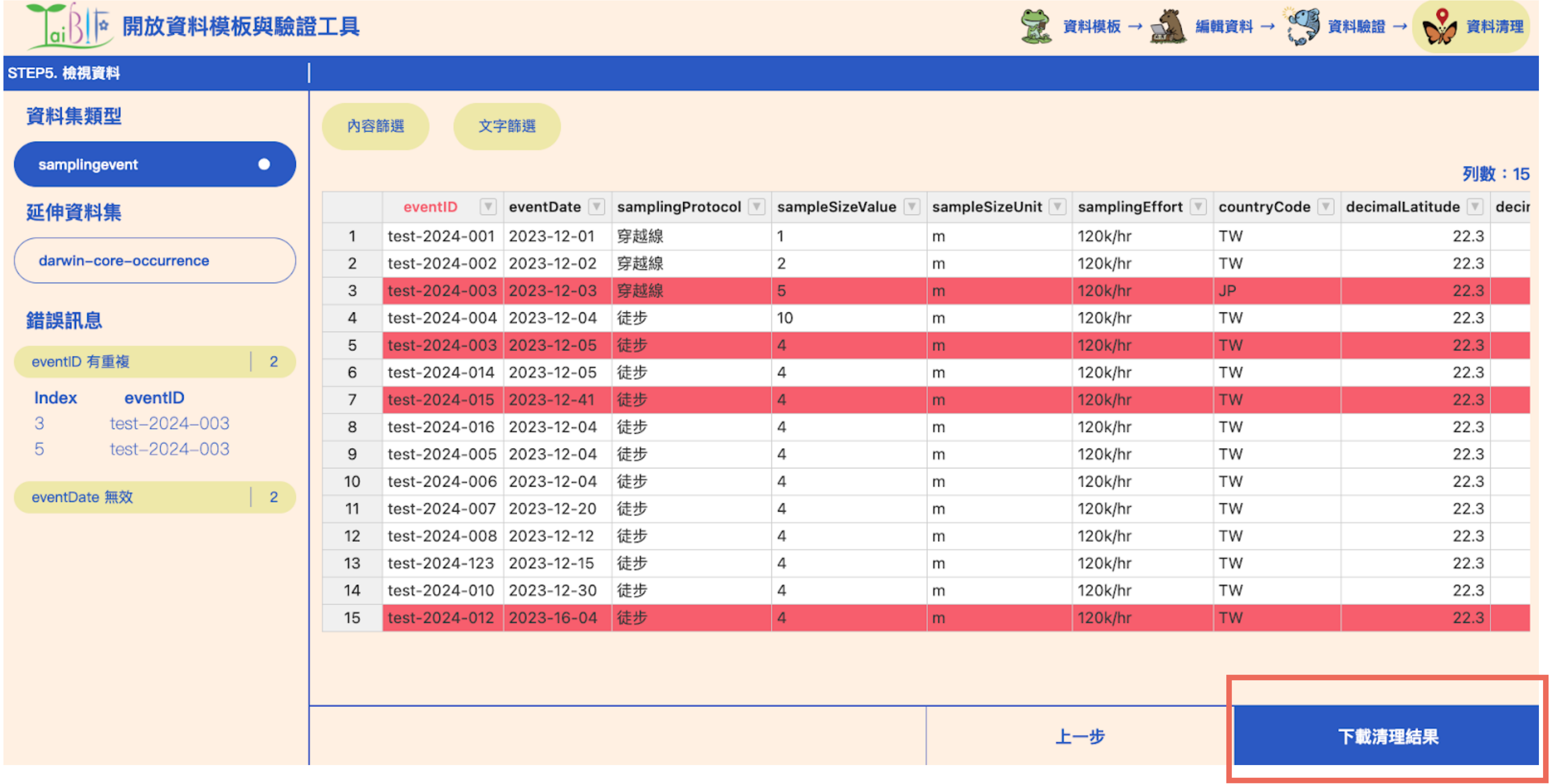
錯誤訊息

eventID 有重複	5
decimalLongitude 有空值	15
decimalLatitude 有空值	15
eventDate 無效	2

Index	eventDate
7	2023-12-41
15	2023-16-04

countryCode 有空值	10
geodeticDatum 有空值	15
sampleSizeUnit 有空值	15
sampleSizeValue 有空值	10

第四步：清理資料與打包



資料模板 → 編輯資料 → 資料驗證 → 資料清理

資料集類型

samplingevent

延伸資料集

darwin-core-occurrence

錯誤訊息

eventID 有重複 | 2

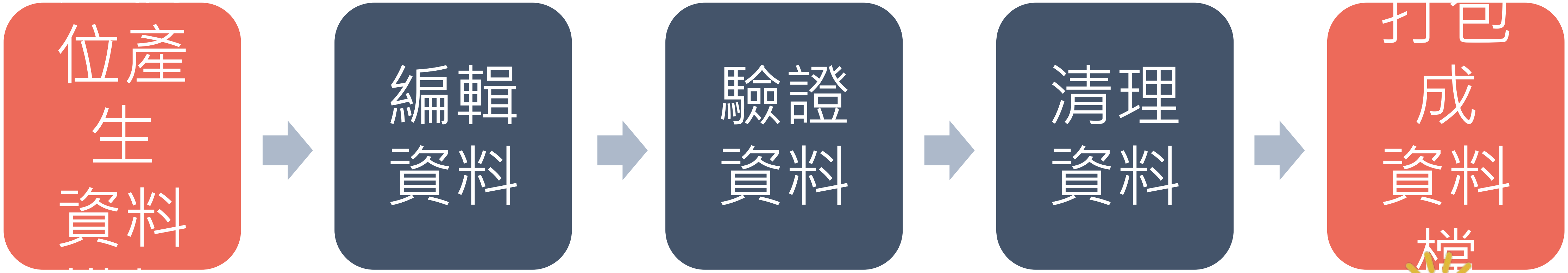
Index	eventID
3	test-2024-003
5	test-2024-003

eventDate 無效 | 2

列數：15

	eventID	eventDate	samplingProtocol	sampleSizeValue	sampleSizeUnit	samplingEffort	countryCode	decimalLatitude	decir
1	test-2024-001	2023-12-01	穿越線	1	m	120k/hr	TW	22.3	
2	test-2024-002	2023-12-02	穿越線	2	m	120k/hr	TW	22.3	
3	test-2024-003	2023-12-03	穿越線	5	m	120k/hr	JP	22.3	
4	test-2024-004	2023-12-04	徒步	10	m	120k/hr	TW	22.3	
5	test-2024-003	2023-12-05	徒步	4	m	120k/hr	TW	22.3	
6	test-2024-014	2023-12-05	徒步	4	m	120k/hr	TW	22.3	
7	test-2024-015	2023-12-41	徒步	4	m	120k/hr	TW	22.3	
8	test-2024-016	2023-12-04	徒步	4	m	120k/hr	TW	22.3	
9	test-2024-005	2023-12-04	徒步	4	m	120k/hr	TW	22.3	
10	test-2024-006	2023-12-04	徒步	4	m	120k/hr	TW	22.3	
11	test-2024-007	2023-12-20	徒步	4	m	120k/hr	TW	22.3	
12	test-2024-008	2023-12-12	徒步	4	m	120k/hr	TW	22.3	
13	test-2024-123	2023-12-15	徒步	4	m	120k/hr	TW	22.3	
14	test-2024-010	2023-12-30	徒步	4	m	120k/hr	TW	22.3	
15	test-2024-012	2023-16-04	徒步	4	m	120k/hr	TW	22.3	

TaiBIF Open Data Toolkit
開放資料整合工具出爐！



位產生資料

編輯資料

驗證資料

清理資料

打包成資料檔

預計今年優化完成後，年底上線

下個月記得交作業喔～

Open Refine / DwC 作業繳交 - https://forms.gle/zVCr1Ujeq3971TSs6

需要認證證書的人一定要交喔！

Thank you!

Q & A 時間