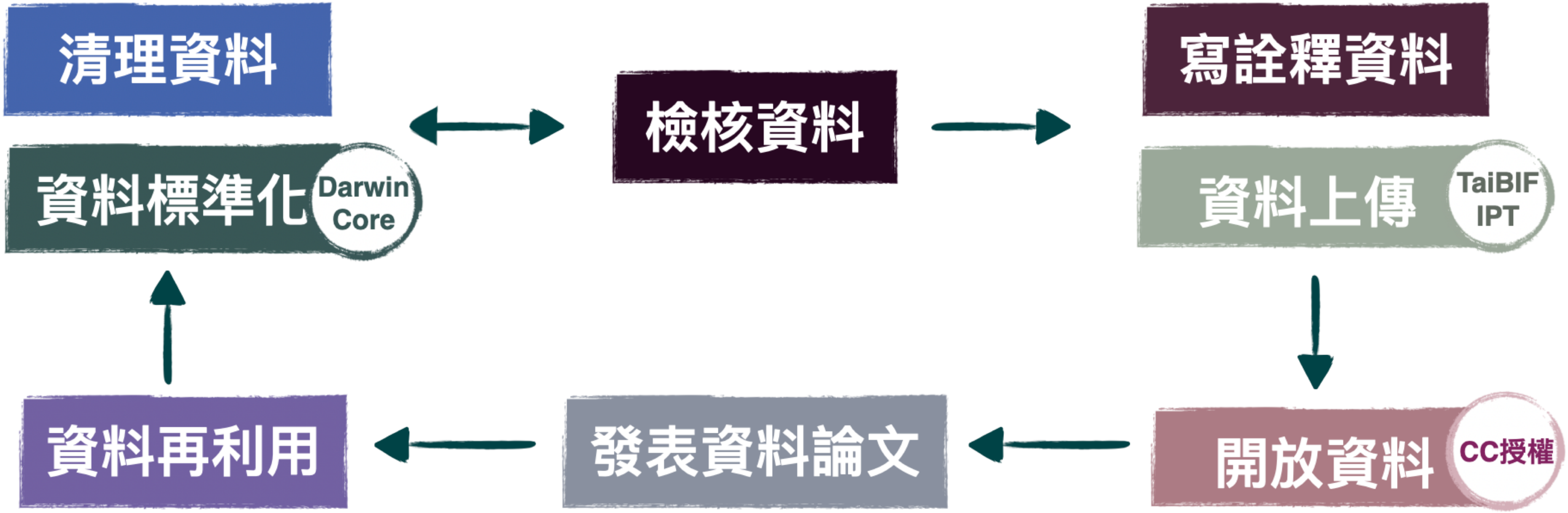
開放資料的前置準備

TaiBIF 內容經理 劉環儀

上傳資料前...
你應該準備好這些事



清理資料

資料標準化 Darwin Core

檢核資料

寫詮釋資料

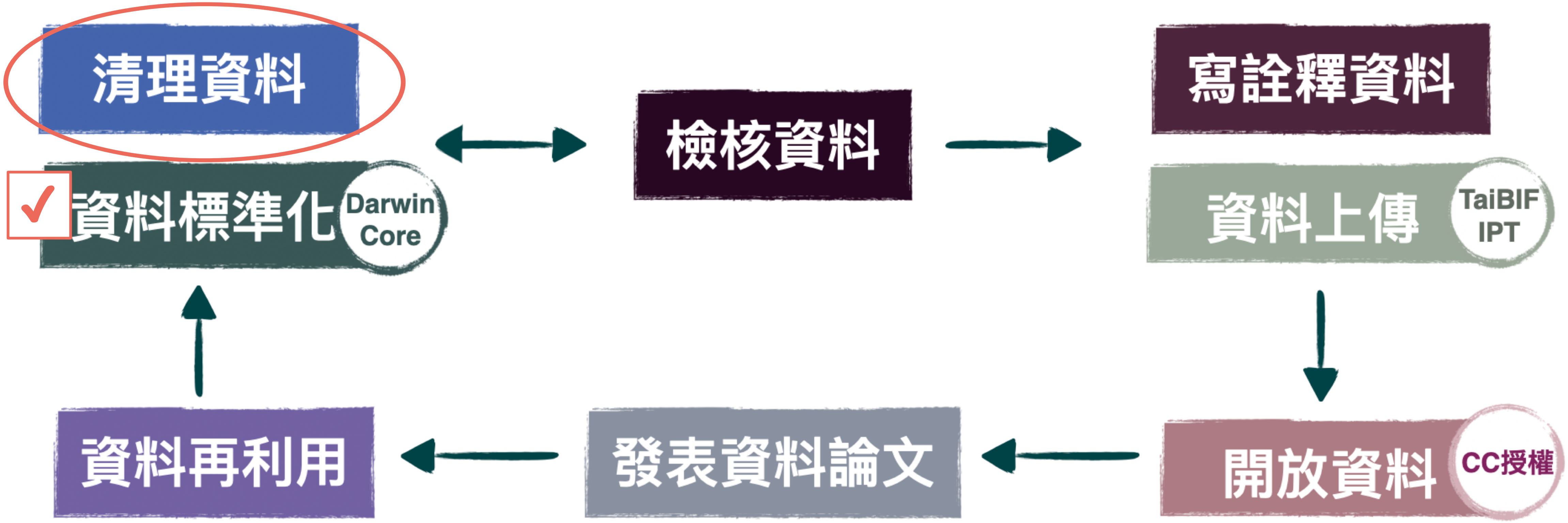
資料上傳 TaiBIF IPT

開放資料 CC授權

發表資料論文

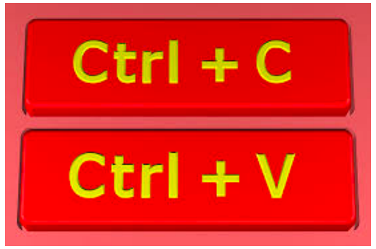
資料再利用

上傳資料前...
你應該準備好這些事



什麼是資料清理？

你可能以為的資料清理...



把別人的資料弄成自己看得懂的樣子

把資料表 A 貼到資料表 B

~~把不顯著的資料刪掉~~ 刪除明顯錯誤的資料

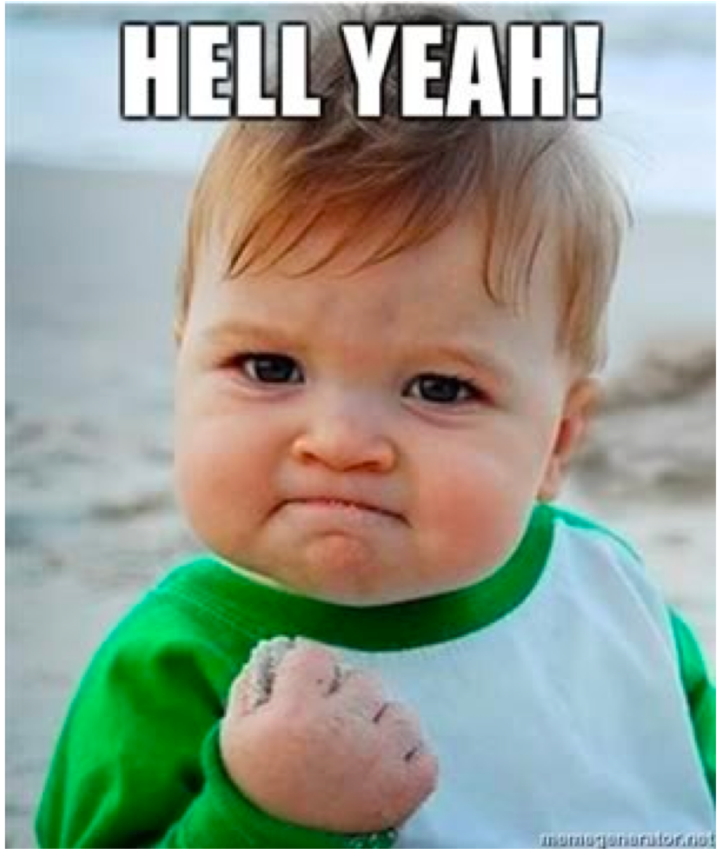
什麼是資料清理？

透過找出資料的錯誤、缺漏並修正，進而提升資料品質的過程

什麼是資料清理？

透過找出資料的錯誤、缺漏並修正，進而提升資料品質的過程

讓資料適合被利用

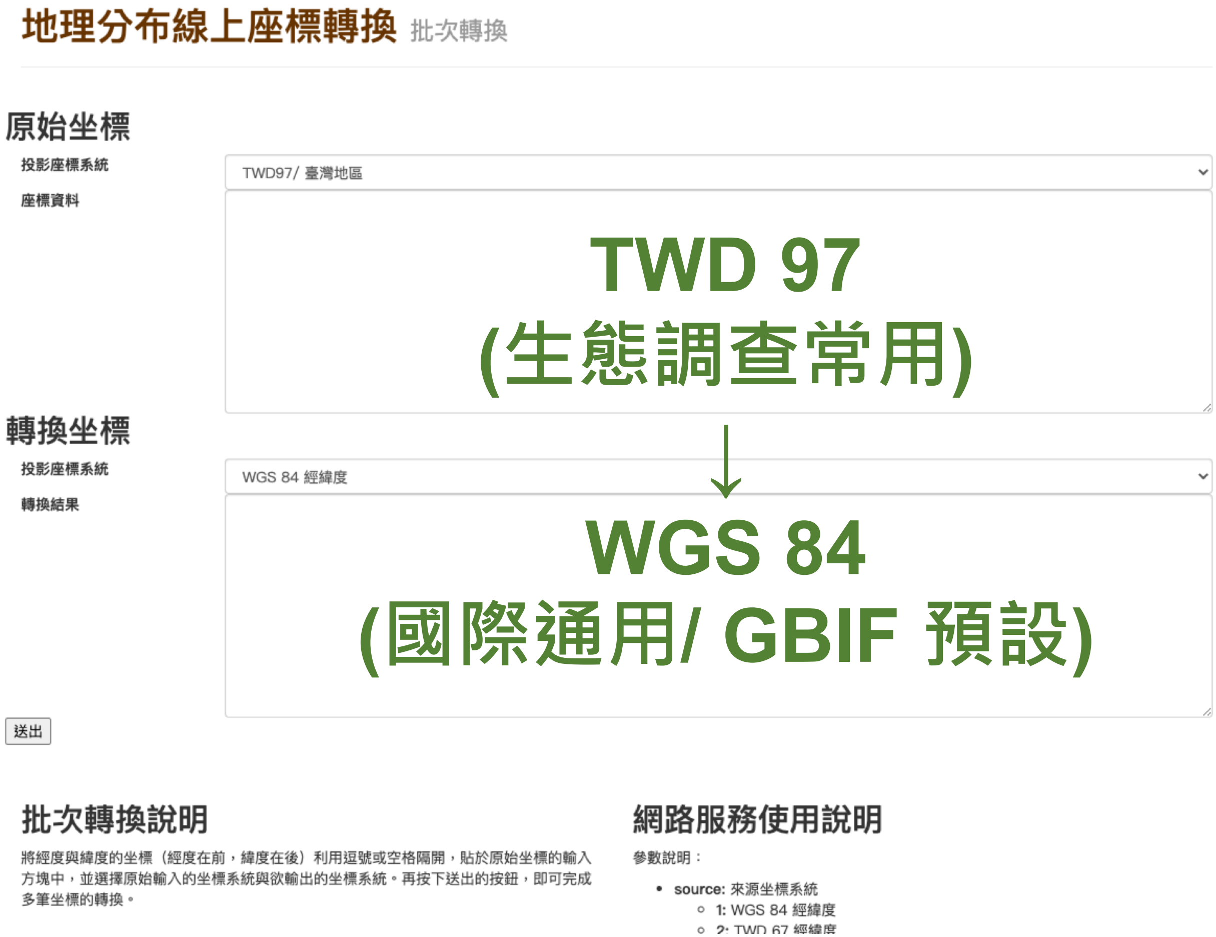


常見的資料錯誤

- 格式不一致/錯誤
- 拼字錯誤
- 資料缺漏
- 範圍錯誤
- ID/編碼重複

資料清理小工具

- 地理分布線上座標系統轉換 https://portal.taibif.tw/coordinateConverter.php



地理分布線上座標轉換 批次轉換

原始坐標

投影座標系統: TWD97/ 臺灣地區

座標資料

TWD 97
(生態調查常用)

↓

轉換坐標

投影座標系統: WGS 84 經緯度

轉換結果

WGS 84
(國際通用/ GBIF 預設)

批次轉換說明

將經度與緯度的坐標（經度在前，緯度在後）利用逗號或空格隔開，貼於原始坐標的輸入方塊中，並選擇原始輸入的坐標系統與欲輸出的坐標系統。再按下送出的按鈕，即可完成多筆坐標的轉換。

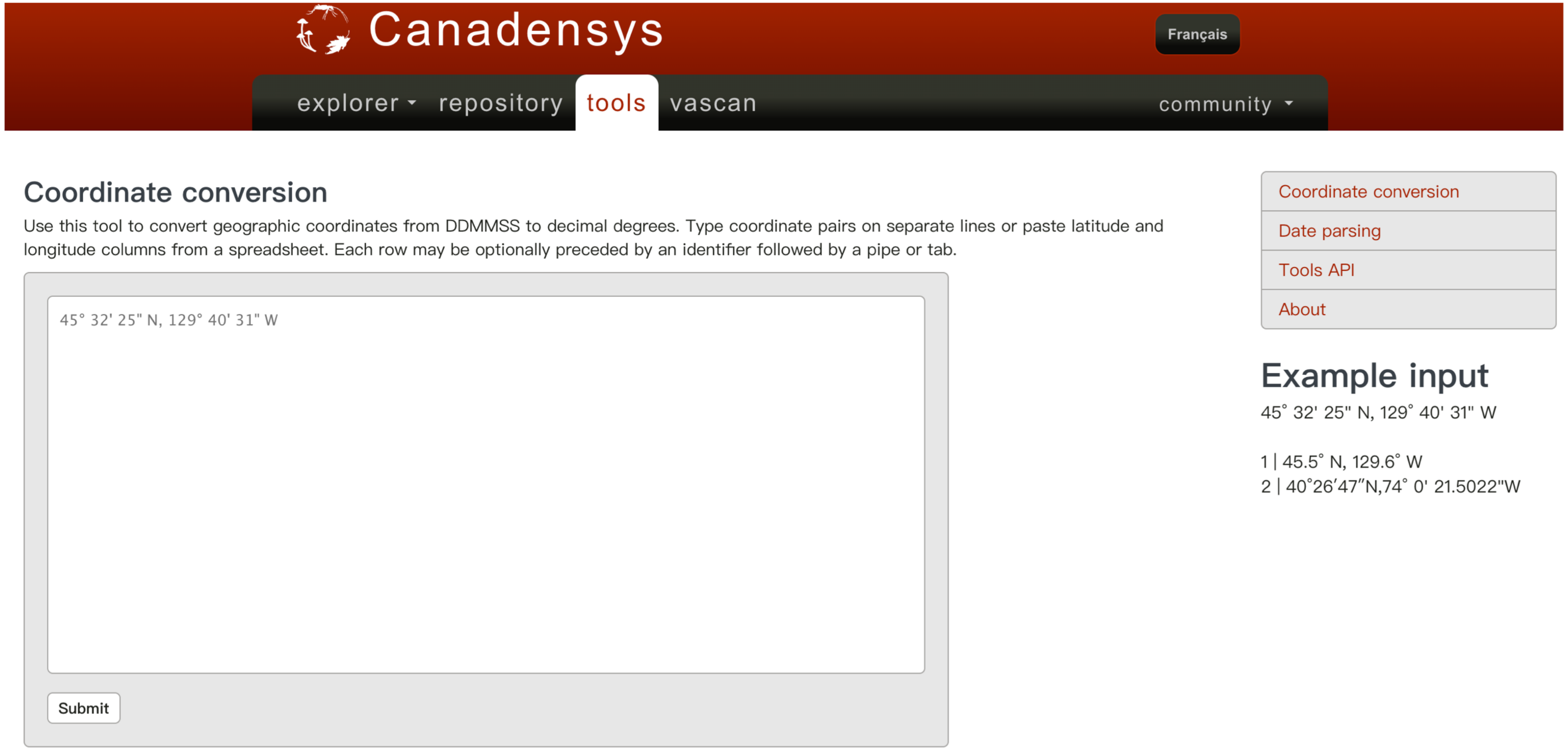
網路服務使用說明

參數說明：

- **source:** 來源坐標系統
  - **1:** WGS 84 經緯度
  - **2:** TWD 67 經緯度

資料清理小工具

- 座標轉換（度分秒- 十進位） https://data.canadensys.net/tools/coordinates



Coordinate conversion

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

45° 32' 25" N, 129° 40' 31" W

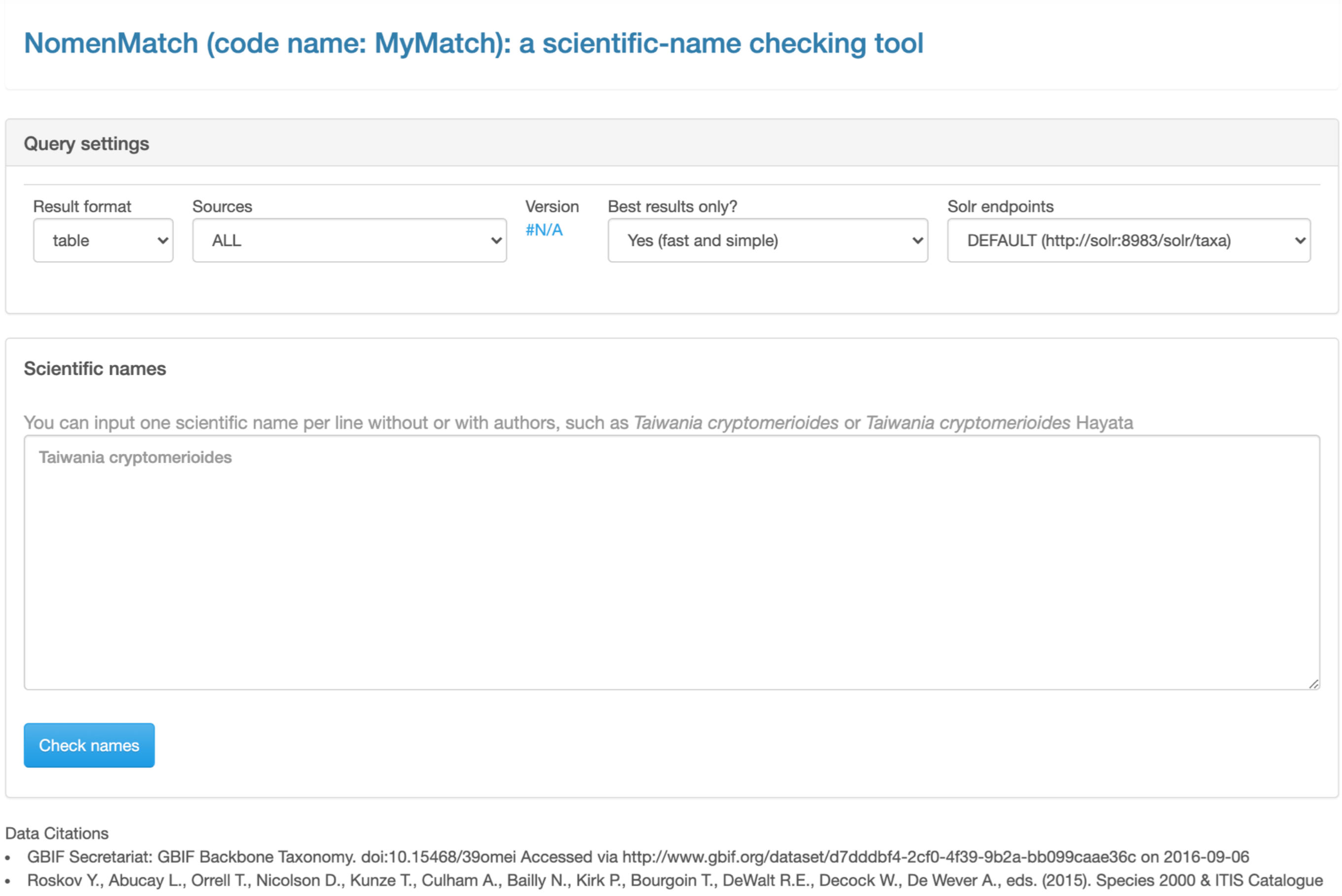
Example input

45° 32' 25" N, 129° 40' 31" W

1 | 45.5° N, 129.6° W
2 | 40°26’47”N,74° 0' 21.5022"W

資料清理小工具

- 有效學名比對 NomenMatch http://match.taibif.tw/



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

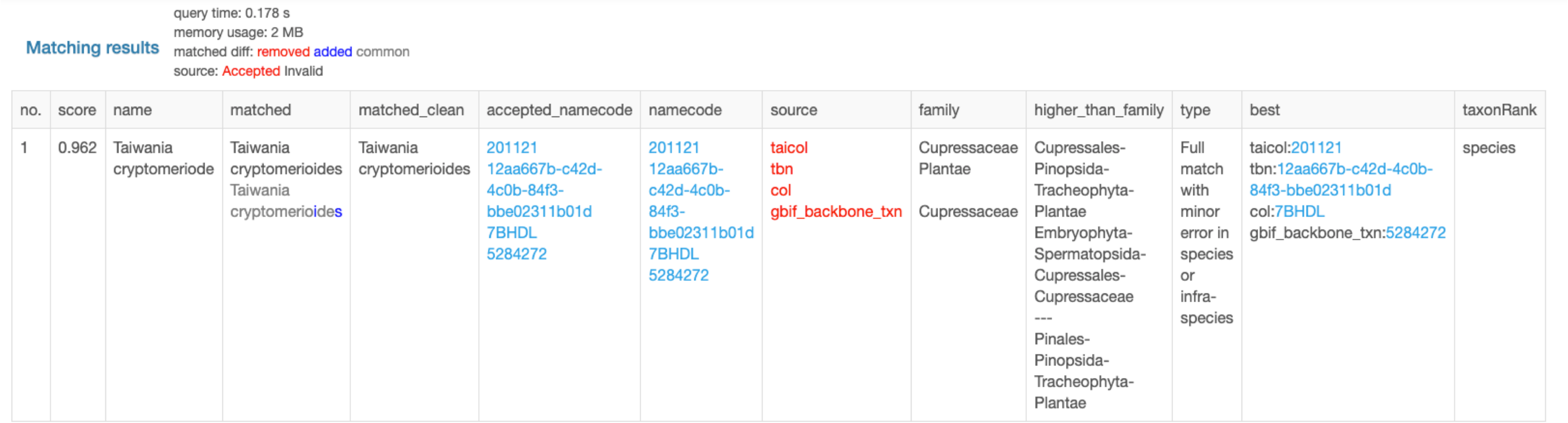
Taiwania cryptomerioides

Data Citations

- GBIF Secretariat: GBIF Backbone Taxonomy. doi:10.15468/39omei Accessed via http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c on 2016-09-06
- Roskov Y., Abucay L., Orrell T., Nicolson D., Kunze T., Culham A., Bailly N., Kirk P., Bourgoin T., DeWalt R.E., Decock W., De Wever A., eds. (2015). Species 2000 & ITIS Catalogue

資料清理小工具

- 有效學名比對 NomenMatch http://match.taibif.tw/



Matching results

query time: 0.178 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	taxonRank
1	0.962	Taiwania cryptomeriode	Taiwania cryptomerioides Taiwania cryptomerioides	Taiwania cryptomerioides	201121 12aa667b-c42d-4c0b-84f3-bbe02311b01d 7BHDL 5284272	201121 12aa667b-c42d-4c0b-84f3-bbe02311b01d 7BHDL 5284272	taicol tbn col gbif_backbone_txn	Cupressaceae Plantae Cupressaceae	Cupressales-Pinopsida-Tracheophyta-Plantae Embryophyta-Spermatopsida-Cupressales-Cupressaceae --- Pinales-Pinopsida-Tracheophyta-Plantae	Full match with minor error in species or infra-species	taicol:201121 tbn:12aa667b-c42d-4c0b-84f3-bbe02311b01d col:7BHDL gbif_backbone_txn:5284272	species

資料清理小工具

OpenRefine

不是資料庫

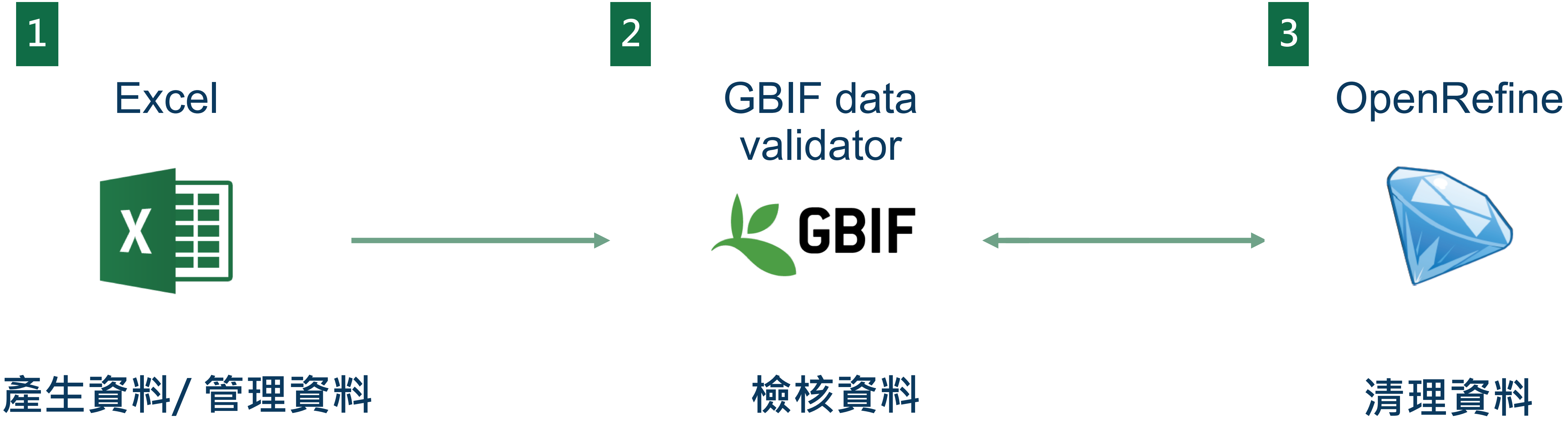
(無法儲存資料)

與 Excel 的使用方式不同

（只能清理資料）

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

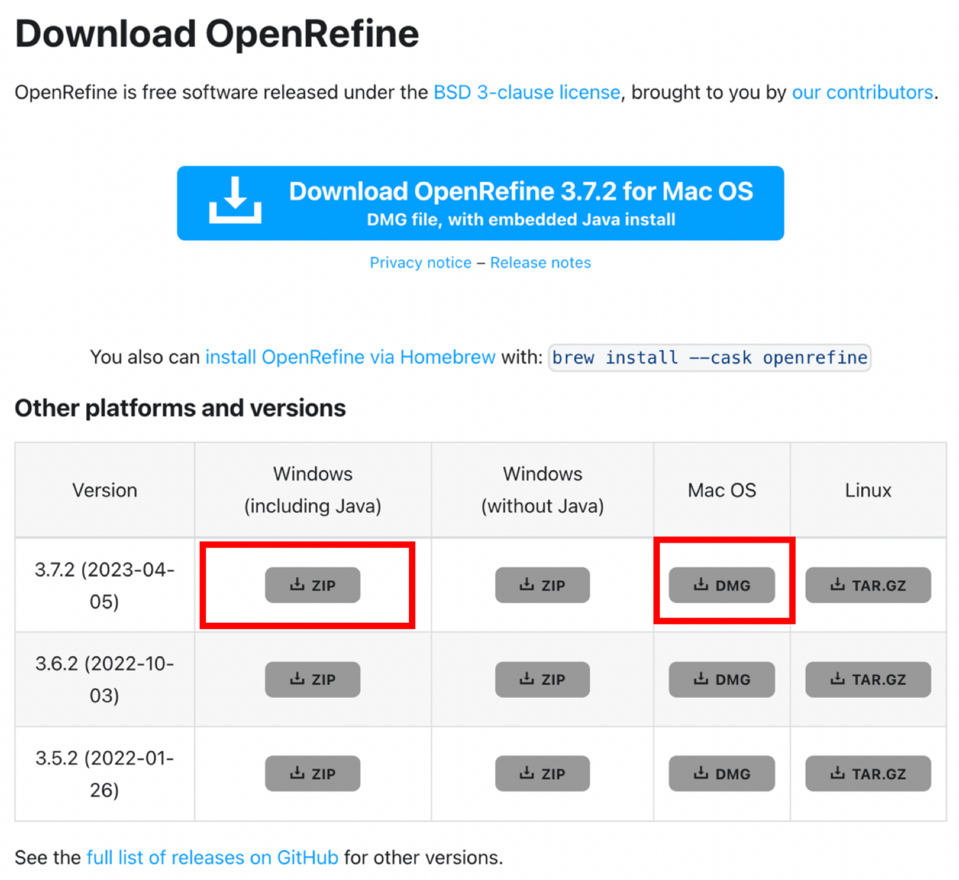
3

OpenRefine

清理資料

資料清理小工具

用 OpenRefine 清資料 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 3.7.2 for Mac OS
DMG file, with embedded Java install
Privacy notice – Release notes

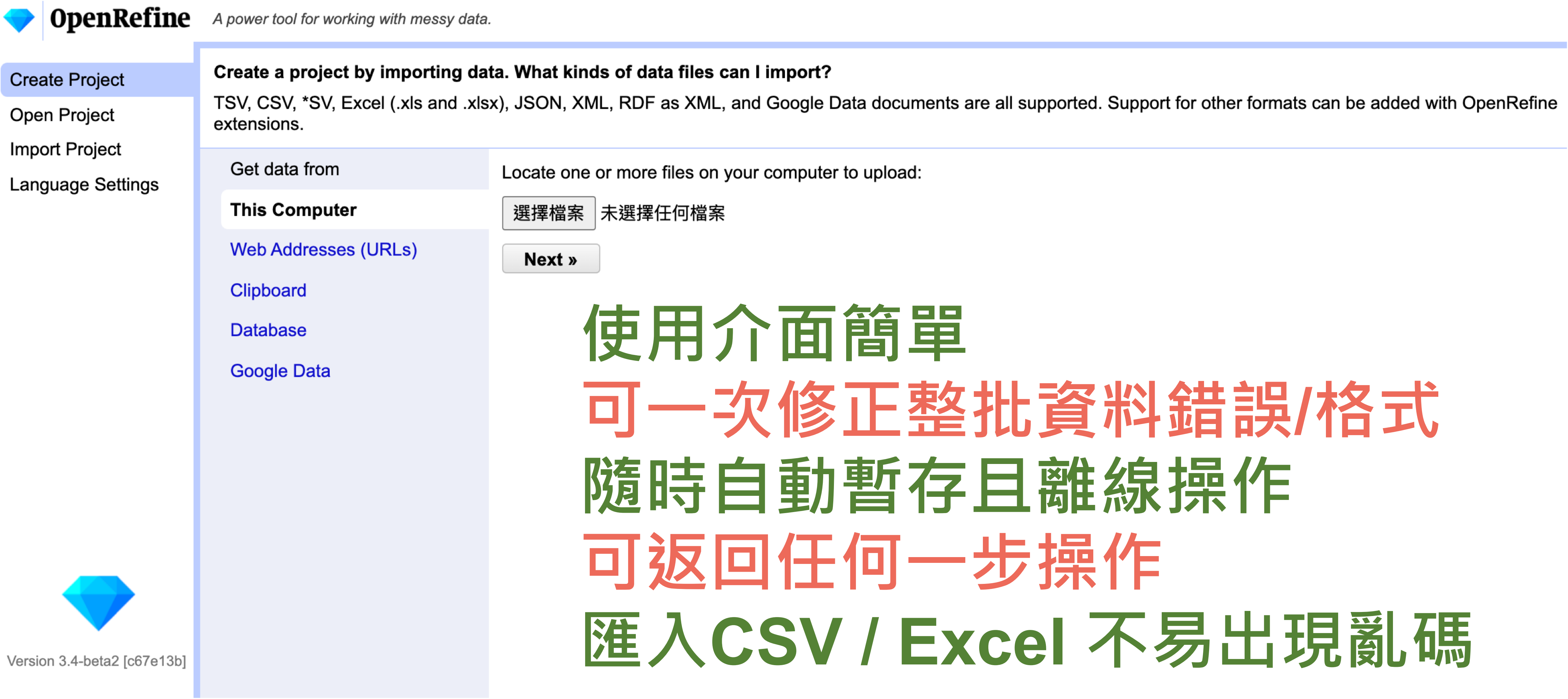
You also can install OpenRefine via Homebrew with: `brew install --cask openrefine`

Other platforms and versions

Version	Windows (including Java)	Windows (without Java)	Mac OS	Linux
3.7.2 (2023-04-05)	ZIP	ZIP	DMG	TAR.GZ
3.6.2 (2022-10-03)	ZIP	ZIP	DMG	TAR.GZ
3.5.2 (2022-01-26)	ZIP	ZIP	DMG	TAR.GZ

See the full list of releases on GitHub for other versions.

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:

使用介面簡單

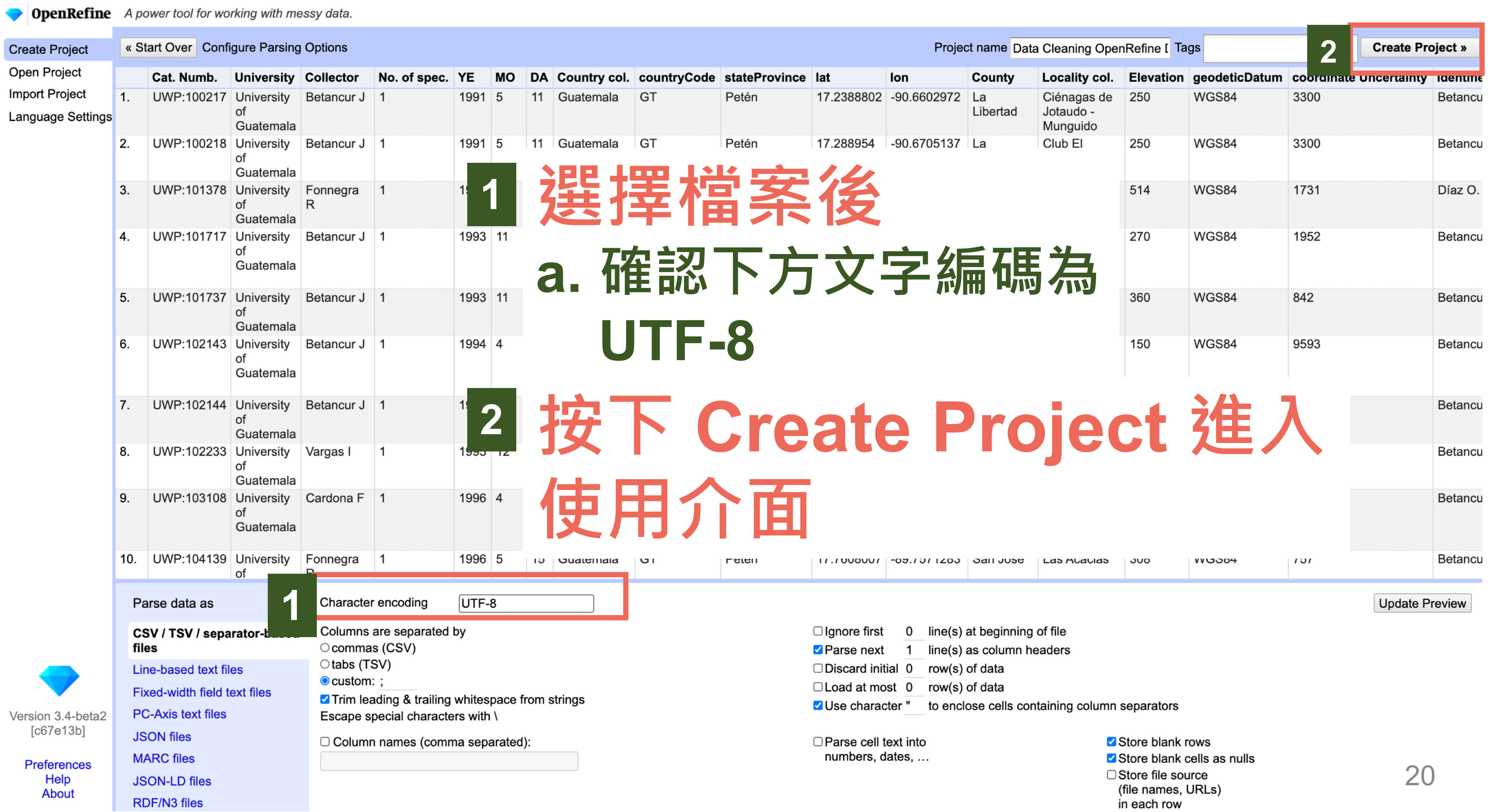
可一次修正整批資料錯誤/格式

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

用 OpenRefine 清資料



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

Project name: Data Cleaning OpenRefine [

	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	1										514	WGS84	1731	Díaz O.
4.	UWP:101717	University of Guatemala	Betancur J	1	1993	11									270	WGS84	1952	Betancu
5.	UWP:101737	University of Guatemala	Betancur J	1	1993	11									360	WGS84	842	Betancu
6.	UWP:102143	University of Guatemala	Betancur J	1	1994	4									150	WGS84	9593	Betancu
7.	UWP:102144	University of Guatemala	Betancur J	1	1													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 José	Las Acacias	308	WGS84	757	Betancu

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, ...

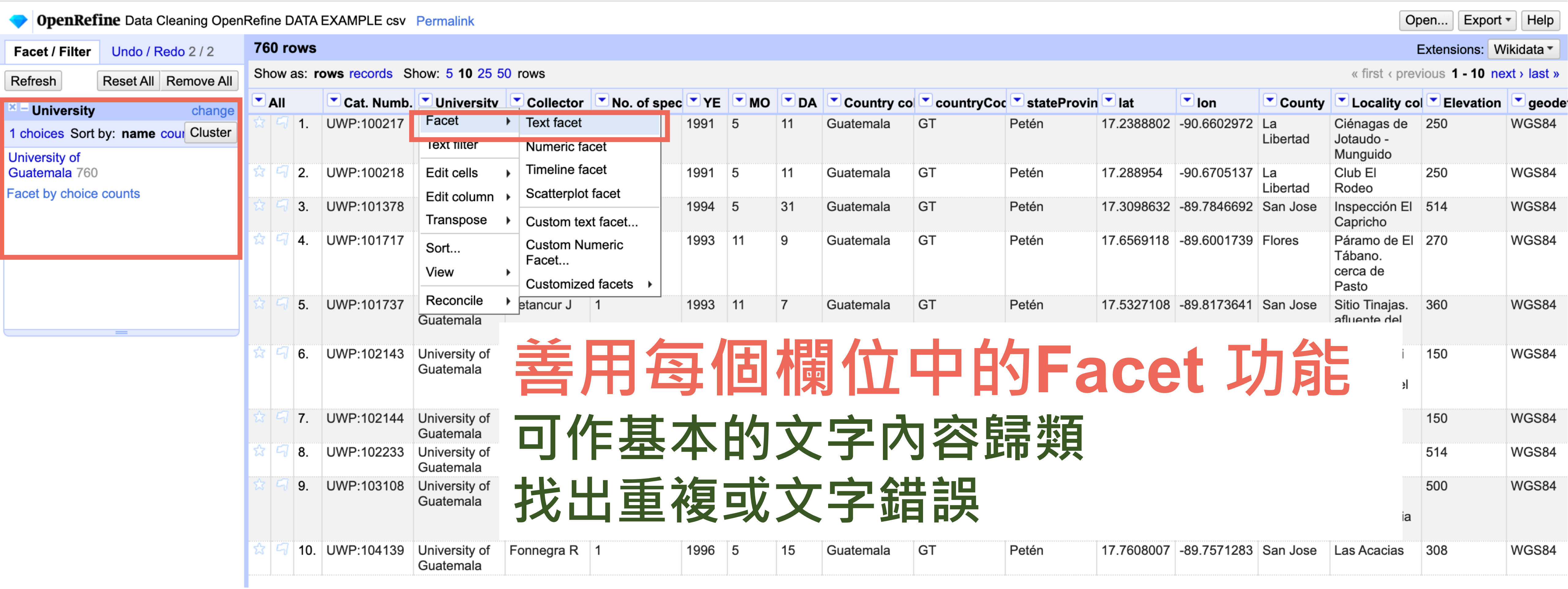
☑ Store blank rows
☑ Store blank cells as nulls
☐ Store file source (file names, URLs) in each row

1 選擇檔案後

a. 確認下方文字編碼為 UTF-8

2 按下 Create Project 進入使用介面

用 OpenRefine 清資料 範例1



Data Cleaning OpenRefine DATA EXAMPLE csv Permalink

Facet / Filter Undo / Redo 2 / 2

Refresh Reset All Remove All

University change

1 choices Sort by: name count Cluster

University of Guatemala 760

Facet by choice counts

760 rows

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

« first ‹ previous 1 - 10 next › last »

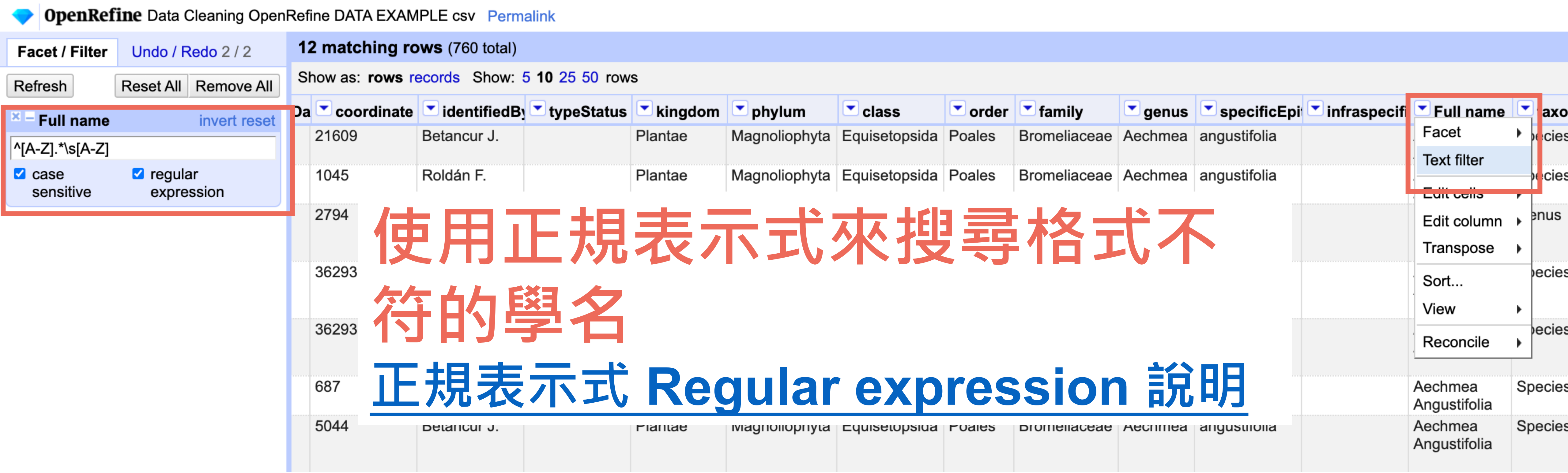
All	Cat. Numb.	University	Collector	No. of spec	YE	MO	DA	Country co	countryCoc	stateProvin	lat	lon	County	Locality col	Elevation	geode
1.	UWP:100217				1991	5	11	Guatemala	GT	Petén	17.2388802	-90.6602972	La Libertad	Ciénagas de Jotaudo - Munguido	250	WGS84
2.	UWP:100218				1991	5	11	Guatemala	GT	Petén	17.288954	-90.6705137	La Libertad	Club El Rodeo	250	WGS84
3.	UWP:101378				1994	5	31	Guatemala	GT	Petén	17.3098632	-89.7846692	San Jose	Inspección El Capricho	514	WGS84
4.	UWP:101717				1993	11	9	Guatemala	GT	Petén	17.6569118	-89.6001739	Flores	Páramo de El Tábano. cerca de Pasto	270	WGS84
5.	UWP:101737	Guatemala	etancur J	1	1993	11	7	Guatemala	GT	Petén	17.5327108	-89.8173641	San Jose	Sitio Tinajas. afluente del	360	WGS84
6.	UWP:102143	University of Guatemala												i el	150	WGS84
7.	UWP:102144	University of Guatemala													150	WGS84
8.	UWP:102233	University of Guatemala													514	WGS84
9.	UWP:103108	University of Guatemala												ia	500	WGS84
10.	UWP:104139	University of Guatemala	Fonnegra R	1	1996	5	15	Guatemala	GT	Petén	17.7608007	-89.7571283	San Jose	Las Acacias	308	WGS84

善用每個欄位中的Facet 功能

可作基本的文字內容歸類

找出重複或文字錯誤

用 OpenRefine 清資料 範例2



OpenRefine Data Cleaning OpenRefine DATA EXAMPLE csv

Full name

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

☑ case sensitive ☑ regular expression

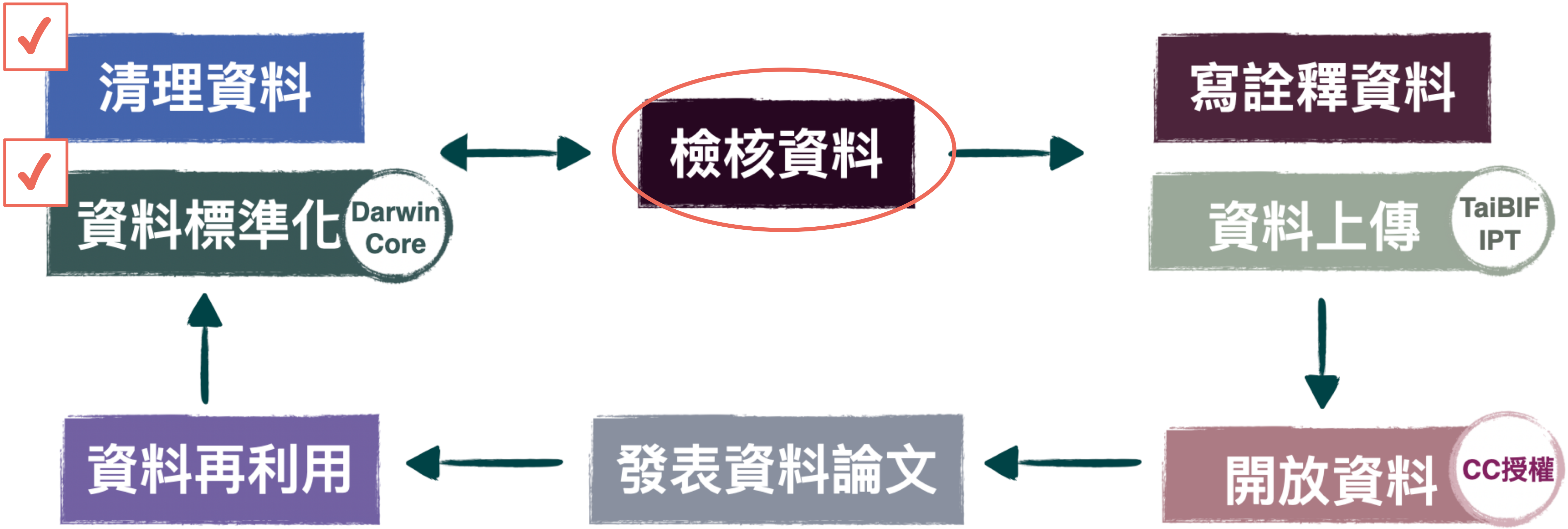
12 matching rows (760 total)

Da	coordinate	identifiedBy	typeStatus	kingdom	phylum	class	order	family	genus	specificEpi	infraspecifi	Full name	axo
	21609	Betancur J.		Plantae	Magnoliophyta	Equisetopsida	Poales	Bromeliaceae	Aechmea	angustifolia			ecies
	1045	Roldán F.		Plantae	Magnoliophyta	Equisetopsida	Poales	Bromeliaceae	Aechmea	angustifolia			ecies
	2794												enus
	36293												pecies
	36293												pecies
	687											Aechmea Angustifolia	Species
	5044	Betancur J.		Plantae	Magnoliophyta	Equisetopsida	Poales	Bromeliaceae	Aechmea	angustifolia		Aechmea Angustifolia	Species

使用正規表示式來搜尋格式不符的學名

[正規表示式 Regular expression 說明]

上傳資料前...
你應該準備好這些事



☑ 清理資料

☑ 資料標準化 Darwin Core

檢核資料

寫詮釋資料

資料上傳 TaiBIF IPT

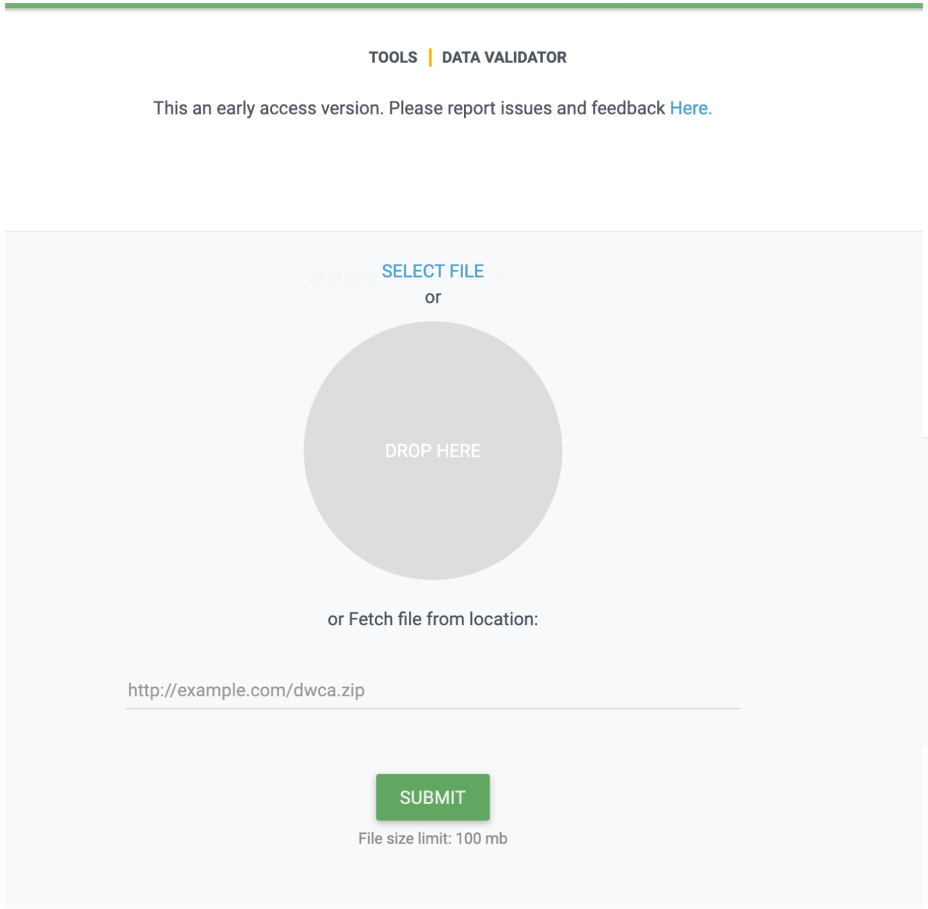
開放資料 CC授權

發表資料論文

資料再利用

檢核小工具

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



TOOLS | DATA VALIDATOR

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

SELECT FILE

or

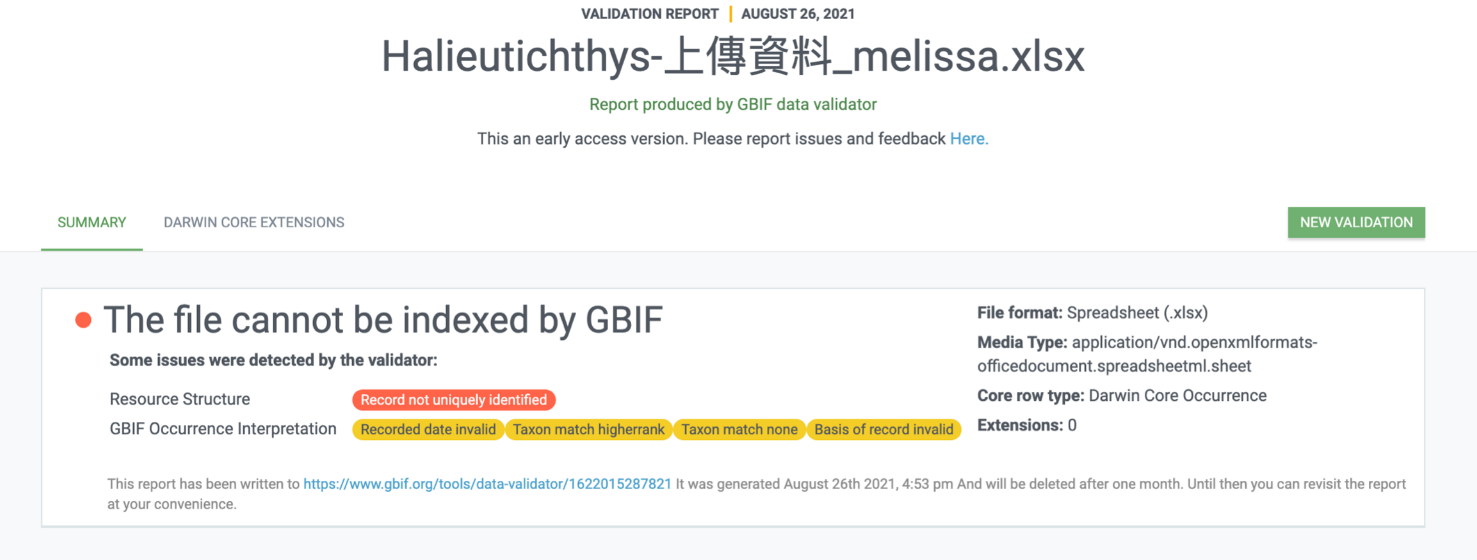
DROP HERE

or Fetch file from location:

http://example.com/dwca.zip

SUBMIT

File size limit: 100 mb



VALIDATION REPORT | AUGUST 26, 2021

Halieutichthys-上傳資料_melissa.xlsx

Report produced by GBIF data validator

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

SUMMARY   DARWIN CORE EXTENSIONS   NEW VALIDATION

● The file cannot be indexed by GBIF

Some issues were detected by the validator:

Resource Structure	Record not uniquely identified
GBIF Occurrence Interpretation	Recorded date invalid Taxon match higherrank Taxon match none Basis of record invalid

File format: Spreadsheet (.xlsx)

Media Type: application/vnd.openxmlformats-officedocument.spreadsheetml.sheet

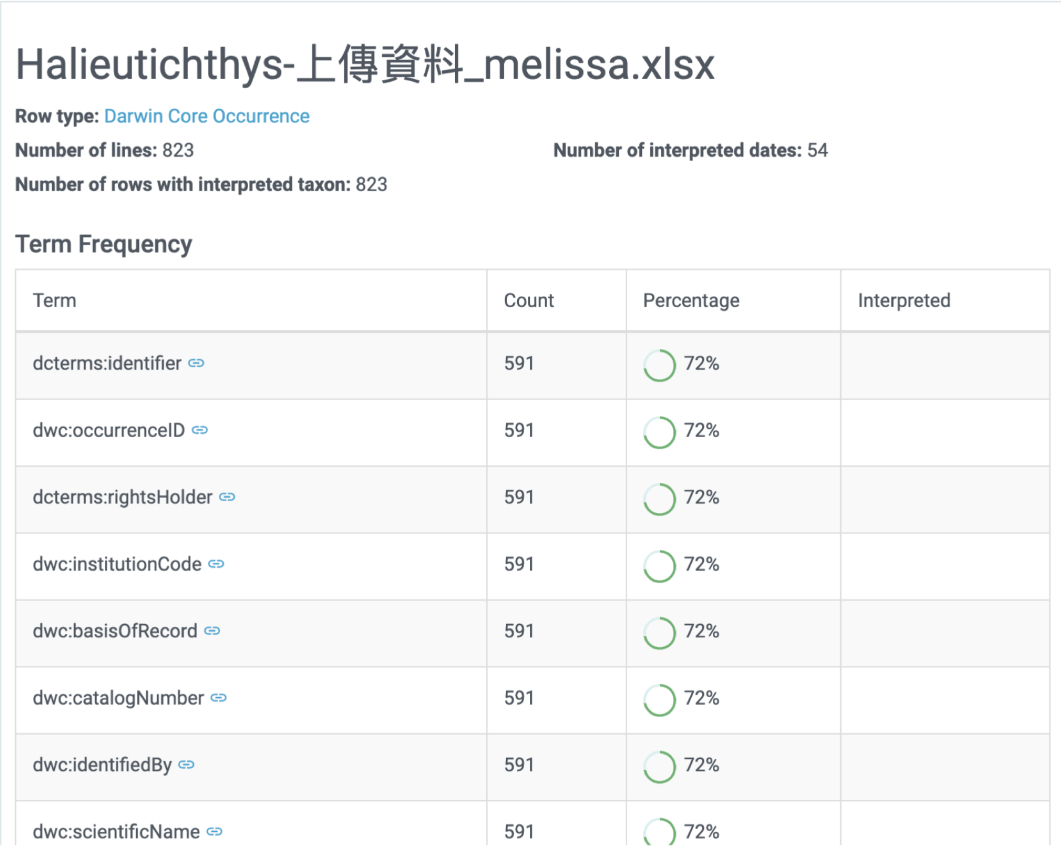
Core row type: Darwin Core Occurrence

Extensions: 0

This report has been written to https://www.gbif.org/tools/data-validator/1622015287821 It was generated August 26th 2021, 4:53 pm And will be deleted after one month. Until then you can revisit the report at your convenience.

檢核小工具

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



Halieutichthys-上傳資料_melissa.xlsx

Row type: Darwin Core Occurrence

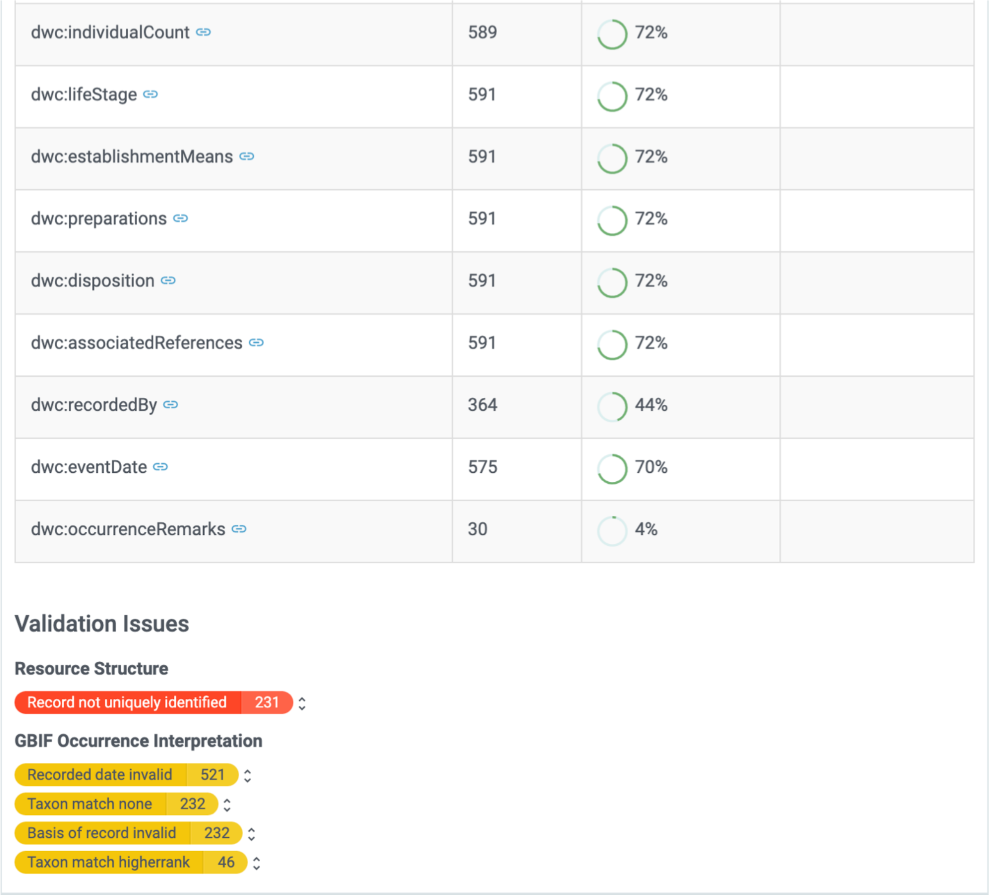
Number of lines: 823

Number of interpreted dates: 54

Number of rows with interpreted taxon: 823

Term Frequency

Term	Count	Percentage	Interpreted
dcterms:identifier	591	72%	
dwc:occurrenceID	591	72%	
dcterms:rightsHolder	591	72%	
dwc:institutionCode	591	72%	
dwc:basisOfRecord	591	72%	
dwc:catalogNumber	591	72%	
dwc:identifiedBy	591	72%	
dwc:scientificName	591	72%	



dwc:individualCount	589	72%	
dwc:lifeStage	591	72%	
dwc:establishmentMeans	591	72%	
dwc:preparations	591	72%	
dwc:disposition	591	72%	
dwc:associatedReferences	591	72%	
dwc:recordedBy	364	44%	
dwc:eventDate	575	70%	
dwc:occurrenceRemarks	30	4%	

Validation Issues

Resource Structure

Record not uniquely identified 231

GBIF Occurrence Interpretation

Recorded date invalid 521

Taxon match none 232

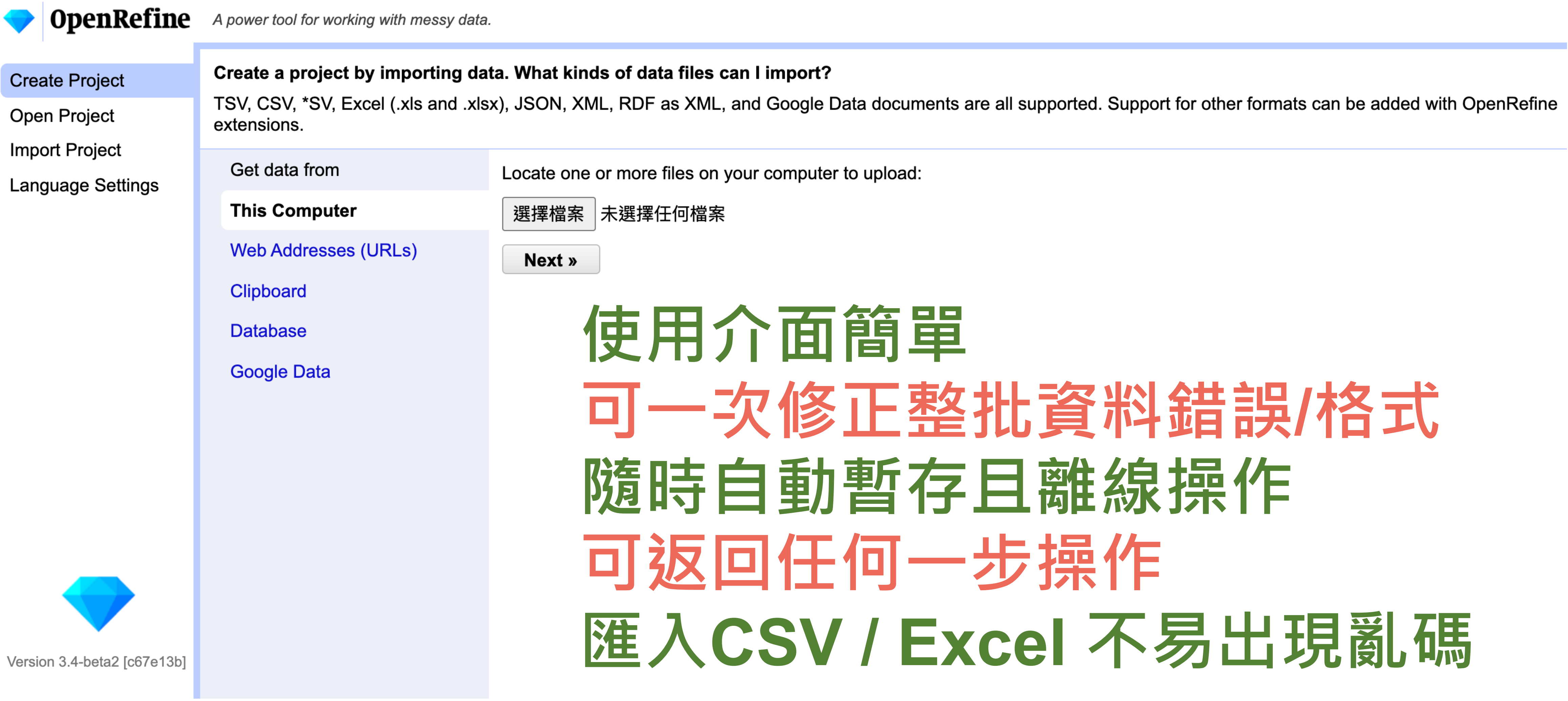
Basis of record invalid 232

Taxon match higherrank 46

用OpenRefine清理資料

TaiBIF 內容經理 劉璟儀

用 OpenRefine 清資料



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 3.4-beta2 [c67e13b]

使用介面簡單

可一次修正整批資料錯誤/格式

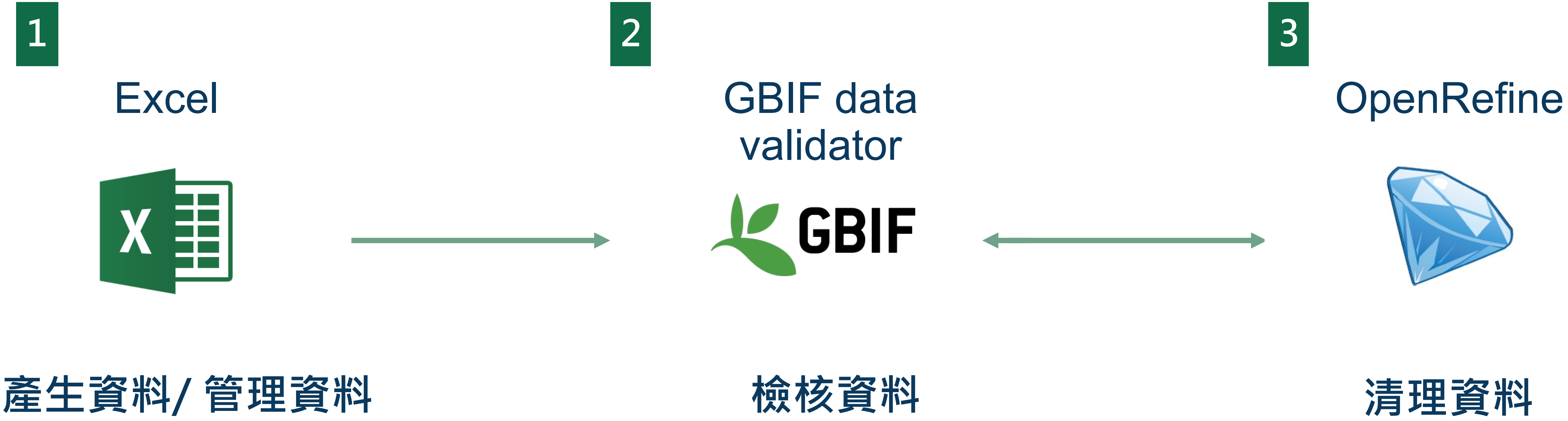
隨時自動暫存且離線操作

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

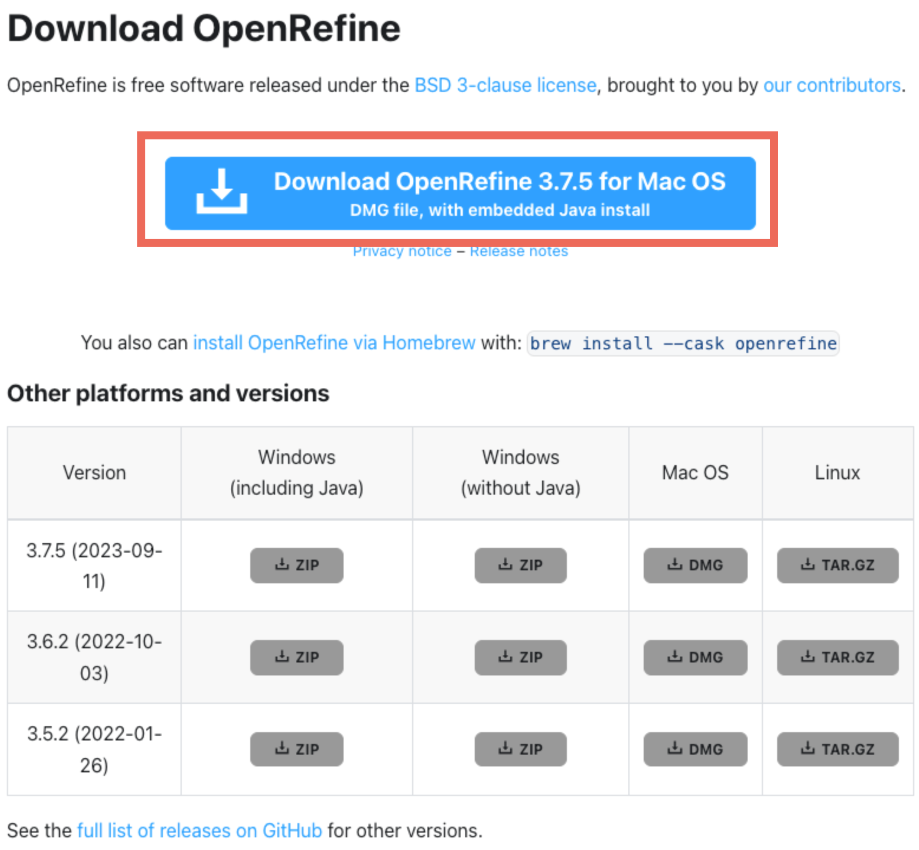
用OpenRefine清理資料

會需要用到的連結

練習檔案下載
GBIF Data Validator
NomenMatch 學名比對
Global Names Resolver

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 3.7.5 for Mac OS
DMG file, with embedded Java install
Privacy notice – Release notes

You also can install OpenRefine via Homebrew with: `brew install --cask openrefine`

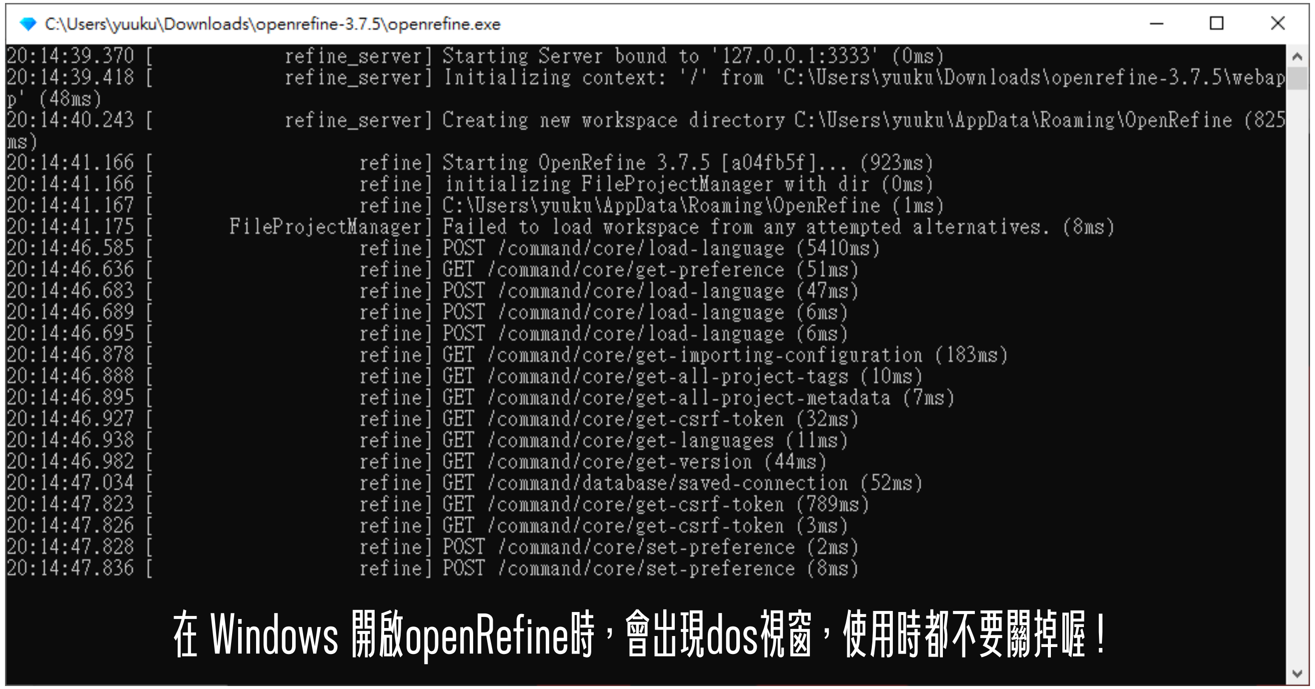
Other platforms and versions

Version	Windows (including Java)	Windows (without Java)	Mac OS	Linux
3.7.5 (2023-09-11)	ZIP	ZIP	DMG	TAR.GZ
3.6.2 (2022-10-03)	ZIP	ZIP	DMG	TAR.GZ
3.5.2 (2022-01-26)	ZIP	ZIP	DMG	TAR.GZ

See the full list of releases on GitHub for other versions.

用 OpenRefine 清資料

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

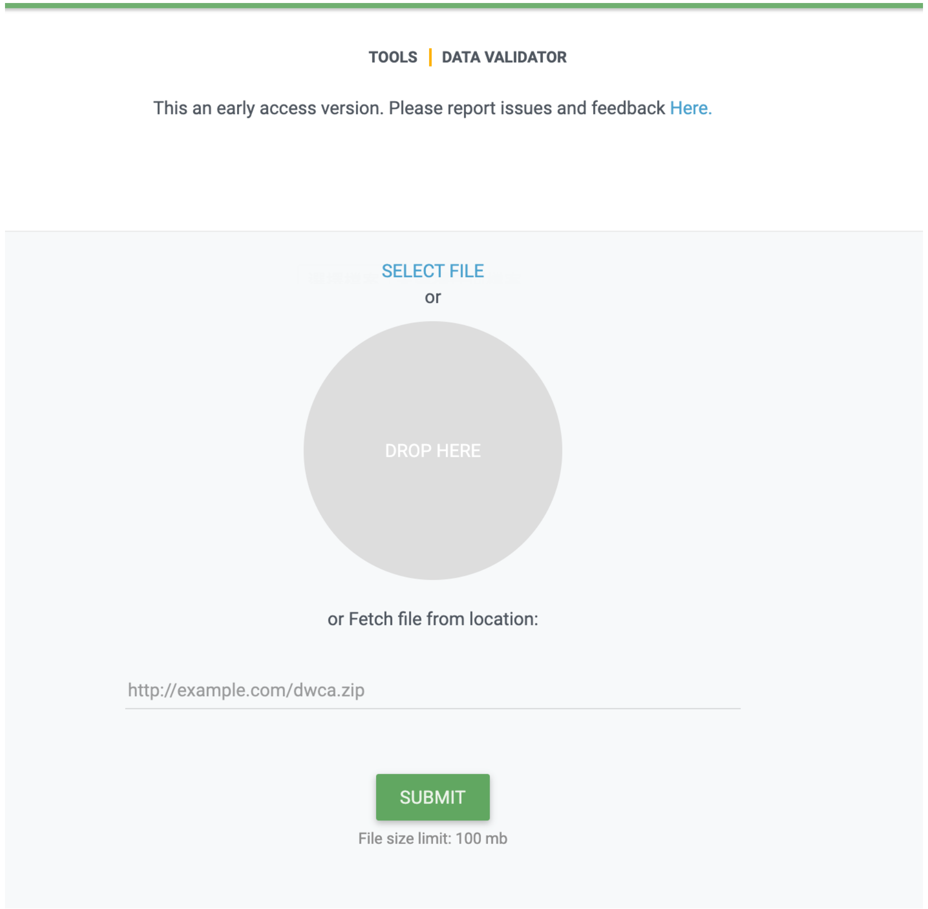


```
20:14:39.370 [            refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
20:14:39.418 [            refine_server] Initializing context: '/' from 'C:\Users\yuuku\Downloads\openrefine-3.7.5\webap
p' (48ms)
20:14:40.243 [            refine_server] Creating new workspace directory C:\Users\yuuku\AppData\Roaming\OpenRefine (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！

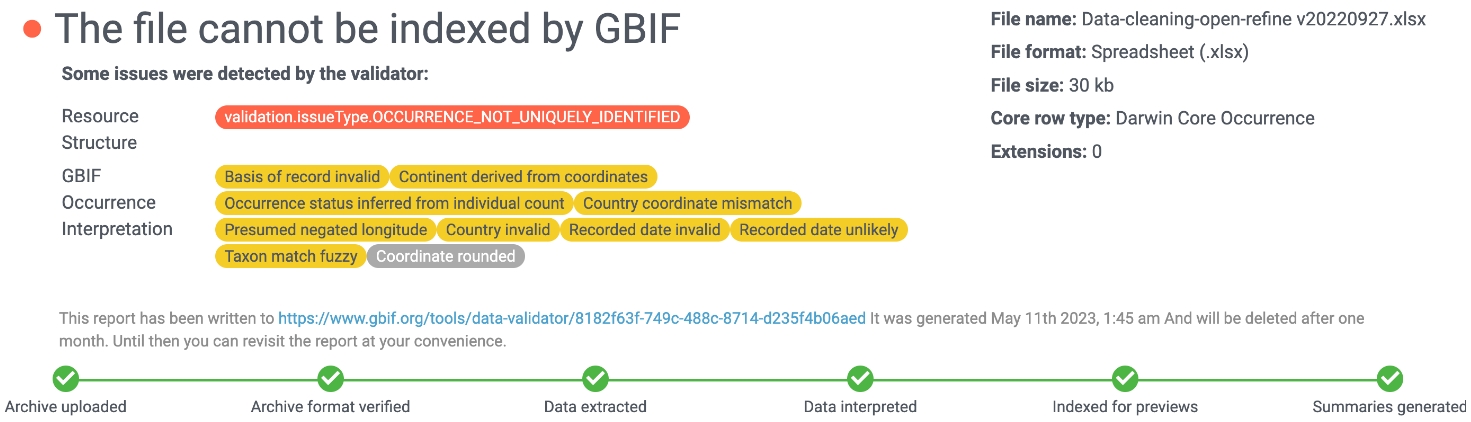
檢核資料—先找出可能的資料錯誤

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



TOOLS | DATA VALIDATOR

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



● 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

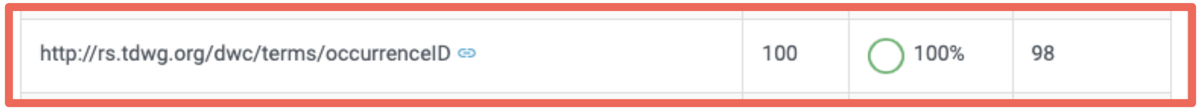
File name: Data-cleaning-open-refine v20220927.xlsx
File format: Spreadsheet (.xlsx)
File size: 30 kb
Core row type: Darwin Core Occurrence
Extensions: 0

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.

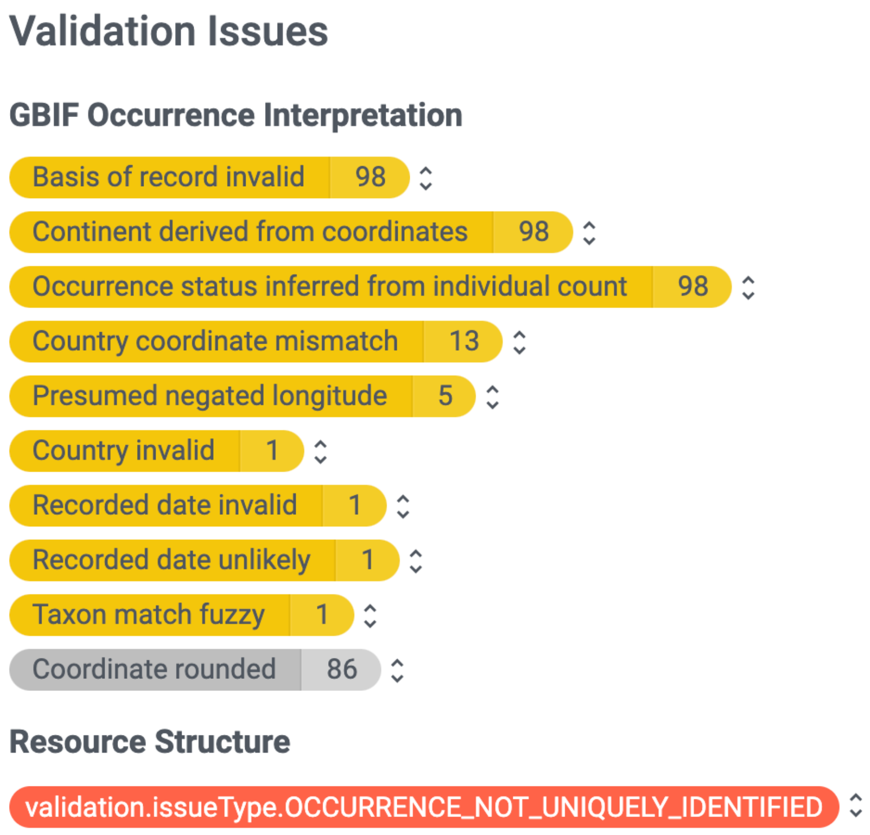
案例練習- 進階作業

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

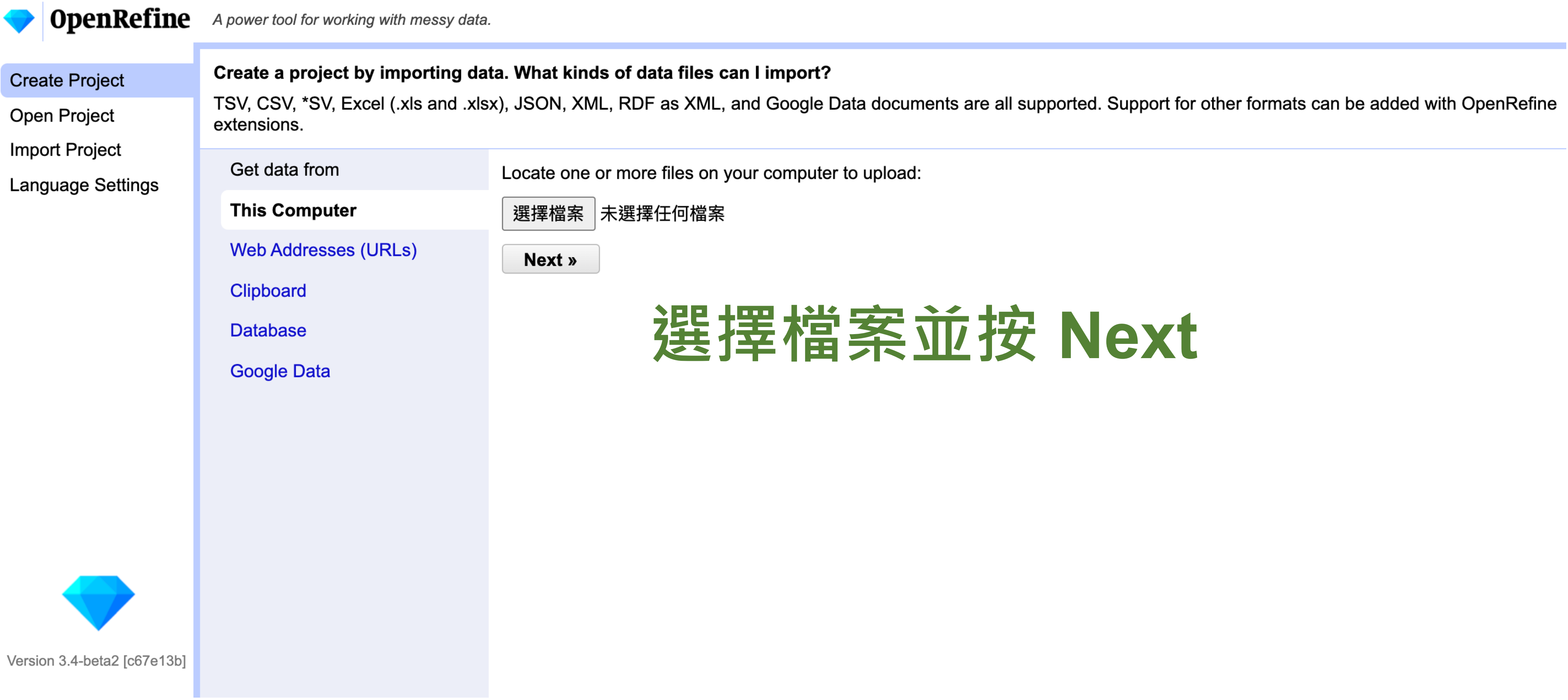
GBIF Occurrence Interpretation

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.*

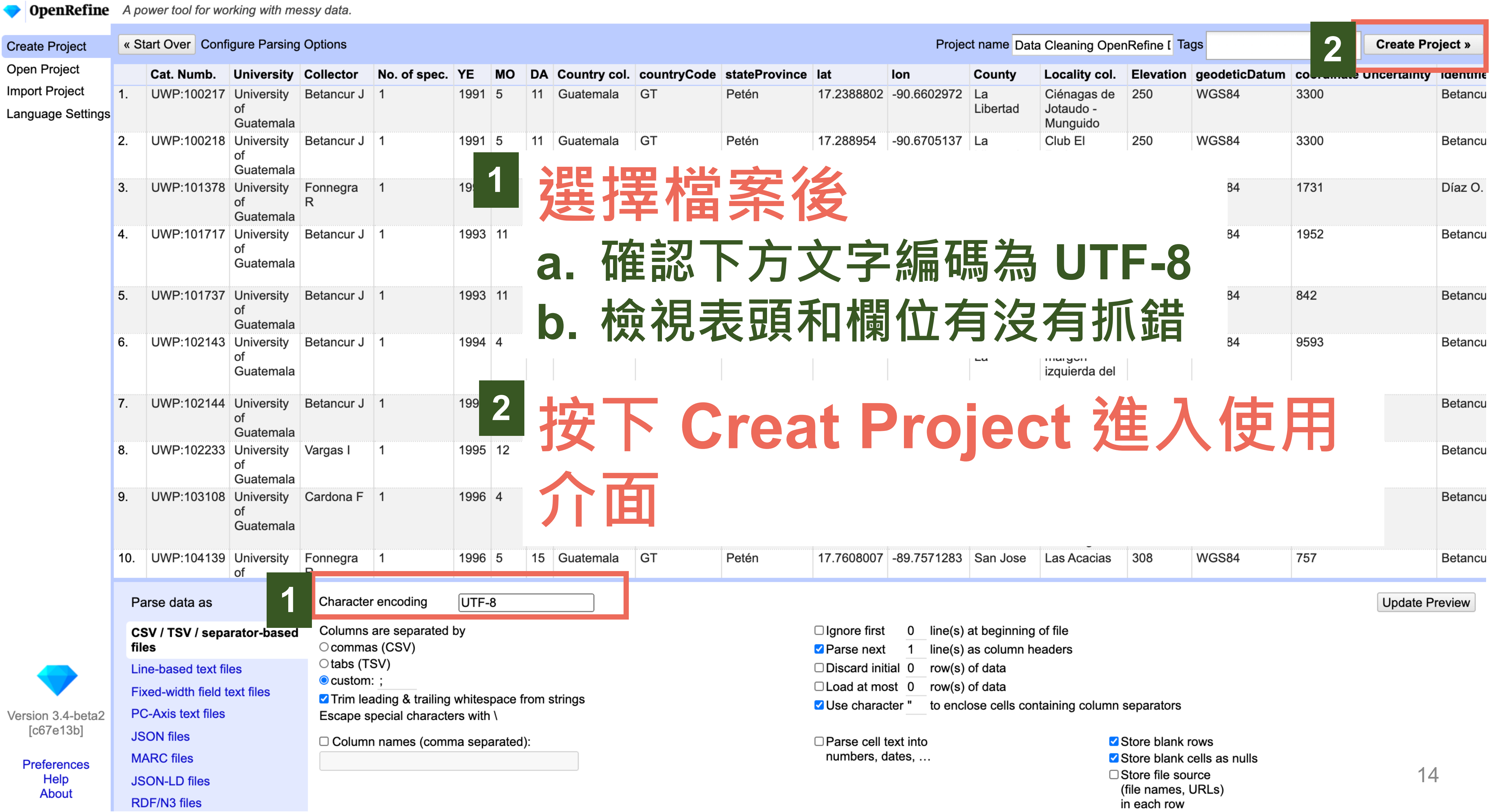
Create a project by importing data. What kinds of data files can I import?

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

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

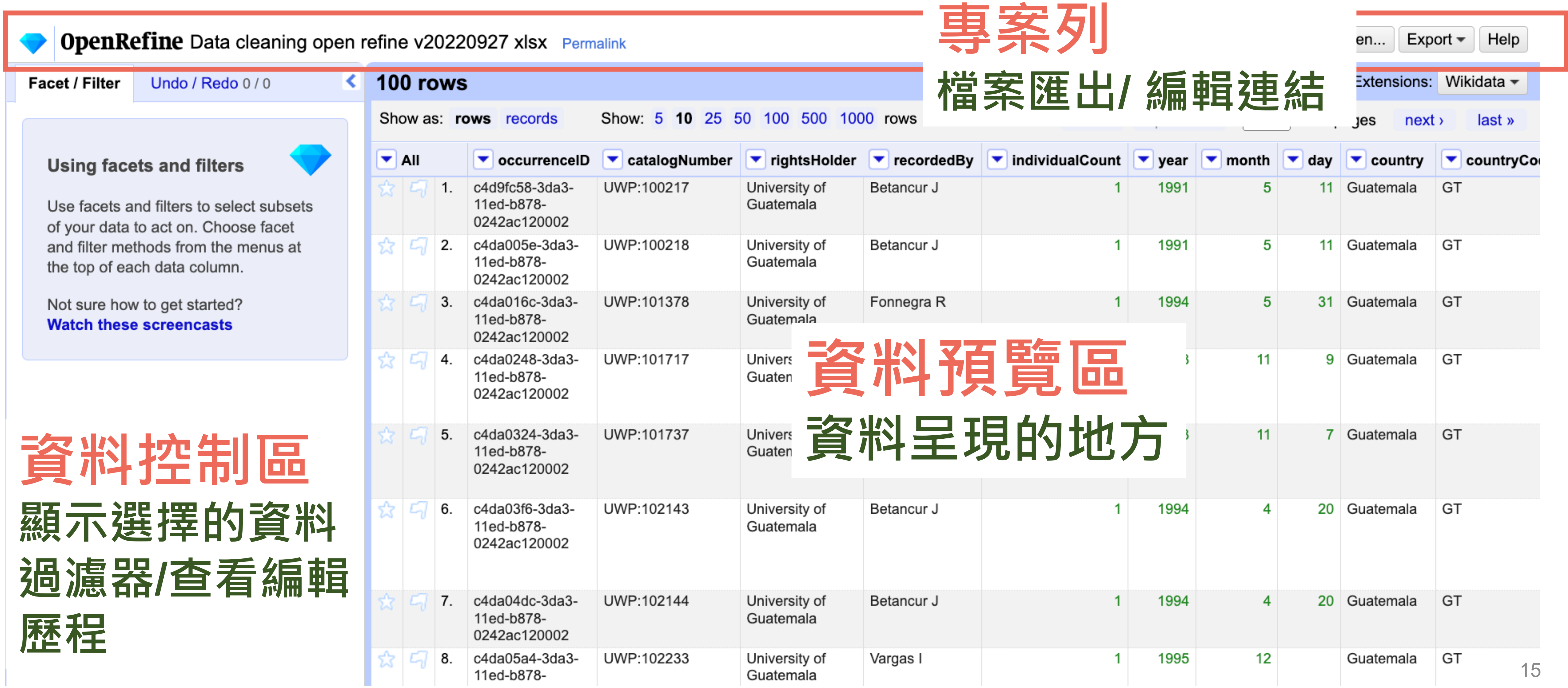
1 選擇檔案後

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 Creat Project 進入使用介面

用 OpenRefine 清資料



OpenRefine Data cleaning open refine v20220927 xlsx Permalink

en... Export Help

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			3	11	9	Guatemala	GT
5.	c4da0324-3da3-11ed-b878-0242ac120002	UWP:101737	Univers Guaten			3	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

專案列

檔案匯出/ 編輯連結

資料預覽區

資料呈現的地方

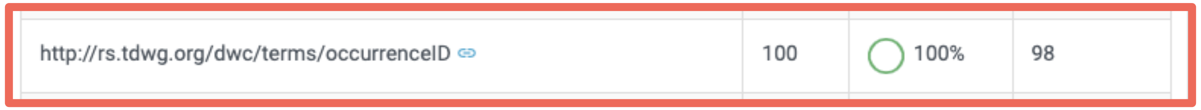
資料控制區

顯示選擇的資料過濾器/查看編輯歷程

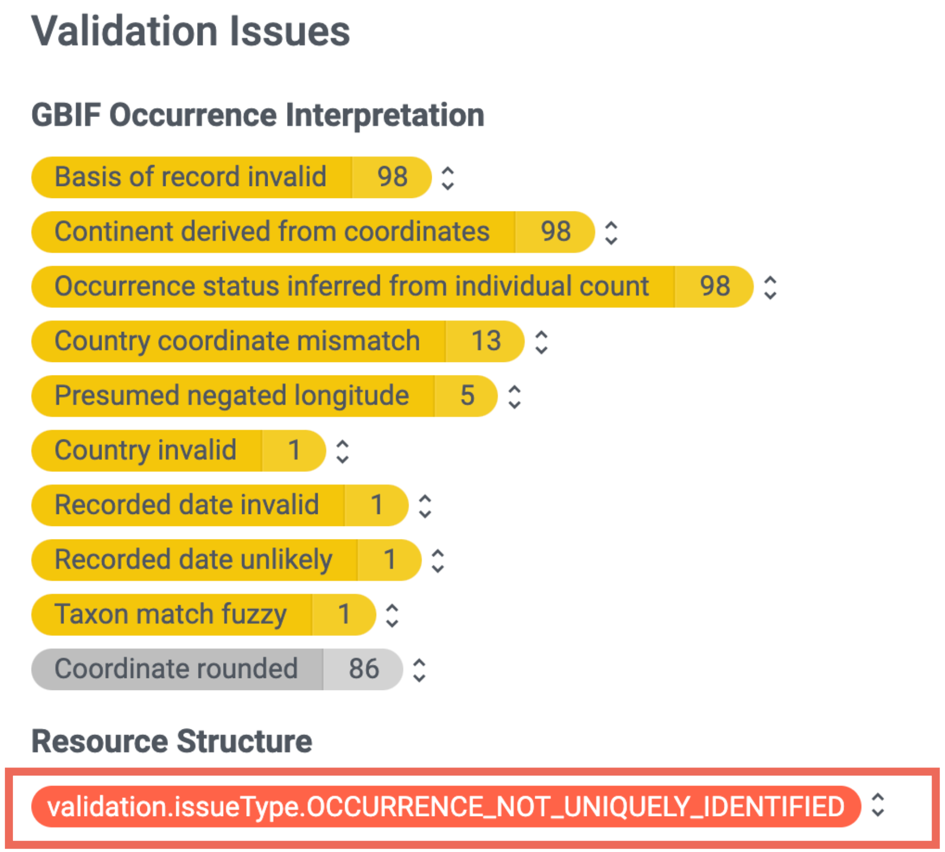
案例練習- 進階作業

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

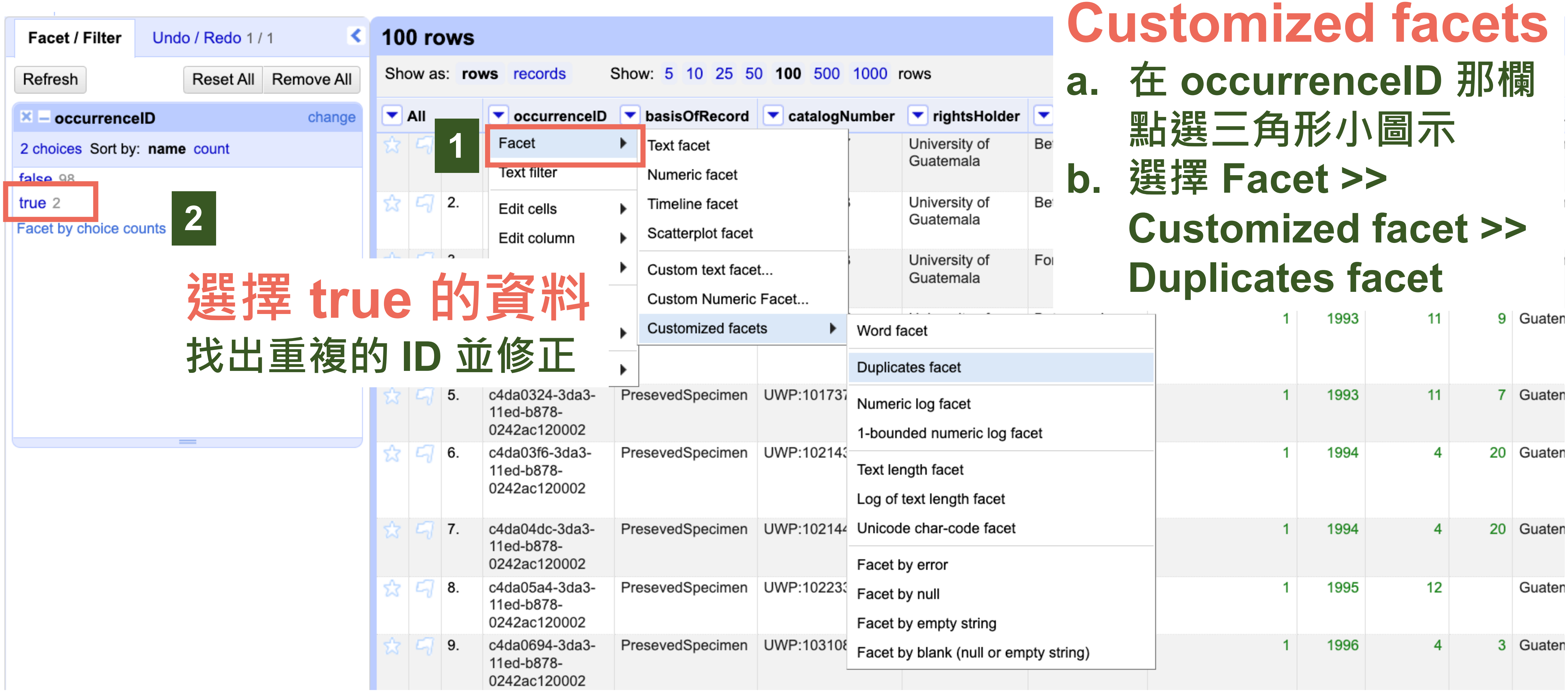
Customized facets

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



Facet / Filter | Undo / Redo 1 / 1

Refresh | Reset All | Remove All

occurrenceID change

2 choices Sort by: name count

false 98

true 2

Facet by choice counts

100 rows

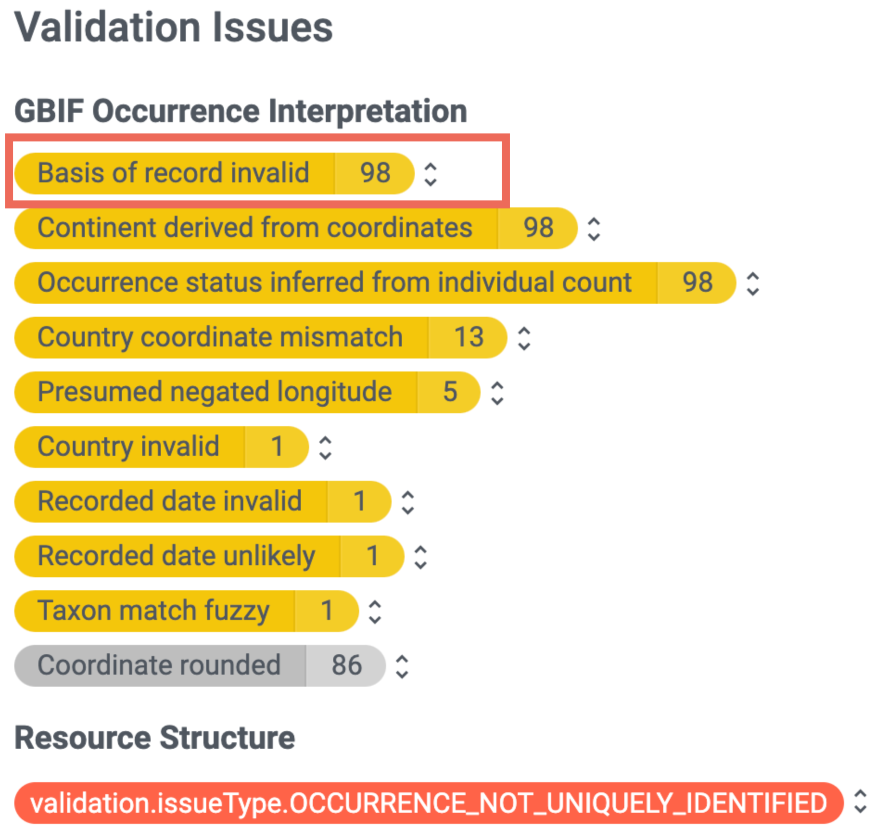
Show as: rows records Show: 5 10 25 50 100 500 1000 rows

All	occurrenceID	basisOfRecord	catalogNumber	rightsHolder						
1.				University of Guatemala	Be					
2.				University of Guatemala	Be					
				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

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

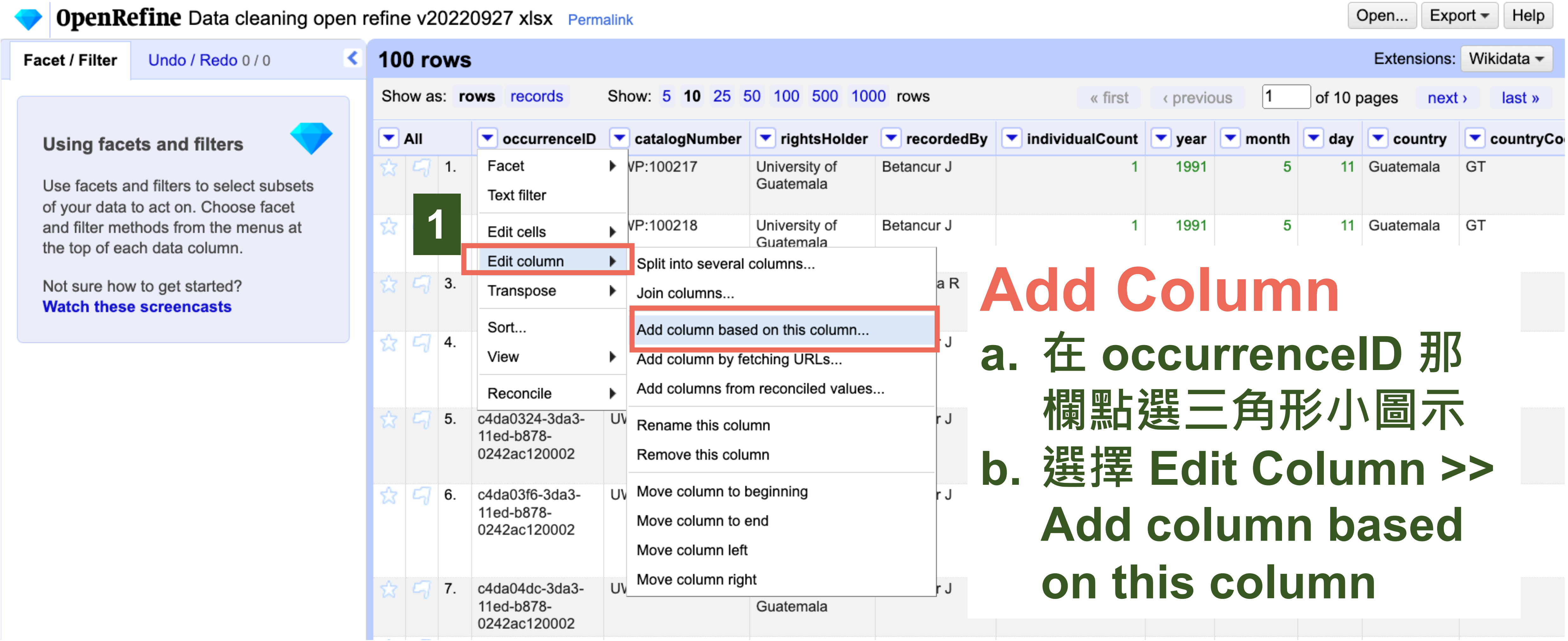
GBIF Occurrence Interpretation

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



OpenRefine Data cleaning open refine v20220927 xlsx Permalink

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 column.

Not sure how to get started?
Watch these screencasts

100 rows

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

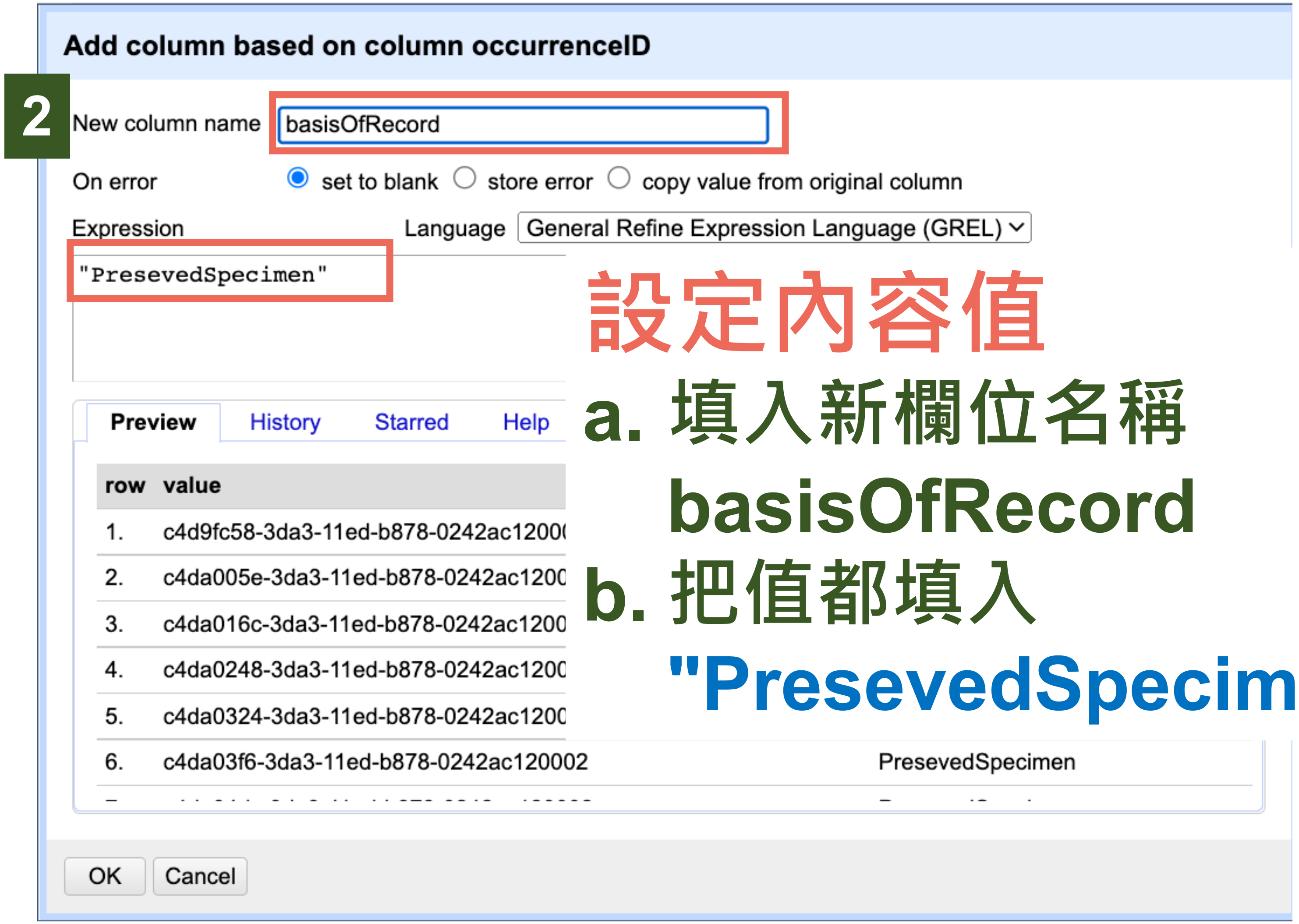
All	occurrenceID	catalogNumber	rightsHolder	recordedBy	individualCount	year	month	day	country	countryCo
1.		VP:100217	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT
2.		VP:100218	University of Guatemala	Betancur J	1	1991	5	11	Guatemala	GT
3.				a R						
4.				r J						
5.	c4da0324-3da3-11ed-b878-0242ac120002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

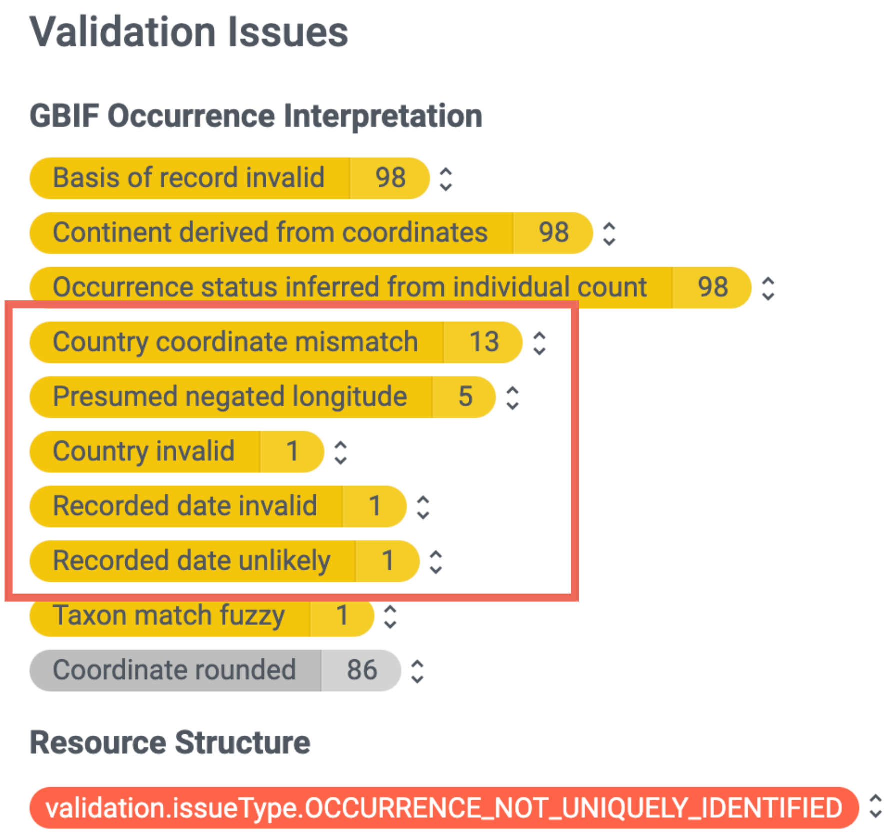
a. 填入新欄位名稱
basisOfRecord

b. 把值都填入
"PresevedSpecimen"

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

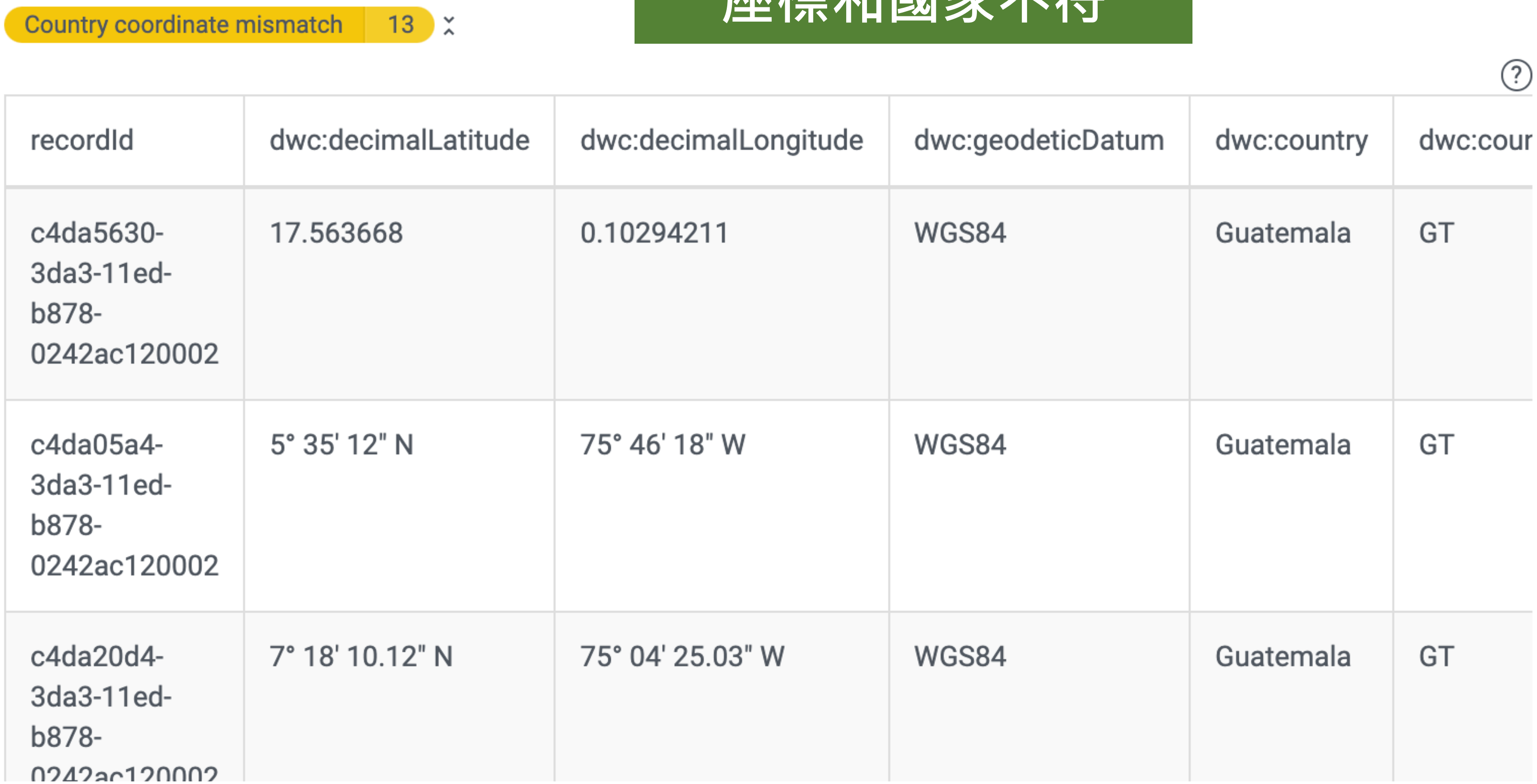
Resource Structure

validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

案例練習- 進階作業

資料問題

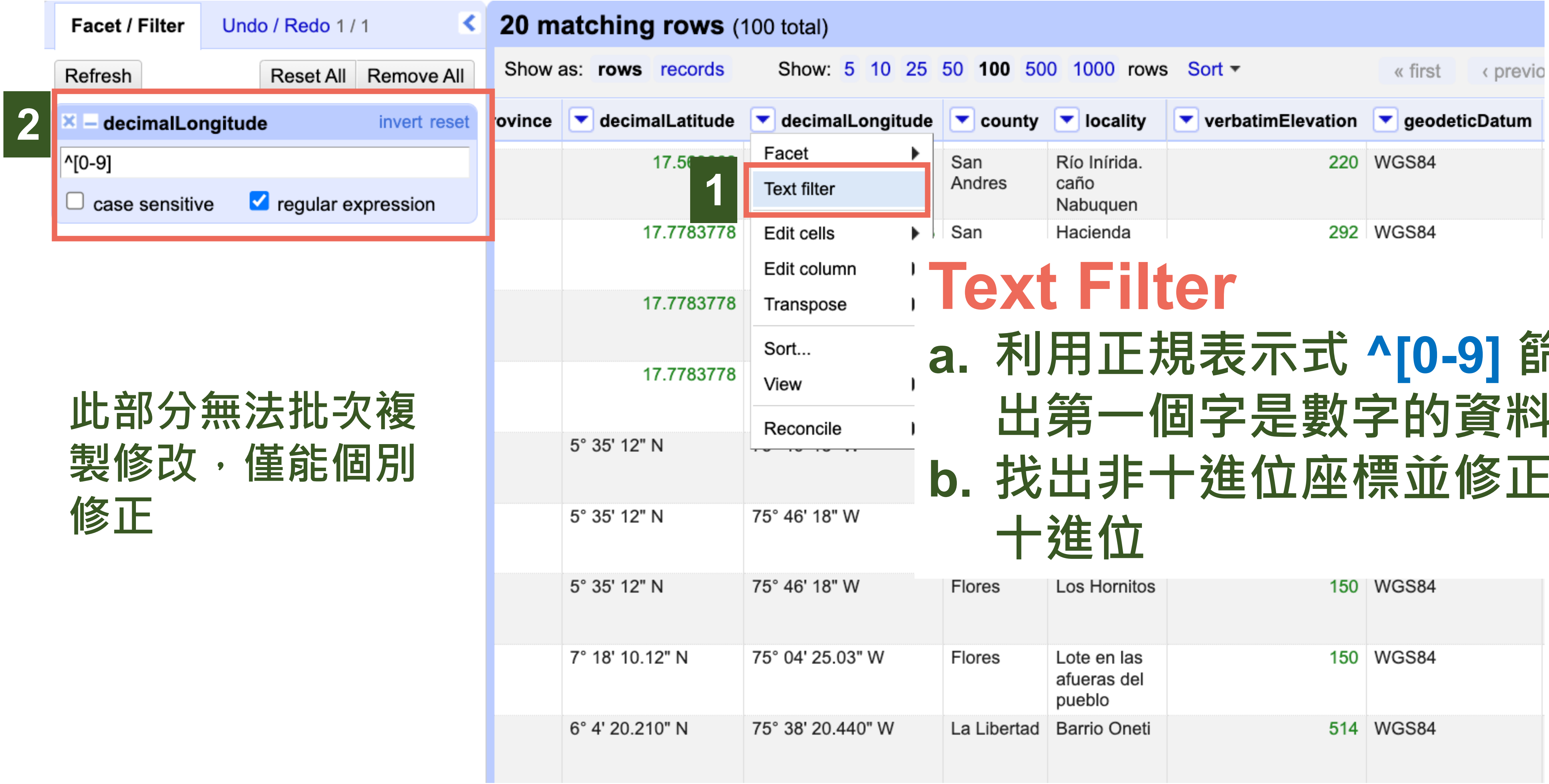
座標和國家不符



Country coordinate mismatch 13

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)



decimalLongitude

^[0-9]

☐ case sensitive ☑ regular expression

20 matching rows (100 total)

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

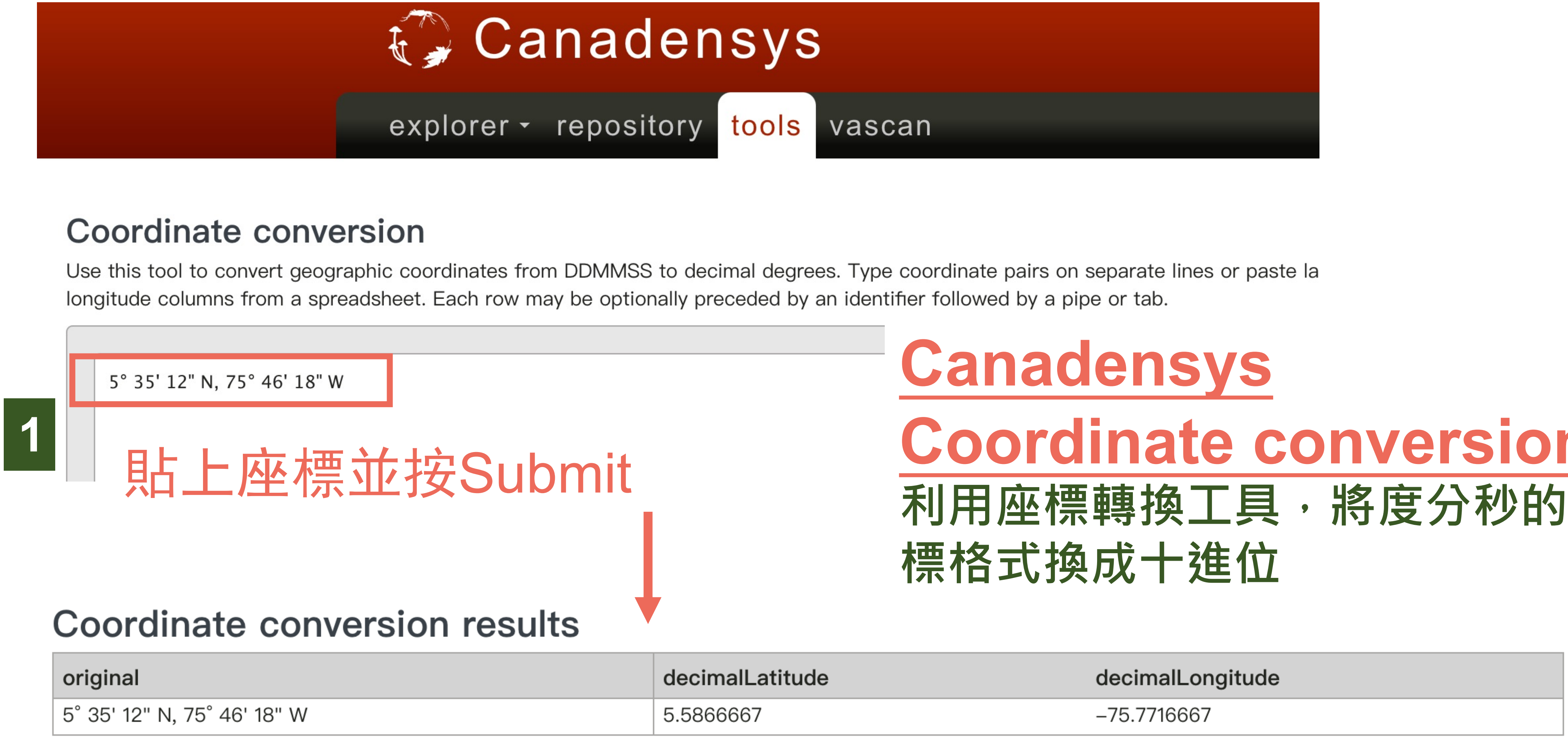
此部分無法批次複製修改，僅能個別修正

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

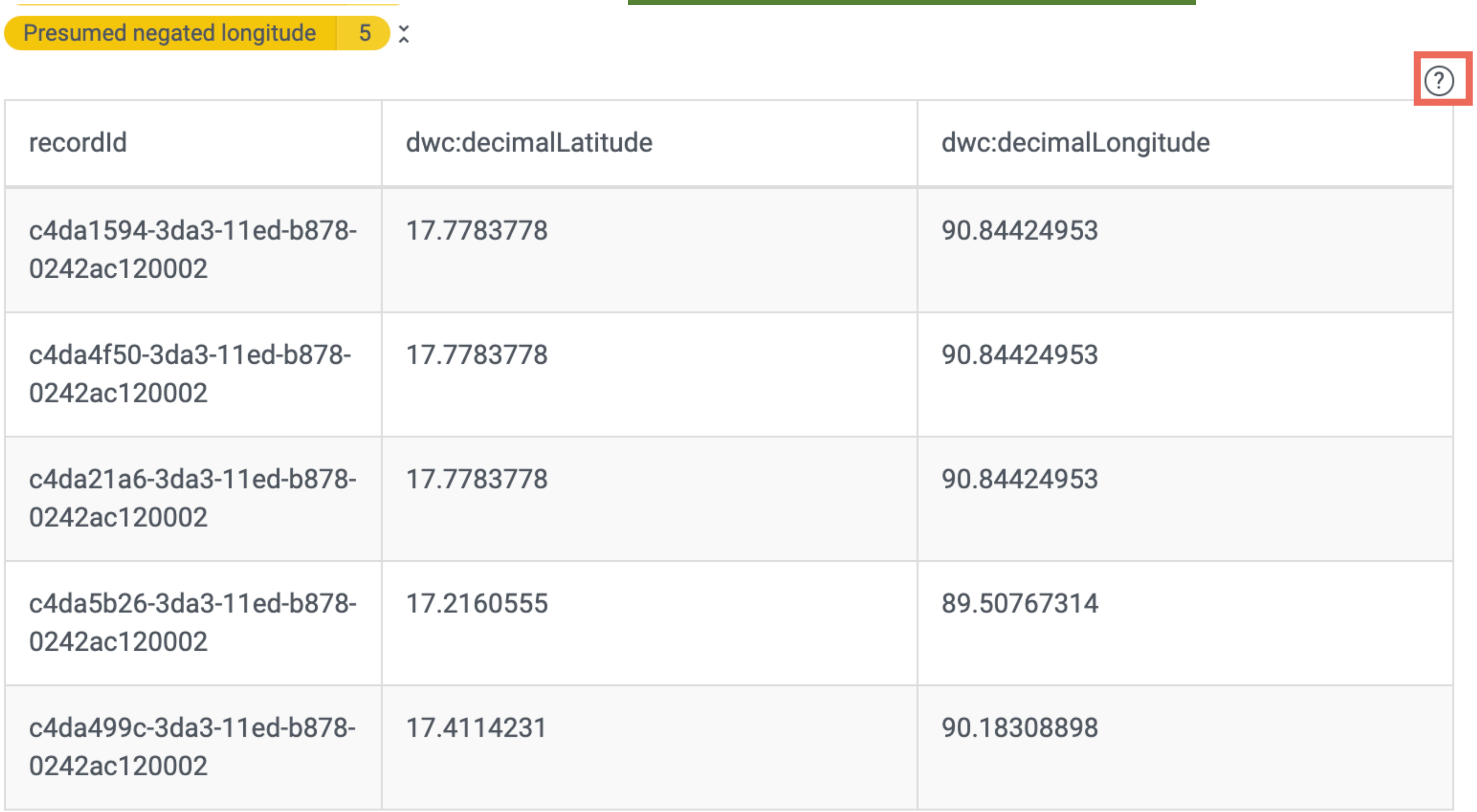
Canadensys Coordinate conversion

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

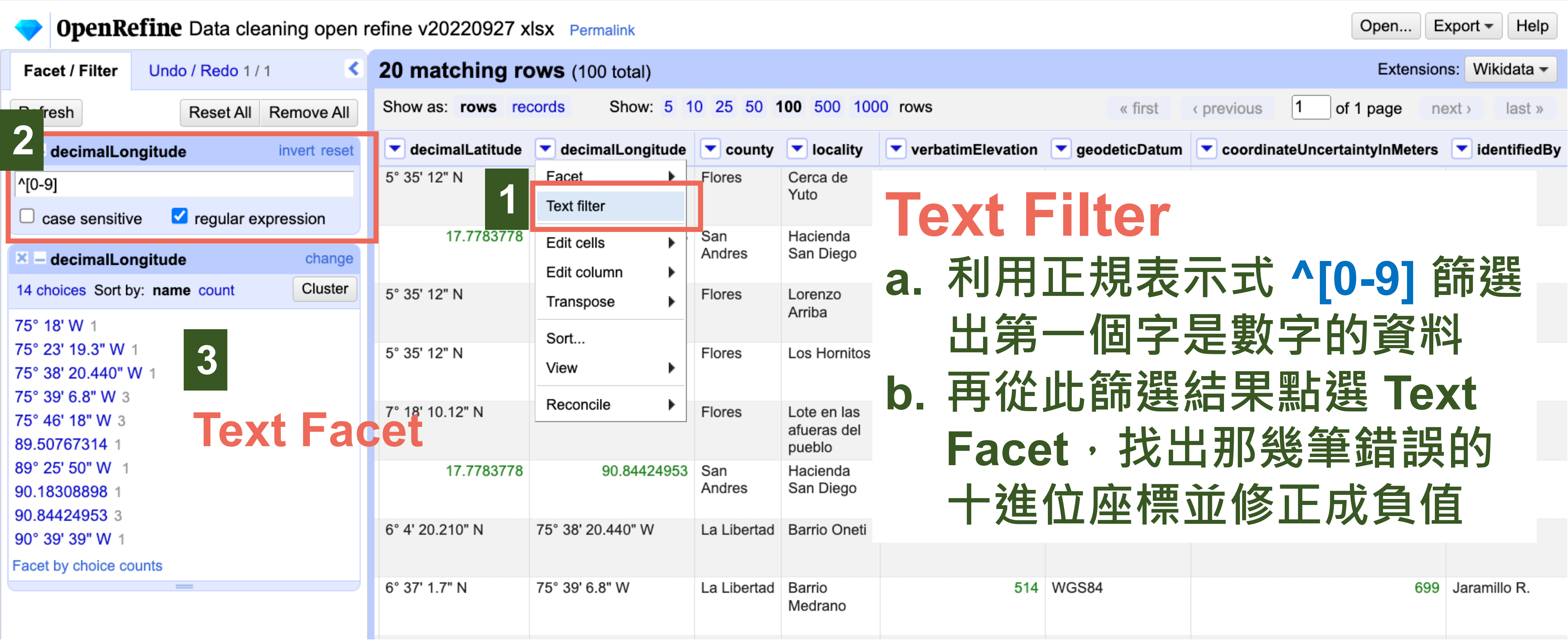
推定經度應為負值



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)



20 matching rows (100 total)

2 decimalLongitude invert reset

^[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

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

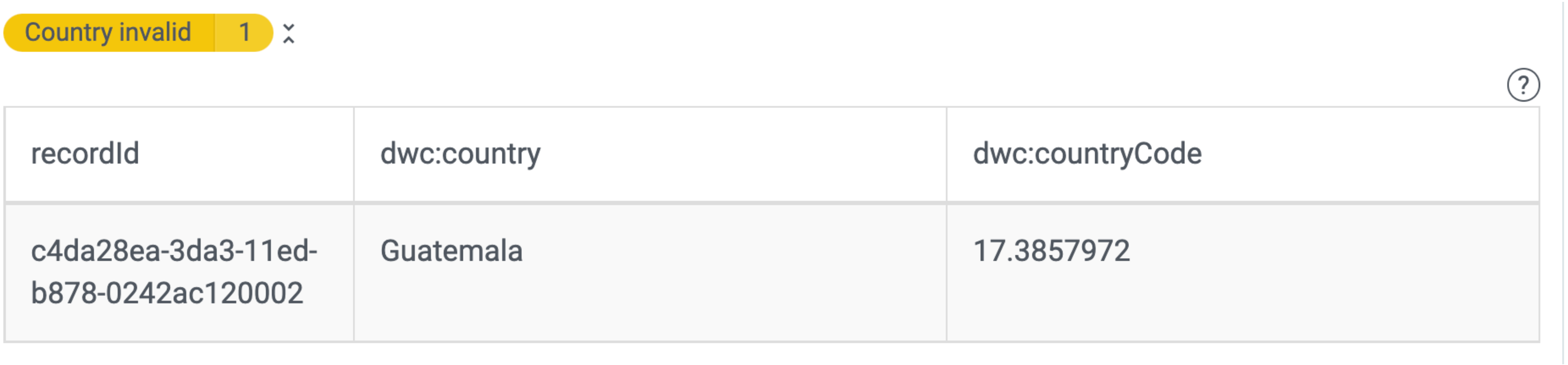
a. 利用正規表示式 ^[0-9] 篩選出第一個字是數字的資料

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

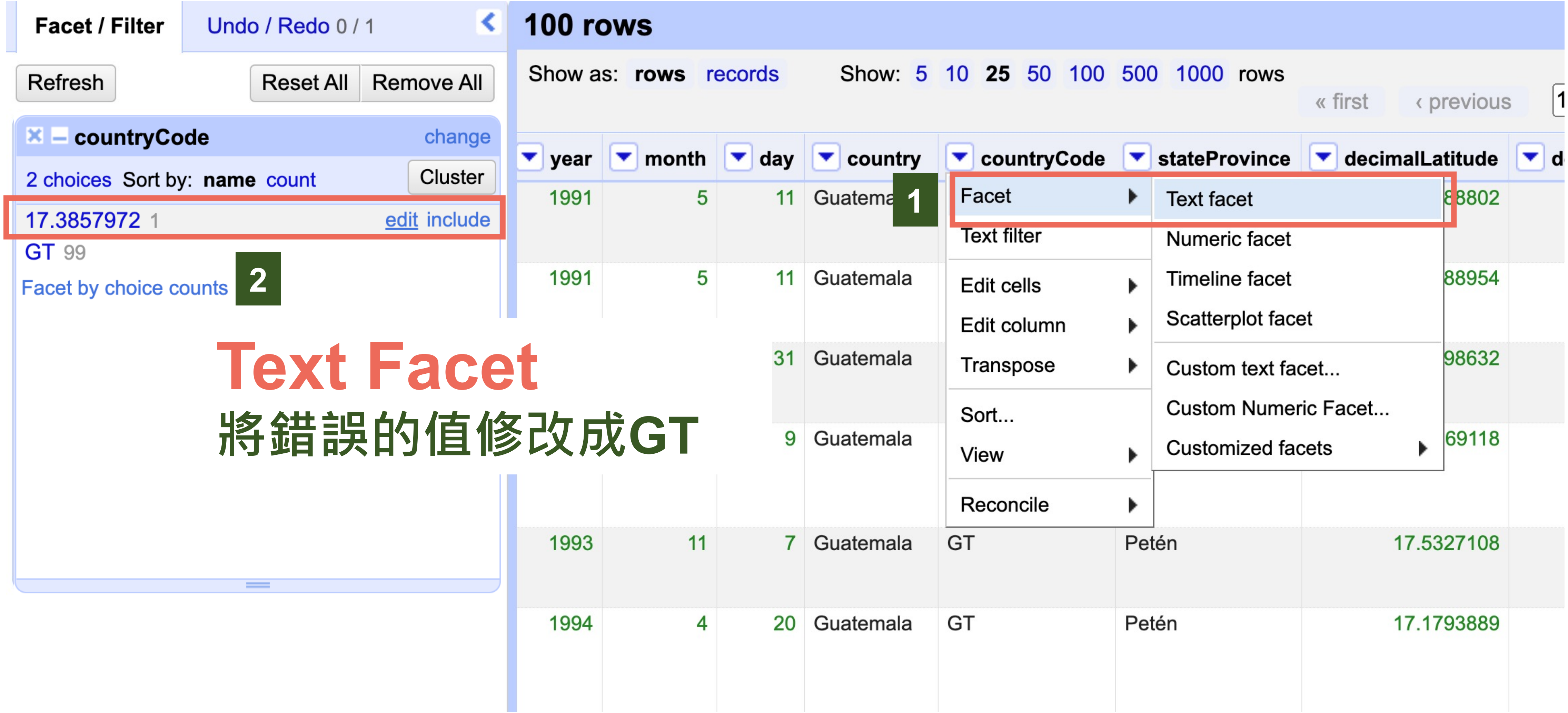
國家代碼無效



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

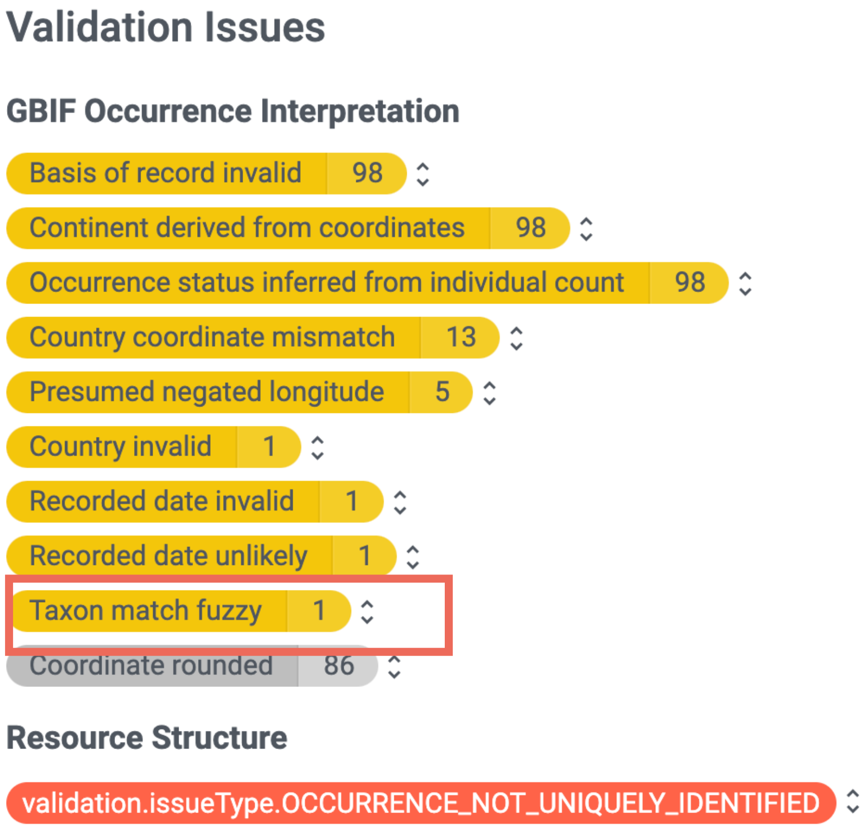
Text Facet

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

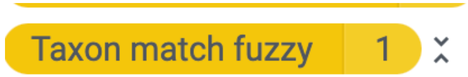
GBIF Occurrence Interpretation

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

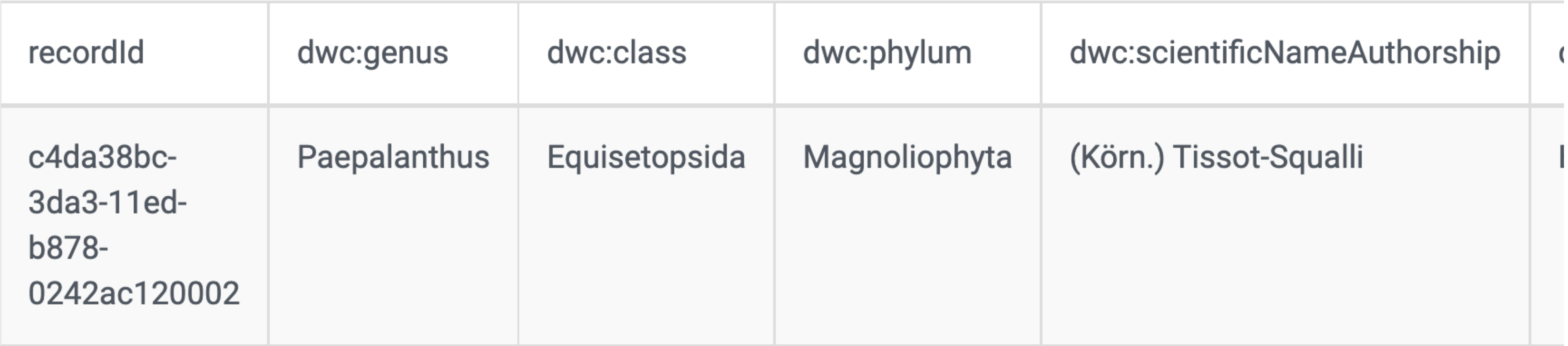
validation.issueType.OCCURRENCE_NOT_UNIQUELY_IDENTIFIED

用 OpenRefine 清資料- 學名比對



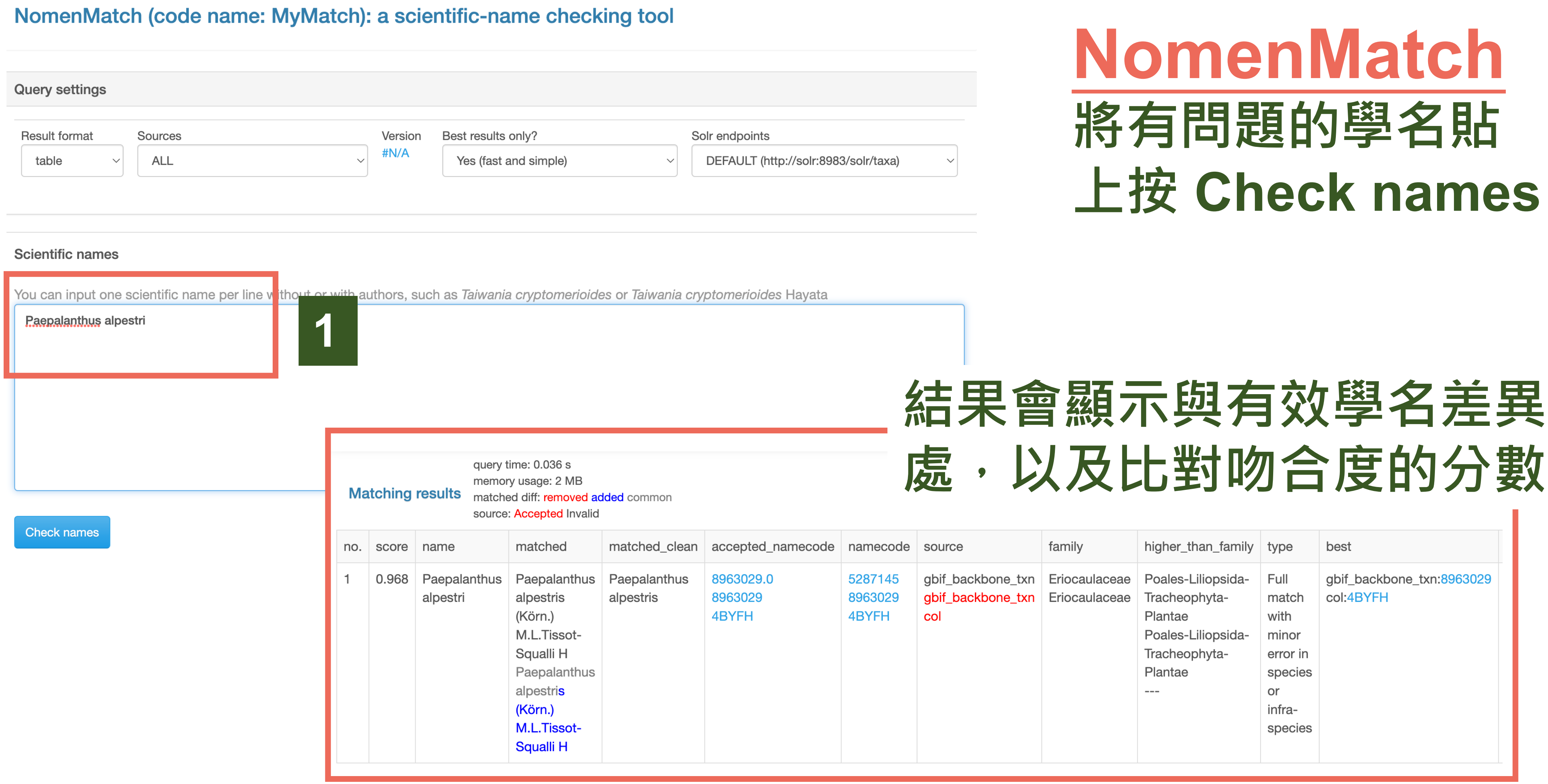
Taxon match fuzzy 1

分類未對應 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
```

1

Matching results

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

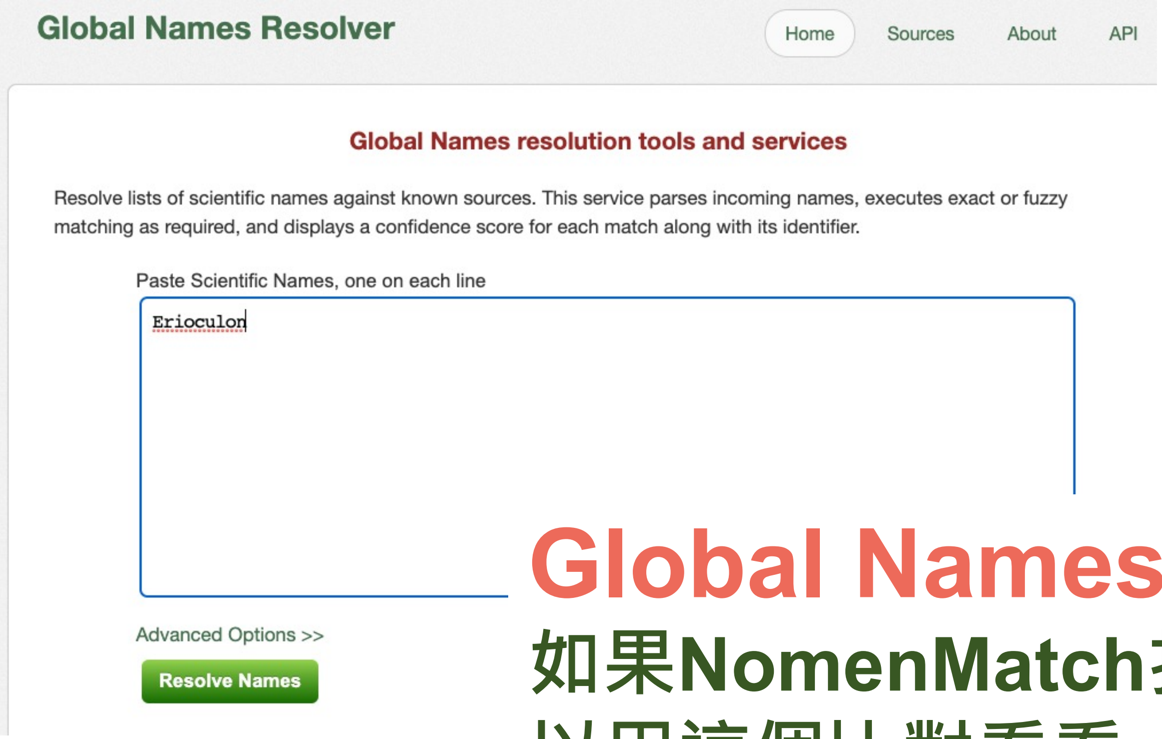
no.	score	name	matched	matched_clean	accepted_namecode	namecode	source	family	higher_than_family	type	best
1	0.968	Paepalanthus alpestri	Paepalanthus alpestris (Körn.) M.L.Tissot-Squalli H Paepalanthus alpestris (Körn.) M.L.Tissot-Squalli H	Paepalanthus alpestris	8963029.0 8963029 4BYFH	5287145 8963029 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

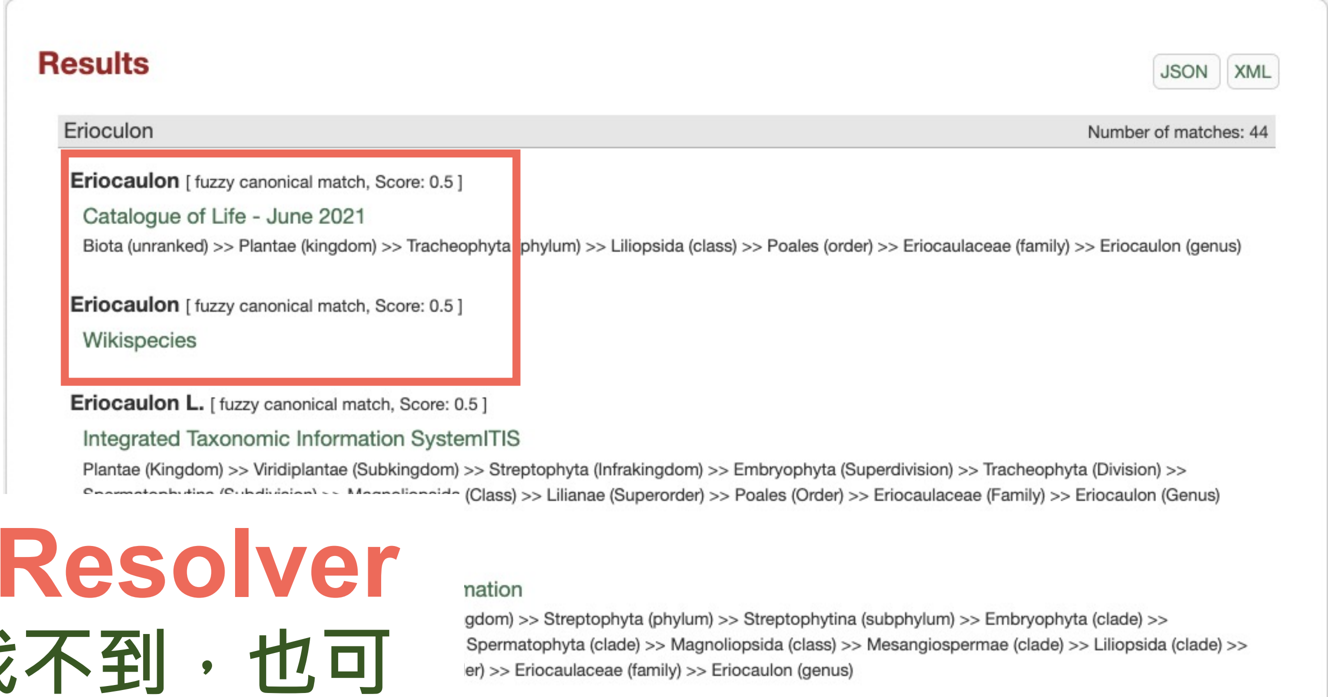


Global Names resolution tools and services

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



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) >> (Class) >> Lilianae (Superorder) >> Poales (Order) >> Eriocaulaceae (Family) >> Eriocaulon (Genus)

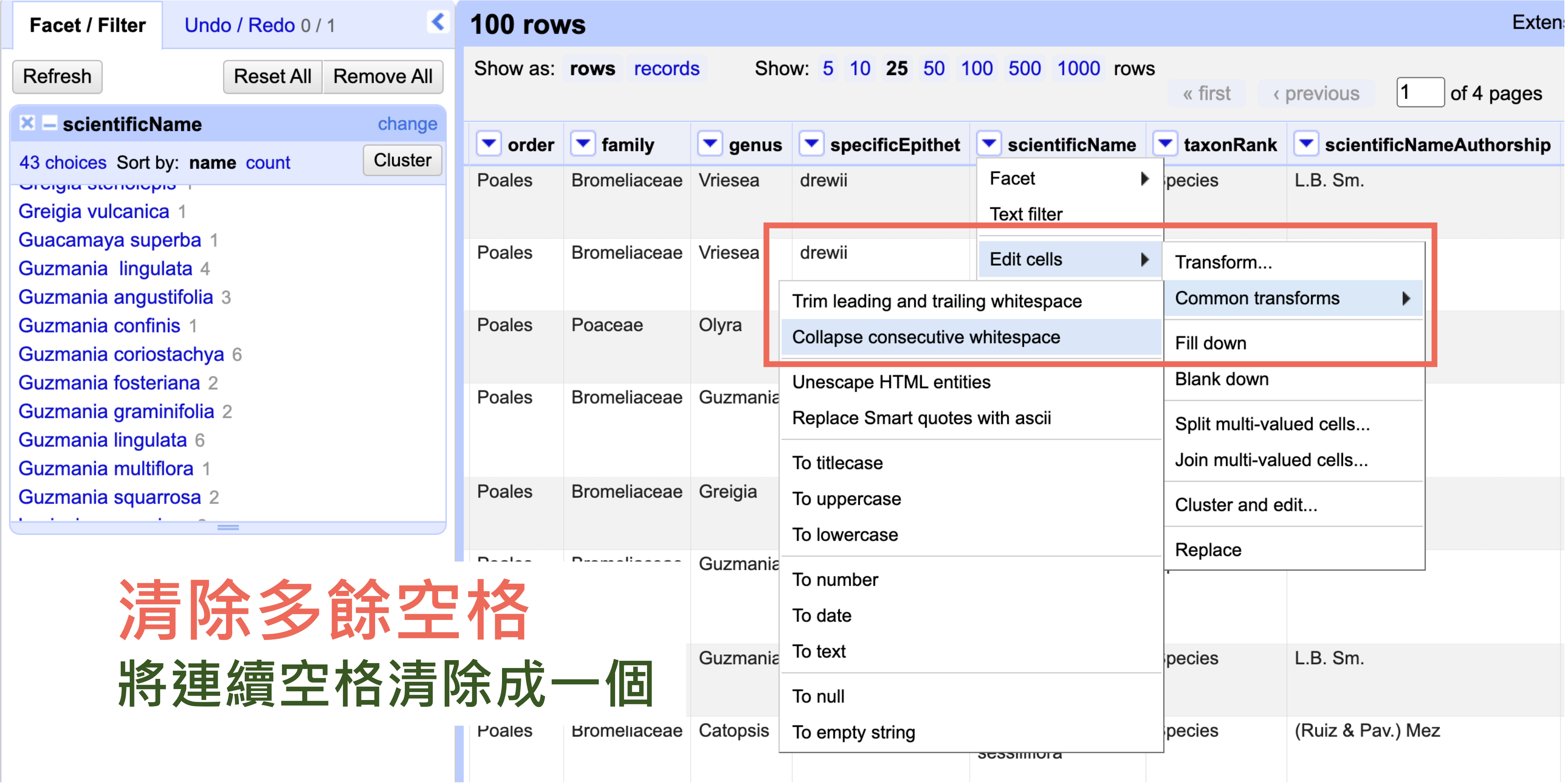
nation

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



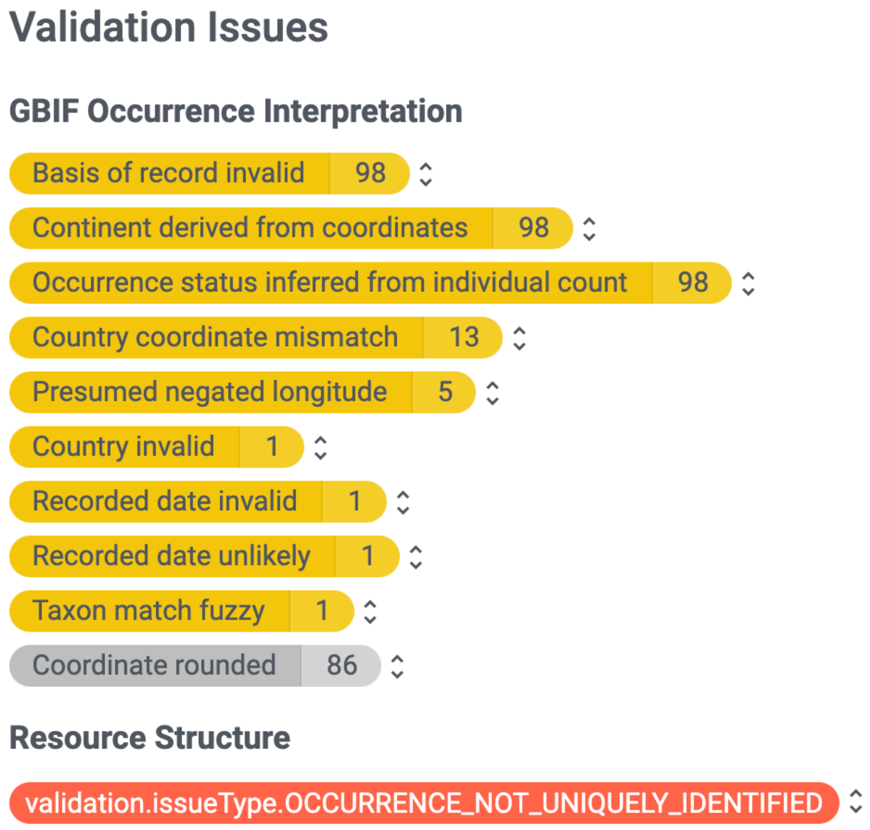
清除多餘空格

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

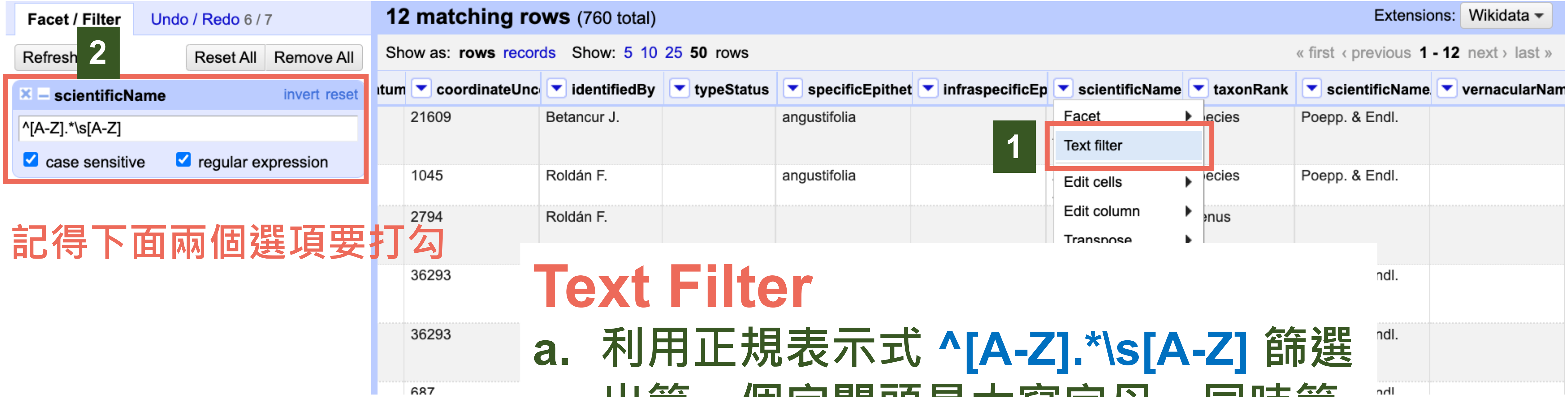
GBIF Occurrence Interpretation

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



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

scientificName

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

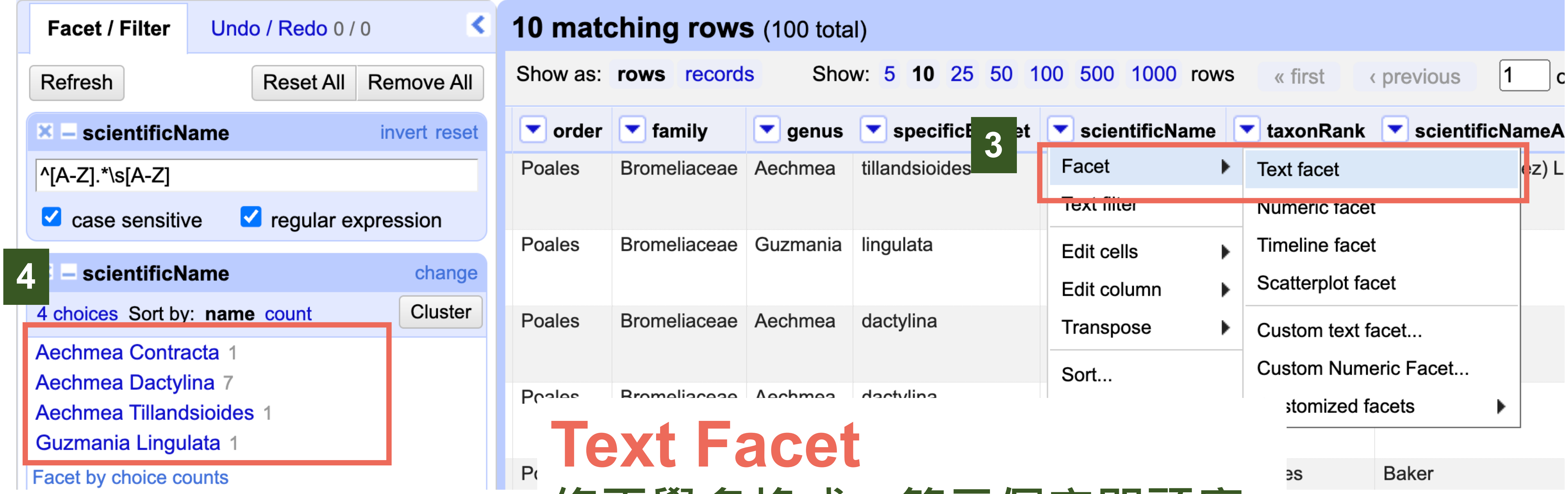
☑ case sensitive ☑ regular expression

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



Facet / Filter　Undo / Redo 0 / 0

Refresh　Reset All　Remove All

☒ scientificName　invert reset

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

☑ case sensitive　☑ regular expression

4 scientificName　change

4 choices Sort by: name count　Cluster

Aechmea Contracta 1
Aechmea Dactylina 7
Aechmea Tillandsioides 1
Guzmania Lingulata 1

Facet by choice counts

10 matching rows (100 total)

Show as: rows records　Show: 5 10 25 50 100 500 1000 rows　« first　‹ previous　1

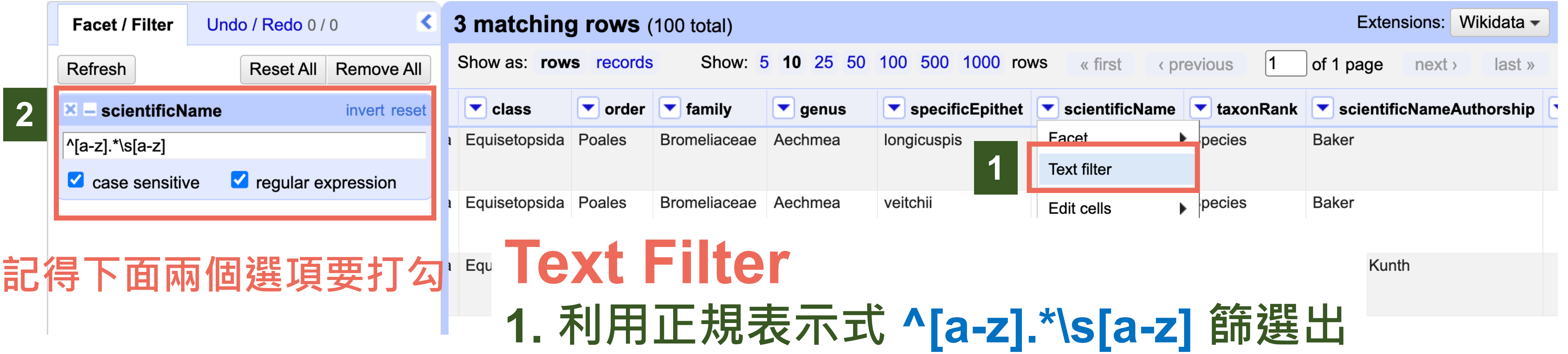
order	family	genus	specificEpithet	scientificName	taxonRank	scientificNameA
Poales	Bromeliaceae	Aechmea	tillandsioides			z) L
Poales	Bromeliaceae	Guzmania	lingulata			
Poales	Bromeliaceae	Aechmea	dactylina			
Poales	Bromeliaceae	Aechmea	dactylina			
Po					es	Baker

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



Facet / Filter　Undo / Redo 0 / 0

Refresh　Reset All　Remove All

2

scientificName　invert　reset

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

☑ case sensitive　☑ regular expression

3 matching rows (100 total)　Extensions: Wikidata

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

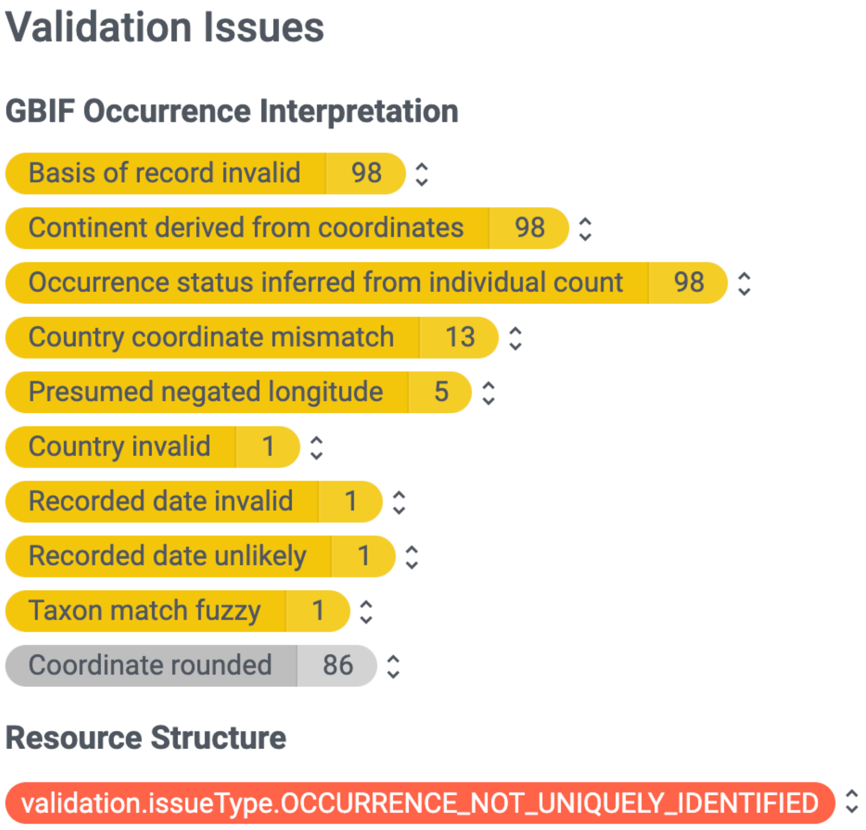
Text Filter

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

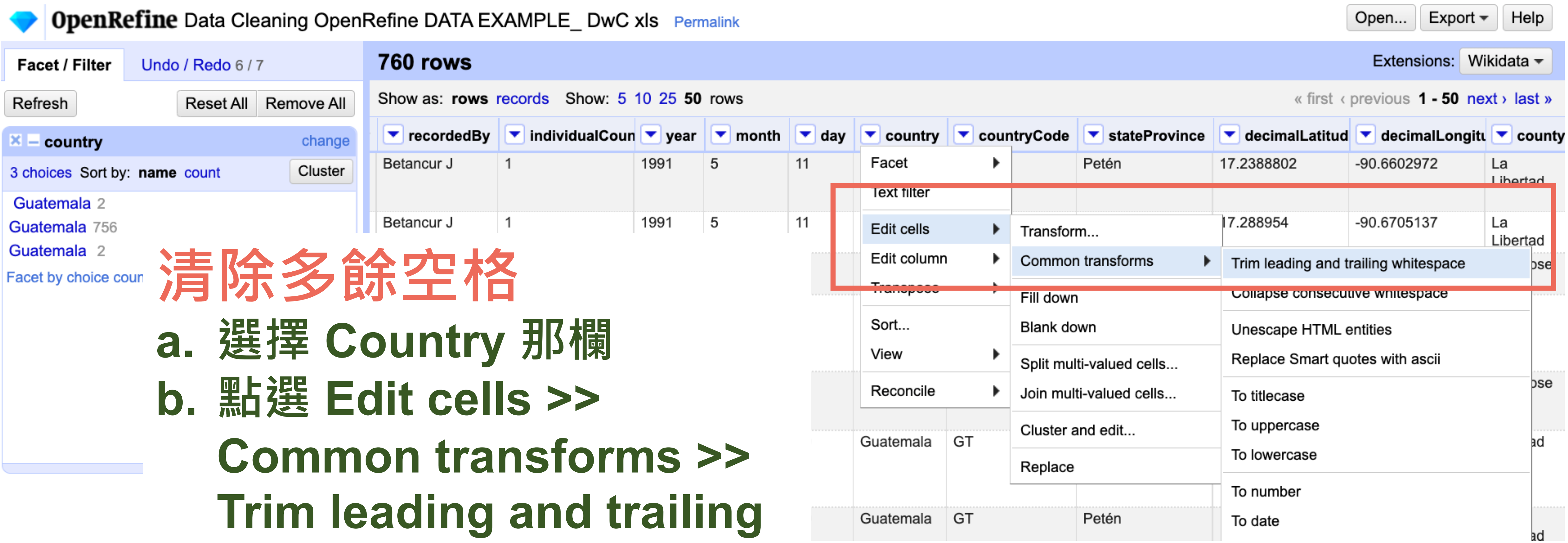
GBIF Occurrence Interpretation

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



recordedBy	individualCoun	year	month	day	country	countryCode	stateProvince	decimalLatitud	decimalLongitu	county
Betancur J	1	1991	5	11			Petén	17.2388802	-90.6602972	La Libertad
Betancur J	1	1991	5	11				17.288954	-90.6705137	La Libertad
					Guatemala	GT				
					Guatemala	GT	Petén			

清除多餘空格

a. 選擇 Country 那欄

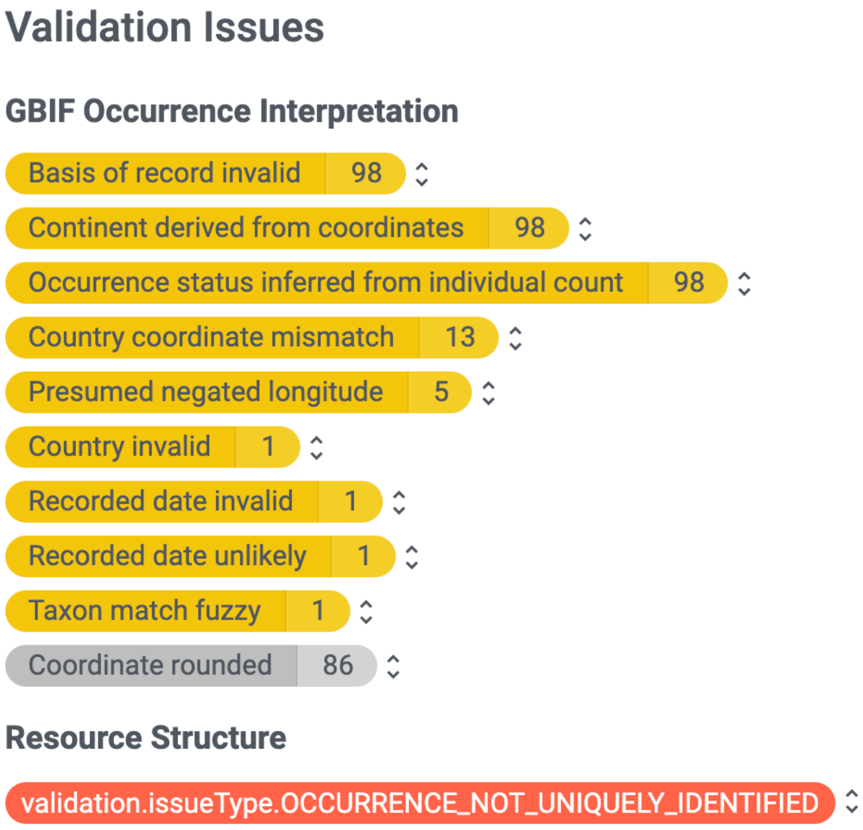
b. 點選 Edit cells >> Common transforms >> Trim leading and trailing whitespace

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

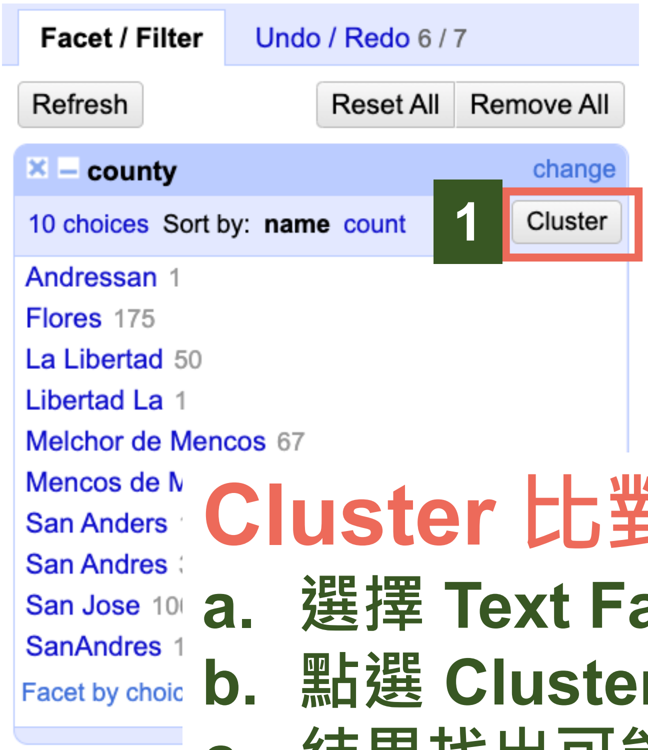
GBIF Occurrence Interpretation

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

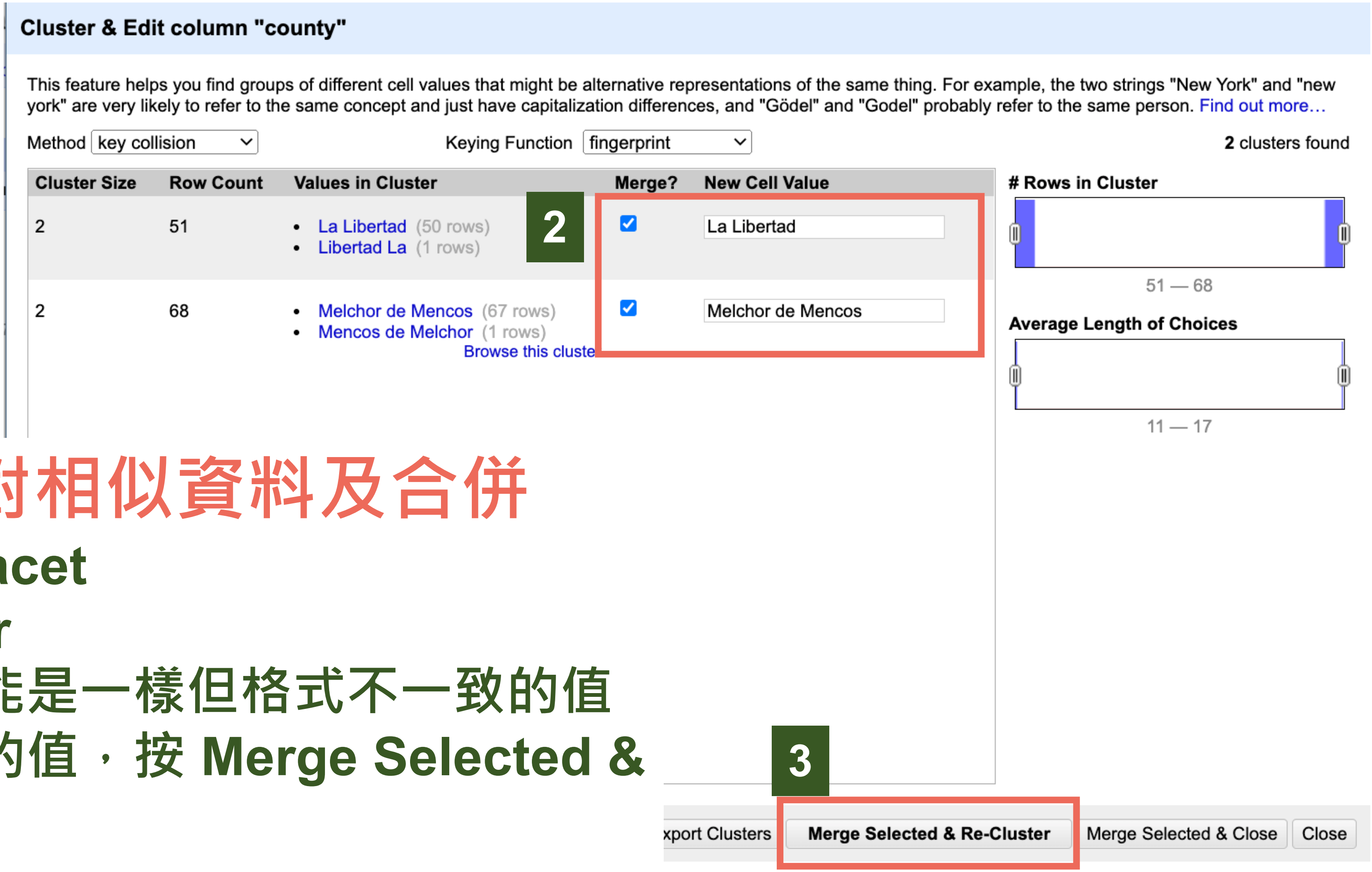
用 OpenRefine 清資料- 統一資料格式



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

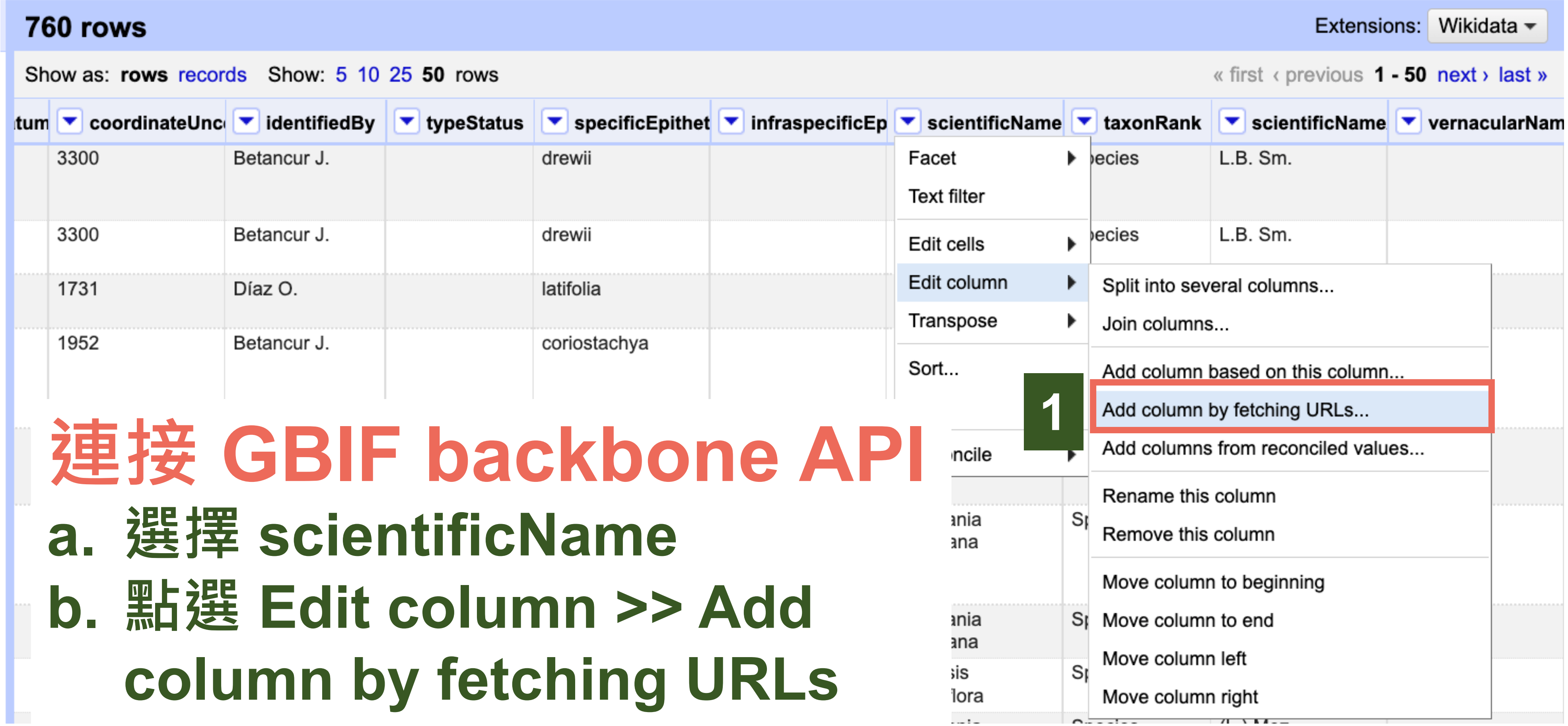
Average Length of Choices

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

« first ‹ previous 1 - 50 next › last »

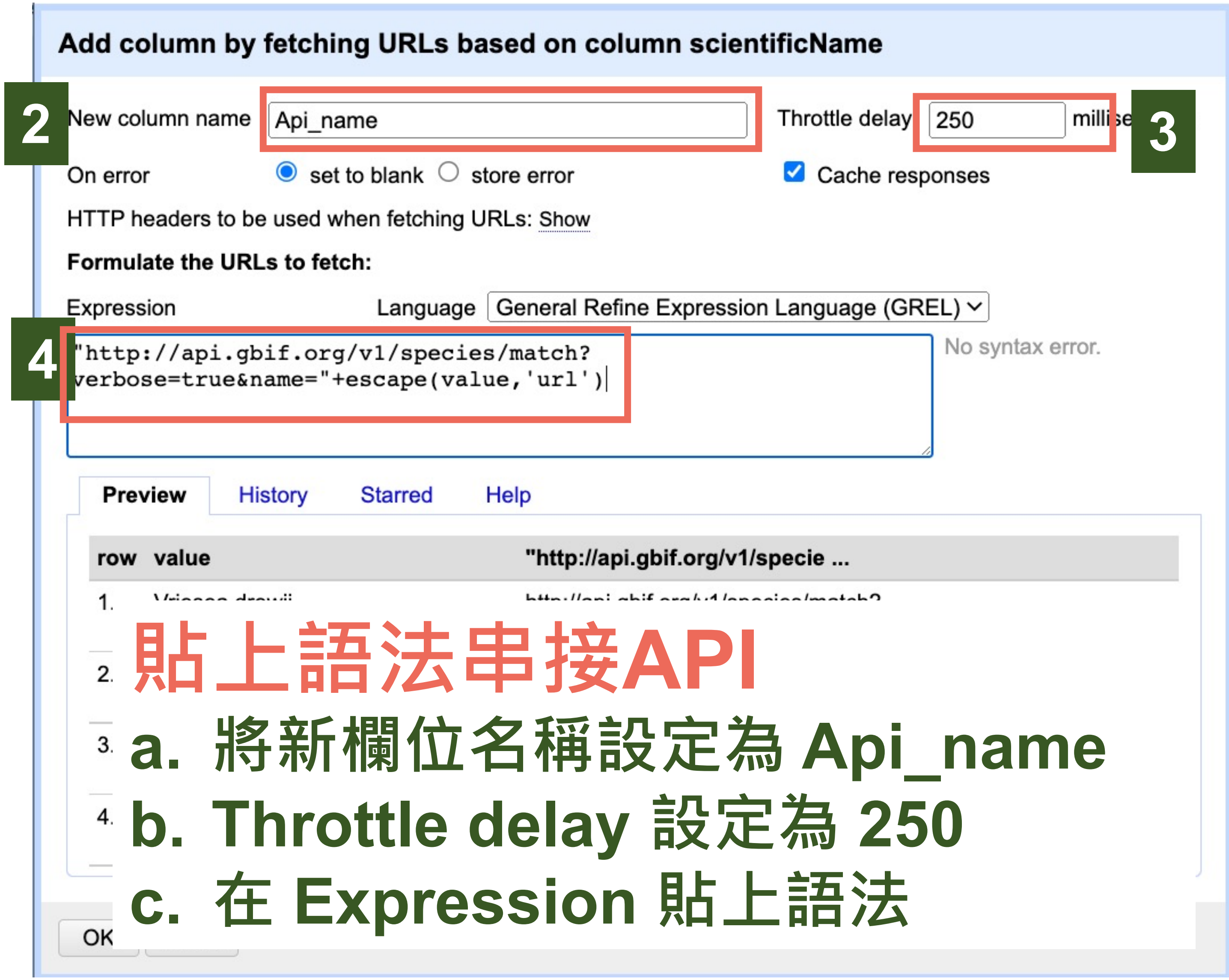
tum	coordinateUnc	identifiedBy	typeStatus	specificEpithet	infraspecificEp	scientificName	taxonRank	scientificName	vernacularNam
	3300	Betancur J.		drewii			ecies	L.B. Sm.	
	3300	Betancur J.		drewii			ecies	L.B. Sm.	
	1731	Díaz O.		latifolia					
	1952	Betancur J.		coriostachya					

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.

Preview | History | Starred | Help

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

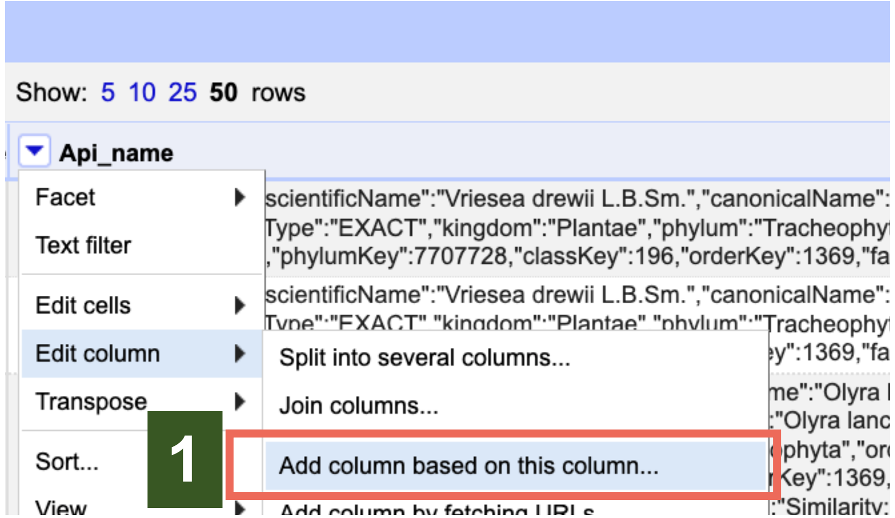
進階題-自動匯入高階層分類欄位

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

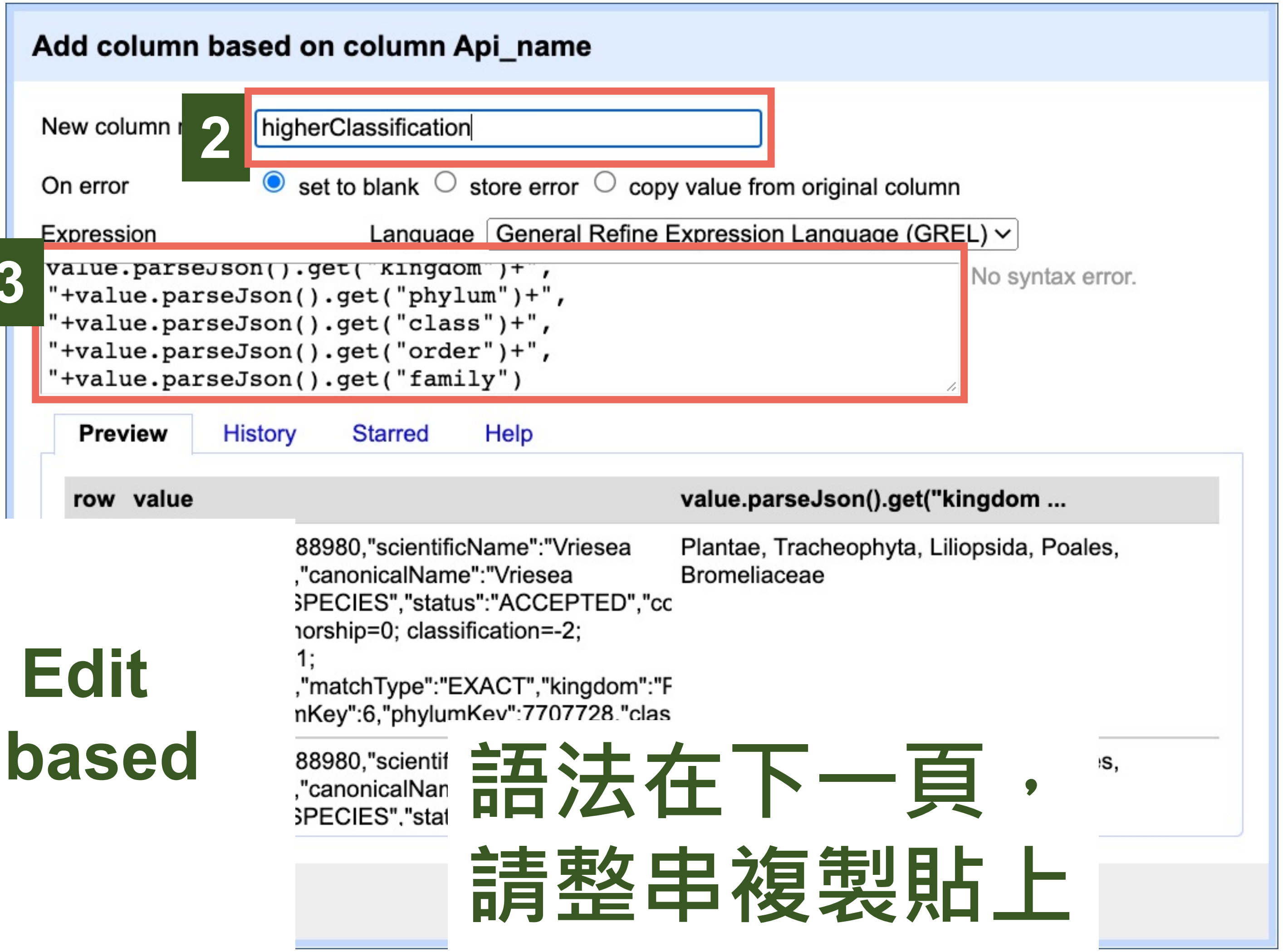


Show: 5 10 25 50 rows

Api_name

```
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:
```

Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Split into several columns...
Join columns...
Add column based on this column...
Add column by fetching URLs
1



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               Plantae, Tracheophyta, Liliopsida, Poales,
     ,"canonicalName":"Vriesea                     Bromeliaceae
     SPECIES","status":"ACCEPTED","cc
     horship=0; classification=-2;
     1;
     ,"matchType":"EXACT","kingdom":"F
     nKey":6,"phylumKey":7707728,"clas
     88980,"scientif                               s,
     ,"canonicalNan
     SPECIES","stat
```

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

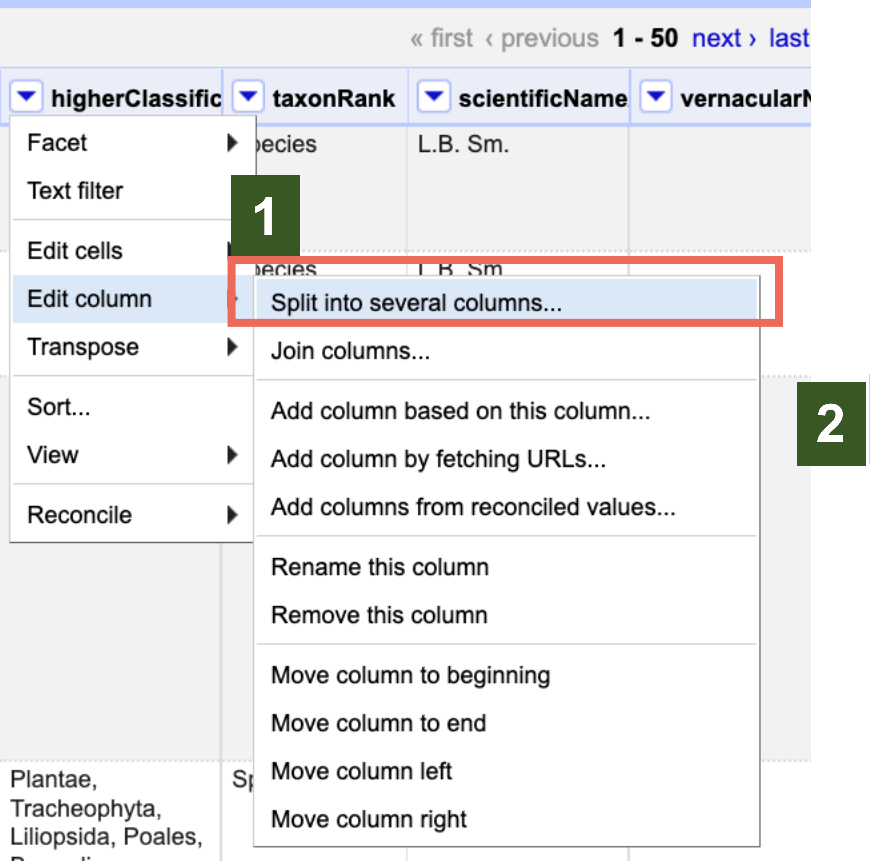
進階題-自動匯入高階層分類欄位

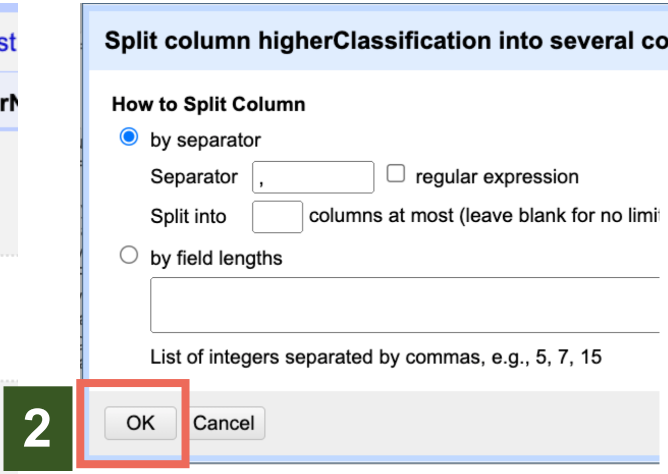
語法在此，請整串複製貼上

```
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

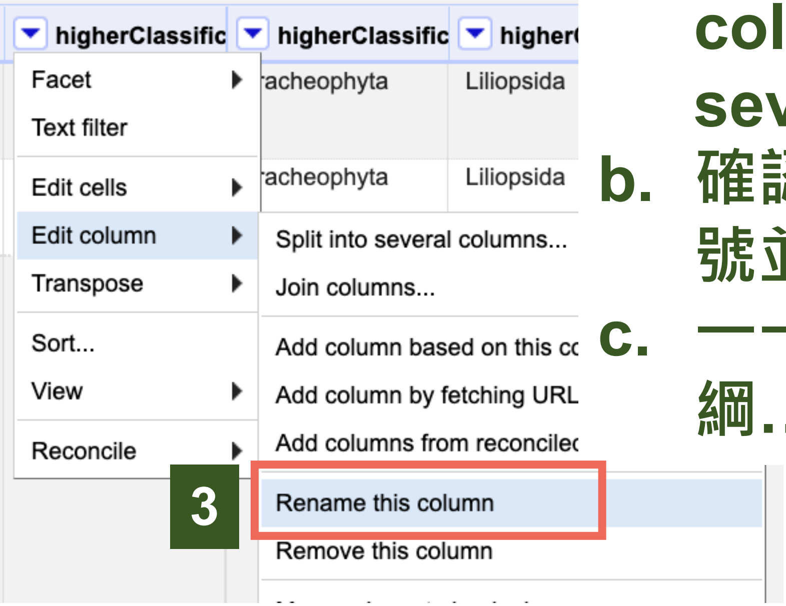
Separator , □ regular expression

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



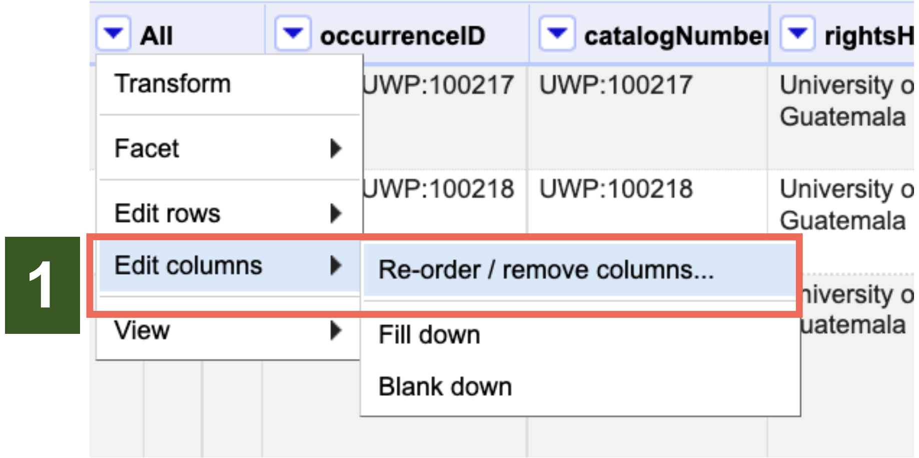
將一個欄位中的值分成不同欄位

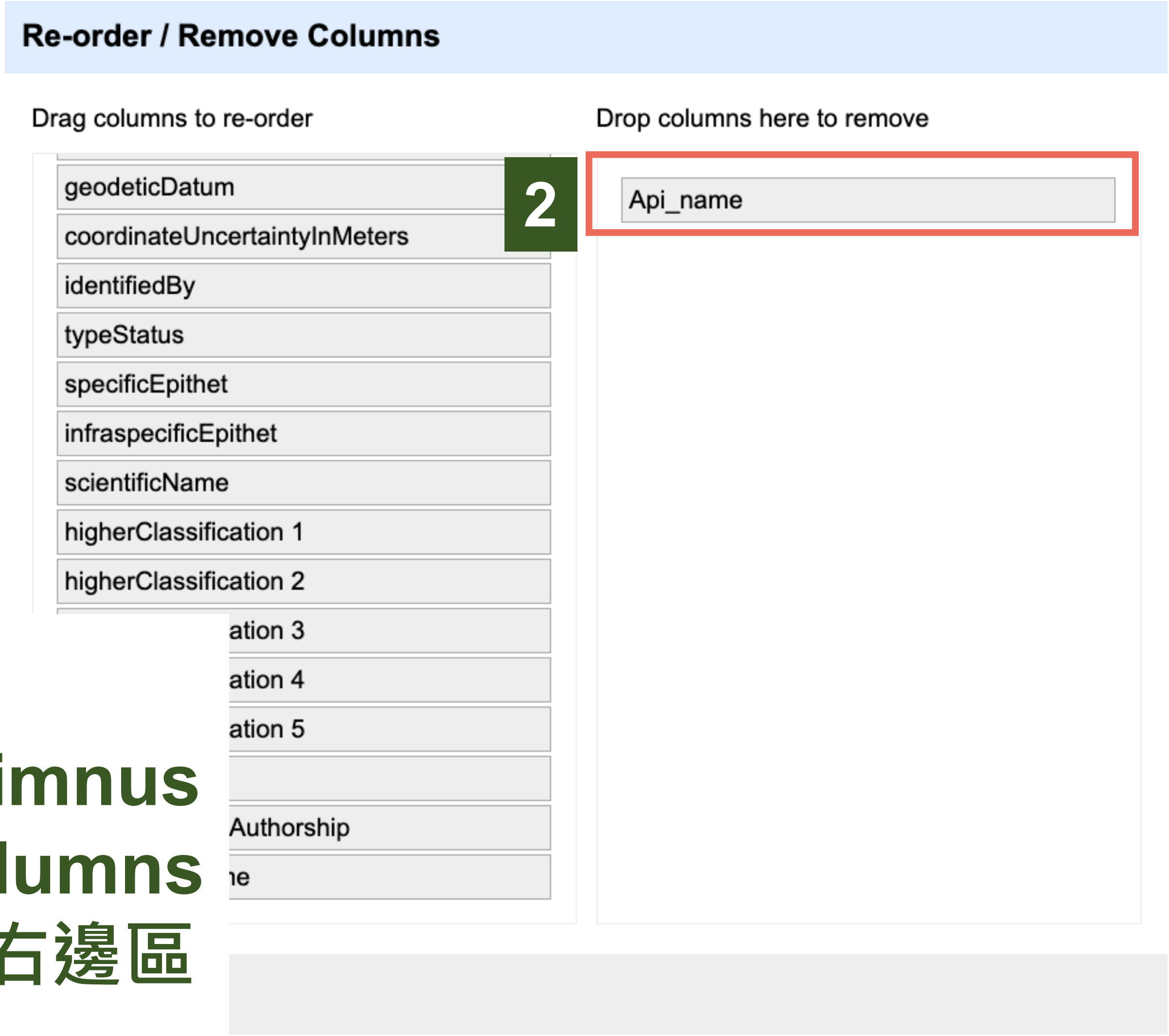
a. 到 higherClassification 欄位並選擇 Edit column >> Split into several columns

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





將不要的欄位刪除

a. 到 All欄位並選擇Edit colimnus >> Re-order/ remove columns

b. 拖曳左邊不想要的欄位到右邊區域並按 OK

Thank you!