開放資料的前置準備與清理

楊富鈞

2024生物多樣性資料發布與應用工作坊

關於我

自然史博物館典藏管理
生物多樣性資料開放 (與應用)

楊富鈞 Fu-Chun Yang

典藏管理組 助理研究員

yuukumo0312@gmail.com

文化部　國立臺灣博物館

動機/誘因

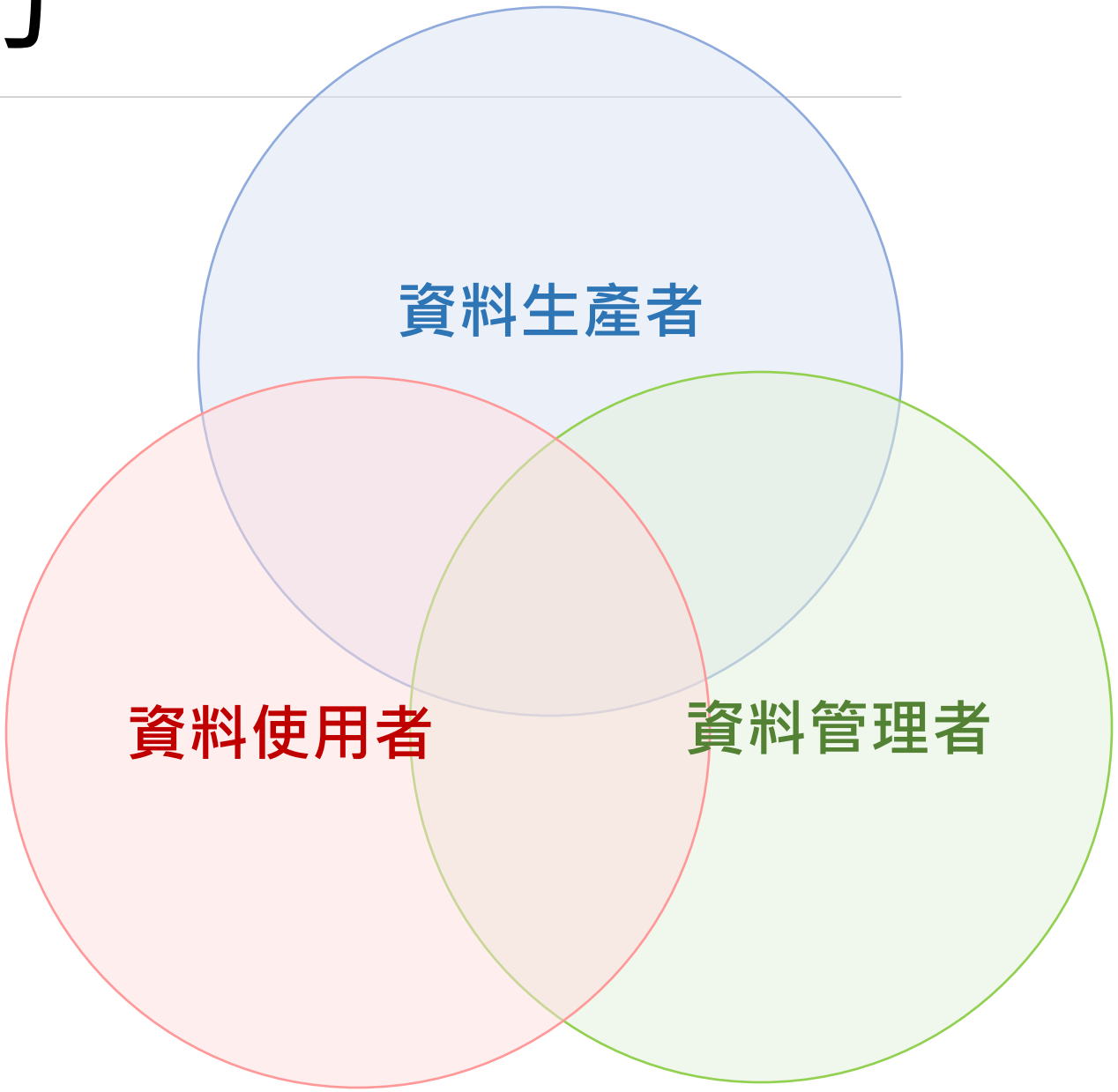
為何要花時間整理資料？

讓資料變得更好用

動機/誘因 + 角色/人力

讓**誰**好**用**？為**誰**整理？

- 1 週後的自己
- 1 月後的自己
- 1 年後的自己
- 長官或老闆
- 同行好友們（專業同溫層）
- 非同行好友們（人生同溫層）
- 同行冤家們
- 社會大眾（終極目標）



我想使用

或

由我負責管理的資料

最**完美**的資料狀態

已足夠**合理**的資料狀態

一千零一夜後...

@ Luis Prado from Noun Project

開放資料

已足夠合理成為開放資料的資料狀態

- 內容維護具可行性與持續性(重要)
- 資料集內部邏輯一致具整體性(強烈建議)
- 合法且合乎倫理(重要)

...


一筆快樂的原始資料
！！請先備份！！
雲端1份、本機1份、外接硬碟1份 (這3個大家都有吧)
確認對上的共通欄位
Darwin Core Quick Reference Guide
確認：
(1) Core & extension分配
(2) 各表單格式、單位、表示法
(3) 有沒有還可以填的欄位
(定下所有的資料表格式)
建議：留下文件給內部共通用
可藉此寫下MetaData初稿
(防止自己金魚腦，做事好利利)
資料清理
Excel、OpenRefine、R
目標：符合定下的DwC格式
學名對不對
Nomenmatch
ID生成&配對
自設ID 或 GUID
引入分類API
GBIF Backbone
座標轉換
TaiBIF上有小工具、R 或
Canadensys coordinate conversion
敏感資料處理
可跟著同步處理MetaData、
與coordinateUncertaintyInMeters
抓錯字、多餘空格、ID重複
記得要做版本控制喔
GBIF Data Validator
檢核完成 & 蓄勢待發
別忘了還有MetaData喔

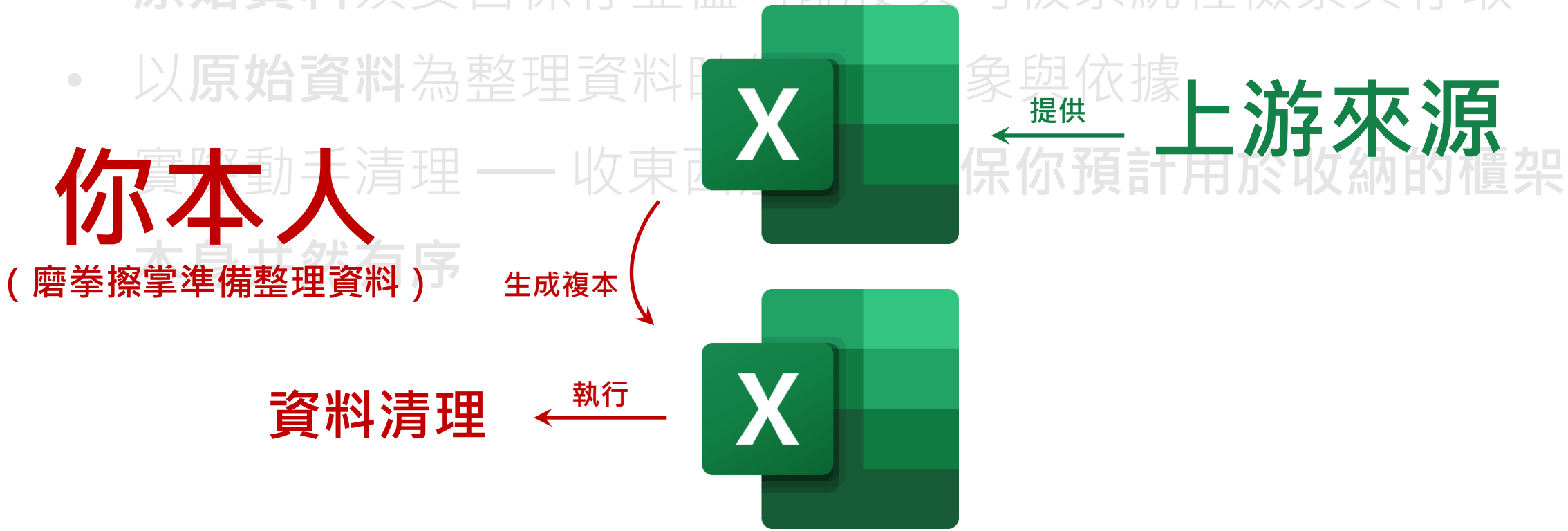
開始處理資料前的重要概念

開始處裡資料前的重要概念

- 資料**備份**與**版本維護**永遠重要
- **原始資料**須妥善保存並儘可能使其可被系統性檢索與存取
- 以**原始資料**為整理資料時的根本對象與依據
- 創建**適用的表單**

開始處裡資料前的重要概念

- 資料**備份**與**版本維護**永遠重要

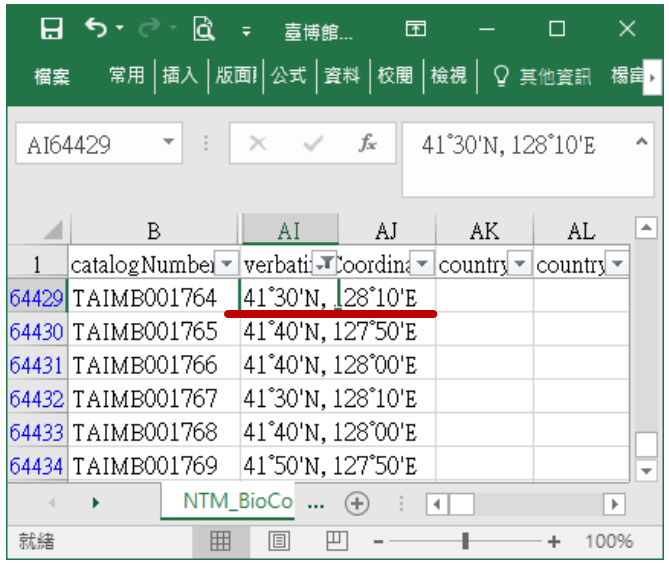


概念/專業知識

資料處理前重要的基本概念

- 資料備份與版本維護永遠重要
- **原始資料**須妥善保存並儘可能使其可被系統性檢索與存取
- 以**原始資料**為整理資料時的根本對象與依據
- 實際動手清理 —— 收東西前應先確保你預計用於收納的櫃架本身井然有序

養成思考哪一套內容版本才是
「原始資料」的習慣！


臺博館...
檔案 常用 插入 版面i 公式 資料 校閱 檢視 其他資訊 楊
AI64429
41°30'N, 128°10'E
B AI AJ AK AL
1 catalogNumbe verbati coordina country country
64429 TAIMB001764 41°30'N, 128°10'E
64430 TAIMB001765 41°40'N, 127°50'E
64431 TAIMB001766 41°40'N, 128°00'E
64432 TAIMB001767 41°30'N, 128°10'E
64433 TAIMB001768 41°40'N, 128°00'E
64434 TAIMB001769 41°50'N, 127°50'E
NTM_BioCo ...
就緒
100%

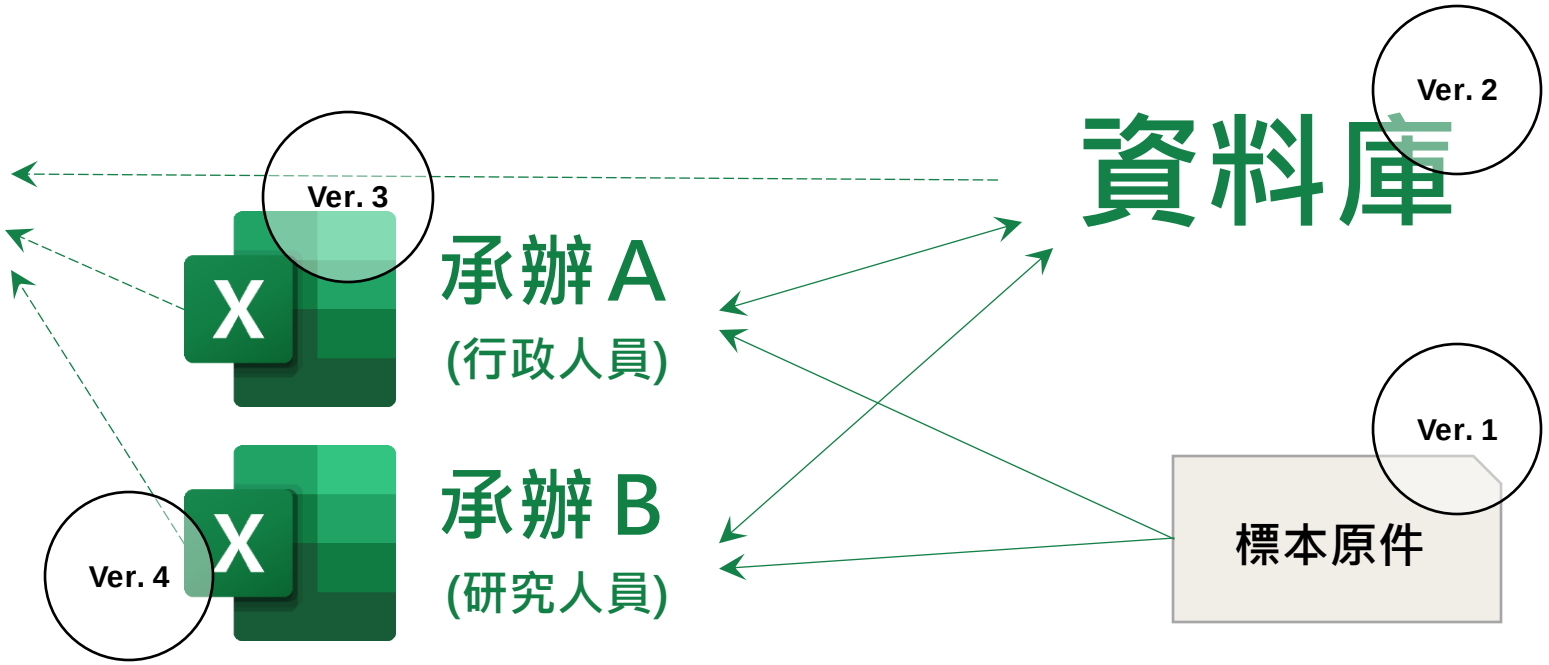

919.
TMBAE001764
BOTANICAL MUSEUM, UNIVERSITY OF HELSINKI
Helsinki, Finland
Mniaceae
Mnium ambiguum H. Mill.
det. T. Koponen -82
CHINA. Jilin (Kirin) Prov. An-tu Co.:
Mt. Chang Bai, Valley of R. Er-do
Bai-hu. Upper oroboreal Betula ermanii
forest at small lake, alt. 1700 m,
41°30'N, 128°10'E, on boulder
22 September 1981 Timo Koponen 36870
(collection site no. 6)

數位化很重要！

標本原件

你本人

（磨拳擦掌準備整理資料）

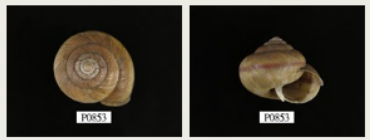


個案說明與分析

CASE: P0853

典藏資源檢索系統
國立臺灣博物館

貝類/P0853江西巴蝸牛



江西巴蝸牛 詮釋資料

藏品記錄
- 編目號：P0853
- 中文名稱：江西巴蝸牛
- 數量：1
- 單位：件
- 來源說明：碧山巖寺99.11.23捐贈

採集地理資訊
- 採集地：廣州市廣西壯族自治區

物種分類資訊
- 綱名：腹足綱 Gastropoda
- 目名：柄眼目 Stylommatophora
- 科名：扁蝸牛科 Bradybaenidae
- 學名：*Bradybaena kiangsinensis*
- 命名者：(Martens)
- 命名年代：1875

藏品描述

屬較大型的蝸牛。螺殼約有6個螺層、右旋，殼體呈圓球形，殼質厚且堅固，螺塔呈圓錐形，殼頂尖，各螺層略膨脹，體層膨大，縫合線深。殼面具有許多細密的生長紋。殼表呈黃褐色具光澤，體層的周緣有1條紅褐色的寬色帶。殼口呈橢圓形，外唇緣薄而銳利，臍孔深、部分被反捲的的軸唇所遮蓋。本種為中國大陸的特有物種，目前記錄於黑龍江、北京、河北、河南、湖南、湖北、四川、江西、廣西。主要棲息在樹林中，喜好陰暗潮濕的環境，也可見於農田田埂、壟間的雜草叢、灌木叢或亂石堆里、甚至住家公園也可發現到他們的蹤跡。本科物種的卵屬於小型且柔軟，產於鬆軟的泥土中或是地表上。

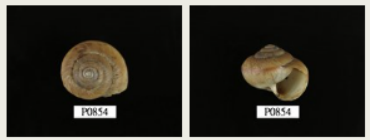
殼頂附近的殼面輕微磨損、殼皮脫落，殼體有幾處裂損，殼表和殼口內有髒污，殼唇緣有缺損。

廣州市廣西壯族自治區???

個案說明與分析

CASE: P0854

典藏資源檢索系統
國立臺灣博物館



貝類/P0854江西巴蝸牛

江西巴蝸牛 詮釋資料

藏品記錄
- 編目號：P0854
- 中文名稱：江西巴蝸牛
- 數量：1
- 單位：件
- 來源說明：碧山巖寺99.11.23捐贈

採集地理資訊
- 採集地：廣州市廣西壯族自治區

物種分類資訊
- 綱名：腹足綱 Gastropoda
- 目名：柄眼目 Stylommatophora
- 科名：扁蝸牛科 Bradybaenidae
- 學名：*Bradybaena kiangsinensis*
- 命名者：(Martens)
- 命名年代：1875

藏品描述

屬較大型的蝸牛。螺殼約有6個螺層、右旋，殼體呈圓球形，殼質厚且堅固，螺塔呈圓錐形，殼頂尖，各螺層略膨脹，體層膨大，縫合線深。殼面具有許多細密的生長紋。殼表呈黃褐色具光澤，體層的周緣有1條紅褐色的寬色帶。殼口呈橢圓形，外唇緣薄而銳利，臍孔深、部分被反捲的的軸唇所遮蓋。本種為中國大陸的特有物種，目前記錄於黑龍江、北京、河北、河南、湖南、湖北、四川、江西、廣西。主要棲息在樹林中，喜好陰暗潮濕的環境，也可見於農田田埂、壟間的雜草叢、灌木叢或亂石堆里、甚至住家公園也可發現到他們的蹤跡。本科物種的卵屬於小型且柔軟，產於鬆軟的泥土中或是地表上。

殼頂附近的殼皮磨損脫落，殼體有幾處裂損，殼表和殼口內有髒污，殼唇緣有缺損。

廣州市廣西壯族自治區???

廣州市 - Google 地圖

google.com/maps/place/中國廣東省廣州市/@23.1258868,110.5910576,7z/data=!4m6!3m5!1s0x3402f895a35c2bc7:0xe59e075adeae415!8m2!3d23.1290799!4d113.26436!16zL20vMDM5M2c?hl=zh-TW&entry=ttu

無痕式視窗

中國廣東省廣州市

廣州市

广州市

中國

廣東省

多雲時陰 · 29°C
凌晨1:41

規劃路線 | 儲存 | 附近 | 傳送到手機 | 分享

廣州市，通稱廣州，簡稱廣、穗，別稱羊城，是中華人民共和國廣東省省會、副省級市、首批沿海開放城市。廣州市為中國大陸和廣東對外的商貿中心兼綜合交通樞紐，是中國大陸的一線城市之一，也是粵港澳大灣區的中心城市之一，中國人民解放軍南部戰區聯合指揮部亦駐紮該地。
維基百科

廣州市的代表性景點

已儲存

最近

廣州市

廣西壯族自治區

登入

圖層

地圖資料 ©2023 Google 台灣 使用條款 隱私權 提供意見 100 公里

廣西壯族自治區 - Google 地圖

google.com/maps/place/中國廣西壯族自治區/@23.2774687,105.6145154,7z/data=!3m1!4b1!4m6!3m5!1s0x36a4e4af44bcbf8f:0x61d9b3862087774a!8m2!3d22.8152099!4d108.32754!16zL20vMDE5MjBt?hl=zh-TW&entry=ttu

無痕式視窗

中國廣西壯族自治區

廣西壯族自治區

广西壮族自治区

中國

規劃路線 儲存 附近 傳送到手機 分享

相片

已儲存 最近 廣西壯族自治區 廣州市

登入

六盤水市 貴陽市 黔東南苗族侗族自治州 衡陽市 永州市 安順市 黔南布依族苗族自治州 郴州市 曲靖市 黔西南布依族苗族自治州 桂林市 昆明市 韶關市 河池市 柳州市 賀州市 百色市 來賓市 清遠市 梧州市 紅河縣 文山市 貴港市 肇慶市 廣州市 南寧市 佛山市 東莞市 玉林市 雲浮市 深圳市 欽州市 澳門 防城港市 陽江市 茂名市 北海市 湛江市 海防市 Hải Phòng 南定市 Tp. Nam Định 寧平市 Tp. Ninh Bình 清化 Thanh Hoá 海口市

Google

地圖資料 ©2023 Google 台灣 使用條款 隱私權 提供意見 100 公里

個案說明與分析

CASE: P0853 & P0854

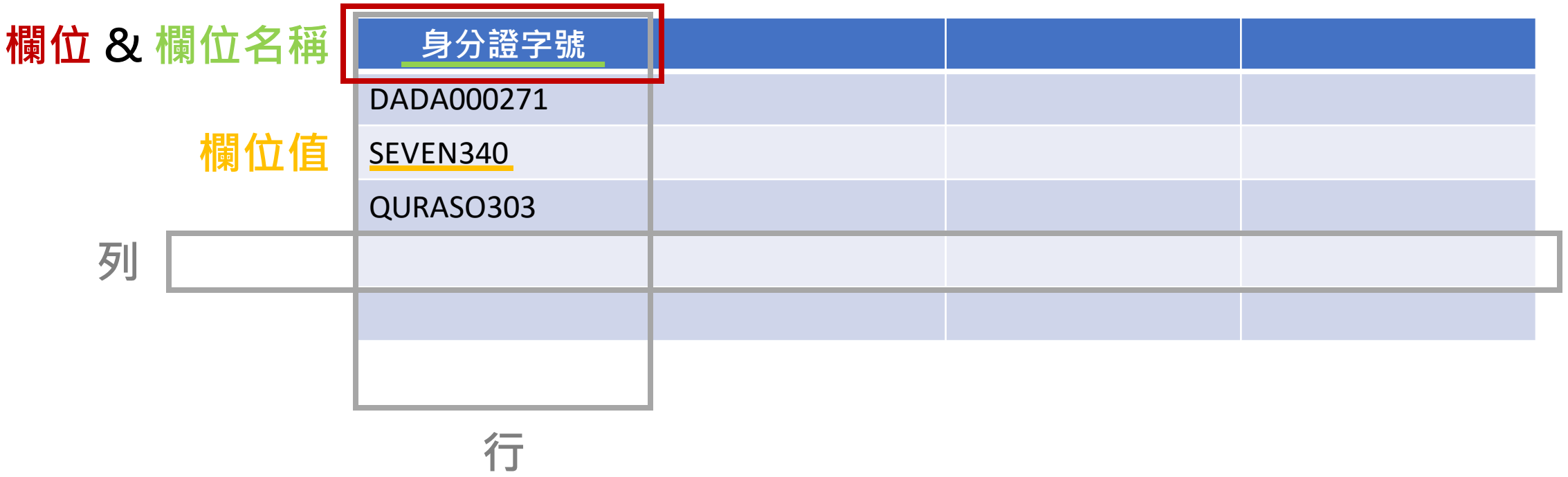
概念/專業知識

資料處理前重要的基本概念

- 資料**備份**與**版本維護**永遠重要
- **原始資料**須妥善保存並儘可能使其可被系統性檢索與存取
- 以**原始資料**為整理資料時的根本對象與依據
- 創建**適用的表單**

資料的櫃架是表格 (Table)

以由**直行 (column)**、**橫列 (row)** 組成，且**遵循一套標準規則**的**表格**來存儲及整理你的資料。



欄位定義與欄位分工

欄位分工

- 所有欄位都有明確功能。
- **欄位之間的功能不重複。**

欄位定義

- 所有欄位都有明確定義（用來載錄什麼資料、資料的值域、是否使用控制詞彙等），**且該定義最好是可被使用者透過網際網路系統性查詢的**。

欄位定義的維護頗麻煩，所以較合理的做法是

引用人家已發展成熟的方案

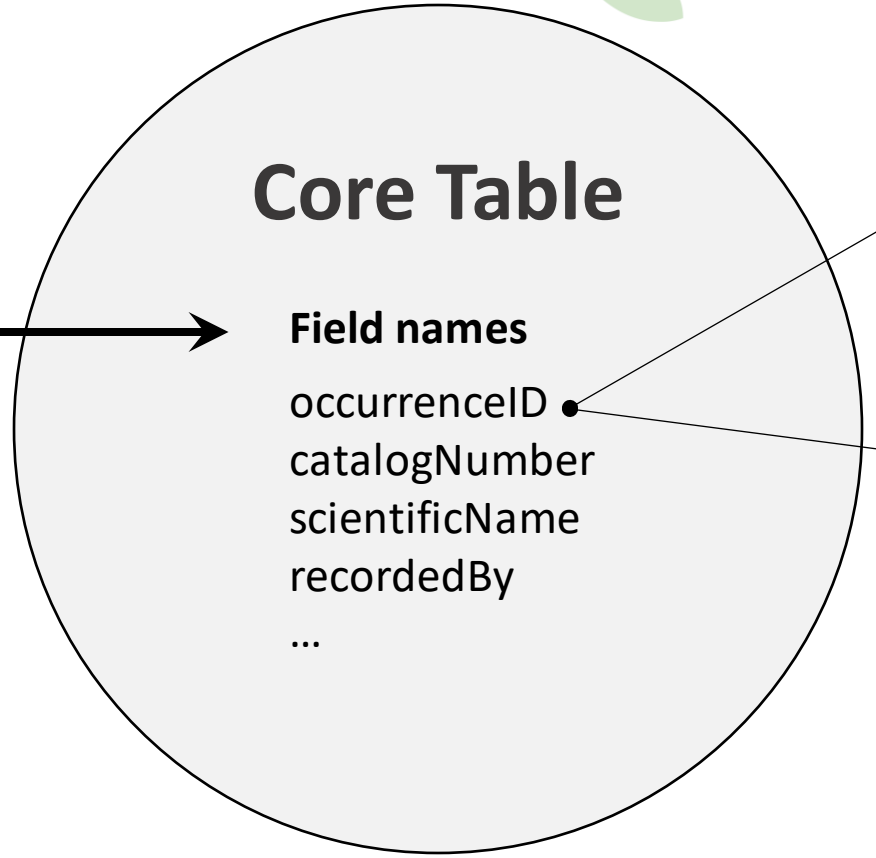
—— 至少以之為發展起點

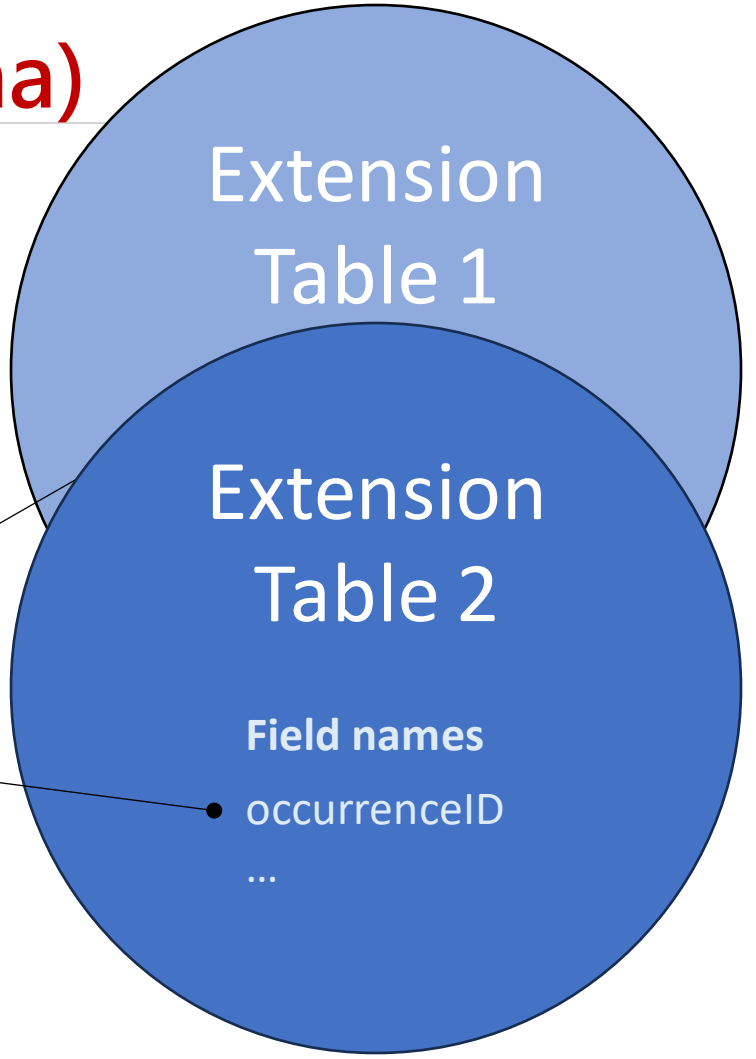
GBIF 及 TBIA 倡議的資料結構 (Data Schema)

DwC

https://dwc.tdwg.org/

cited

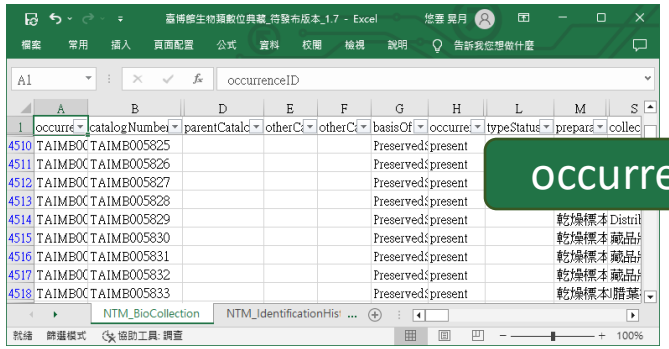




https://rs.gbif.org/extensions.html

概念/專業知識

以臺博館為例



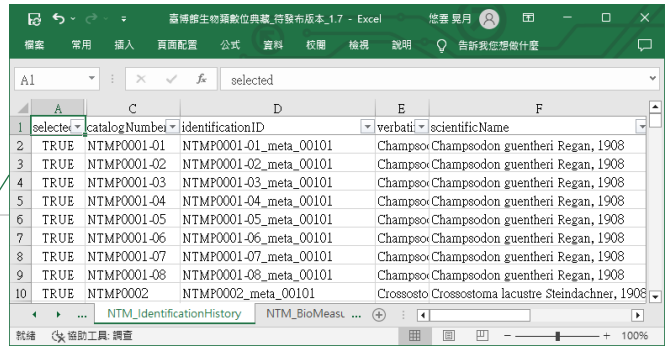
occurrenceID

Darwin Core Occurrence

主表處理館號、採集者、採集日期、採集地、館方選定的鑑定結果等資料

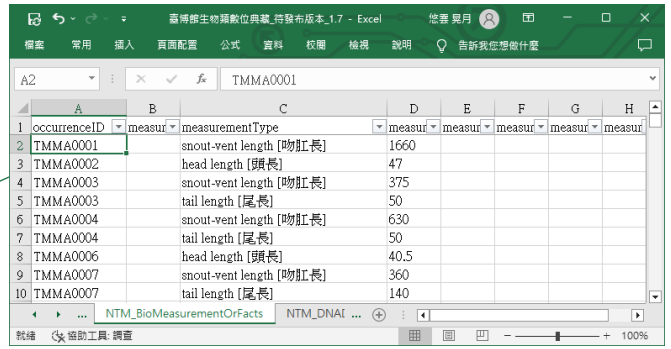
ONLINE

(in OneDrive)



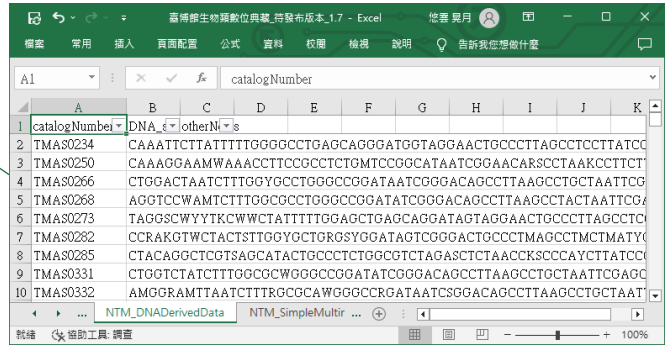
Darwin Core Identification History

處理各標本的鑑定脈絡



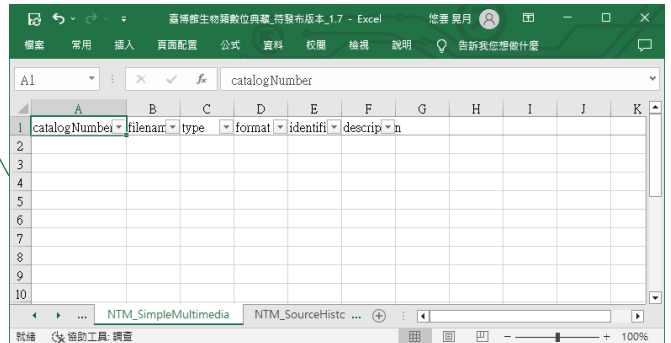
Extended Measurement Or Facts

處理各標本的形質測量資料



DNA derived data

處理各標本的定序資料

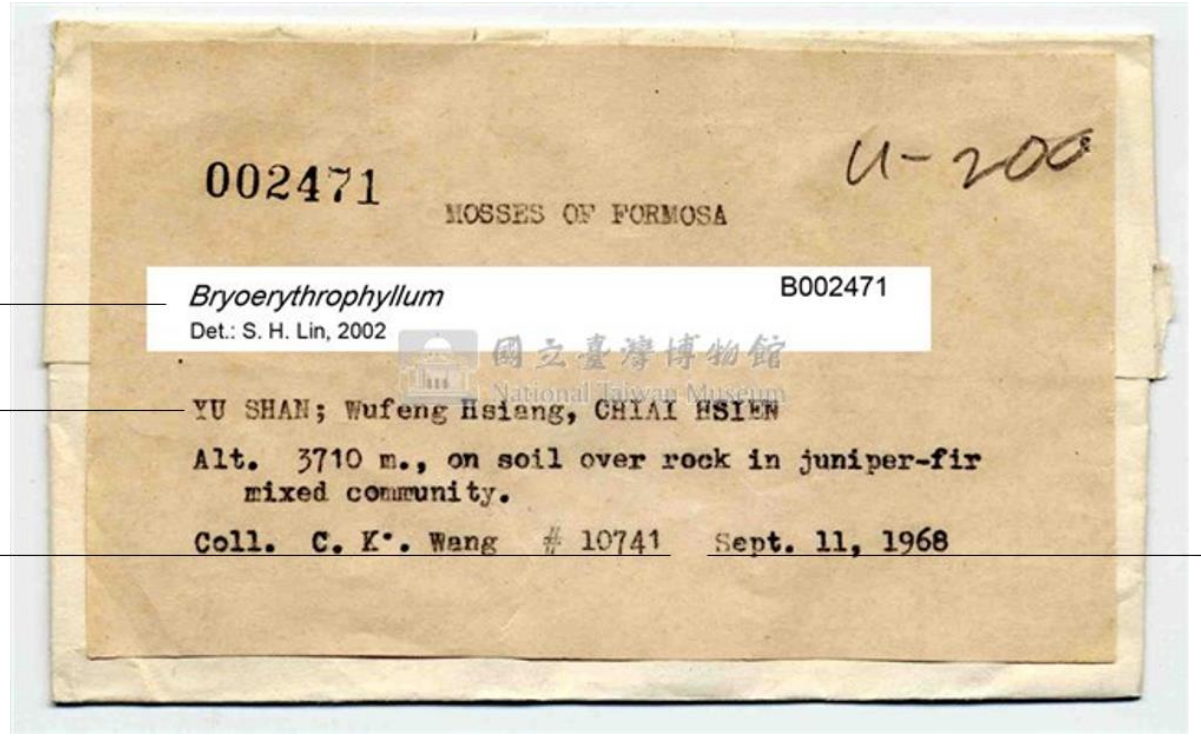


Simple Multimedia

處理各標本的影像資料

我對欄位的分類邏輯

基於內容類型



分類鑑定資訊

採集地資訊

採集者資訊

採集日期資訊

我對欄位的分類邏輯

基於長期維護的方式

- 辨識碼 (ID)
- 逐字稿欄位 ← 原封不動地將內容轉錄進去！
- 標準化內容欄位 ← 基於特定規則之內容；字數一般較少。
  - 根源於逐字稿欄位
  - 獨立於逐字稿欄位
- 註記欄位 ← 用於收納任意型式的註記；字數可多可少。
- 其他

表格型制穩定後

再來就是正式面對欄位值清理

生物多樣性資料有哪些常見的待清理樣態？

以「欄位值」錯誤樣態舉例：

- 違反資料結構（如欄位值未落於該欄位值域範圍、資料型態錯誤）
- 不合理的資料重複（如 unique field 出現重複值）
- 拼字錯誤
- 缺失值
- 異常值
- 冗餘的「空格」
- 欄位值缺乏一致性或整體性
- 相關欄位彼此矛盾或關聯性出現斷裂

…

概念/專業知識 + 技巧/工具

條列一些個人對資料清理的"原則性建議或經驗"

- 善用各種線上服務與開源軟體
- 根基於常識/專業知識的聯想力或想像力
- 先建立權威檔
- 善用自己已建立的資料
- 把握與每個錯誤樣態的不期而遇
- 與 Excel 函式 & 內建功能當好朋友
- 透過與其他資料庫相互比對來校正內容
- 學習正規表達式 (regular expression)

Biological-Collection-Integrati | 臺博館生物類數位典藏_待發布版 | 2029091401_TaiBIF-Workshop | GBIF | Data validator

gbif.org/zh-tw/

2,575,559,246
出現紀錄

89,923
資料集

2,129
發布機構

9,316
使用資料的同儕審查論文

從案例中學習

概念/專業知識 +技巧/工具

- 善用各種線上服務與開源軟體
- **根基於常識/專業知識的聯想力或想像力**
- 先建立檢感機
- 善用自
- 把握與每個錯誤樣態的不期而遇
- 與 Excel 函式 & 內建功能當好朋友
- 透過與其他資料庫相互比對來校正內容
- 學習正規表達式 (regular expression)

catalogNumber	eventDate
NTMP0479-13	1977-06-40

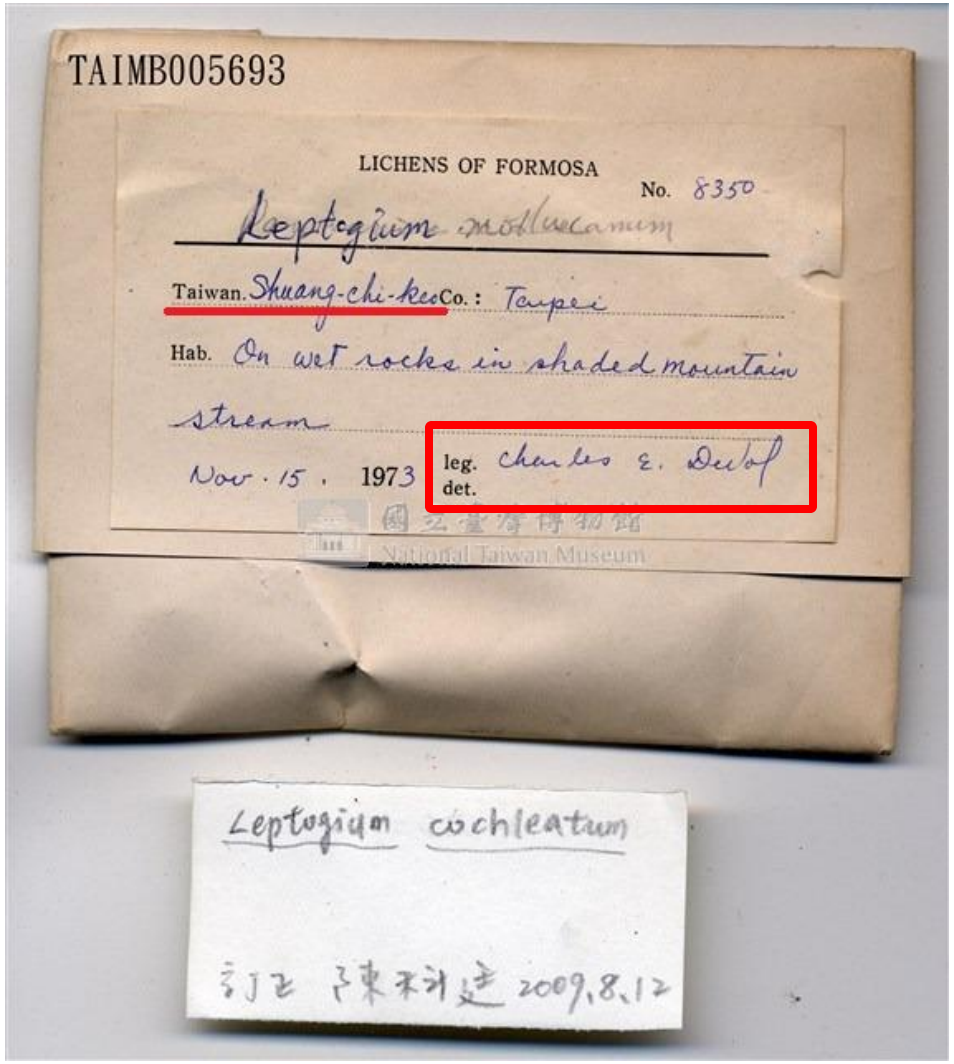
概念/專業知識 +技巧/工具

- 善用各種線上服務與開源軟體
- 根基於常識/專業知識的聯想力或想像力
- 先建立權威檔
- 善用自己已建立的資料
- 把握與每個錯誤樣態的不期而遇
- 與 Excel 函式 & 內建功能當好朋友
- 透過與其他資料庫相互比對來校正內容
- 學習正規表達式 (regular expression)

概念/專業知識 +技巧/工具

- 善用各種線上服務與開源軟體
- 根基於常識/專業知識的聯想力或想像力
- 先建立權威檔
- 善用自己已建立的資料
- 把握與每個錯誤樣態的不期而遇
- 與 Excel 函式 & 內建功能當好朋友
- **透過與其他資料庫相互比對來校正內容**
- 學習正規表達式 (regular expression)

CASE: TAIMB005693



verbatiumLocality

Taiwan. Shuang-Chi-{Keo[?]} Co.: Taipei. Hab. On wet rocks in shaded mountain stream

Shuang-Chi-{Keo[?]} 在哪？

Shuang-Chi = 雙溪？哪個雙溪？

Charles E. DeVol

棣慕華 (1903-1989)

tai2.ntu.edu.tw/search/2

台灣植物資訊整合查詢系統 | Plants of Taiwan | English

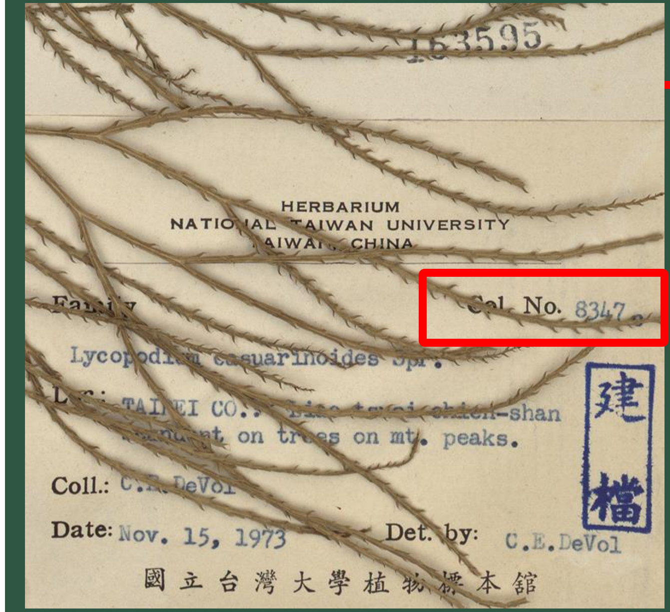
搜尋

找植物

找標本

電子書

如頁面顯示有問題，請填寫回報表單
如有任何意見也請留言，謝謝。



Plumbaginaceae 藍雪科	Limonium sinense 石苙蓉	Statice arbuscula Maxim.
1973/5/17	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spr.
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Lycopodiaceae 石松科	Lycopodiastrum casuarinoides 木賊葉石松	Lycopodium casuarinoides Spring
1973/11/15	C.E.DeVol 棣慕華	
Gleicheniaceae 裏白科	Diplopterygium chinensis 中華裏白	Diplopterygium chinensis (Rosenst.) DeVol
1973/11/15	C.E.DeVol 棣慕華	
Marattiaceae 觀音座蓮舅科	Angiopteris lygodiifolia 觀音座蓮	Angiopteris lygodiifolia Rosenst.
1973/12/17	C.E.DeVol 棣慕華	
Marattiaceae 觀音座蓮舅科	Angiopteris lygodiifolia 觀音座蓮	Angiopteris lygodiifolia Rosenst.

WHCM+CM 新店區 新北市

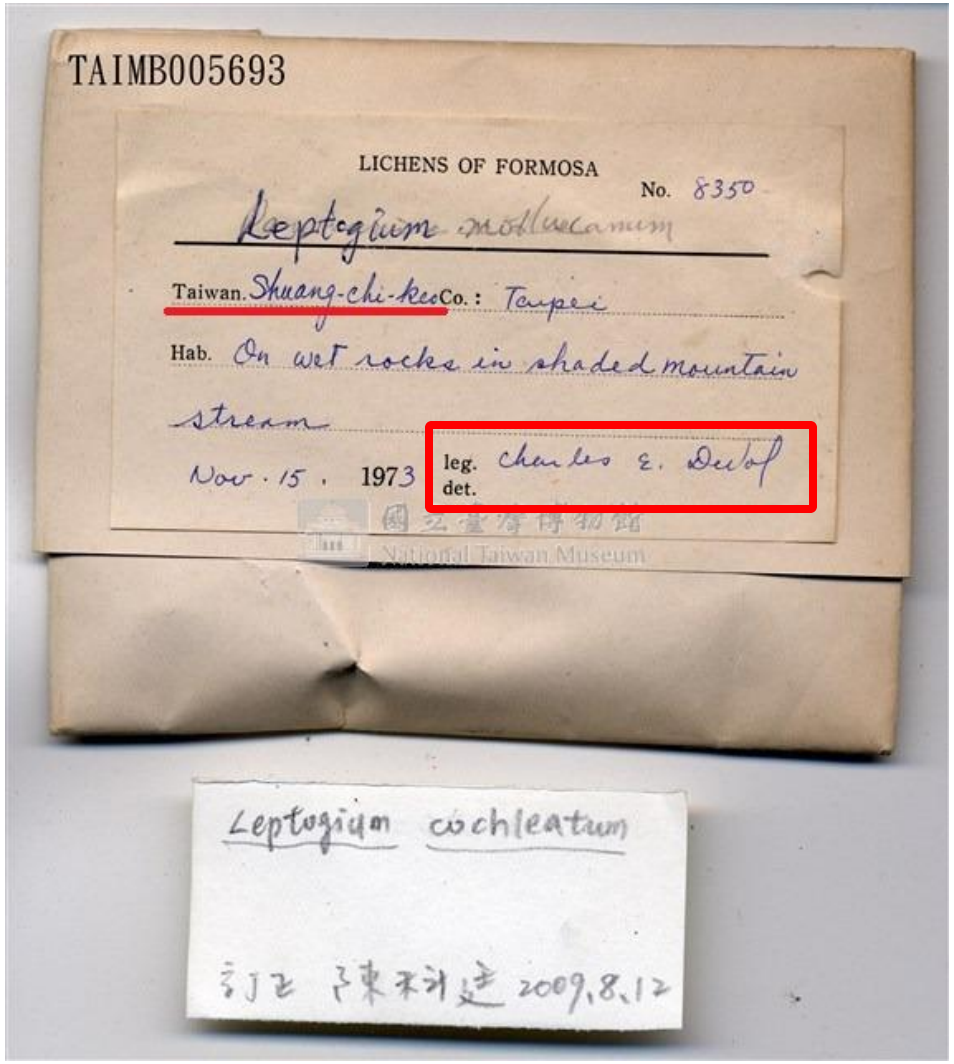
相片

新增相片

找不到所需的答案嗎？
社群通常會在 20 分鐘內回答問題。

詢問社群成員

CASE: TAIMB005693



verbatiumLocality

Taiwan. Shuang-Chi-{Keo[?]} Co.: Taipei. Hab. On wet rocks in shaded mountain stream

locality

New Taipei City|Xindian District|Shuangxikou, near Niaozuijian Mountain [新北市|新店區|雙溪口，鄰近鳥嘴尖山]

結語

經你整理的資料不再是檔案而是品牌

身為資料管理者的自尊

就是時刻以中立之姿幫助使用者取得好資料

加油！敬祝工作順利！