

資料標準及共通欄位

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

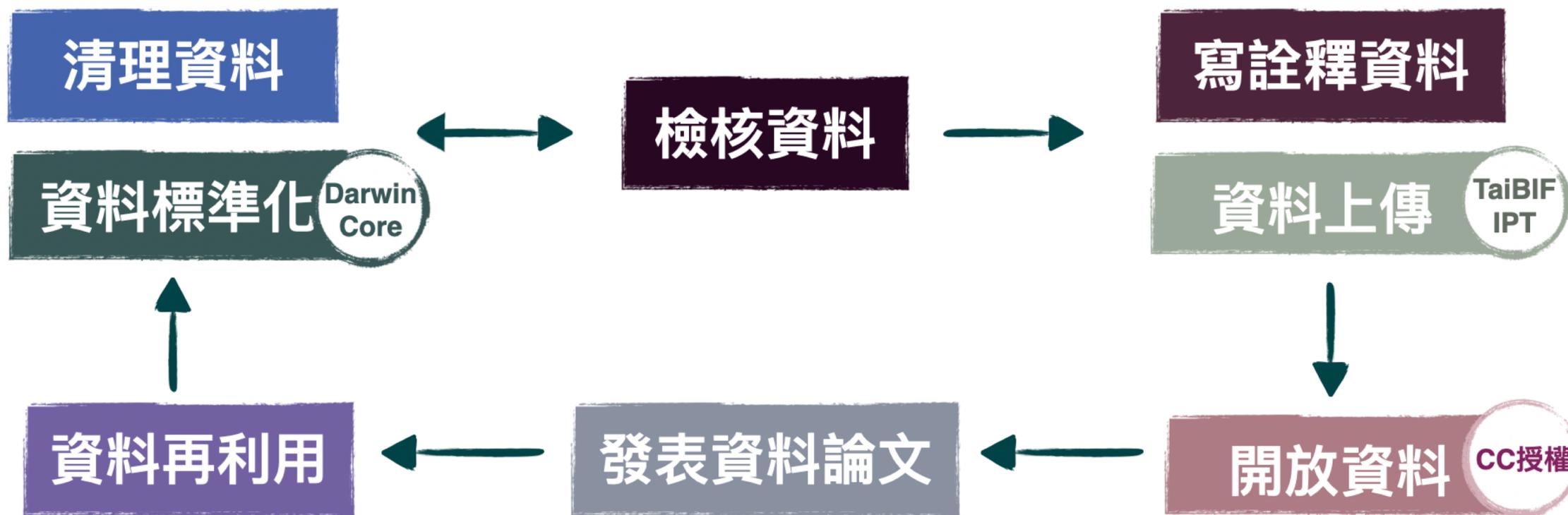


- 推動及宣傳國內生物多樣性資料標準、資料開放
- 國內生物多樣性資料庫相關單位合作
- TaiBIF 資料庫管理諮詢

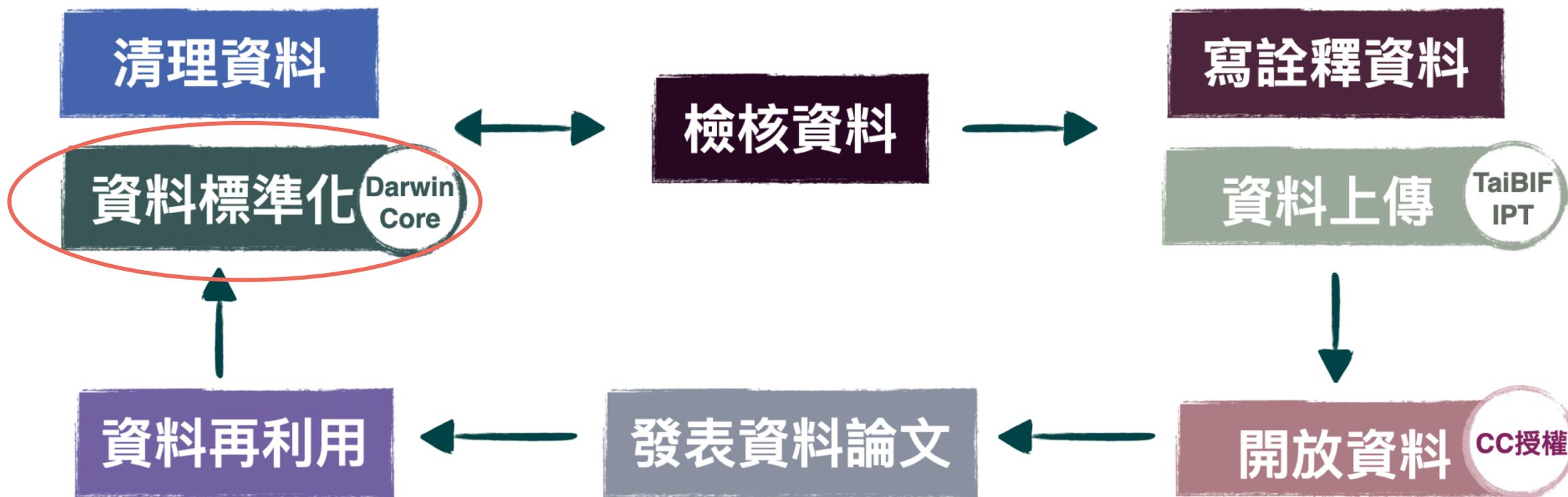
- 亞洲計畫團隊開放資料諮詢
- 推動亞洲國家節點合作
- 促進亞洲區域開放資料



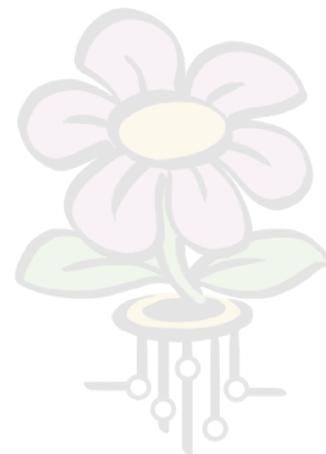
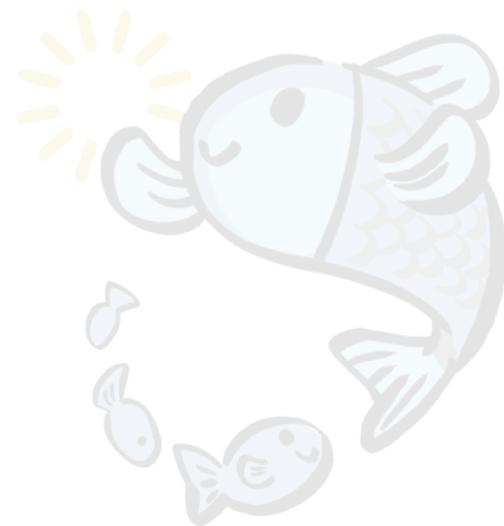
理想的資料開放流程



理想的資料開放流程



什麼是 資料標準



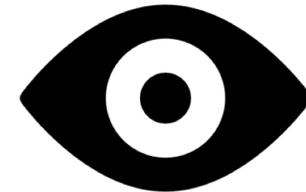
簡單來說 就是資料欄位的共通格式

學名 / 物種名
↓
scientificName

讓大家的資料欄位
名稱統一

日期 = yyyy-
mm-dd

讓同一個資料欄
位名稱的內容有
個格式規範



讓任何看到這個
資料的人，都可
以看得懂

資料集類型 再複習一下

Checklist (taxon)

名錄

屬於某一類別的物種清單

- 以物種學名為單位
- 不需要有數據、時間、詳細地點
- 常為地區性調查資料
- 如圖鑑、動物誌、植物誌

Occurrence

出現紀錄

帶有時間和空間資訊的物種出現紀錄

- 要有時間、地點
- 物種不一定要鑑定到種階層
- 大多研究資料屬於此類

Sampling event

調查活動

經過取樣設計的生物多樣性調查資料

- 要有時間、地點、取樣方法
- 基於一套試驗設計所做的調查

Metadata

詮釋資料

用來說明資料的背景資訊，但不包含在資料中

- 如計畫名稱、計畫目的、計畫描述、參與人員、授權方式等

Darwin Core 標準

EML 標準

說明

資料集類型 再複習一下

Checklist (taxon)

名錄

屬於某一類別的物種清單

- 以物種學名為單位
- 不需要有數據、時間、詳細地點
- 常為地區性調查資料
- 如圖鑑、動物誌、植物誌

Occurrence

出現紀錄

帶有時間和空間資訊的物種出現紀錄

- 要有時間、地點
- 物種不一定要鑑定到種階層
- 大多研究資料屬於此類

Sampling event

調查活動

經過取樣設計的生物多樣性調查資料

- 要有時間、地點、取樣方法
- 基於一套試驗設計所做的調查

Metadata

詮釋資料

用來說明資料的背景資訊，但不包含在資料中

- 如計畫名稱、計畫目的、計畫描述、參與人員、授權方式等

Darwin Core 標準

EML 標準

說明

⚙️ Tools

Over the years, many tools have been developed to facilitate effective data management, archiving content, and retrieving data for synthetic analysis projects.

📁 EML - Ecological Metadata Language

A metadata standard developed for the earth, environmental and ecological sciences.

It is based on prior work done by the Ecological Society of America and associated efforts. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document scientific research data.

[More about EML »](#)

🔧 KNB Developer Tools

The KNB supports the DataONE REST API for automating the process of uploading, downloading, and searching for data on the KNB using scripted languages such as shell, R, Matlab, and Python, among others.

[Learn more about the KNB Developer Tools and API »](#)

源自美國長期生態研究團隊的資料標準，對於研究資料集(dataset)／研究計畫內容的描述欄位

生態標記語言

Ecological Metadata Language

用機器可讀的檔案格式(XML)，來描述生態資料，讓資料容易被交換共享。

Metadata 詮釋資料 大概有這些



基本

Title 資料集題目
Type 資料類型
Organisation 單位
Data License 授權
Description 摘要
Contacts 作者資訊

標準

Geographic Coverage
地理涵蓋範圍
Temporal Coverage
時間涵蓋範圍
Keywords 關鍵字
Associated Parties
機關或個人資訊
Citations 引用方式

進階

Taxonomic Coverage
物種涵蓋範圍
Project Data
計畫資訊
Sampling Methods
採樣方法
Collection Data
典藏資訊

Metadata 詮釋資料 長這樣

簡單來說就像名片
讓別人對這個資料集有初步概念

調查活動 | 註冊於 2017年10月4日

The Taiwan Breeding Bird Survey Data

發布者為 [Taiwan Endemic Species Research Institute](#)

✉ Jerome Chie-Jen Ko

[資料集](#) [指標](#) [活動](#) [↓ 下載](#) [🏠 首頁](#)

3,120 出現紀錄

22 引用書目

The Taiwan Breeding Bird Survey (BBS Taiwan) is a national monitoring project with the aim of establishing a breeding bird population index for conservation. It was established in 2009 as a citizen science project with a cross-organizational working group. The BBS Taiwan has obtained 423,119 records of bird sampling data from 480 sampling sites during the breeding season in Taiwan from 2009 to 2016.

最後修改的詮釋資料: 2020年6月19日

託管者 : [Taiwan Biodiversity Information Facility \(TaiBIF\)](#)

授權: [CC BY-NC 4.0](#)

[如何引用](#) [DOI](#) [10.15468/qqkhd5](#)

 3,120
出現紀錄

 99.4%
有對應的分類群

 99.9%
含座標

 100%
含年份

臺灣繁殖鳥類大調查 BBS Taiwan

Metadata 詮釋資料 長這樣

簡單來說就像名片
讓別人對這個資料集有初步概念

調查活動 | 註冊於 2017年10月4日

The Taiwan Breeding Bird Survey Data

發布者為 [Taiwan Endemic Species Research Institute](#)
✉ Jerome Chie-Jen Ko

資料集 指標 活動 下載 首頁

3,120 出現紀錄 22 引用書目

The Taiwan Breeding Bird Survey (BBS Taiwan) is a national bird population index for conservation. It was established in 2009 by an organizational working group. The BBS Taiwan has obtained 423,119 records from 423 sites during the breeding season in Taiwan from 2009 to 2016.

最後修改的詮釋資料: 2020年6月19日
託管者: : Taiwan Biodiversity Information Facility (TaiBIF)
授權: CC BY-NC 4.0
如何引用 DOI 10.15468/qqkhd5

IPT 資料上傳重點!

 3,120 出現紀錄	 99.4% 有對應的分類群	 99.9% 含座標	 100% 含年份
--	---	---	--

[臺灣繁殖鳥類大調查 BBS Taiwan](#)

資料集類型 再複習一下

Checklist (taxon)

名錄

屬於某一類別的物種清單

- ❑ 以物種學名為單位
- ❑ 不需要有數據、時間、詳細地點
- ❑ 常為地區性調查資料
- ❑ 如圖鑑、動物誌、植物誌

Occurrence

出現紀錄

帶有時間和空間資訊的物種出現紀錄

- ❑ 要有時間、地點
- ❑ 物種不一定要鑑定到種階層
- ❑ 大多研究資料屬於此類

Sampling event

調查活動

經過取樣設計的生物多樣性調查資料

- ❑ 要有時間、地點、取樣方法
- ❑ 基於一套試驗設計所做的調查

Metadata

詮釋資料

用來說明資料的背景資訊，但不包含在資料中

- ❑ 如計畫名稱、計畫目的、計畫描述、參與人員、授權方式等

Darwin Core 標準

EML 標準

說明

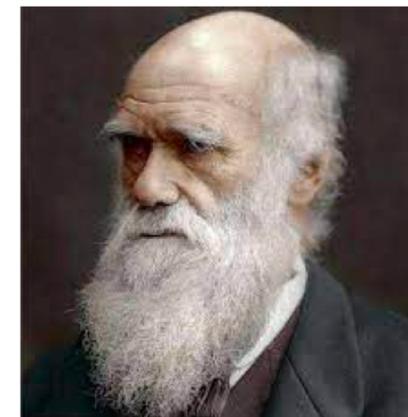
GBIF/ TaiBIF & 國際生物多樣性資料庫常在使用...

達爾文核心標準

Darwin Core

由國際的資料標準工作小組
T D W G 訂定。

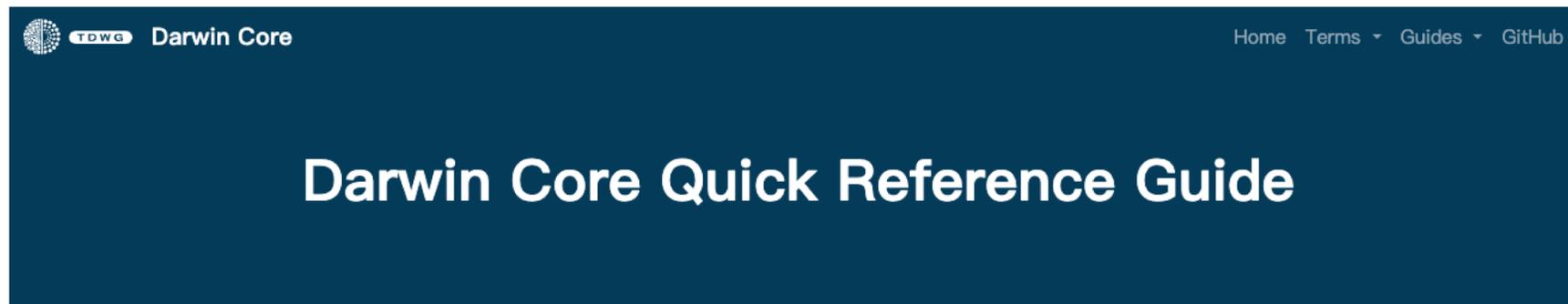
Biodiversity
Information
Standards
T D W G



嗯？不是我訂的

達爾文核心標準-詞條快速查詢檢索表

如何快速找到想要的詞條？



Record-level

This category contains terms that are ge

- type
- modified
- language
- license
- rightsHolder
- accessRights
- bibliographicCitation
- references
- institutionID
- collectionID
- datasetID
- institutionCode
- collectionCode
- datasetName
- ownerInstitutionCode
- basisOfRecord
- informationWithheld
- dataGeneralizations
- dynamicProperties

type

Identifier	http://purl.org/dc/elements/1.1/type
Definition	The nature or genre of the resource.
Comments	Must be populated with a value from the DCMI type vocabulary (http://dublincore.org/documents/2010/10/11/dcmi-type-vocabulary/).
Examples	StillImage MovingImage Sound PhysicalObject Event Text

2- 個別類別下的詞條內容
(意即可用的欄位名稱)

3- 每個詞條的定義
格式和範例說明

On this page

- Record-level
- Occurrence
- Organism
- MaterialSample
- Event
- Location
- GeologicalContext
- Identification
- Taxon
- MeasurementOrFact
- ResourceRelationship
- UseWithIRI
- LivingSpecimen
- PreservedSpecimen
- FossilSpecimen
- MaterialCitation

1- 詞條類別區



<https://dwc.tdwg.org/terms/>

Darwin Core Category 詞條類別

注意：這裡的taxon, occurrence, event只是一種詞條類別，並非指資料集類型



各類型資料 必填欄位

Checklist (taxon)

名錄

Occurrence

出現紀錄

Sampling event

調查活動

Metadata

詮釋資料

Checklist (taxon)	Occurrence	Sampling event	Metadata
名錄 ID taxonID	出現紀錄 ID occurrenceID	調查活動ID eventID	基本詮釋資料 Basic Metadata
學名 scientificName	紀錄類型 basisOfRecord	調查日期 eventDate	計畫資訊 Project Data
分類階層 taxonRank	紀錄日期 eventDate	調查方法 samplingProtocol	
	學名 scientificName	採樣量/範圍 sampleSizeValue	
	個體隻數 individualCount	採樣量單位 sampleSizeUnit	
	地點 locality	採樣努力量 samplingEffort	
	十進位座標 decimalLatitude decimalLongitude		
	大地基準 geodeticDatum		
	國家代碼 countryCode		

Darwin Core 標準

EML 標準

說明

各類型資料 必填欄位

Checklist (taxon)

名錄

Occurrence

出現紀錄

Sampling event

調查活動

Metadata

詮釋資料

名錄 ID **taxonID**

學名 **scientificName**

分類階層 **taxonRank**

出現紀錄 ID **occurrenceID**

紀錄類型 **basisOfRecord**

紀錄日期 **eventDate**

學名 **scientificName**

個體隻數 **individualCount**

地點 **locality**

其他座標

verbatimLatitude

verbatimLongitude

座標系統 **verbatimCoordinateSystem**

大地基準 **geodeticDatum**

國家代碼 **countryCode**

調查活動ID **eventID**

調查日期 **eventDate**

調查方法 **samplingProtocol**

採樣量/範圍 **sampleSizeValue**

採樣量單位 **sampleSizeUnit**

採樣努力量 **samplingEffort**

基本詮釋資料 **Basic Metadata**

計畫資訊 **Project Data**

Darwin Core 標準

EML 標準

說明

有沒有發現
有個欄位很重要？

三種類型的資料都有的

有沒有發現
有個欄位很重要？

沒錯，就是 ID！

為什麼 ID 重要？

- 避免資料重複發布
- 資料庫系統識別資料的依據

taxonID, occurrenceID, eventID



設定ID原則：

- 在資料集中不能重複
- 可設定有意義的序號 or 全球唯一識別碼

設定ID原則：

- 可設定有意義的序號-範例

典藏資料

urn:catalog:UWBM:Bird:89776

機構代碼+類群+館藏號

生態調查資料

TPE-230913-01

樣區/地點名稱+日期+調查次

設定ID原則：

- 全球唯一識別碼

沒有強制要什麼格式，但如果你對於設定什麼格式的序號沒特別想法的話.....

設定ID原則：

- 全球唯一識別碼

推薦 GUID/UUID !

- 隨機亂數產生
- 重複機率趨近於零

最主要目的：
讓資料使用者取得資料
時可自行判斷重複資料

線上UUID 產生器

<https://www.uuidgenerator.net/>

線上GUID 產生器

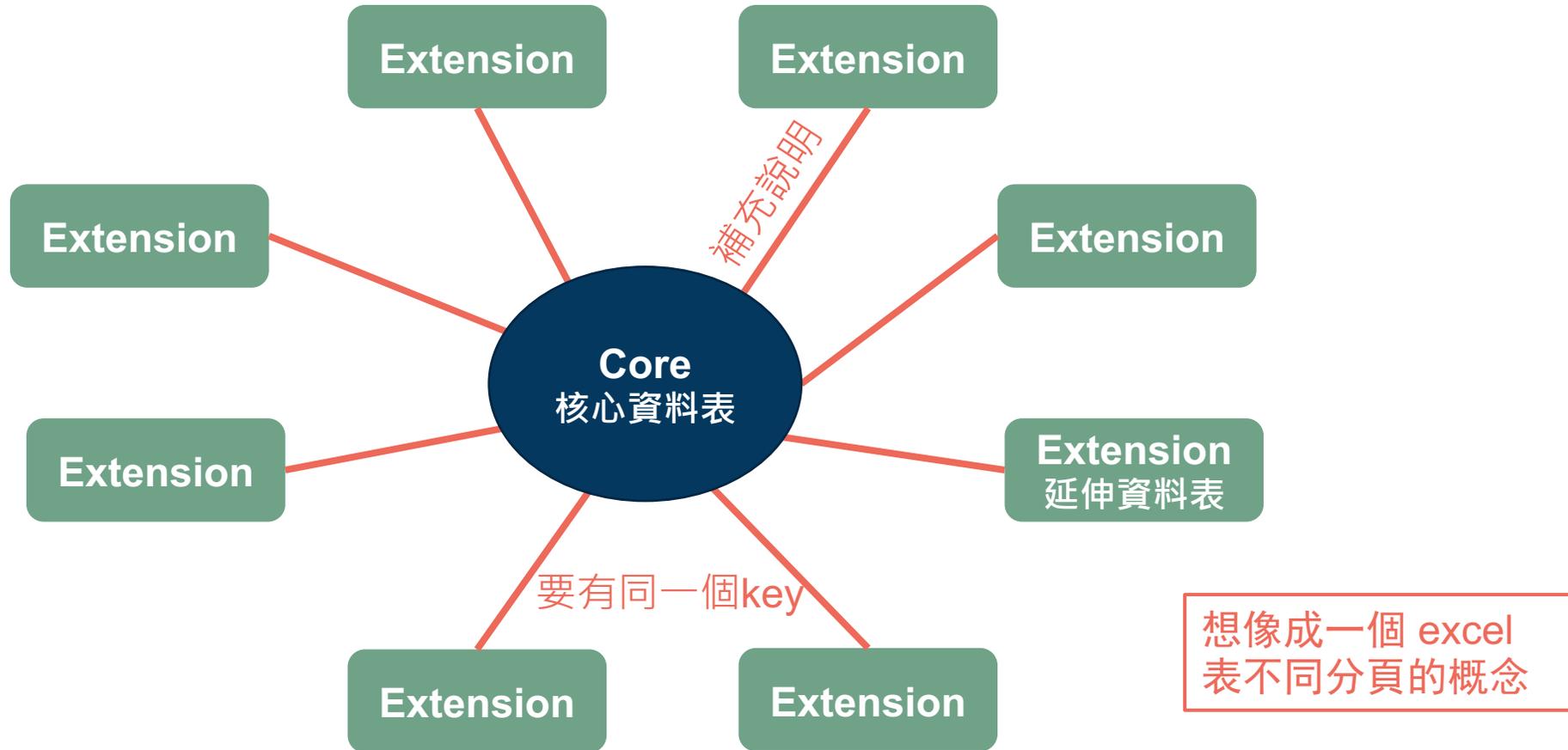
<https://www.guidgenerator.com/>

當我選好了某個資料類型
也填了必填欄位
那我剩下其他的資料怎麼辦？

GBIF的資料結構：Star schema 星芒架構

GBIF 資料結構

Star schema 星芒架構



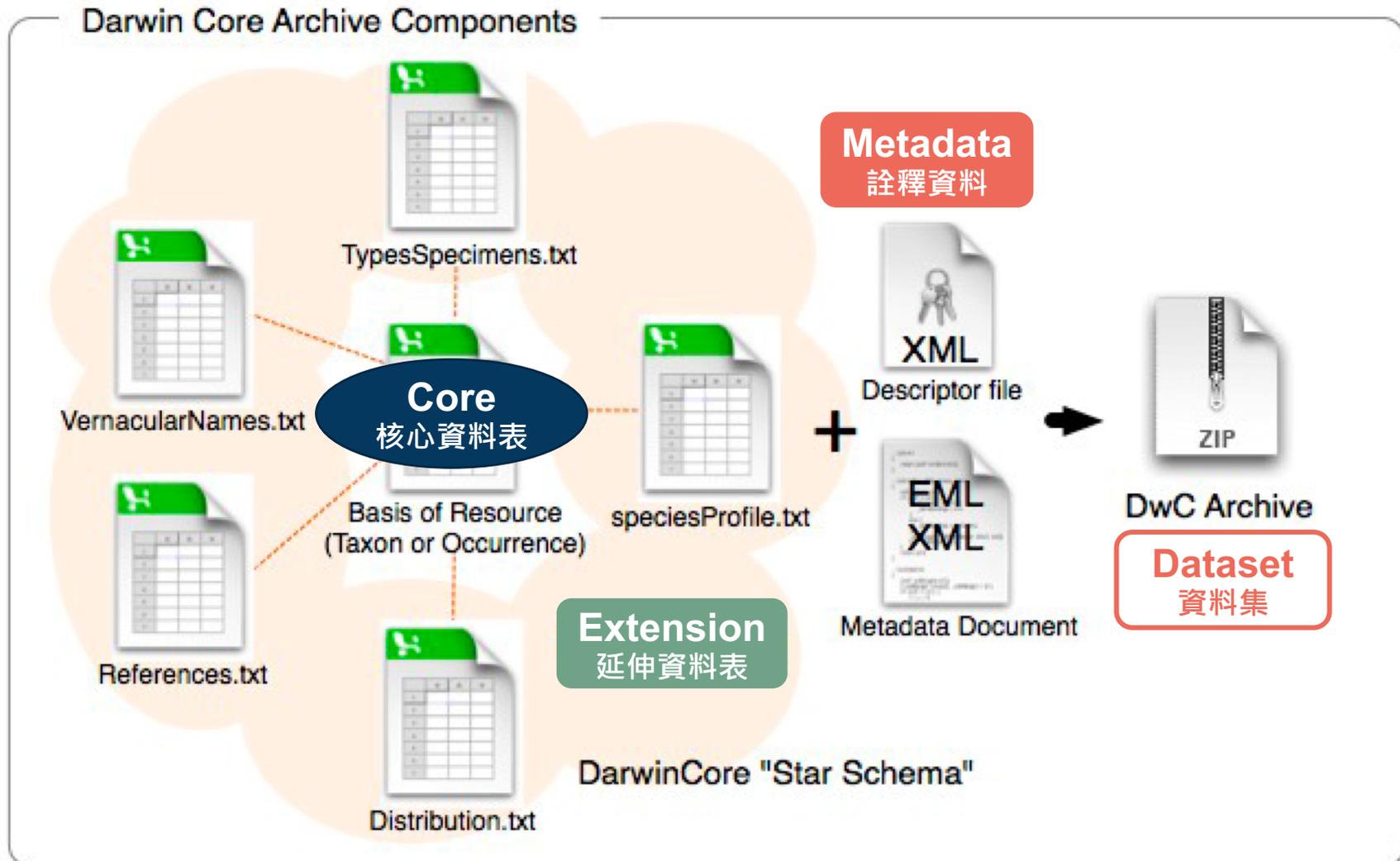
GBIF 資料結構

Star schema 星芒架構



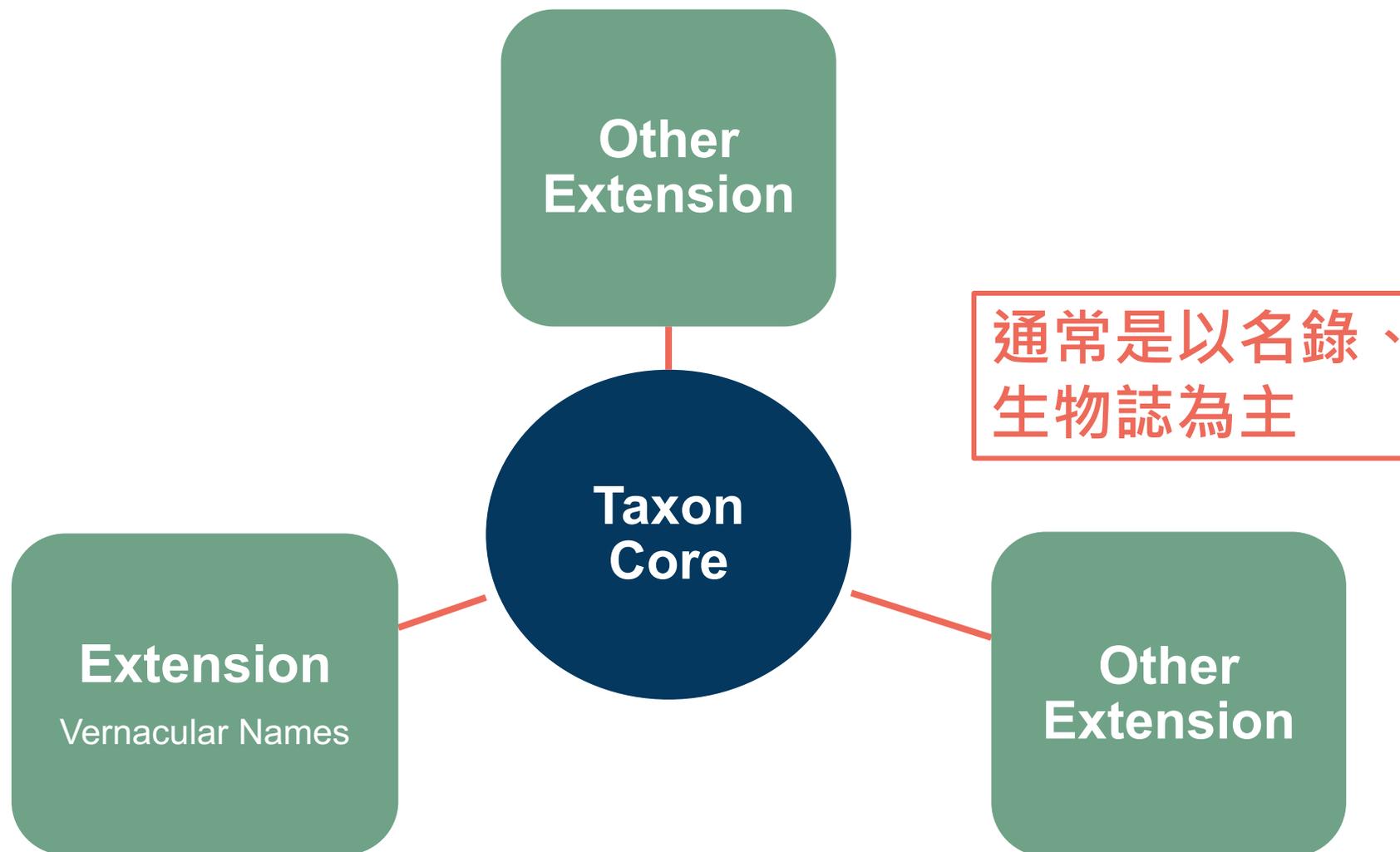
GBIF資料結構

Star schema 星芒架構



GBIF資料結構

Checklist



GBIF資料結構

Checklist

Core

學名

界

門

綱

taxonID	scientificName	kingdom	phylum	class
311809	<i>Geothelphusa luto</i>	Animalia	Arthropoda	Malacost
311808	<i>Geothelphusa lanyu</i>	Animalia	Arthropoda	Malacost
311542	<i>Somanniathelphusa taiwanensis</i>	Animalia	Arthropoda	Malacost
311823	<i>Geothelphusa wutai</i>	Animalia	Arthropoda	Malacost
310883	<i>Birgus latro</i>	Animalia	Arthropoda	Malacost
310823	<i>Caridina pseudodenticulata</i>	Animalia	Arthropoda	Malacost
311822	<i>Geothelphusa wangi</i>	Animalia	Arthropoda	Malacost

Extension

俗名

taxonID	vernacularName
311809	綠島澤蟹
311808	蘭嶼澤蟹
311542	臺灣束腰蟹
311823	霧臺澤蟹
310883	椰子蟹
310823	假鋸齒米蝦
311822	王氏澤蟹

Vernacular Names Extension:

<https://rs.gbif.org/extension/gbif/1.0/vernacularname.xml>

GBIF資料結構

Occurrence Data

e.g. 標本紀錄、野外隨機觀察紀錄、自動相機紀錄

**Occurrence
Core**
(occurrence)

Extension
MeasurementOrFact

GBIF資料結構

Occurrence Data

Core

occurrenceID	rightsHolder	basisOfRecord	scientificName	individualCount	lifeStage	eventDate	locality
AMNH16324	Hsuan-ching Ho	PreservedSpecimen	Halieutichthys intermedius	2	Adult	1944	Near Chauvin, Terre
AMNH16460	Hsuan-ching Ho	PreservedSpecimen	Halieutichthys bispinosus	1	Adult	1944	Near Chauvin, Terre

一筆紀錄可對應不同類型的量測值，
如一隻魚的體長、體重等

Extension

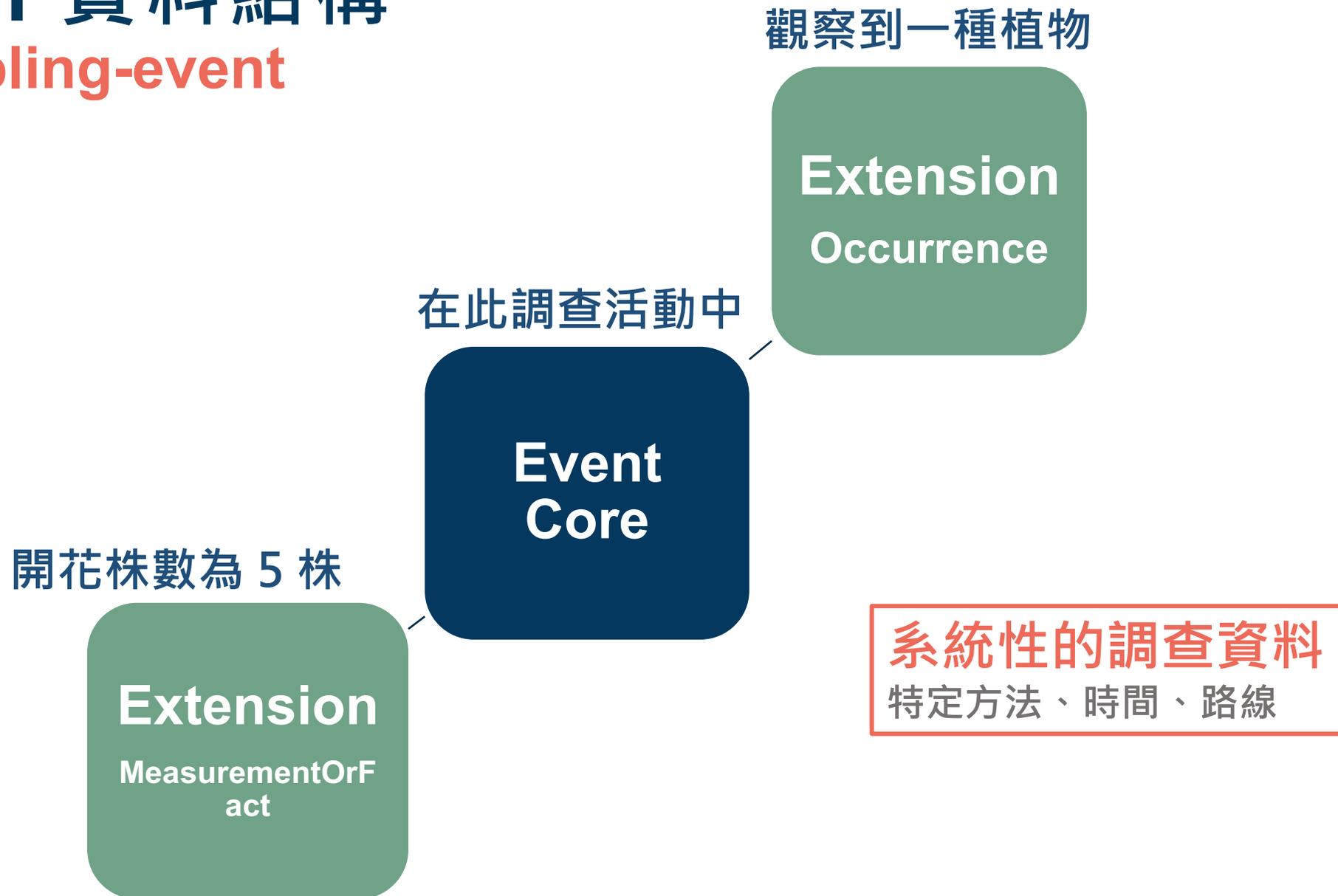
occurrenceID	measurementID	measurementType	measurementValue	measurementUnit
AMNH16324	m001	standard length	45.2-48.3	mm
AMNH16324	m002	total weight	1650	g

Darwin Core Measurement or Facts Extension:

https://rs.gbif.org/extension/dwc/measurements_or_facts_2022-02-02.xml

GBIF資料結構

Sampling-event



GBIF 資料結構

Sampling-event

Core

調查日期 調查方法 採樣量/範圍 採樣量單位

eventID	event Date	sampling Protocol	sample SizeValue	sample SizeUnit
TW_20210827-01	2021-08-27	回播法	200	m
TW_20210827-02	2021-08-27	沿線調查	5	square metre
TW_20210827-03	2021-08-27	自動相機監測法	10	days

Extension

出現紀錄ID

紀錄類型

學名

個體隻數

eventID	occurrenceID	basisOfRecord	scientificName	individual Count
TW_20210827-01	TW_20210827-01_occ001	HumanObservation	<i>Pitta nympha</i>	2
TW_20210827-01	TW_20210827-01_occ002	HumanObservation	<i>Urocissa caerulea</i>	4
TW_20210827-02	TW_20210827-02_occ001	HumanObservation	<i>Macaca cyclopis</i>	6
TW_20210827-02	TW_20210827-02_occ002	HumanObservation	<i>Callosciurus erythraeus</i>	2

一次的調查活動（同日期同樣區同方法），會包含數筆物種出現紀錄，將與物種相關的紀錄分成另一張延伸資料表

Darwin Core Occurrence:

https://rs.gbif.org/core/dwc_occurrence_2022-02-02.xml

Extension 有哪些？

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

GBIF Repository of Schemas

GBIF Registered Extensions

The following extensions are the latest versions of those registered with GBIF for production use.
The list was last updated at 19:31:11 on 21 February 2022, and is also available in [JSON format](#).

Darwin Core Event

The category of information pertaining to an action that occurs at some location during some time.

Name Event
Namespace <http://rs.tdwg.org/dwc/terms/>
RowType <http://rs.tdwg.org/dwc/terms/Event>
Issued 2022-02-02
Keywords

Darwin Core Taxon

The category of information pertaining to a group of organisms (sensu http://purl.obolibrary.org/obo/OBI_0100026) considered by taxonomists to form a homogeneous unit.

Name Taxon
Namespace <http://rs.tdwg.org/dwc/terms/>
RowType <http://rs.tdwg.org/dwc/terms/Taxon>
Issued 2022-02-02
Keywords None



怎麼把現有的資料 對應 Darwin Core 呢？

簡單步驟說明

1

自己的
資料欄位

2

搜尋符合的
DwC欄位名稱

參考 [DwC quick reference guide](#)

3

把欄位名稱換掉
& 加上 ID

注意是否包含必要欄位

4

分成 Core
和 Extension

可拆成不同分頁或不同檔案

5

確認兩者可以
對應到同一個ID

Extension 那裡最好也要有 Core的ID

6

完成！

怎麼把現有的資料 對應 Darwin Core 呢？

來看看實際例子

Checklist (Taxon)

台灣454種鳥類特徵資料庫

Occurrence

台博館生物典藏資料集

台灣海龜目擊紀錄

Sampling event

臺灣農業長期生態監測生物資料集

玉山國家公園SILIC聲音辨識鳥類資料集

生態調查監測資料

覺得剛剛說的不容易記得嗎？ 這裡有小抄！

各類型資料的達爾文核心標準 (DwC) 必填欄位彙整說明
<https://hackmd.io/TmsAwdC6TaGr-Icilwxh3g?view>



Q & A

DwC 常見問題羅列

假設第一次發布資料時選用了 Occurrence Core，後續可再更改為
Sampling-Event Core 嗎？

不行，選了以後不能更動，只能刪除重來

需要為了配合 GBIF 的資料標準而徹底改變目前所用 (且有可能是使用已久) 的表單格式嗎？

能塞的先塞進去

只要不要用 PDF 或 word 來儲存 data，都有機會可以慢慢對應完全

不要用合併儲存格 (很重要！！)

一個資料集中的 taxonID、occurrenceID、eventID 所採用的數值能相同嗎？

不建議，容易miss掉或容易出錯

更新已發布至 IPT 上的資料集時可新增額外的 Extension(s) 嗎？

可以，一個核心資料集(Core)可對應多個延伸資料集(Extension)

當資料中涉及基因資料時應如何處理？

用 DNA 的 extension，或使用associatedSequences這個詞條來附上已開放至NCBI(或其他基因資料庫)的基因序列URL連結

捕捉到蜥蜴後也發現其身上有寄生蟲，此為某研究之重要資訊，該如何記錄寄生蟲之數量與形質資料？

可以用 `ResourceRelationship` 這個 `Extension` 來記錄

兩者間的關係(寄生、被寄生)可用 `relationshipOfResource` 來描述

數量可以用 `individualCount` 記錄

寄生蟲的相關量測數值可以用 `MeasurementOrFact`

要如何保存特定兩筆 `occurrence` 間的關聯性？比如說，目擊一隻蝶類幼蟲正在取食一株植物的葉片，則兩者間「取食者」和「被取食者」間的關聯性可被如何記錄？

回答同上一題

若 **occurrence** 存在與其相應的多媒體資料（圖片、影片、錄音），應如何記錄？

可考慮利用 **Simple Multimedia** 這個 **Extension** 所提供的詞條

(<https://tools.gbif.org/dwca-validator/extension.do?id=gbif:Multimedia>)

Darwin Core Resource Relationship

occurrenceID	type	identifier
occ1	StillImage	https://flic.kr/p/24m4EbL
occ2	StillImage	https://flic.kr/p/EupFem

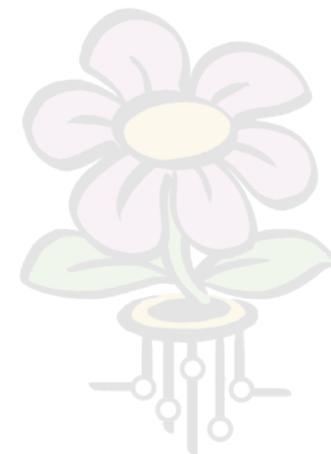
或用 **associatedMedia** 這個詞條來記錄其URL連結

注意，GBIF平台不提供多媒體的倉儲，所以要另外存放，再提供每張照片或影片等的URL連結。免費的多媒體倉儲地方：[depositar 研究資料寄存所](#)、[Zenodo](#)



實作練習時間

每組可互相討論 可請教助教
最後請一位組員分享
到 15:40



示範案例

Occurrence_case1

① 序號	② 紀錄者	③ 目擊的日期	④ 學名	⑤ 描述	⑥ 島嶼	⑦ 影像紀錄	⑧ 緯度	⑨ 經度
1	Sophie Parmelon	2010-01-12	Rhizophora mangle	In bloom	Big Island		20.05529	-154.838
2	Sharon Grant	2010-01-12	Rhizophora mangle	In bloom	Big Island		20.05531	-154.838
3	Sharon Grant	2010-01-26	Clidemia hirta	full bush	Big Island		19.2863	-154.95



① occurrenceID	③ basisOfRecord	② eventDate	⑧ recordedBy	⑨ verbatimLatitude	⑨ verbatimLongitude	⑧+⑨ verbatimCoordinateSystem	④ geodeticDatum	⑤ scientificName	⑤ reproductiveCondition	⑥ island	⑦ associatedMedia
occ_2010-01-12_1	HumanObservation	2010-01-12	Sophie Parmelon	20.05528747	-154.837712	decimal degrees	WGS84	Rhizophora mangle	In bloom	Big Island	
occ_2010-01-12_2	HumanObservation	2010-01-12	Sharon Grant	20.05530765	-154.8379572	decimal degrees	WGS84	Rhizophora mangle	In bloom	Big Island	
occ_2010-01-26_3	HumanObservation	2010-01-26	Sharon Grant	19.2863	-154.9504167	decimal degrees	WGS84	Clidemia hirta	full bush	Big Island	
occ_2010-01-26_4	HumanObservation	2010-01-26	Sharon Grant	19.28633333	-154.9504167	decimal degrees	WGS84	Clidemia hirta		Big Island	
occ_2010-01-27_5	HumanObservation	2010-01-27	Sophie Parmelon	19.44618471	-156.03645	decimal degrees	WGS84	Cryptostegia madagas	purple flowers	Big Island	https://www.flickr.com/photos

要自行新增，
Occurrence 必填欄位

案例練習

1. 選擇 **Occurrence_Luku_sample** 或 **SamplingEvent_bbstaiwan_練習_v202305**
2. 試將這份表單中的資料欄位對應“mapping”到 **Darwin Core (DwC)的欄位名稱**
3. 確認是否需要分 **core** 和 **extension**



我要如何知道我原始資料表單中的欄位要對應到 DwC 的哪個詞條？

1. 趕快趁現在與助教討論
2. 快速查詢 DwC 資料欄位標準英文定義
<http://rs.tdwg.org/dwc/terms/index.htm>
3. 查閱 DwC 資料欄位延伸資料集的內容
<https://rs.gbif.org/extensions.html>