

台灣生物多樣性資料整合之前瞻性研究

賴昆祺¹ 邵廣昭¹ 柯智仁¹ 林永昌¹ 吳信輝³ 陳麗西¹ 許正欣²
中研院生物多樣性中心¹ 中研院物理所² 美國聖路易(斯)大學³

摘要

大量生物多樣性資料轉換成一般常見格式並經由網路被存取已成為未來趨勢，其中又以全球生物多樣性機構（Global Biodiversity Information Facility，GBIF）推動的資料整合最具有代表性，到目前為止已累積全球 1.57 億資料。數位典藏計畫中生物主題小組將資料轉換成 Dublin Core 後設資料，匯入聯合目錄約 24 萬筆；台灣生物多樣性資訊網（Taiwan Biodiversity Information Facility，TaiBIF）整合台灣標本及觀測分布資料達 28 萬筆資料，兩者同為台灣具代表性整合網站。在推動整合台灣生物多樣性資源於內容方面，除繼續整合及提供 GBIF 台灣之標本及其分布資料外，並建議未來參與計劃採用國際上主要之 Darwin Core 及 EML 之 metadata 定期提供原始圖文資料，包括名錄、物種解說(形態、生態、影音、文獻)、觀測分布等資訊。為更有效整合分散於不同單位資料庫，及有效的利用整合後的資源並應用於生態研究、管理及保育甚至供永續利用參考，一套網路資訊系統可以進行跨資料庫展示與分析為重要關鍵因素，本文將生物多樣性資訊學角度，利用 XML 資料交換特性、地理資訊系統之分析 (analysis) 與即時製圖(mapping)應用、與科學工作流程 (scientific workflow) 方法將生態、環境、資訊學家的專業知識且將其流程化自動化，利用地理分布及學名（包含同種異名）整合分散於不同環境及不同格式資料庫，未來對於決策者可容易取得資料物種分布與資料整合結果。

關鍵字 空間決策支援系統、生物多樣性資訊學、數位典藏、

台灣生物多樣性資訊網

一、前言

就台灣特殊的地理環境而言，腹地面積雖不大卻擁有豐富的生物多樣性資源，且特有生物種類亦多，因此需妥善管理及運用這些資源，並建立完善的生物多樣性基本資料庫提供大眾及研究單位人員能更方便地作資料整合及檢索，且未來將可提供予生物研究、教育與產業經濟等等之加值應用，滿足推動生物多樣性工作之需求，冀期達到永續發展及利用。

在過去數十年間，政府部門積極推動各種生物多樣性工作，並投入大量的人力、物力在各式各樣的生物資源的評估、調查與管理模式之開發。然而，多年來累積之各項研究資料與成果，大多以紙本報告的格式散佚在個別研究計畫主持人或各經費贊助單位手中，研究資料保存的永續性與存取的方便性都未臻理想；即使少數近期研究資料已經強調數位化，但或因研究主題相異、調查方法、資料格式不同，導致研究成果無法相互比較，更遑論資料的共享與成果的整合。

據估計全球約有大約 10 億筆以上陸域、水域及海洋生物標本與觀測資料（含未數位化資料），而在 GIBF 在 2010 年工作目標中欲整合 10 億全球生物多樣性資料，而台灣估計也有上百萬筆資料有待整合（邵廣昭，2005）。因此處理此大量生物多樣性資料的工作，逐漸成爲一門學門與共同挑戰，稱爲「生物多樣性資訊學（Biodiversity Informatics）」。生物多樣性的資訊其實涵蓋的範疇很廣，可從基因、物種到生態系。不但資料之屬性、格式與形式多樣化，且使用之語言不同，時間與空間尺度亦不一。再加上缺失資料及智財權之問題複雜，故其資料之整合相當困難且具挑戰性，相較於以基因序列爲主的生物資訊 (Bioinformatics)，只有 A、T、C、G 等核苷酸資料之整合要複雜甚多。

生物多樣性資訊學第一次被提出大約 1992 年¹，主要結合 GIS、GPS、資料庫管理、環境管理、博物館分類系統。後續亦有學者提出爲利用運用資訊工具與資訊技術於生物多樣性，特別是在生物體方面層級；或將物種的原始資料透過資訊技術協助完成資料管理、分析、運算；或將生物個體和群聚的分類資料、生態及基因資料，透過任何形式資訊分享 (Jim Edwards, 2007)。生物多樣性資訊學其精神爲運用資訊工具與資訊技術協助物種的原始資料，達成資料管理、分析、運算，最後再透過網路或其他媒介完成資訊分享。而地理資訊系統正可協助生物多樣性資料進行資料管理、分析，甚至資料分享，因此本文將以生物多樣性資訊學及地理資訊角度出發，將整合生物多樣性資料途徑提出可能

¹ www.bgbm.org/BioDivInf/TheTerm.htm

最佳方法。

二、數典一期的資料整合--聯合目錄

第一期五年的數典計畫是由聯合目錄(<http://catalog.ndap.org.tw>)來整合，所收集之資料來源，除台大、中研院、科博館三個機構外，還有 21 個公開徵選的計畫，收集了很多動植物標本及影像的資料，每一個計畫即有一個網頁，如中研院之魚類、貝類、植物資料庫，科博館的網頁等，總計給聯合目錄整合生物類群資料有 22 萬筆，影像資料大約 24 萬筆。然而國內除數典計畫外，其他單位所建置與累積數位化資料包括標本、分布、文獻及影音資料等，估計有上百萬筆的資料。

在聯合目錄的網頁上雖然可以用類群或字串(種名)來查物種資訊，但其格式設計主要是爲了典藏品的項目(collections)如標本，而非其它更進一步的物種或生物資訊；且其呈現方式並不符合生物領域的習慣。這主要是因爲聯合目錄是要整合「人文」與「生物」兩種屬性很不同的資料，而採用有限的 Dublin Core 的欄位的結果，因此很多欄位無法做適當的對應，如只有標本資訊(含圖片)而沒有 GIS 的分布等資料，也沒有辦法和國際上所用的 Darwin Core 作接軌及整合。聯合目錄的原始格式可維持適用在特色藏品集(collection-level)的呈現上，而無法對應到 GBIF 所用之 Darwin Core。爲解決此問題，我們建議聯合目錄之後設資料應加入 Darwin Core 之欄位(圖一)，也就是去建立一個檢索平台，以便把聯合目錄所蒐集的生物資料可以經由 TaiBIF 提供到 GBIF 上。

```

<MetaDesc>
<Title field="現用組合" dwc="ScientificName">Enolatia moorei (Hutton, 1865)</Title>
<Title field="原組合" dwc="Remarks">Ocinarra moorei Hutton, 1865</Title>
<Title field="原標籤書寫學名" dwc="Remarks">Ocinarra moorei</Title>
<Creator field="鑑定者" dwc="IdentifiedBy">Hutton</Creator>
<Creator field="標籤判讀者" dwc="Remarks">顏聖紘</Creator>
<Subject field="門" dwc="Phylum">Arthropoda</Subject>
<Subject field="亞門" dwc="SubPhylum">Mandibulata</Subject>
<Subject field="綱" dwc="Class">Insecta</Subject>
<Subject field="目" dwc="Order">Lepidoptera</Subject>
<Subject field="科" dwc="Family">Bombycidae</Subject>
<Description field="模式地位" dwc="TypeStatus">Syntype</Description>
<Description field="性別" dwc="Sex">Female</Description>
<Description field="標本模式地位認證者" dwc="Remarks">W. Dierl</Description>
<Contributor field="採集者" dwc="Collector">Hutton</Contributor>
<Type field="紀錄類型" dwc="BasisOfRecord">昆蟲標本影像</Type>
<Format field="影像類別" dwc="Remarks">標本背頁</Format>
<Identifier field="影像代碼" dwc="CatalogNumberNumeric">103</Identifier>
<Source field="館藏地點" dwc="Remarks">大英自然史博物館昆蟲系 NHM</Source>
<Source field="數位化單位" dwc="Remarks">國立中山大學生物科學系 NSYSU</Source>
<Language field="語言" dwc="Remarks">中文</Language>
<Language field="語言" dwc="Remarks">英文</Language>
<Relation field="學名發表文獻" dwc="Remarks">
  Hutton, 1865 Trans. ent. Soc. Lond. (3) 4: 326</Relation>
<Coverage field="模式產地" dwc="Locality">Himalaya: Masuyi</Coverage>
<Rights field="授權單位" dwc="Remarks">國立中山大學生物科學系</Rights>
</MetaDesc>

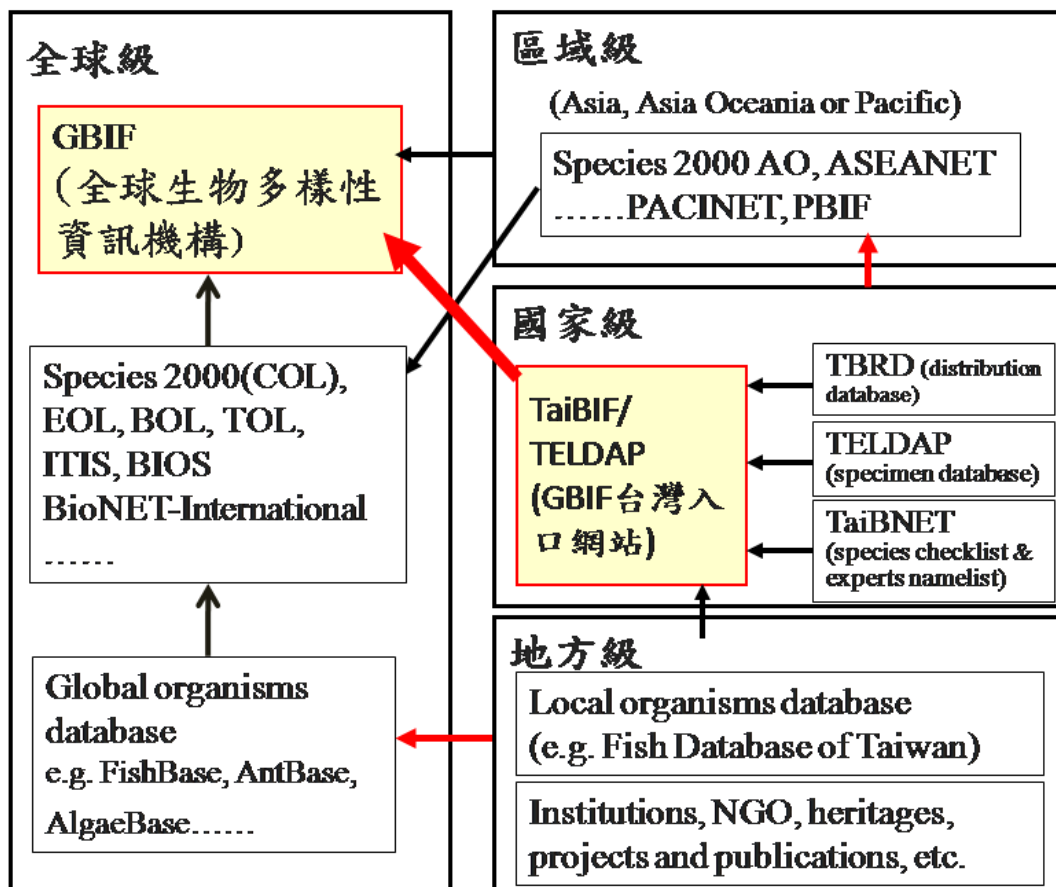
```

marking-up with
DwC attributes

圖一、聯合目錄之後設資料應加入 Darwin Core 之欄位

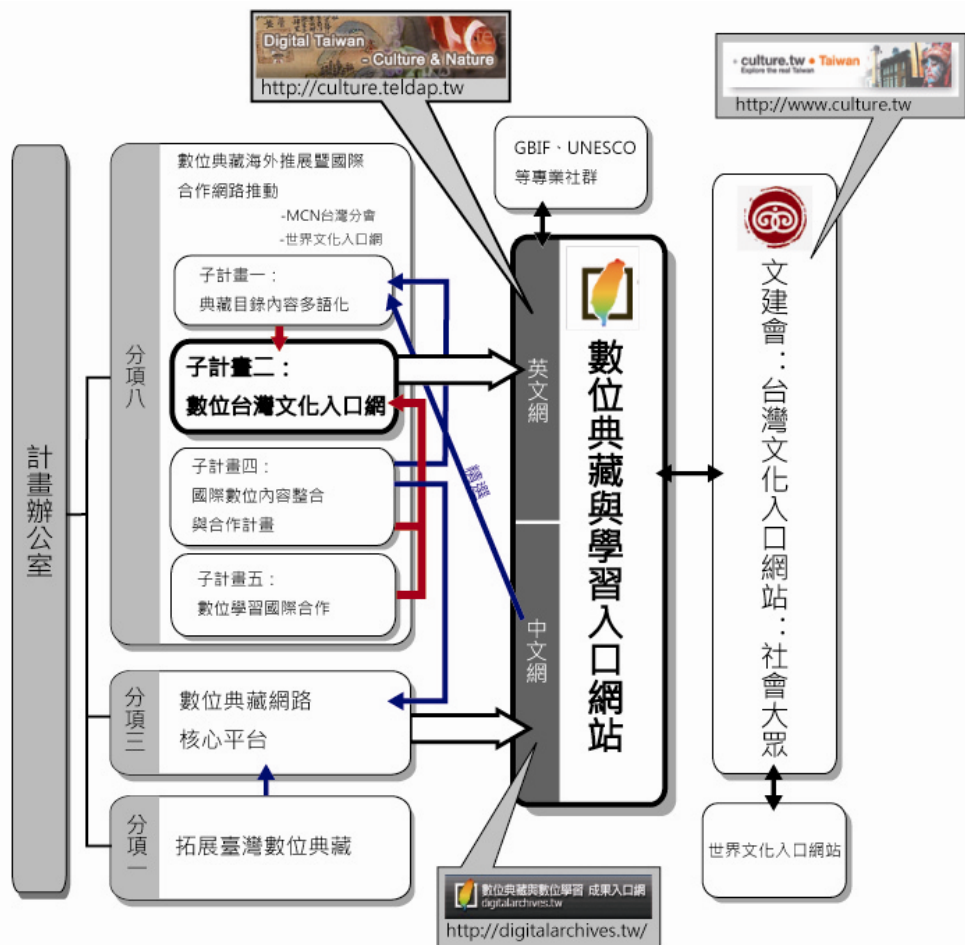
三、 數典二期資料之整合—英文國家入口網

為加強國際合作，我們在二期數典計畫中的生物部份把 TaiBIF 的工作系統納入，並同時納入 TaiBIF 所要求的 Darwin Core 之必填欄位，以便與國際上之 GBIF 或 EOL 等接軌。事實上，除了由 TaiBIF 轉到 GBIF 的第一條路外，也可以透過第二條路即區域網路和國際合作，這些都是 GBIF 的副會員；或是透過第三條路，即與不同生物類群和各類群之全球資料庫如 FishBase 或 AntBase 等合作，經由這些全球物種資料庫(GSD)而把台灣本土或特有種之資料送到國際(圖二)。



圖二、整合台灣生物多樣性資訊網並與國際接軌

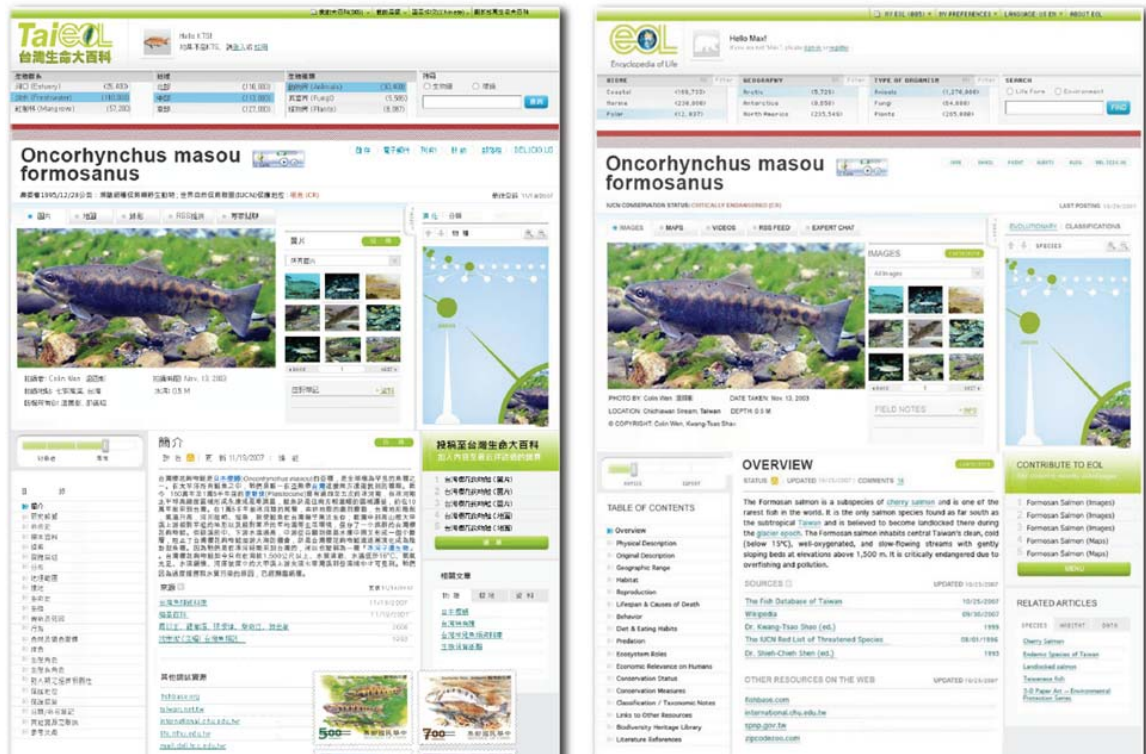
圖三是數典二期計畫的架構及流程圖。二期計畫共有八個分項計畫，國合是第八分項，下有3個子計畫，我們負責生物資源之整合及國際英文版網站之建置。所有數典計畫的成果會先彙整到聯合目錄，但大多是中文的資料，其文化部份的資料除了另有「數位島嶼」及「數位101」的展示外，還會經由國合分項子一計畫來篩選並做翻譯後，再交給子二計畫納入「數位台灣-文化與自然」之網站，即以專業社群為對象的國家入口網來與國際接軌。又文建會另有預算在推動對象為一般民眾之「數位台灣文化入口網」，亦為英文版，未來會與數典之國際入口網互相合作分享資料。



圖三、數典二期計畫的架構及流程圖

四、「線上生命大百科」(Encyclopedia of Life, EOL)

EOL 已在 97 年 2 月 26 日正式公布其首批 3 萬種，每種一頁，共 3 萬頁的資料，預計在 2017 年完成全球 180 萬種之電子百科全書。首批主要是魚類，此乃因 FishBase 已是全球最成功、資料最完整之物種資料庫，做起來最容易。我們已與 FishBase 合作多年，是 FishBase 的重要合作伙伴。我們也打算與 EOL 合作，提供台灣生命大百科中台灣特有種的網頁。譬如「台灣櫻花鉤吻鮭」之樣張，左邊是中文，右邊是英文，內有生態照、原始描述、文獻、相關報告、甚至鈔票及郵票之圖案等(圖四)。



圖四 中英文版台灣生命大百科示範頁面

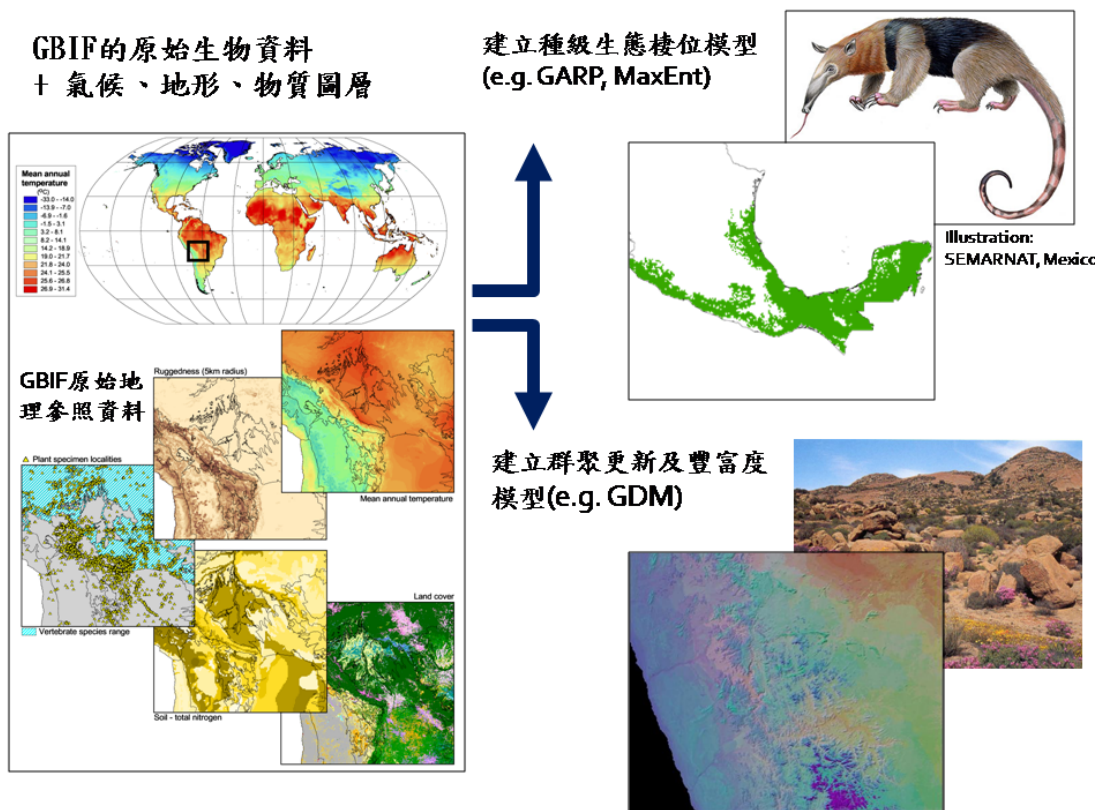
EOL 之資訊架構主要區分為兩個部分，第一部份為資訊科技開發部分，利用各類最新的網路應用科技，例如：Web2.0 相關概念與 AJAX 等。以 EOL 入口前端呈現部分為例，EOL 即是利用 Ruby on Rails 與其他開放原始碼的軟體進行開發。第二部份為物種資料整合部份，EOL 基於共享的精神，並無限制合作之對象，但因合作單位之資料類型迥異，EOL 也提供了相關的資料整合機制，依據合作對象的資料蒐集程度提供相對應的解決方案。

表 1、EOL 建議之解決方案

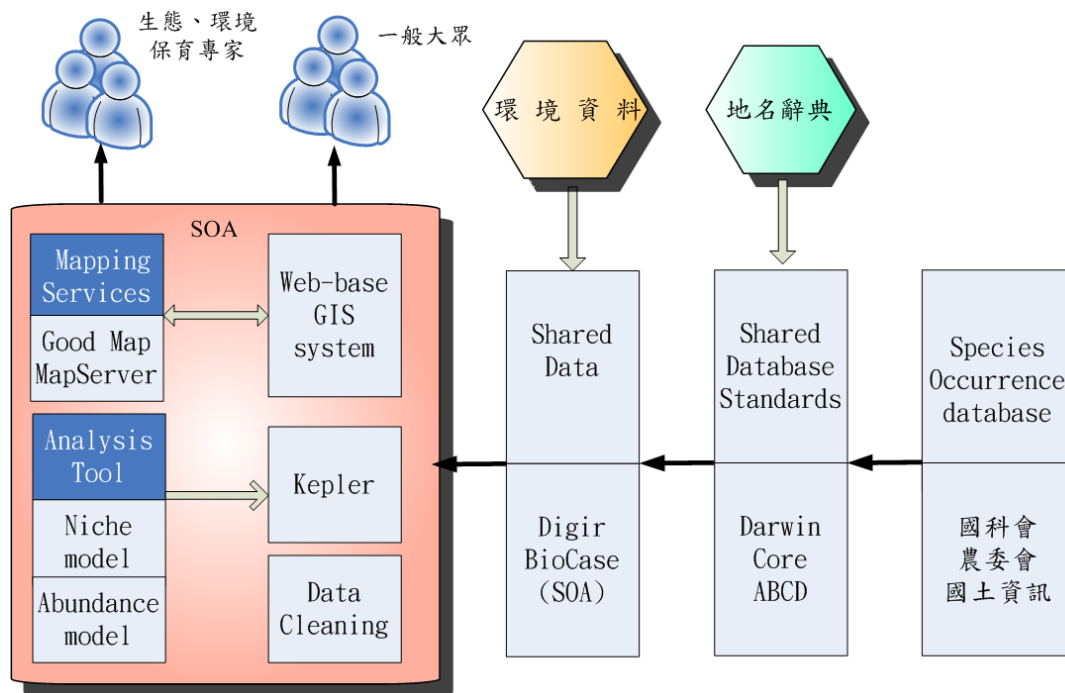
資料有無	尚未有資料	已有資料
蒐集平台有無		
無資料蒐集平台	建議以開放原始碼之 PHP 內容管理系統 Drupal 為資料蒐集平台，EOL 亦會提供相關建置之協助與支援	交由 EOL 資訊小組中之內容管理與整合小組進行轉換
有資料蒐集平台	利用該蒐集平台將資料蒐集完成後，依據 EOL Taxon Resource Transfer Schema 轉換成 EOL 需要之資料格式	依據 EOL Taxon Resource Transfer Schema 轉換成 EOL 需要之資料格式

五、生物多樣性整合應用

GBIF 國際上 1.57 億筆原始分布資料，運用 GIS 及一些國際上正在開發運用之模式，由物種群聚之層次結合環境及地理資訊因子資料來作分析及預測，使資料倉儲能真正應用到產生知識及制訂政策等，以達到物種減少喪失、資訊永續利用的保育目標（圖五），這些也是目前生物多樣性公約（CBD）及 GBIF 正在大力推廣與努力的目標。台灣是生物多樣性之島，人為破壞因素亦多，可說是國際上一個很好的研究案例地區，正好符合 GBIF 在推動 CBD 2010 目標之 campaign。以此為整合台灣生物多樣性資料為出發點，研究架構如圖六所示，其過程中可能會面臨問題與挑戰，及擬解決方法如下：



圖五、GBIF 資料應用方向



圖六、研究架構圖

(1) 大量資料整合問題

在全球約 10 億筆以上生物多樣性資料（含大部分資料未數位化），除了加快數位化流程外，另一項重大問題為地理性空間資料，據估計大約有 5% 的數位化資料具有空間性資訊（Beaman *et al*, 2003）。整合不同來源之資料，針對地理性資料欄位亦可能出現不同格式資料，如測站名稱可能包含古地名與現今地名、或採用不同座標系統，依調查資料會有不同比例尺及解析度，如小從單一物種的調查，一個小地點、生態棲地、縣市、國家，甚至到大到如大洋洲...等，面對不同的解析度會有不同調查結果資料。

物種發生資料（species occurrence data）大部分為分散在各研究計畫、單位。這些資料必須透過單一的資料格式如 Darwin core 或 ABCD 等慣用格式將其初步整合，再透過分享工具如 DiGIR（Distributed Generic Information Retrieval）或 TAPIR（TDWG Access Protocol for Information Retrieval）將其資料轉換成 XML 格式，以達資料交換與分享的目的。

(2) 具科學 workflow 生態棲位模擬

1980 年代中期，隨著電腦軟體如 BIOCLIM 的開發，利用氣候等環境資料進行物種分布模式的概念逐漸實現。自此出現許多模式方法及軟體，包括廣義線

性模式、廣義累加模式(GAM, Generalized additive model)、遺傳演算法(GARP, Genetic Algorithm for Rule-set Production) ...等等 (Chapman, A. D., 2005)。由於過去環境資料有限，早期的研究只能探討某類動植物的大尺度分布狀況，例如澳洲眼鏡蛇類的研究(Longmore, 1986)，侷限於可用的軟體及環境資料，導致研究進度緩慢，單單為某一物種進行模式就耗時數個月，也因研究範圍較大，結論常過於廣泛。

隨著新的軟體開發後，加上環境資料圖層品質的改善，因此有更多時間深入研究單一物種或研究更多物種。分布模式主要為利用物種原始發生資料搭配物種、環境模型，可以辨認出更多的物種分布地。物種分布模式也利用博物館藏及新的調查資料來預測馬達加斯加的爬蟲類多樣性，並成功預測新的變色龍出現地點(Raxworthy *et al.*, 2003)，諸如此類也越來越常見。本文將有效利用空間資料特性，將生物資料與環境資料進行連結與分析，利用多種生物分布模式建立生物的時空分布分析模組。

生態調查或分布資料通常具有空間與時間性，如空間上的點、線、面（調查點、線段與分布範圍）的特性。早期的生態資料蒐集，對調查對象的時空記錄通常較為粗糙，以描述性質為主；近年來全球定位系統（GPS）之普及，大大地提升了生物空間分布資料的精準度。而另一方面，遙測與生物觀測技術的精進，再加上現有的環境資料庫與資訊系統更趨完善，使得生物分布資料日趨豐富。

生態棲位模擬（ecological niche modeling）為結合生態學與棲地的環境因子，利用地理資訊系統及空間統計方法，預測物種分布可能性。而生態棲位原理為生態學家設定棲地環境條件，進行物種的可能性分布預測，近年來逐漸成為分析軟體之一，GBIF 從 2005 年至今共舉辦 4 場研討會以推廣，有此可見其重要性，而 MAXENT 更是 GBIF 大力推廣軟體。

工作流的定義：「工作流為描述一群流程之中，每一個個別的工作任務以及各個任務之間的相互關係，(Ailamaki *et al.*, 1998)」。具體而言一個工作流就是一連串的個別工作任務的組合，形成單一的流程，可降低成本與可以隨需求應變的彈性之優點。從早期使用資訊技術使工作流程自動化的概念開始到目前，工作流則是著重在與網際網路的標準結合上，如：Web Services Flow Language 標準（吳信輝，2006）。

科學工作流程為將科學研究過程流程化，以標準語法建立檔案，而工作流程技術之應用可幫助科學家分析數據資料。目前支援地理資訊之科學工作流程軟

體，以 ESRI model builder，及 open-source 且專注於生態領域的 Kepler 為多人使用，本文亦透過地理資訊服務標準與生態分析模組，利用科學工作流的概念加以整合，簡化過去需要複雜操作程序及專業知識才可完成研究。另一個導入科學工作流的目的，在於利用知識管理的方法將重複的一些流程重新設計成單一的元件，利用組合的方式來形成新的知識流程，減少重複流程降低人力負荷。

Kepler 為一套科學數據分析軟體，利用圖形化模式及網路服務（web service）取得資料，並將專家的知識融入行程工作流，透過科學工作流程，將生態、環境、地理科學家針對分析預測工作流程以標準的語法建立程序，讓系統自動化完成資料分析與整合功能，且這樣工作是可以重覆與分享的，可達成資料共享、資訊整合與知識管理之最終目的。

（3）地理資料網路服務（Geospatial Web Service）

在執行生物多樣性分析時，除了提供正確物種分布資料外，另一個重要資料為相對環境資料，然這部分地理性資料，則必須透過國土資訊系統 Web service。技術配合服務導向架構（Services Oriented Architecture，SOA）作為環境資料整合資料提供者。

服務導向架構是一種新興的系統架構模型，針對企業需求組合而成的一組軟體元件，透過服務註冊程序將各單位既有空間資訊資料分享，建立服務並公開服務以供資料使用者可快速取得所需資料，利於資料之加值應用（衷嵐焜）。

過去GIS圖資多半以檔案的型式存在，在資料傳遞過程中，無法使用如Web Service的機制加以擷取，這對資料的流通相當不便；另一方面，許多地理資訊系統必須在其特定的環境下執行，且不同軟體所需之格式竟不相同，然在使用這些不同格式空間資料時，必須透過轉檔的程序，而轉檔的過程中經常發生漏失與錯誤訊息，OGC的成立便是要解決此類問題。OGC成立至今訂定與多標準，在眾多其制定之標準格式中，本文建議以WFS及WMS作為整合不同空間資訊的標準。下表為簡易對照表(賴昆祺，2002)：

表二、WFS與WMS比較表

	WFS	WMS
Input	Vector(點、線、面)	Raster or Vector(點、線、面)
Output	XML為主資料（GML）	Image(jpeg、GIF...等)

環境資料除了可從國土資訊系統中獲得外，地球觀察組織(Earth Observations)裡的團隊正努力提供一個全球性的平台(The Global Earth Observation System of Systems, GEOSS)，其中收集環境資料，可以用來支持地

球永續經營，未來大量環境資料亦可同樣透過 Web service 方式獲得。

進行生態模式分析時，使用 SOA 架構分別從國土資訊統及 GEOSS 獲取資料外，亦會採用同樣的架構，將分析結果同樣使用 SOA 提供外界參考使用。

六、結論

數位典藏一期計畫中(2002-2006)共有台大、科博館及中研院三個機構及 21 個公開徵求計畫，聯合目錄中已蒐集及累積約 22 萬筆動植物之圖文資料，透過本文研究將 Dublin Core 欄位加上 GBIF 及國際慣用之 Darwin Core，並透過數位典藏英文入口網將資料整合勾勒出完整途徑。GBIF 成功整合國際上 1.57 億筆原始分布資料後，開始運用 GIS 及模式運用，由物種群聚之層次結合環境及地理資訊因子資料來作分析及預測，使資料能真正應用到產生知識及制訂政策等，以達到物種減少喪失、資訊永續利用的保育目標，本文亦在相同目的下，提出結合模式及科學工作流，簡化過去需要複雜程序及專業知識才可完成分析結果，正好符合 GBIF 在推動 CBD 2010 目標之 campaign。

參考文獻

1. 賴昆祺、廖滋銘、范毅軍(2005)，開放式 GIS 標準於數位典藏整合之應用與前瞻，第三屆兩岸三院資訊技術與應用交流研討會，中國大陸 海拉爾。
2. 邵廣昭、彭鏡毅、嚴漢偉、賴昆祺、王明智、林永昌、李瀚、陳元憲(2005)，台灣生物多樣性資料庫之整合及與全球資訊接軌，2005 年 9 月，2005 自然物標本與生物多樣性資料庫整合國際研討會，台中科學博物館。
3. 衷嵐焜，導入 SOA 於國土資訊系統之規劃架構與實作示範，http://ngis.moi.gov.tw/get_file.aspx?file_name=20070528100539.pdf&folder=edu_train/2007050117064500/&file_id=20070528100539763。
4. 吳信輝，馮正民（2006），應用 Workflow 概念與 Web Services 技術於國土資訊系統資訊整合，2006 年台灣地理資訊學會年會學術研討會，台北
5. Ming-hsiang Tsou (2005), The Big Changes in Internet GIS, GIS development, http://www.gisdevelopment.net/magazine/years/2005/oct/webgis_tsou44_1.htm
6. Paul Flemons, Robert Guralnick, Jonathan Krieger, Ajay Ranipeta and David Neufeld(2007), A web-based GIS tool for exploring the world's biodiversity:

The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA), *Ecological Informatics*, Vol. 2, Issue 1, 1 January 2007, Pages 49-60.

7. Benjamin D. Best, Patrick N. Halpin, Ei Fujioka, Andrew J. Read, Song S. Qian, Lucie J. Hazen and Robert S. Schick(2007), Geospatial web services within a scientific workflow: Predicting marine mammal habitats in a dynamic environment, *Ecological Informatics*, Vol. 2, Issue 3, October 2007, Pages 210-223.
8. Kwang-Tsao Shao, Ching-I Peng, Eric Yen, Kun-Chi Lai, Ming-Chih Wang, Jack Lin · Han Lee(2007), Yang Alan, Shin-Yu Chen, Integration of Biodiversity Database in Taiwan and Linkage to Global Databases, *Data Science Journal*, Vol. 6 2007 Pages. S2-S10.
9. Chapman, A. D. (2005) Uses of Primary Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
10. Chapman, A. D. (2005 a). Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
11. Robert P Guralnick, David Neufeld (2005) , Challenges Building Online GIS Services to Support Global Biodiversity Mapping and Analysis: Lessons from the Mountain and Plains Database and Informatics project. Vol. 2, 2005, *Biodiversity Informatics*.
12. Longmore, R. (1986) , Atlas of Elapid Snakes of Australia. Australian Flora and Fauna Series No. 7. Canberra: Australian Government Publishing Service.
13. Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Otrega-Huerta, M.A. and Peterson, A.T. (2003), Predicting distributions of known and unknown reptile species in Madagascar. *Nature*. 426: 837-841
14. Jim Edwards (2007), The future of biodiversity informatics: GBIF, the Encyclopedia of Life and beyond, GBIF Science Symposium.