林業「大」數據—以植物標本館資料為例

陳建文[1]

大數據有多大?

「大數據」(big data)一詞起源於1990年代，二十幾年來已成為大家耳熟能詳的一個資料科學名詞。它被應用於各個層面領域，無論是經濟學、社會學、資訊科學、醫學，當然也包括自然科學。大數據的誕生，可說是起因於資訊設備的進展與普及速率。由於各類感測器 (sensor) 等資料觀測裝置的廣泛佈建，透過這些裝置長期且持續地收集累積，所得的資料逐漸累積成為各種巨量資料集。

一般來說，能被稱為「巨量資料」的資料集，往往並非幾百MB或是幾GB的容量，而是動輒數百TB (Terabyte, 1 TB = 1,000 GB)，甚至於是PB (Petabyte, 1 PB = 1,000,000 GB)、EB (Exabyte, 1 EB = 1,000,000,000 GB)以上的規模。這麼龐大的資料量已非一般電腦軟體所能處理或運算的了，而是必須透過更高階的資訊科學技術及軟體，去剖析這些巨量資料集的內涵，此即為巨量資料科學的任務。在林業科學領域中，研究人員長期有計畫地進行資料調查收集，乃至後來透過各式感測器或感測/遙測技術，進行持續性的精密記錄所獲得的生物及棲地環境資料，其規模雖然不及前面所提到的資訊容量等級，但這些透過研究計畫所取得的資料集，其中同樣地蘊含著大量的資訊脈絡，等待著研究人員去進行分析解讀，產出科學成果。

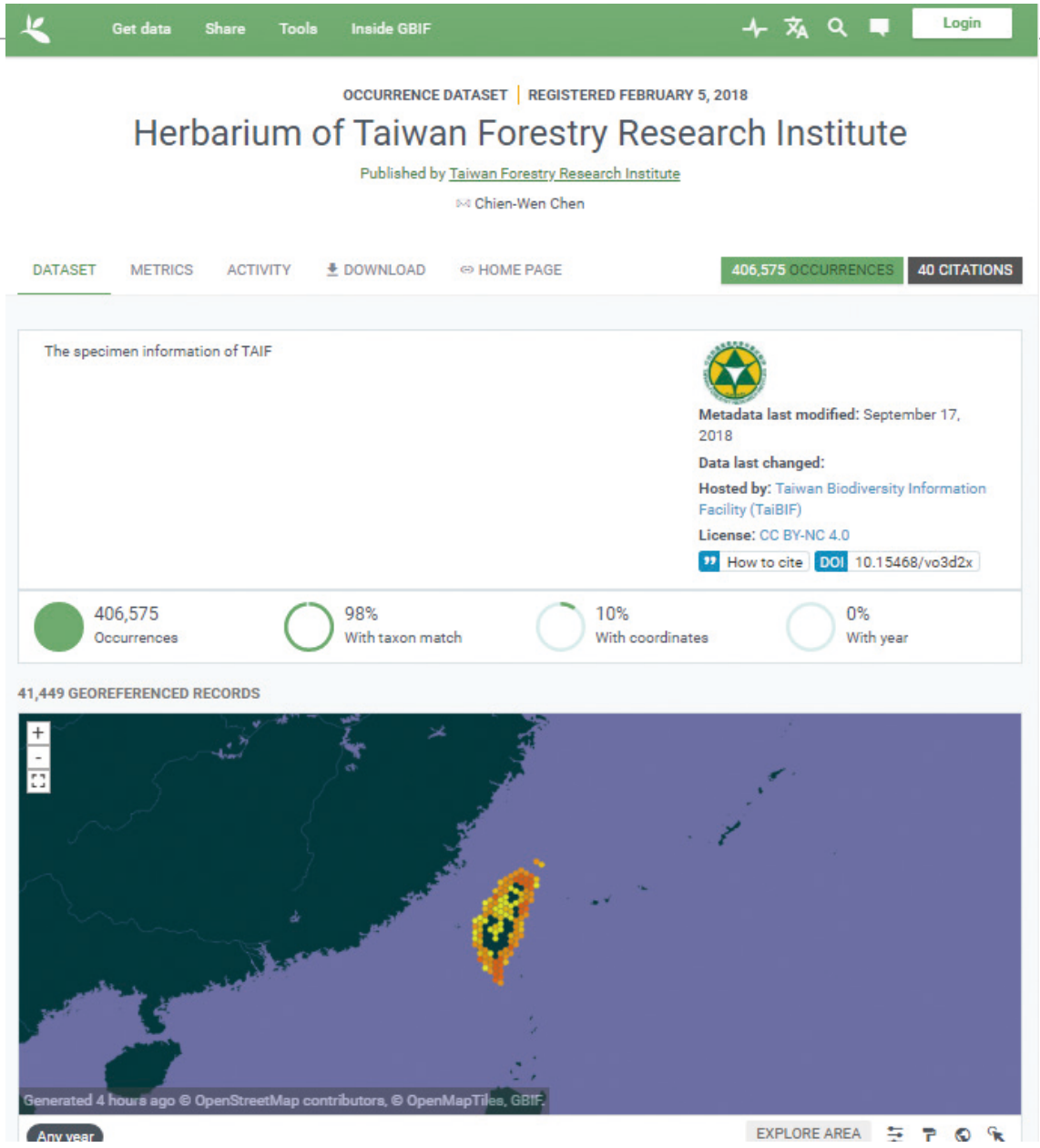
林業研究中的大型資料

其實，林業科學領域中也還是有著資料容量可觀的巨量資料集。例如林地航照圖，就是一種資料量十分龐大，需要具備高速運算能力的電腦才能進行處理的巨量資料集。此外，林業試驗所植物標本館(以下簡稱標本館)藉由標本資料庫所產製出的生物觀測資料(occurrence data)，透過國際生物多樣性資料機構的彙整，至今也逐漸成長為頗具規模的大型資料集。

林業試驗所植物標本館的資料

植物標本(plant specimen)是植物學研究工作上不可或缺的基礎材料，標本館的任務即為持續地收集、典藏並管理植物標本，這些標本是研究材料的存證，並供後續的分類學與植物多樣性保育研究使用。而為了長期而有效地保存標本資料，標本館從1990年代開始嘗試著手植物標本的數位化作業。一開始是先從標本採集資訊的建檔開始進行，針對採集人員所記錄下來的採集及棲地資訊進行完整輸入，後來歷經網路資料庫改版，以及參與國家型數位典藏計畫……等階段，至今已完成超過43萬筆標本資訊的建檔。每一筆植物標本紀錄，包括：館號、鑑定學名、採集人員、日期、採集地點、各項生育環境……等欄位。此外，像是標本學名的每一筆訂正歷程、採集者額外的紀錄事項，或是

[1] 林業試驗所・植物園組



林業試驗所植物標本館於國際生物多樣性資料機構GBIF的資料提供頁面。截至今年為止，共有406,575筆資料已經發佈公開於網際網路上供各界瀏覽。

館方為了管理所做的備註文字等，也都有對應的欄位可供建檔人員進行記錄。

除了標本採集資訊之外，另一項標本館所進行的數位化作業為植物標本影像的數位建檔。由於植物標本影像的數位化，必須採取平放且非接觸的方式進行，且成像規格

(影像解析度及畫質等)也不能太低，需要倚靠規格較為特殊的影像數化設備才能達成。因此，影像數位化作業是從2004年才開始進行。影像掃描一開始採用了相當於解析度300 DPI 的專業影像設備來進行影像數位化。直至今年，標本館影像掃描設備進行了提升，改以解析度600 DPI 的掃描器進行影像掃描。由於解析度提升為原來2倍，因此資料擷取量也增加為原來的4倍，也就是說，每一張標本影像的原始圖檔大小，由原本的約50 MB增加到超過200 MB。而為了因應圖檔資料容量的提升，標本館也必須持續擴充資料倉儲設備(現階段已擴充至220 TB)，用以滿足未來持續成長的數位影像資料容量。此外，標本館網站中的標本數位影像瀏覽，也將由原本的小尺寸圖檔，改為直接提供原寸大小的圖檔，並可於網站介面中進行放大、縮小及平移檢視，以提升使用便利性。

植物標本館資料的應用與流通

典藏標本的資料建檔作業，除了提供資料保全以外，其實也具有便於進行資料發布流通的優點。而促進標本資料的交流及使用率，也是標本館一項十分重視的目標。前面所提到的國際生物多樣性資料機構GBIF (Global Biodiversity Information Facility)，自1999年創始以來，至今已累積來自1,471個資料提供機構(data publisher)、超過13億筆的生物觀測資料，其中也包括了標本館所上傳的40多萬筆標本採集資料(主要為台灣的採集紀錄)。藉由如GBIF這樣的資料彙整機構，研究人員可以免去分散蒐集資料所花費的時間與精力，只需在單一網站中，搜尋符合研究目標的生物種群或分布區域，即可快速而便利地取得所需資料集以進行後續研究，對於研究能量及資料流通率的提升皆有助益。

除了標本採集/分布地點資訊的資料集以外，標本數位影像的應用，也是標本館經常對外提供的服務。藉由高解析度的數位影像，研究人員可直接在電腦螢幕上，針對如枝條上的細毛或是蕨類植物葉柄的微小鱗片等構造進行觀察。此外，大量的標本數位影像，也可以提供做為電腦程式的植物特徵影像分析研究，或是人工智慧的深度學習訓練材料之用。學術界其實已有許多實際的應用案例，都是使用各大植物標本館的標本影像作為研究的基本資料來源。

結語

隨著各項科學研究的進行，產出的研究資料也持續地累積成長。要如何應用這些產出的科學研究資料，除了倚賴研究人員的頭腦及強大的資料分析工具之外，完善的資料倉儲管理，也是提升研究資料使用率，甚至是延續資料壽命不可或缺的機制與措施。資料倉儲管理並非只是提供電子資料儲存空間供人存放資料而已，其精髓在於研究資料內容的詳盡描述，也就是後設資料(metadata)的建立。完備的後設資料加上完整的原始資料(raw data)，才可以稱得上是一份具可用性及保存性的研究資料集。透過描述完備的研究資料集，研究人員才可以將更多來源的研究資料交叉整合，用以發掘更新面向的科學問題解答，如此才得以發揮大型數據科學的威力，創造更多的研究機會。