大數據時代的林業研究新典範

林朝欽[1]

前言

林業是一門生物學與生態學為主的應用學科，也是一門長期經營的自然資源管理學門，所以林業一直以來就是一個以數據為根據的學門。二十世紀以來因電腦與資訊科技的發展，數位數據呈現爆炸式的成長速度，所謂的「數據密集型科學(Data-Intensive Science)」成了數位時代科學典範。數據密集型科學的特徵是數據量與種類的不斷成長，以及從數據蒐集、獲取、分析(量大、內容複雜、流量高)而衍生出的科學研究新方法，來源包括衛星和航遙測、各式儀器、感測網以及人為的觀察。由於數據本身獲取容易與快速性，往往超過我們處理(驗證、儲存、分析、歸檔)的能力，因此，必須依賴數據科學的各種工具及技術來協助處理。除此，21世紀以來「大數據(big data)」這個新名詞出現後，已經影響到各行各業，甚至與國家的經濟發展、健康照顧、能源永續、公共安全、及國家安全連結在一起。大數據指的是數據量(volume)、速度(velocity)與複雜性(variety)的多維數據內容。

因為全球氣候變遷產生的環境惡化和改變問題，生態學的基本理論成為解決環境問題必須依賴的工具，林業本身是應用生態學之一，長期以來已累積大量且極具價值的數據，但很明顯的仍無法有效面對氣候變遷與森林生態系變化間不可預測的關係，因為傳統林業研究過於分散、異質性高、短期性過多，使得大多數的數據都無法整合，加上以往對數據的管理缺乏系統，造成很多數據描述不足難以重新使用，或未進行倉儲而不斷流失。因此，在大數據興起後，林業必須加以改變，以大的時間與空間尺度、多團隊與長期性，以及網絡與網絡的連結合作方式，才能探討物種、環境及物種與環境間的關係。因此，林業要真正成為深、廣、遠的學科，必須連結大數據與新的資訊工具。而林業研究者不但要接受這個挑戰，而且在研究方法與資訊使用和管理上要作出調整，建立使用巨量、多樣、與快速累積數據之能力。

林業研究新典範－資訊學方法

大數據與數據密集型科學的興起促使環境相關研究產生變革，因此，林業研究者必須把資訊科技視為與林業研究有密切關連，並將兩者結合成為一個新的典範(new paradigm)，這個新典範稱之為「資訊學方法(informatics)」。對林業研究者來說，資訊科技已不只是數據存取或資訊查詢而已；讓研究數據獲得可持續再利用及創新用途的價值，才是林業研究未來解決氣候變遷與森林生態系間不可預測的困難。例如，每年颱風頻率與強度改變，涉及傳統研究與新科技介入的擴大調查能力，尤其在數據收集與獲取上更突破以往無法達到的優勢，數據蒐集的正確性也大幅提高。又如感測器網路擴展了傳統森林昆蟲生態研究調查的能力，一項東

[1] 林業試驗所・森林保護組退休研究員

方蜜蜂與虎頭蜂的關係研究，顯示了資訊學方法在大數據影響下的變革，以往以人力觀察東方蜜蜂與虎頭蜂關係，現改以無線網路攝影機，每分鐘取得一張的影像比傳統每一天取得幾張影像來得精細，一天可以蒐集840張照片，一年即可累積達30萬張影像數據，這是人力無法進行的，再加上自動判視軟體協助，得出傳統方法無法發現的結果。

資訊學方法的核心所在是什麼？其一，即是透過政策促進數據分享與管理，以增加數據的價值及瞭解，與大數據蒐集、分享與使用所涉及安全與學術倫理。若要促進數據管理，研究數據必須以任何軟體都可以讀取的文字檔格式(text format)，及詳細描述數據的標準與編輯工具。例如：當今使用最普遍的文字檔格式XML(Extensible Markup Language)，就是一種穩定且一致性強的標準，不會因軟體改變或不同而有讀取的問題。另外，數據管理必須能長期保存，資訊學方法宜採用公共數據倉庫及交流中心。這些數據倉庫及交流中心的目的是為長期保存研究數據而設，研究人員可以將他們的數據傳輸到這些倉庫中。這些系統也保證長期開放、異地備份與更新，所以數據不會因為軟體更新而不能讀取。例如：林業試驗所已建立15年的研究資料目錄即是一個可與世界接軌的數據倉庫。另外資訊學方法強調數據倫理，當使用別人的數據時需堅守某些原則，就像我們處理自己所蒐集的研究數據一樣。最後資訊學方法強調建立新的文化，主張研究者有分享數據的文化，如此才能在社群內立足與獲得良好的名聲。

資訊學方法的第二個核心是林業研究者須具備處理與大數據有關的新技能與知識。根據2012~2016年美國對環境相關科系研究生的調查發現，80%的研究生沒有受過或修習過正式的資訊處理及使用的課程，74%的研究生沒有任何電腦程式語言編寫能力，72%的研究生不知道描述數據的元數據(metadata)是甚麼。在線的資深或資淺林業研究人員是否已具備資訊的新技能與知識，影響林業研究至為巨大。2005年林業試驗所開始進行研究數據倉儲計畫，當時大多數研究人員是抗拒的，十幾年下來資深研究員有多少是自己描述研究數據更是一個疑問。更甚者農委會、科技部至今還沒有研究數據倉儲系統，更別說研究人員普遍使用開源軟體R或Python來進行數據整合分析。因此，如何提供工具、協助研究人員處理與整合各種得到的數據能力是未來應努力的方向。

為了彌補與提升科學研究人員的資訊新技能與知識缺陷，美國國家生態分析整合中心(NCEAS)自2014年起定期開辦一系列的資訊新技能與知識訓練，從2天到3個星期，針對年輕學者、研究生提供課程與實作的工作坊。表1是過去5年間提供的資訊新技能與知識內容。

新典範案例

新典範相關的林業研究的第一個例子是森林動態樣區的設立。森林生態系的變動除了涉及樹種與樹種的關係，1980年代開始有熱帶森林的大樣區研究，由美國史密斯研究所(Smithsonian Tropical Research Institute)所領導的森林動態研究，其在巴拿馬建立了第一個50公頃的森林樣區進行研究，之後這項研究組織成為國際研究網。2009年，林業試驗所邀集4個國家來臺灣，探討森林動態樣區數據如何完整建檔、倉儲、存取、與分析使

表1 美國國家生態分析整合中心所提供的資訊新技能與知識訓練

資訊技能	軟體使用	應用範圍
研究歷史記錄保存	Git	分散式版本控制系統
命令式作業系統	Linux(Ubuntu, Fedora…)	研究數據分析與整合
程式語言	R, Python, Java, XML, SQL	研究數據管理、運算、分析
資料庫	MySQL, Postgresql	研究數據管理
伺服器	Apache, Tomcat, LDAP,	研究數據倉儲、分享
科學工作流程	Kepler, Apache Taverna	研究數據整合、建模、分析

用。利用臺灣、馬來西亞、日本、波多黎各等四個國家的動態樣區數據，產出包括森林動態樣區數據管理的觀念性架構外，並建立了資料庫、認證界面、元數據(metadata)查詢網頁與三個科學工作的分析流程。這個例子展示了資訊學方法的兩個核心，因為利用資訊學方法倉儲管理數據，所以研究人員可以進行更深層的科學研究；並訂定數據分享政策，嘗試建立研究團隊的數據倫理文化。此外，也展示了可以讓更多樣區加入共同倉儲與分享的資訊架構，例如：使用此一架構，整合臺灣已有的16個不同大小森林樣區，再一起分享數據，圖1是2012年建立的臺灣森林動態樣區倉儲與分析的測試網站。

新典範相關的林業研究的第二個例子是植物的分布預測。林業試驗所植物標本館典藏的臺灣植物標本成立於1904年，至今已有百餘年歷史，為臺灣最古老的標本館，目前館藏標本總數約43萬號，其中包括日治時代典藏至今的標本近30,000餘份，以及植物命名發

圖1 新典範的林業研究案例-臺灣森林動態樣區倉儲與分析的測試網站。

表文獻所引證之模式標本1,800餘份。這些豐富的數據是物種分布預測模式不可少的基礎數據。預測物種分布的模式是從自然史博物館所發展出來的，模式與分布預測的方法使用到物種調查或標本採集資料，及環境地理資訊圖層加以建構。使用資訊學方法，將數據由倉儲到整合，透過網路服務可以方便的結合，並加入地理資訊的環境圖層，只要建立整個運算流程，系統可以自動完成使用者所要求的預測結果。2008年林業試驗所利用國土地理資訊系統計畫完成此一應用，圖2是此新典範的成果網站(http://ngis.tfri.gov.tw)。

結語

林業研究在大數據新文化的影響下，面對挑戰已出現新的研究典範資訊學方法，其核心強調數據管理，及開發可以幫助林業研究者詳細描述其數據的標準、分析、整合的資訊新技能與知識，並且有許多工具可資運用；另外，資訊學方法強調建立的分享文化，主張林業研究者必需有分享數據的研究倫理，以在社群內立足與獲得良好的名聲。更甚者，在大數據的快速發展影響下，林業也正在改變傳統的研究方法，朝向更廣的尺度、更整合的內容、更依賴大型的數據倉庫、自動化數據蒐集的方向前進，參與在這樣改變下的林業研究者必須要採取行動來因應新的挑戰，才不會被時代淘汰。

(參考文獻請逕洽作者林朝欽，email: linchauchin@gmail.com)

圖2 新典範的林業研究案例-物種預測的數據整合。