生態學在巨量資料下的新視野

林朝欽[1, 2]，陸聲山[1]

[1]林業試驗所森林保護組；[2]通訊作者 E-mail: *chin@tfri.gov.tw*

[摘要] 巨量資料的概念起源於 1990 年代，並成為 21 世紀以來的新名詞，巨量資料快速的發展，已與國家的經濟發展、健康照顧、能源永續、公共安全、及國家安全連結在一起，巨量資料的興起，使得我們必須面對未來會影響我們日常生活、工作環境與思想的實質轉變。在巨量資料興起後，生態學的研究趨勢也改變為大時間與空間尺度、多團隊與長期性，生態學者不但要接受這個挑戰，而且在研究方法與資訊使用、管理上要作出調整。生態學家必須把資訊科技視為與生態學研究有密切關連，並將兩者結合成為一個新的典範，這個新典範稱之為生態資訊學方法，強調數據管理與分享。

關鍵字：生態、數據、資訊、系統

New Perspective in the Ecology of Big Data

Chau-Chin Lin[1, 2] and Sheng-Shan Lu[1]

1Division of Forest Protection, Taiwan Forestry Research Institute; 2Corresponding author E-mail: *chin@tfri.gov.tw*

ABSTRACT First coined in the 1990s, “big data” has become a phenomenal concept as the world entered 21th century. The rapid development of big data has influenced the modern society, linking to national economic development, healthcare, energy sustainability, public safety, and national security. Big data is a revolution that influences our daily living, thinking, and working. In big data environment, ecological researches have shifted to large spatial scale, multiple team works, and long-term scale. Ecologists have to accept this challenge to change their research methodology as well as the usage and management of information to adapt in the big data era. Ecologists need to incorporate ecological research with informatics to a new paradigm which is called ecoinformatics, emphasizing data management capacity and sharing.

Keywords: ecology, data, information, system

前言

巨量資料(big data, 亦稱大數據)的概念起源於 1998 年的數據探勘討論，2001 年多維的(three-dimensional (3D))概念被具體的提出來，所謂的多維數據指的是量、速度與複雜性(volume, velocity, and variety) (Diebold 2012)。從此巨量資料成為 21 世紀以來的新名詞並快速的發展。為因應這個新的領域，2012 年 3 月美國歐巴馬政府啟動巨量資料研究與發展政策，並宣佈投入 2 億美金執行這個政策，也成立了網路與資訊技術研究發展(The Networking and Information Technology Research and Development Program, NITRD)

計畫，其目的是為了不斷成長的數位數據所帶來的許多挑戰，例如國家安全。2013 年 11 月美國的公私團體在這個政策的支持下，提出「從數據到知識到行動」(data to knowledge to action)運動，把巨量資料與國家的經濟發展、健康照顧、能源永續、公共安全、及國家安全連結在一起，這個運動歷經兩年的討論與規劃。2016 年 5 月 NITRD 提出美國聯邦政府巨量資料研究發展策略計畫(The Federal Big Data Research and Development Strategic Plan)，標舉這個策略計畫的主要目的是為迎頭趕上國家在明日世界仍然保持有創新的競爭力(NITRD 2016)。可見巨量資料是進入 21 世紀以來很重要的一項革命，巨量資料的興起，使得我們必須面對未來會影響我們日常生活、工作環境與思想的實質轉變，巨量資料一詞已不再只是資訊科技界的專有名詞，它成為當今社會大眾朗朗上口的詞彙，也影響了各行各業(Mayer-schonberger and Cukier 2013)。

生態學在生物學領域裡算是年輕的學門，但因為全球氣候變遷產生的環境惡化、改變問題，生態學的基本理論成為解決環境問題必須依賴工具，生態學原本就是數據科學之一，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生態系變化間可預測的關係，因為傳統生態學研究過於分散、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(Michener and Jones 2012, Michener 2015)。因此，在巨量資料興起後，生態學的研究趨勢也改變為大時間與空間尺度、多團隊與長期性，例如長期生態研究(ILTER)、國際森林動態學研究(CTFS-ForestGEO[1])、國際湖泊生態研究(GLEON[2])、全球生物多樣性資訊機構(GBIF[3])網絡興起，以及網絡與網絡的連結如國際地球觀測系統之系統(GEOSS[4])，不論這些系統是針對環境還是生物或是生態系，它們都屬於生態學範圍，探討物種、環境及物種與環境間的關係。因此，生態學將真正成為深、廣、遠的環境問題研究基礎，並連結巨量資料成為環境保育與永續發展不可少的工具。

一、數據密集型科學

巨量資料為什麼影響到各行各業？在第四典範《The Fourth Paradigm: Data Intensive Scientific Discovery》這本書中清楚告訴我們這個時代是以數據為中心的時代(Hey *et al.* 2009)，所謂的數據密集型科學(Data-Intensive Science)就是這個時代科學典範，它的特徵是數位數據呈現爆炸式的成長速度，我們以 Megabyte (MB)來看微軟的辦公室軟體(MS Word)編輯的一篇論文的長短或數位相機拍的一張照片的解析度；網際網路發展後我們在網路上可以看影片並以 Gigabyte (GB)來衡量影片的播放時間；我們目前用 Terabyte (TB)來看個人電腦的硬碟容量，但這些都還不足以描述數據密集型科學的面貌，國際數據資訊(International Data Corporation, IDC)在 2013 年發表全球的數位數據量時用的是 Zetaabyte (ZB)，2013 年全球數位數據量約有 4.4ZB，IDC 預測 2020 年時全球數位數據量會達 44ZB。一 ZB 如果換算成個人電腦硬碟容量等於一兆多 GB (1ZB=1,099,511,627,776 GB)。數據密集型科學除了上述的特徵外，它也是一種從數據蒐集、獲取、分析(量大、內容複雜、流量高)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，及獲取容易與快速，並超過我們處理(驗證、儲存、分析、歸檔)能力。我們可以用不同的生態學研究來說明數據密集型科學的模樣，以及它與巨量資料的關係。

以美國長期生態研究網(US LTER[5])為例，美國長期生態研究網為了解美國生態環境在多重因子影響下，大陸性與區域性尺度變化的趨勢、如何解釋區域性的變異、以及這些影響因子對未來所產生的結果所進行的 EcoTrends[6]計畫(Michener and Jones 2012)。此計畫針對分布於全美國的 28 個野外站、大尺

度的國家氣候數據中心、國家海洋與大氣觀測中心歷年所蒐集的數據進行整合，他們以生態元數據語言(Ecological Metadata Language, EML)作為數據描述的共同標準，引用語意聯結(Ontology of Ecology Observation, OBOE)進行分析(Madin *et al*. 2007)，把不同來源的數據加以連結後，彙整成圖形化來解釋想探究的問題，例如海平面上升與海面溫度變化的全國性趨勢均呈上升，但在此趨勢下的區域性的變異。從這個例子，我們看到生態學研究面對巨量資料與數據密集科學的年代，生態學者不但要接受這個挑戰，而且在研究方法與資訊使用與管理上要作出調整，建立使用巨量、多樣、與快速累積數據之能力。

二、生態研究新典範－生態資訊學方法

巨量資料與數據密集型科學的興起促使生態學研究產生變革，因此，生態學家必須把資訊科技視為與生態學研究有密切關連，並將兩者結合成為一個新的典範 (new paradigm)，這個新典範稱之為生態資訊學方法(林朝欽等 2008)。對生態學家來說，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，才是生態學研究未來解決氣候變遷與生態系間不可預測的困難(Rüegg *et al*. 2014)。例如，感測器網路擴展了傳統生態研究調查的能力，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。一項東方蜜蜂與虎頭蜂關係的研究，顯示了生態資訊學方法在巨量資料影響下的變革，以往以人力觀察東方蜜蜂與虎頭蜂關係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(陸聲山等 2009)。

生態資訊學方法的核心所在是什麼？2016 年美國聯邦政府巨量資料研究發展策略計畫提出七個策略(NITRD 2016)，其中的第四與第五策略為：透過政策促進數據分享與管理以增加數據的價值，及瞭解與巨量資料蒐集、分享與使用所涉及的隱私、安全與倫理；這兩個策略正是生態資訊學方法的核心。

首先要促進數據管理，生態資訊學方法採用任何軟體都可以讀取的文字檔格式(text format)，及開發可以幫助生態學家詳細描述他們的數據的標準與輸入工具，這些工具還具有防呆功能以減少輸入錯誤。例如，當今使用最普遍的文字檔格式稱為 XML(eXtendible Markup Language)就被生態資訊學方法所取用，這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，生態資訊學方法採用開發共用的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，它們以 20-100 年的保存目標作為系統的設置基準，不同的倉儲系統可能因應不同領域或特殊計畫而設，但都有共同的特點，就是系統間可以透過開放的標準互相分享，研究人員可以將他們的數據加上描述數據的元數據送到這些倉庫中。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。除此這些系統也提供搜尋功能。倉儲系統通常是一個獨立系統，當許多獨立系統出現後，為了擴大分享，有一種只屬於資訊交流的系統被建構出來，這些系統把分散的倉庫透過網路服務(web service)功能串聯起來。例如，建立在美國的地球觀測數據網(DataONE[7])就是串接全世界生態研究的數據交流中心。又如全球生物多樣性資訊機構(GBIF)是一個全球性的數據倉庫，於 2001 年成立，迄目前為止共有 816 個資料發布單位，累計蒐集超過 6.5+ 億筆物種分布資料[8]。

其次，要促進數據分享。生態資訊學方法強調數據倫理(Hampton *et al*. 2013)，數據倫理主張使用別人的數據需堅守某些原則，就像我們處理自己所蒐集的研究數據一樣，我們不希望自己的數據未經我們授權同意的狀況下被

拷貝、抄襲或偷竊發表；所以，當你使用別人分享出來的數據時，應該引用來源；這也是許多期刊目前處理分享數據的一個作法。另外也強調建立新的文化，主張生態學者必需有分享數據的研究倫理才能在社群內立足與獲得良好的名聲，如果不能做到，除了可能面臨法律問題外，經費補助機構也可能給予處罰。美國長期生態研究網設置了生態資訊管理委員會(Ecological Information Management Committee)，負責規劃設計各研究站的數據倉儲與分享規範。目前，美國長期生態研究網的數據入口網有 41,556 份研究數據分享(Porter 2010)。經由生態資訊學方法所產生的巨量資料相關的研究，在國際長期生態研究網已經出現，以下是兩個很典型的例子。

三、巨量資料相關的生態研究

巨量資料相關的生態研究的第一個例子是森林(Lin *et al.* 2011)：森林是一個動態的系統，森林生態系的變動除了涉及樹種與樹種的關係，也涉及族群內與族群間的競爭，另外傳統生態學的生態棲位(niche)理論，認為不同物種因分化而有分布的不同假說，但在熱帶雨林中生態棲位理論受到挑戰，還有為最有效檢定森林生產問題也須要知道森林的變化。於是，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，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，迄今有 60 個大型森林樣區參與，觀測 6 百萬棵樹，一萬個樹種，並且每個樣區每 5 年要更新一次觀測資料。

在國際長期生態研究網的東亞太平洋地區(EAP-ILTER)及美國長期生態研究網有四個這樣的森林動態樣區，原本各自分析調查數據，並沒有共同的標準進行倉儲與分享，為了促進森林動態樣區數據管理與應用，2009 年在臺灣舉辦了森林動態樣區數據管理與應用的國際研討會。探討森林動態樣區數據如何完整建檔、倉儲、存取、與分析使用。研討會利用臺灣、馬來西亞、日本、波多黎各等四個動態樣區的數據，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、元數據查詢網頁與三個科學工作的分析流程。

這個例子展示了生態資訊學方法的特點，因為利用生態資訊學方法倉儲管理數據，所以生態學家可以進行更深層的科學研究。但這個研究如果沒有訂定數據分享政策，加上研究團隊遵守數據倫理恐怕是無法達成的。這個例子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，讓森林動態樣區與巨量資料結合。

巨量資料相關的生態研究的第二個例子是湖泊(Porter *et al.* 2012)：全球的湖泊共同的面臨取用水資源、集水區調整，優氧化、魚類捕撈、外來入侵種壓力。在人類人口不斷增加的趨勢下，這些壓力不大可能減少；加上氣候變變遷改變湖泊生態系統的衝擊，監控、分析各別的湖泊，提供了預測未來湖泊變化的最好機會；再擴大到全球湖泊，我們就可以比較了解湖泊在人類與自然的影響下的反應。但如何才能監控湖泊的變化？湖泊學最基本的監測是透過感測儀(sensor)量測水溫、溶氧等物理與化學的因子。既然是儀器，就可以很密集，很多點，也很快速的蒐集到數據；這正是巨量資料的所謂多維面向數據的意義。2004 年美國維斯康辛大學的長期生態研究站與臺灣鴛鴦湖長期研究站首先展開了一項先驅研究，分別在兩地架設感測網，並透過網際網路連結，並利用生態資訊學方法，數據以每分鐘的頻率紀錄、傳送、倉儲到分析與分享。迄今它成為國際湖泊生態研究網(Global Lake Ecological Observatory Network)，六大洲有 34 個國家 60 座湖泊參加網絡，它不只是湖泊網絡，也是研究人員網絡，全球有 500 個生態學家參與其中。

從這個例子顯示了生態資訊學方法的另一個特性，它可以結合不同領域的研究人員，

共同設計出來的分享數據的模式，使感測器快速、密集、大量的數據能真正結合生態研究與資訊科學來滿足研究的需要。

四、巨量資料與未來的生態學研究

在巨量資料的影響下生態學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、自動化的數據蒐集的方向前進。例如，康乃爾大學鳥類研究所啟動的「公民科學」－eBird (NABCI 2011)，每天全世界自認為公民科學家的鳥類愛好者即時的把觀測到的鳥類數據透過網際網路送進資料庫，這些數據配合遙測技術可以讓生態學者進行預測鳥類分布，及鳥類在全球暖化影響下的遷徙型態與改變。雖然，公民科學所得到的數據可能有瑕疵，但這種打破傳統的數據分享，讓前面所提到的生態資訊學方法更具有發展的潛力；因此，我們更需要工具來確認公民科學的數據正確性與可用性，我們也更需要分析巨量資料的軟體來處理這種不斷湧入資料庫的即時數據。又例如，美國所啟動的國家生態觀測網(National Ecological Observatory Network, NEON)，這種進行大尺度的觀測的數據不但公開，並且還提供工具，協助研究人員處理與整合各種觀測儀器所得到的數據，這也是傳統生態學研究所沒有的。

生態學除了上述屬於社群內的改變，在社群外的其他領域的研究者，在大量公開分享的數據資源提供下，也不斷的增加生態學議題的探討。地球物理學者逐漸增加生態研究，他們藉由公開分享的環境及物種數據，運用遙測技術探討生態過程與樣態的研究，已經打破了以往只有生態學家才進行的課題。例如物候(phenology)研究就是一個明顯的例子(Hampton *et al.* 2013)，美國地球物理聯合會(American Geophysical Union, AGU)年會中物候的論文高達 40%。同樣的，生態學家也加入跨領域的研究中，這也是拜巨量資料新的文化之賜。

巨量資料所開創的新文化已促使生態學進入資訊時代，參與在這樣的時代生態學者必須要採取下列的行動來因應新的挑戰(Hampton *et al.* 2013)：

為後人整理、記錄、保存數據，如此才能讓自己的研究數據有效的再使用。

把數據送進倉庫或交流中心分享，只有數據分享才能將不同來源的數據整合起來，使得新型態的生態研究變得可能，更一般性的生態結論的獲得及更容易了解大尺度的生態現象。

合作進行研究，如此才能探討更深、廣與長期的生態問題。

認真看待數據管理的重要性，並與同儕切磋數據管理實務，學會使用數據管理軟體。

結論

生態學在大尺度環境問題解決的知識空缺上，是有很重要貢獻的領域，尤其是在巨量資料的新文化影響下，面對巨量資料的挑戰，生態學研究已出現生態資訊學提供新的典範。生態資訊學方法的核心強調數據管理，並採用任何軟體都可以讀取的文字檔格式，及開發可以幫助生態學家詳細描述他們的數據的標準與工具；另外，生態資訊學方法也強調建立新的文化，主張生態學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。在巨量資料的影響下，生態學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、自動化的數據蒐集的方向前進，巨量資料已開創新的文化，參與在這樣的改變下，生態學者必須要採取行動來因應新的挑戰。

引用文獻

林朝欽、鄭美如、陸聲山。2008。生態資訊學之發展與應用回顧。台灣林業科學 23(Supplement):S1-10。

陸聲山、林文智、陳永修、林朝欽。2009。無線感測網應用於動物行為研究。國家公園

學報 19(1):1-8。

Diebold FX. 2012. On the origin(s) and development of the term ‘big data’. PIER Working Paper No. 12-037. Available at SSRN: http://ssrn.com/abstract=2152421 or http://dx.doi.org/10.2139/ssrn.2152421.

Hampton SE, CA Strasser, JJ Tewksbury, WK Gram, AE Budden, AL Batcheller, CS Duke and J Porter. 2013. Big data and the future of ecology. *Frontier Ecological Environment* 11(3):156-162.

Hey T, S Tansley and K Tolle (eds.). 2009. The fourth paradigm: data-intensive scientific discovery. Microsoft Corporation.

Lin CC, AR Kassim, K Vanderbilt, D Henshaw, EC Melendez-Colom, JH Porter, K Niiyama, T Yagihashi, SA Tan, SS Lu, CW Hsiao, LW Chang and MR Jeng. 2011. An ecoinformatics application for forest dynamics plot data management and sharing. *Taiwan Journal Forest Science* 26(4):357-69.

Madin JS, S Bowers, MP Schildhauer and MB Jones. 2007. Advancing ecological research with ontologies. *TREE* 23(3):159-168.

Mayer-Schonberger V and K Cukier. 2013. Big data: a revolution that will transfer how we liv, work, and think. Eamon Dolan/Houghton Mifflin Harcourt.

Michener WK and MB Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *TREE* 27(2):85-93.

Michener WK. 2015. Ecological data sharing. *Ecological Informatics* 29: 33-44.

NABCI (North American Bird Conservation Initiative). 2011. The state of the birds 2011 report on public lands and waters. US Department of the Interior.

NITRD (Networking and Information Technology Research and Development Program). 2016. The federal big data research and development strategic plan. The Networking and Information Technology Research and Development Program. Available at https://www.nitrd.gov/Publications/PublicationDetail.aspx?pubid=63 Accessed on June 12 2016.

Porter JH. 2010. A Brief History of Data Sharing in the U.S. Long Term Ecological Research Network. *Bulletin of the Ecological Society of America* 91:14-20.

Porter JH, PC Hanson and CC Lin. 2012. Staying afloat in the sensor data deluge. *TREE* 27(2):121-129.

Rüegg J, Gries C, Bond-Lamberty B, Bowen GJ, Felzer BS, NE McIntyre, Soranno PA, Vanderbilt KL and Weathers KC. 2014. Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment* 12:24–30.

註腳

[1] http://www.forestgeo.si.edu/

[2] http://gleon.org/

[3] http://www.gbif.org/

[4] http://www.earthobservations.org/geoss.php

[5] https://www.lternet.edu/

[6] https://www.ecotrends.info/

[7] https://www.dataone.org/

[8] 2016 年 6 月 26 日 GBIF 網站紀錄