生態大數據哪裡來？
—生態感測網絡之資料收集、應用與管理

鄭美如[1]、張哲彰[1]、林書瑄[1]、蘇大舜[2]

什麼是大數據？

近幾年很夯的科技名詞，如大數據(Big Data)分析、人工智慧(Artificial Intelligence, AI)、物聯網(Internet of Things, IoT)及區塊鏈（block chain）等，大家應該都耳熟能詳，這些技術也應用在非常多的產業上，大幅提升商業、醫學、科學研究的進展及與日常生活的連結。

大數據一詞的概念相對於資料(data，本文中除big data翻成大眾認同的大數據外，其他data一詞均以資料稱之)而言較新，但大數據的起源應可以追溯剛開始使用第一個資料中心和開發關聯資料庫時的1970年代左右，只是相較於現在的大數據資料庫，規模非同日而語。大數據的概念是在2001年提出，但經過10年的時間，直到2012年大數據才被大肆探討，似乎任何產業只要運用到大數據分析，都可以提升營運效率。

什麼是大數據？字面上解釋即是「巨大的資料量」，又因為資料量非常巨大又多元，所以有些特性並非傳統資訊處理技術可以歸納分析的，故需要新的技術，所以大數據也可以說不單指規模大的資料，而是一種分析處理龐大資料的技術。

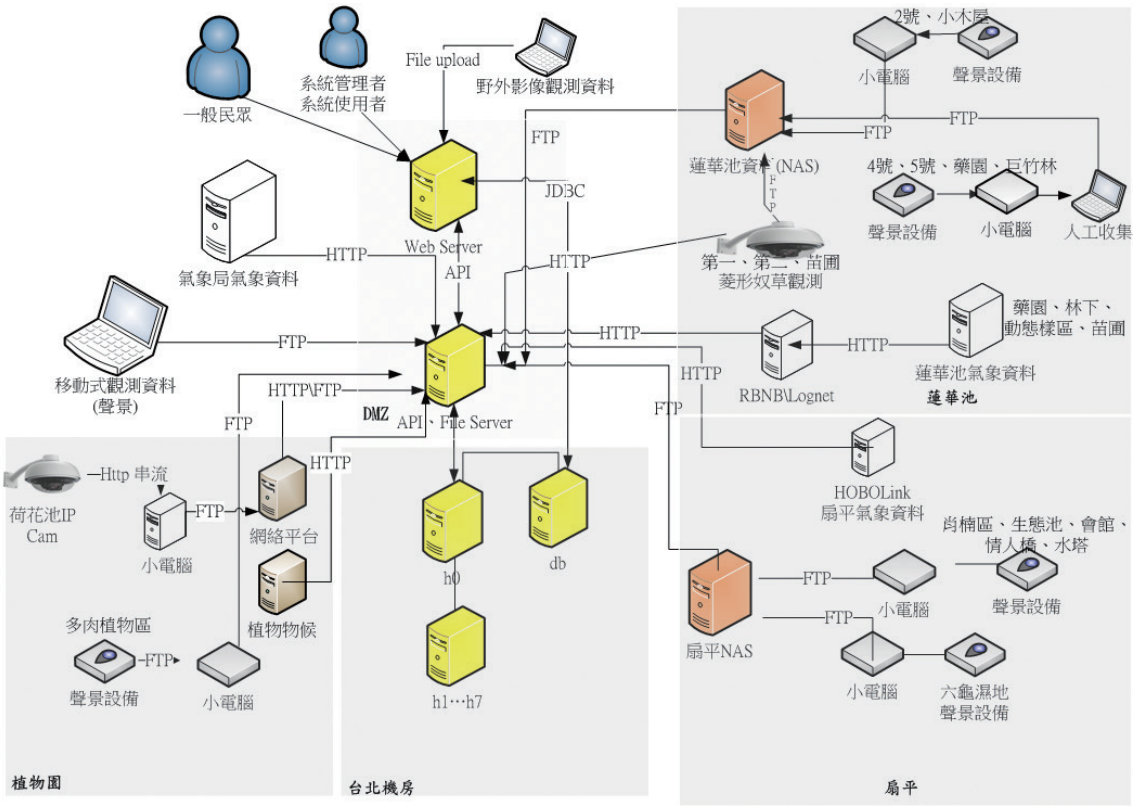
那麼，大數據中所謂的資料特性指的又是什麼呢？2001年麥塔集團(META Group)的分析師萊尼(Doug Laney)在一份報告中對大數據提出「3-D資料管理」，即資料量的規模(Volume) 、資料格式的多樣性(Variety)及資料即時處理的速度(Velocity)等3個面向發展，亦稱為「3V」或「3Vs」。之後，隨著資訊科技不斷地進步，大數據的複雜程度愈來愈高，2012年在3V之外，增加了「準確性」(Veracity)的特色，成為4V。之後甚至有人提出5V、6V的看法，即增加了「可視性」(Visualization)及「合法性」(Validity)等。

另一方面，原本的資料伺服器也無法因應這樣的需求，於是在2005年，Hadoop(巨量倉儲與分析系統)被開發出來，而NoSQL(非關連式)資料庫在這段時間也開始流行起來。什麼是Hadoop？與一般的伺服器有什麼不同？Hadoop是專門用於倉儲和分析大數據的開源(open source)框架，可以從單一的伺服器到擴展到上千台伺服器，每台伺服器都可以提供運算與倉儲。而什麼是NoSQL資料庫？與傳統式的SQL(Structured Query Language，結構化查詢語言)資料庫又有什麼不同？簡單來說，NoSQL資料庫是為了非結構化的資料(如文件、圖形等)產生出來的資料庫，可以承受高進出流量的工作負載，讓大量資料的查詢獲得更快的回應。

讓生態資料自己來

自2013年起，我們執行了一項為期4年的計畫—「智慧生態計畫-生態及生物多樣性資訊基礎建設與應用計畫」，主要目標是利用網路及資訊設備等基礎建設，透過資料倉儲與網路應用系統的開發，建構一個整合性的資訊管理系統，以提供分眾、主動、全程的生態知識服務。

[1、2]林業試驗所・技術服務組、蓮華池研究中心

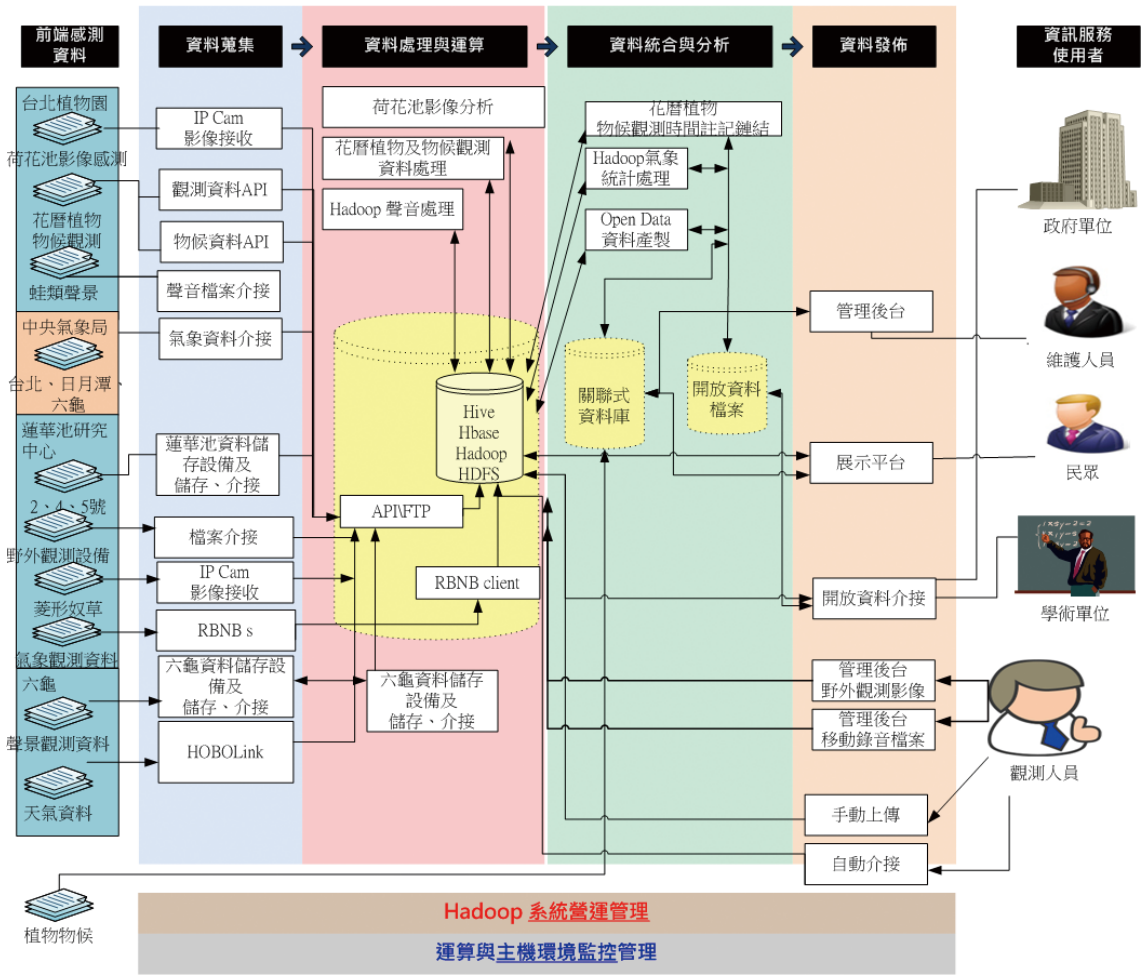


生態感測資料收集與傳輸架構圖。

計畫一開始，我們便希望將林業試驗所原有的生態監測資料進行倉儲與整合，所以從六龜及蓮華池兩個研究中心既有的網路基礎建設及生態聲景與影像設備著手，進行改善與監測點的增設，並將臺北植物園也納入監測範圍，逐年增加各站的監測點。另外也增設了一些監測點增設氣象監測儀器，亦將蓮華池動態樣區中2個原需要人員定期上山手動存取資料的氣象站，改以透過網路自動傳輸資料。所以持續不斷自動收集與透過網路傳輸來的資料包含了4個影像監測點(臺北植物園荷花池及蓮華池菱形奴草生育地)、11個聲景監測點(以蛙類聲景為主)及6個氣象監測點，另外還有2項由人工定期收集的資料(臺北植物園花曆植物開花的物候調查及影像)。因此，在我們的智慧生態子計畫中，總共收錄了3大類型(文字、聲音及影像)的巨量資料，這些檔案多是非結構化的資料檔，所以我們建置了Hadoop來倉儲這些資料，截至目前為止，收集了超過35TB，450萬個檔案。

生態資料的展演台

我們為這些資料檔規劃建置了「生態感測展示平台」(https://iesn.tfri.gov.tw)及「生態感測資料開放平台」(http://iesn.tfri.gov.tw/forestDW/OpenData)，將所有收集來的原始資料全部放在資料開放平台中，並提供農委會資料開放平台及政府資料開放平台以應用程式介面(API, Application Programming Interface)方式進行介接，可即時獲取最新上架的資料。而在展示平台上，我們希望傳達生態知識，服務社會大眾，所以展示平台上，加入了圖像化的介面，並將所收集來的資料進行初步的分析，提供視覺化的生態知識服務。



生態感測資料處理架構圖。

如臺北植物園花曆植物的物候調查，我們以花開程度及百分比的資料，自動算出開花階段，並於平台上顯示相對應的深淺顏色，讓民眾一目了然。而荷花池開花情形的監測，我們利用定時攝影機轉動連拍，再以程式自動拼接的方式，完成一張大景照片，讓民眾可以看到荷花池每小時的全景。另外如稀有的寄生性植物-菱形奴草的監測，在非生長期時，我們每小時拍攝1張照片，在生長期，則每10分鐘拍攝1張照片，這些照片除倉儲於Hadoop外，系統每半小時將13張照片即時上架於展示平台上。生長期同時提供24小時直播，但因考量錄影檔案過大，所以沒有執行錄影程序。

有關蛙類錄音檔的部分，已在前期〈隨時隨地聽見遠方森林裡的蛙鳴！〉一文中有詳盡的介紹了。另一個自動收集但以人工定期取回進行辨識與倉儲的是對天然林中野生動物的種類與活動進行紀錄，此一部份是以紅外線自動相機進行30秒的錄影。這些錄影檔經過人工辨識物種後，都上架於展示平台，其中有許多是石虎、穿山甲及麝香貓的活動身影，難得一見，而使用者亦可隨時下載應用。

設備與資料的管理

因為生態感測網絡平台所涉及的層面廣，包括了網路基礎設施、野外設備、資料傳輸與倉管、資料展示平台等等，為了有效的管理，

在生態感測展示平台上，可以清楚的辨識各種花曆植物從2014年開始的開花月份、開花階段。(http://iesn.tfri.gov.tw/forestDW/Flower/Calendar/Monthly)

在菱形奴草的照片上，套疊中央氣象局日月潭氣象站的即時資料。(http://iesn.tfri.gov.tw/forestDW/Grass/SlideShow)

我們在計畫的最後一年，與維護廠商花了數月時間，為整個感測網絡平台各項設備及作業程序進行詳細的盤點，訂定了維護管理流程。

我們將整個維運團隊成員分成現場測站維護人員、測站管理員、總管理員及維護廠商，每個角色都有應負的權責與執行的工作項目。在維護管理的10張紀錄表中，包含了每個設備的進出與維護記錄、監測點每天的資料與設備檢查紀錄、測站維護報告、系統與資料處理問題追蹤紀錄等等。定期的資料與設備狀況的確認，可降低資料的遺失及提升資料的正確性。例如：在每天例行的資料檢查中發現某個數值有異，便可以及時查核問題的發生原因，避免錯誤資料的持續收集。另一方面，各項維護紀錄能提供後續系統或設備發生問題時的處理參考，縮短處理時程。

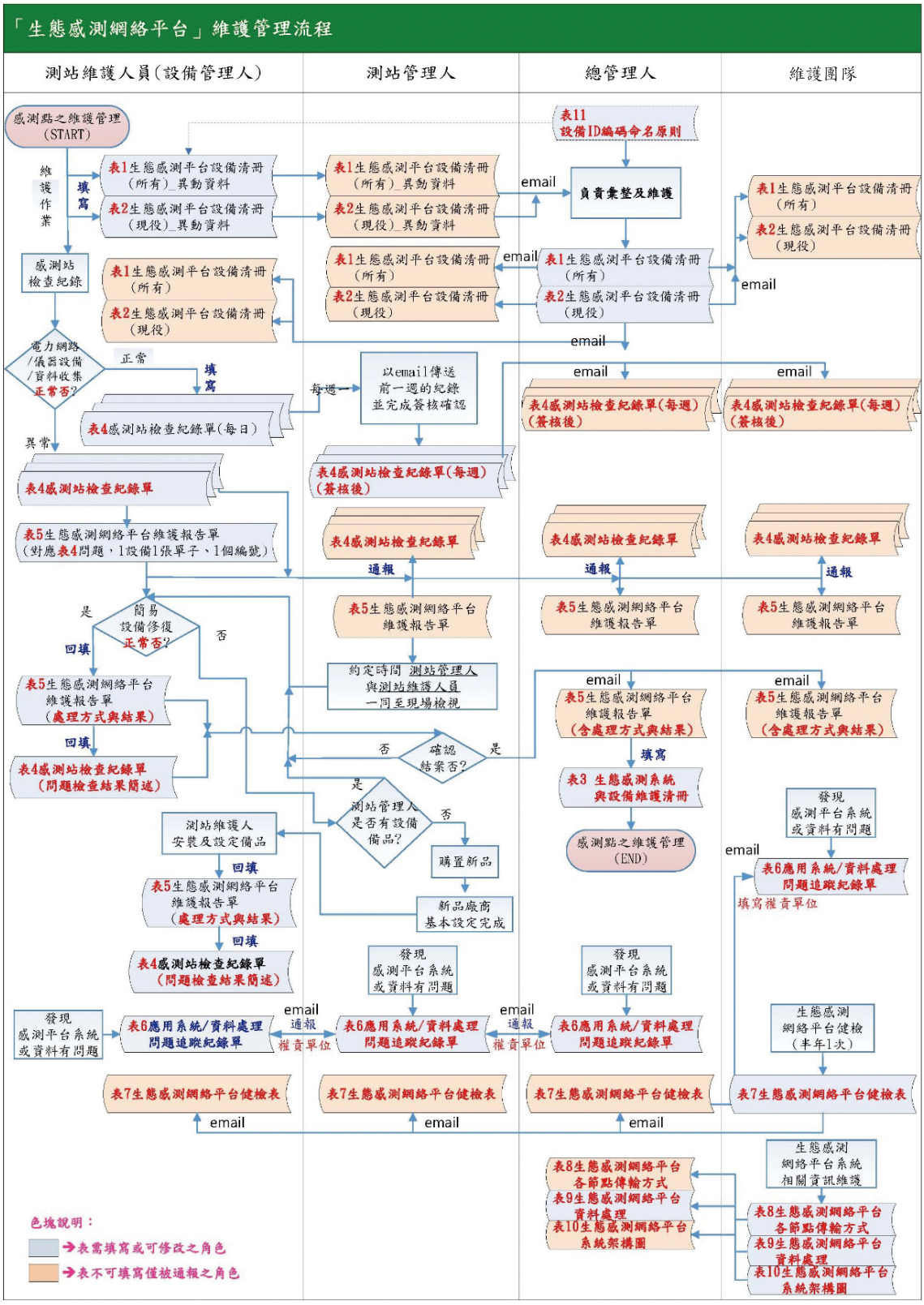
除了這些紙本紀錄與掃描檔，在生態感測展示平台上，我們闢了「工作足跡」專區，將較為重大的維護工作以文字與照片的方式進行網頁紀錄，不僅分享我們的工作歷程，其實也是分享生態感測網絡平台的架設及維運經驗，貫徹我們公開分享的精神。

繼續向前行的挑戰

大數據有許多特點，其中有一項為使用的資料幾乎是全體資料，也就是母體資料，而不是抽樣過的資料。以往礙於人力、物力與技術，資料的取得僅能以抽樣為之，既然是抽樣，就一定會有誤差，若改以母體資料進行分析，則分析結果幾乎就是事實了。

目前我們仍因野外環境、電力與網路技術及倉儲設備等因素，在生態監測資料的收集上，仍以取樣方式為之。如人工收集的野生動物監測影像檔，因考量同一動物在監測地點逗留，若連續錄影，將可能因類似檔案過大，記憶卡空間不足而無法錄製後續其他動物的活動，但若設定5分鐘內不啟動錄影程序，這樣便可能犧牲了動物其他行為表現的紀錄。

另一方面，雖然聲音與影像分析已有長足的進步，但對於物種聲音辨識，因無法過濾各種大自然的環境聲音，形成辨識上一大障礙，而影像的自動辨識，也因物種特徵、形態與行為等的多變，仍需要更大量的資料進行智能學習，這也是我們未來前行的目標，達成聲景及影像的自動化物種辨識及數位物候分析。



生態感測網絡平台維護管理流程。