從研究流程談整合型計畫的資料蒐集與管理

⊙林業試驗所森林保護組・林朝欽　⊙林業試驗所林業推廣組・陸聲山

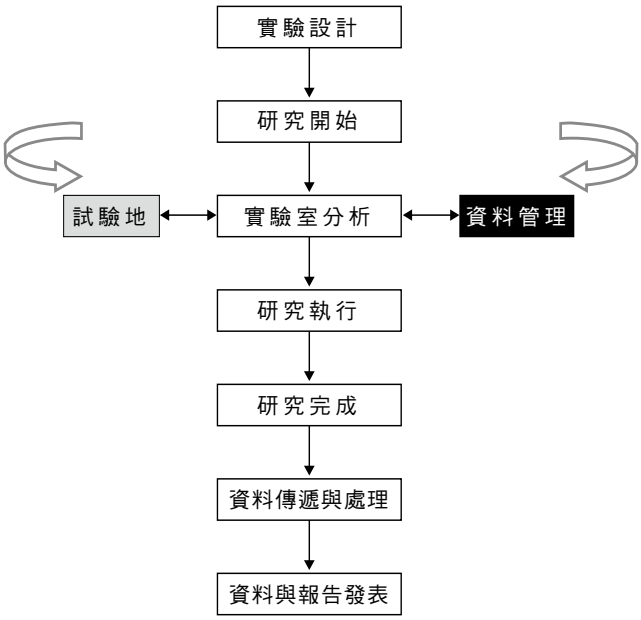
前言

傳統生態學研究中，研究者經常只是各自進行小規模或尺度的觀察和試驗，所蒐集到的資料往往也只有一位或少數幾位研究者進行分析，以這樣的尺度與方式收集到的資料，很難用於推導出共同或通用的生態理論。然而現今環境的許多重大議題，例如全球暖化、生物多樣性流失、外來種的入侵或自然資源耗竭等，越來越需要瞭解在大規模的空間和時間尺度下複雜生態系統的運作，所需要的資料也早已超過針對單一對象或地點蒐集資料的範圍，越來越迫切要求各類資料的共享、整合。研究人員無法獨立完成這種大規模空間和時間尺度下的生態研究，例如本期「人工林生態系經營及生物多樣性保育研究之因應策略」這種跨領域的整合型計畫，這種需求也導致了研究人員意識到必須增加對資料的分享與取得的能力，因此強化資料保存、累積、使用、分析與整合的資訊管理工作日形重要。本文將從研究流程觀點切入，列出資料蒐集與儲存有關的準備工作，供研究人員做為參考，以促進研究資料建檔、流通、共享與應用。

結合資訊管理的研究流程

傳統上研究計畫的執行，研究流程從研究設計、資料收集、資料整理、資料分析到撰寫報告、各階段環環相扣，每一階段仍需再度評估是否符合原先設計目標，便可據以修正研究步驟或需收集之資料。但往往未將資料管理視為研究的一重要環節，造成日後資料整合與分享的困難。如今結合資訊管理的研究流程已備受重視，資料管理不再與研究工作分離，而成為研究工作的一重要部分。

為達成資料的整合與分享，尤其是跨領域的整合型計畫，最有效的方式是在研究過程的一開始，資訊管理人員便參與研究的設計，這能促進資管人員與研究人員的對話與相互了解，對資料使用或檔案管理的現況與可能面臨的障礙能確實掌握。此外，透過蒐集各研究人員計畫之調查資料表單，或研究人員提供他們所需蒐集及量測的資料，以利做為資料倉儲或歸檔用，將可減少日後資料整合實行上的困難。



結合資訊管理的研究流程

有效的資料庫整合方式是在研究過程的一開始，各個研究人員當提供他們的資料或檔案做為資料倉儲或歸檔用時，便能符合研究流程原則與規範。本文先列出我們認為與資料蒐集與儲存有關的準備工作，供研究人員做為參考。這樣的指導原則使得研究人員能有所遵循，而達到未來擴大資料的利用性。

資料蒐集與儲存的原則

一、有意義的資料檔案名稱

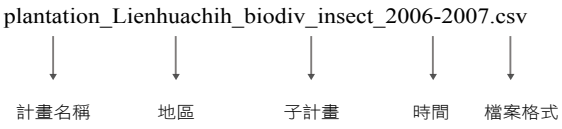
檔案名稱應能反映出資料內容及充足資訊以作為未來檔案的識別，檔案名稱最好能縮寫涵蓋計畫代號、計畫題目、地點、主持人、年度等的識別，大型計畫的檔案名稱甚至應思考系統性的涵蓋所有子計畫。以下幾個為較適當與較不適當的檔名範例：

較不適當的檔名：

mydata.dat

1998.dat

較適當的檔名：



檔案名稱雖然以上述的充足資訊較佳，但亦須配合資料庫管理系統的規則，例如字元長短、特別字型使用、以及存檔的目錄等。

以大型整合型計畫而言，如「人工林生態系經營及生物多樣性保育研究之因應策略」為例，其下共有10項子計畫，則宜考慮本計畫內的任何檔案名稱使用共同幾項因子來表達計畫內容。

二、使用一致性且穩定的檔案格式

檔案格式的一致性影響未來資料使用難易的關鍵，通常以ASCII格式儲存的檔案格式是最穩定且一致的，但ASCII檔案可能在不同語言上有些位元的差異。雖然如此，ASCII的文字格式檔案全球通用，且不會因為軟體更新或軟體不同而造成未來使用的問題。另外，檔案格式切忌以極特殊的軟體格式作為永久儲存的格式。在文字檔案格式中有不同的分隔方式，例如逗號、分號等，最好亦能有一致性，常用的逗號分格值(csv)即為較普遍接受的一種格式。此外，避免將統計後的數字放入檔案，圖與表亦如此。某些圖形檔(如影像、GIS的網格)雖然沒有特殊的建議，但存成二位元格式(binary format)應該是比較適當的。如果沒辦法統一存成上述比較能永久保存的格式，最好能互相討論後瞭解每個子計畫日後將使用的檔案格式。

三、完整定義研究參數(parameters)

為了以後的資料分享及倉儲，資料檔中的參數必須定義完整，包括參數的名稱、量測的單位、格式(型態)，最好能製作一份編碼簿記錄這些詳細內容，參數名稱最好能簡單易懂，例如溫度(temp)、降雨量(Precip)、經緯度(Lat, Long)，在使用英文字時亦應保持大小寫一致性，例如不要在同一檔案中使用temp、Temp、TEMP三種不同形式的英文字代表同一參數。另外，某些軟體(尤其是DOS

的舊軟體)不見得讀得懂中文字，而英文字母亦有八個字母限制，因此必須注意這類軟體的規定。

參數名字以中文命名只能在中文系統中使用，其未來性以及國際交換易受限，因此非常不建議資料檔中的參數以中文命名。參數所使用的單位必須明白定義及記錄，國際上大部分使用公制單位(SI)，不過各領域中或有特殊的單位，不見得與標準相同，只要說明清楚即可適用。參數的型態(文字、數字)必須一致，許多以試算表(Excel)建立的資料，因其參數之型態不一致而導致資料在未來交換中變成無用的內容。

參數的格式在科學上的使用的通則可如下列的參考例子：

1.日期(Dates)：yyyymmdd，如 January 2 ,1997 應為19970102。
2.時間(Time)：通常使用24小時制，並應記錄地方及全球的時差時間(UTC)，因為這兩個時間可能有一天之差。
3.空間座標：最好遵循共同的座標系統如UTM，否則容易出錯，一般GIS軟體均會要求定義座標系統，台灣地區正更換座標系統，有所謂67二分帶與97二分帶系統，應特別加以標註。
4.海拔(Elevation)：海拔通常以公尺為單位。
5.缺失值(Missing Values)：許多軟體以句點(．)或-9999作為缺失值代碼，但不見得合理，最好自己定義，且避免文字作為缺失值代碼。

四、一致性的資料排列格式

資料排列的格式影響爾後的統計分析，何種排列方式較恰當則因人而異，但建議使用一致的排列方式。不論那一種排列格式，觀測值必須在獨立的欄或列中。一般而言，欄(column)通常是放參數，列(row)放觀測值，這種排列是矩陣似的表單格式，例如：

最後一列的-9999.9為缺失值必須在資料檔中或編碼簿中定義，另外站名的YYL_B若是代碼的話亦應記錄與說明。

另一種資料排列則適用於有許多缺失值的使用，此種格式把參數用欄與列混合。例如上述的資料可排列為：

如果檔案很大，切忌把檔案以某個變數分割成許多小檔案，例如以月份把檔案拆成小檔案或以樣區分割許多樣區。比較理想的做法是在完整的大檔案作邏輯性的分割。

Station	Date	Temp	Precip
Unit	YYYYMMDD	C	MM
YYL_B	20041001	12	0.0
YYL_B	20041010	11	3.3
YYL_B	20041020	10	-9999.9

如果資料蒐集內容有不同的量測型態，例如葉面積指數(leaf area index)及生物量(biomass)，建議依不同量測型態分立檔案。但最好以一致性的資料排列加以建檔。如此使用者才能瞭解相互關係，並利未來資料進資料庫時容易建立聯結。

Station	Date	Parameter	Value	Unit
YYL_B	20041001	Temp	12	C
YYL_B	20041010	Temp	11	C
YYL_B	20041001	Precip	0.0	MM
YYL_B	20041010	Precip	3.3	MM

五、資料檢核與品質控管

除了科學上的品質確認外，就資料檔本身最基本的品質控制可列舉如下：

1.核對資料檔的資料欄位確實分隔清楚。
2.核對檔案內的主要資料欄，如樣區號碼、時間、日期、座標等沒有缺失值或空白。
3.檢視資料內容有無不合理值，例如PH74，最好能印出或排序檢查。
4.用統計的頻度或摘要方式檢查一般結果，例如年平均氣溫36℃可能是不合理的值。
5.如果有座標值，應利用GIS軟體將座標值展示在圖上以做核對。
6.確認由data logger 或野外電腦的檔案轉錄，最好的方法是轉錄前作一份摘要統計，與轉錄後的摘要比對，以確認轉錄無誤。

隨著計畫的實施而逐漸累積可觀的資料，但資料的品質良莠是否能確保，將可利用林試所開發的EML資料分析模組之功能，將資料進行初步的檢核，以取代人為耗時的篩選，而逐步達到自動化檢核的功能。

六、資料集標題

任何資料集應給予一個標題，並在這個標題之下建一份完整的說明文件後設資料(metadata)，研究人員應體認這份資料可能是未來10年或20年其他人必須再使用，因此這份文件以及文件所描述的資料必須易讀易懂，尤其是資料的蒐集方法與使用儀器的描述，因為使用者可能根本不瞭解當時的計畫內容。

後設資料的建立必須有一致的標準，目前生態研究使用EML(Ecological Metadata Language)，可使用一般文字編輯器建立，但一些公用軟體如Morpho則提供了簡單的介面，是不錯的選擇(詳細的後設資料文件敘述於下節)。以下是一些資料標題的較好與較不適當例子：

較好的標題如：

“Lake metabolism data of Yuan-Yan Lake in Taiwan, 20040822-20040828”

這個標題清楚涵蓋了時間範圍、地點以及主題。

較不適當的標題如：

“The lake data set”
“Metabolism Data”
“Yuan-Yan Lake Data”

七、提供後設資料(metadata)文件

所有資料集都必須提供後設資料文件，這份文件不只是讓研究人員自己使用，也必須讓未來資料使用者能看得懂；尤其當未來的使用者完全不清楚這份資料的研究背景與方法。為了能確保後設資料能在未來使用，

以ASCII的文字檔儲存是最理想的，沒有人能保證20年後MS Word還是大部份人用的文字編輯器，如果20年後的人不用MS Word，那麼以MS Word編輯的後設資料就變成無用的文件。

如果資料必須包含圖表、地圖、公式、影像，無法使用ASCII檔案格式，則選用較通用格式如html、gif、jpg、rtf或pdf等。

後設資料檔案必須與資料檔分開，但名稱則可以相似或一致。依EML的規範後設資料可以很簡要但也可以很詳盡，端視資料內容而定，但以下項目是後設資料檔案須包含的內容：資料集的名稱、資料蒐集的目的、什麼資料已(或將)蒐集、使用儀器(應包括型號、序號等)、使用地點(如氣象站的資料應標明來源)、資料蒐集者資訊、經費贊助來源資訊、資料集內各個資料檔的名稱、資料集如何引用、資料蒐集的空間尺度，如果地點以代碼表示則須提供代碼內容、資料蒐集開始與結束時間以及蒐集的頻度、每一個參數的量度方法或推導公式量度單位，有效數字及精度，以及量度的尺度等、環境情況(如颱風或火災前後)、處理資料的軟體、資料是否檢核過、特別編碼或代號說明、資料修改的日期、資料使用限制等等。

結語

林業試驗所政策型計畫「因應京都議定書之林業經營策略」，特別將研究計畫管理與資料保存的工作視為研究的一項重要工作，目前推動中的整合型計畫「人工林生態系經營及生物多樣性保育研究之因應策略」與「環境變遷中的樹木界限之動態學」，亦將資料管理視為研究的重要環節，資料整合與分享已成為這些計畫的首要之務，資料管理不應與研究工作分離，而成為研究工作的一重要部分。本文強調透過研究流程一開始的資訊管理規劃，改善研究人員在資料蒐集與儲存的準備工作，提昇資料品質的控管，使得跨領域研究之資料整合、流通與共享，不再是夢想了。

附註：本文列出的七項原則主要依據Best Practices for Preparing Ecological and Ground-Based Data Sets to Share and Archive (Robert B. Cook et al., 2000)而來，讀者可進一步參考原文。