



資料品質原則

Principles of Data Quality

亞瑟·查普曼

Arthur D. Chapman¹

雖然大多數人均不願見到自己的資料中有錯誤而欲去除之，資料中的錯誤需要公眾的關注和了解，因為錯誤是決定適用性的關鍵。(Chrisman 1991)



¹澳洲生物多樣性資訊服務
(Australian Biodiversity Information Services)
PO Box 7491, Toowoomba South, Qld, Australia
email: papers.digit@gbif.org

(本頁留白)

© 2005, 全球生物多樣性資訊機構 Global Biodiversity Information Facility

轉載本出版品請註明出處。建議引用格式如下：

Chapman, A. D. 2005. *Principles of Data Quality*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

2004 年，全球生物多樣性資訊機構自然史藏品數位化計畫(GBIF DIGIT programme)委託亞瑟·查普曼(Arthur Chapman)撰寫本文，旨在強調物種出現原始資料品質的重要性。近來我們對此議題及資料偵錯與除錯工具不斷有新一層的認識，因此本文可說是 2004 年過渡時期的討論。預期未來會有新的版本出現，也在此感謝資料提供者及使用者社群的貢獻。

意見指教請寄至：

Larry Speers
Senior Programme Officer
Digitization of Natural History Collections
Global Biodiversity Information Facility
Universitetsparken 15
2100 Copenhagen Ø
Denmark
E-mail: lspeers@gbif.org

and

Arthur Chapman
Australian Biodiversity Information Services
PO Box 7491, Toowoomba South
Queensland 4352
Australia
E-mail: papers.digit@gbif.org

July 2005

封面圖片 © Per de Place Bjørn 2005
Amata phegea (Linnaeus 1758)

《資料品質原則》

本文件翻譯的目的，在於臺灣或其它華文地區推展生物多樣性原始資料的提供與整合。如果對於中文內容有任何意見及指教，懇請與出版單位聯繫。

本文件出版經費由「臺灣生物多樣性資訊機構」及「數位典藏暨數位學習國家型科技計畫」提供。



發行人 邵廣昭

審訂 邵廣昭

總監 賴昆祺

策劃 陳麗西

翻譯 陳映均 柯智仁

編輯 柯智仁

校對 王惟芬

出版單位 中央研究院 生物多樣性研究中心

地址：115 臺北市南港區研究院路二段 128 號

電話：(02)27899621 ext.220

全球資訊網：<http://biodiv.sinica.edu.tw>

臺灣生物多樣性資訊機構：<http://www.taibif.org.tw>

出版日期 中華民國九十七年八月十日 初版

目錄

目錄	ii
Introduction 前言	1
Definitions 定義	3
Principles of Data Quality 資料品質原則	9
Taxonomic and Nomenclatural Data 分類及命名資料	24
Spatial Data 空間資料	29
Collector and Collection Data 採集者及採集資料	32
Descriptive Data 描述性資料	34
Capturing Data 資料擷取	36
Data Entry and Acquisition 資料輸入及取得	39
Documenting Data 資料建檔	42
Storage of data 資料儲存	48
Manipulation of spatial data 空間資料運用	52
Representation and Presentation 表示及呈現	53
Conclusion 結論	59
Acknowledgements 致謝	60
References 參考文獻	61
Index 索引	64

(本頁留白)

Introduction 前言



資料品質原則(Data quality principles)早已成為企業界(SEC 2002)、醫療業(Gad and Taulbee 1996)、地理資訊系統(Zhang and Goodchild 2002)及遙測(Lunetta and Lyon 2004)等及其它許多領域的核心業務，但全球的博物館與分類學社群直至近年才體認到其重要性。隨著分類學及物種出現資料的流通、交換量快速增加，當使用者對資訊品質細節的要求提高，考慮資料品質原則即成為重要的議題。的確，有些人認為博物館資料的品質不合格、不足以作為環境保育決策的參考，但這究竟是因為資料本身品質的好壞，還是資料建檔的過程有問題？無論如何這些資料至關重要。由於藏品是長年累積的成果，在人類活動對生物多樣性造成巨大衝擊的同時，它們提供了無可取代的基準資料(baseline data, Chapman and Busby 1994)。這些資料也是環境保育上非常重要的資源，為可能因農業開墾、都市化、氣候變遷或其他原因發生棲地變動的地區，提供唯一、完整的物種出現紀錄(Chapman 1999)。

以下我將詳述一些觀念，並提出一些在博物館及標本館資料開放予更廣大的社群之時，應該作為其業務核心的資料品質原則。

在環境資料庫、模型系統、地理資訊系統及決策支援系統等的使用上，資料品質與資料錯誤的問題常為人所忽略。運用資料時常未考慮到其中的誤差，可能導致錯誤的結果、誤導的資訊、不智的環境決策及徒增成本。

博物館與植物標本館藏的動植物標本提供了大量的資料來源，不但保存了產地目前的資訊，更紀錄該地區數百年來的歷史(Chapman and Busby 1994)。

在處理物種資料時有許多品質原則可以適用，尤其是其中空間方面的資訊。這些原則涵蓋了資料管理過程的所有階段，若在任一步驟中疏忽了資料品質，將會降低資料的適用性，使得資料應用範圍受限。資料處理程序包含：

- 資料擷取與收集時的紀錄，
- 藏品數位化前的資料前置作業（準備標籤、複製資料至總表等），
- 藏品的鑑定（標本、觀測紀錄）及其紀錄，
- 資料數位化，
- 資料的建檔（收集並紀錄後設資料），
- 資料儲存與歸檔，
- 資料呈現與傳播（紙本及電子出版品、線上資料庫等），
- 資料使用（分析與運用）。

以上步驟皆會影響到資料最終的品質及適用性，也適用於資料的各個層面—包括「物」，即分類、命名資料；「地」，空間資料；以及其餘如「人」和「時」等要素(Berendsohn 1997)。

在資料品質的詳細說明及其在物種出現資料的應用之前，我們需先定義與描述許多概念，包括「資料品質」本身、常被混淆的準確度(accuracy)與精確度(precision)，以及物種原始資料及物種出現資料在本文的定義。



別低估品質改良的簡潔效果。它只需要團隊合作、訓練與紀律，不需其它特殊技巧。只要有心，人人都可有所貢獻。

(Redman 2001)

Definitions 定義

Species-occurrence data 物種出現資料

本文所指涉之「物種出現資料」包含標本標籤資料、博物館與植物標本館館藏、觀測資料及環境調查資料。大體而言，這些資料是所謂的「點資料」，但仍包含線（環境調查的樣線資料、沿河流的採集資料）、面（特定區域如國家公園的觀測資料）及網格資料（網格內的觀測或調查資料）。通常，我們指的是地理參照的資料——亦即資料已經具有地理上的參考位置，將它和特定時間（日期、時刻）及空間中的特定地點連結起來，有些具地理參照座標（具經緯度、UTM 二度分帶座標），有些則無（地點、海拔及深度的文字描述）。資料大多有分類學名稱，但也包含尚未鑑定的藏品。本詞有時會與「原始物種資料」(primary species data)交替使用。

Primary species data 原始物種資料

「原始物種資料」係指未經加工的藏品資料及不具空間屬性的資料。包含不具空間資訊的分類及命名資料，例如沒有相關地理參照的名稱（譯按：如學名及俗名）、分類群及分類觀。

Accuracy and Precision 準確度與精確度

準確度(accuracy)與精確度(precision)常被混為一談，圖 1 正好說明兩者的差異。

準確度指的是測量值、觀測結果或估計值與實際值之間的一致程度，此實際值可能是約定值，如測量控制點的座標如圖 1 所示。

精確度（或稱解析度(resolution)）可分為兩大類。統計精確度(statistical precision)是指進行重複觀測，測量值間的相近程度，與實際值毫無關係。精確度高，準確度不一定高，如圖 1a 所示。數值精確度(numerical precision)則指量測紀錄中有效數字(significant digits)的總數。電腦問世後，數值精確度大增。舉例來說，某資料庫標示經緯度至小數點後十位，也就是約 0.1 公釐，在現實世界解析度卻不超過 10-100 公尺（小數點後 3 至 4 位），因此常誤導大家對其解析度及準確度的概念。

準確度及精確度這些詞除可用於空間資料，也可用於非空間資料。例如，某藏品可能經鑑定至亞種（高精確度），卻被置於錯誤的分類單位（低準確度），又或者只能鑑定到科（高準確度，低精確度）。



圖 1. 顯示出準確度與精確度在空間資料內的差異。紅點代表正確的位置，黑點表示收集者紀錄的地點。

- 精確度高，準確度低。
- 精確度與準確度皆低，為隨機誤差。
- 精確度低，準確度高。
- 精確度與準確度皆高。

Quality 品質

就資料的品質而言，定義多有不同，但地理學界普遍接受的定義為「適用性」(fitness for use)(Chrisman 1983)或「有潛在用途」(potential use)，這也是最新的空間資料轉換標準(spatial data transfer standards, ANZLIC 1996a, USGS 2004)所採用的定義。經濟及商業界等非空間領域近來也漸漸應用了這樣的概念。有些人則認為（如 English 1999）「適用性」的定義稍嫌狹隘，應該要涵蓋更多未來及潛在的利用價值。

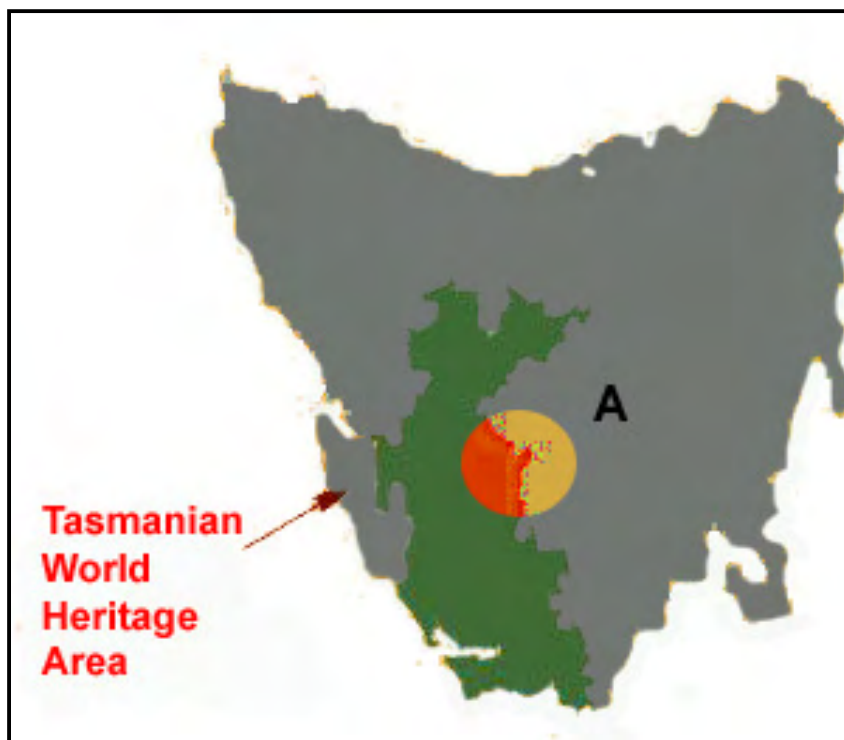


圖 2. 澳洲塔斯曼尼亞地圖。圖中某物種採集紀錄(A)準確度為 0.5 度(約 50 公里)，如圓圈所示。可能的物種分布範圍(依準確值判斷)與塔斯曼尼亞世界遺產區(Tasmanian World Heritage Area)部分重疊。

圖 2 的例子可說明「適用性」概念應用。某特定物種的採集紀錄(標示為 A)準確度為緯度 0.5°(約 50 公里)。若欲製作塔斯曼尼亞島的物種清單、想要瞭解某物種是否出現在該島，這筆資料便適於提供解答—因此該採集紀錄具有「適用性」，就此目的而言可視為高品質。另一方面，若想知道某一物種是否出現在塔斯曼尼亞世界遺產區，從該筆紀錄則無法得知—可能有，也可能沒有。在此種用途上，該筆資料不具「適用性」，故品質不高。資料庫中的經緯度值可能非常精確，看起來準確度很高卻可能誤導使用者，使他們忽略資料準確值不高的事實。

同樣地，資料中的非空間部分也會有類似情形，例如錯誤鑑種便足以破壞資料的價值，成為不適用的資料。研究物種分布(或其生理學、生態學等)時，標本或觀測資料上若標錯了名稱，亦會造成誤導或錯誤的研究結果。

資料品質涵蓋多層面向，牽涉了資料管理、模型建立分析、品質管制與保證、資料儲存及呈現。如同 Chrisman(1991)及 Strong 等人(1997)分別提及的，資料品質與其用途息息相關，無法自外於使用者進行評估。資料庫中的資料沒有實際的品質或價值(Dalcin 2004)，只有潛在的價值，只有在使用該筆資料時，方能發揮其價值。資訊的品質取決於資訊是否能滿足使用者，符合他們的需求(English 1999)。

Redman (2001)認為，適用的資料應該具備親和性(assessible)、準確性(accurate)、時效性(timely)、完整性(complete)、相關性(relevant)、全面性(comprehensive)，且與其他資料來源一致，提供適量的細節，易讀易懂。

另外，資料管理人也需思考該如何提升資料庫的可用性，使其廣為人所用（提升潛在用途或相關性），讓資料更適用於各種用途。在提高可用性的過程中，需投資多少工作在繼續提升資料功能與可用性，得經一番權衡取捨，因為這可能牽涉資料單元的分析，加入地理參照資訊等。



資料若能用於運作、決策及規劃，便是高品質的資料。

(Juran 1964)

Quality Control / Quality Assurance 品質管制/品質保證

品質管制與品質保證的差異並非總是清楚。Taulbee (1996)區分出這兩個用語，並強調若要達到理想品質，兩者缺一不可。依據她的定義，

- 品質管制用以判定品質，乃是以內部為控管品質所訂定之標準、流程及程序為依據；
- 品質保證乃是以流程外部之標準為依據來判定品質。它檢驗了品質管制流程，確保最終的產品達到預設之品質標準。

以更商業取向的說法，Redman (2001)定義品質保證為：

「為製造無瑕疵之資訊產物所設計之活動，期以最低成本達到最重要顧客的最重要需求。」

品管與品保這兩個詞在實務上沒有明顯的區分，但大多時候兩詞互通，用以描述整體的資料品質管理作業。

Uncertainty 不確定性

不確定性可視為「對於知識或資訊的未知量中，其不完整性的量度。若有完美的量測工具，則得確立該未知量之真值」(Cullen and Frey 1999)。不確定性來自於觀察者對資料的了解，與觀察者有關，反倒與資料本身較無關聯。凡資料必有不確定性，但要紀錄、了解、呈現這些不確定因素，使得外人能夠了解，才是最困難的部分。要瞭解風險及風險評估，必先掌握不確定性。

Error 誤差

誤差包含資料中不精確與不準確的部分。有很多原因造成誤差。

一般認為，誤差及不確定性皆是不好的，但事實不盡然如此，因為瞭解誤差及不確定性的成因、管理及減少的方法，可能極有用處。充分瞭解誤差及誤差傳播(error propagation)，才能有效地進行品管(Burrough and McDonnell 1998)。

一般而言，誤差分為隨機的(random)及系統性的(systematic)。隨機誤差通常指偶然發生之誤差，系統誤差或偏差(bias)則為規律的數值變化，在製圖界常被稱作具有「相對準確度」(relative accuracy, Chrisman 1991)。在決定資料是否適用時，系統誤差可能在某些應用上可接受，其餘則不然。以大地基準¹的使用為例，若分析中從頭至尾使用相同的基準，將不會發生重大的問題；但若同一份分析使用不同來源的資料、誤差各異，如使用不同大地基準之資料來源、或以早期的命名系來統鑑定，便會衍生出各種問題。

「由於誤差在所難免，應當作是資料中的一個基礎的面向」(Chrisman 1991)。只有在誤差已包含在資料呈現中，才有辦法了解資料的限制、甚至是當前知識的侷限。三度空間、屬性與時間中若有已知的誤差，需進行測量、計算、紀錄並建檔。

Validation and Cleaning 驗證與清理

驗證(validation)係指判定資料是否準確、完整或合理的過程。過程可能包含格式(format)檢查、完整性(completeness)檢查、合理性(reasonableness)檢查、界線檢查(limit check)、複查資料以挑出極端值（地理、數據、時間或環境值）或其他錯誤，並由專業領域的專家（如：分類學專家）評估資料。經過這些步驟後，通常會標示資料（以進行日後追蹤），或是歸檔、或針對有疑問的紀錄做進一步的檢查。驗證同時也可能檢查了資料是否符合標準、條例及慣例。資料驗證及清理的過程中，最重要的步驟在於找出錯誤的根本成因，避免再度出錯(Redman 2001)。

資料清理(data cleaning)指的是「修正」驗證過程中鑑定出的錯誤。資料清理與「資料淨化」(data cleansing)互為同義詞，不過有時資料淨化會同時指涉資料驗證與清理兩件事。資料清理過程中，需慎防資料遺失，變更現有資料時也須格外謹慎。最好能夠將舊資料（原始資

¹geodetic datum. 不同的大地基準會導致實際位置（在座標系統上）的系統偏移，在地球某些地區這樣的差距可能大至四百公尺。

料)與新資料(修正過之資料)一同保存在資料庫中,若資料清理中發生錯誤,便能復原初始資料。

近來發展出許多工具與手冊來輔助物種資料的驗證與清理,這將在相關文件《資料清理的方法與原則》中詳述。人工清理資料的工作耗時費力,且容易出錯(Maletic and Marcus 2000)。

依據 Maletic 及 Marcus 2000 的說法,資料清理的大致架構為:

- 定義並判定錯誤的種類
- 搜尋並找出錯誤的例子
- 改正錯誤
- 紀錄錯誤例子及種類
- 修正資料輸入的程序,以減少未來的錯誤

Truth in Labelling 真實標示

一般來說,真實標示指的是販售或授予第三方之貨品或產品上的品質標示。在物種出現資料中,真實標示通常包含後設資料(metadata),其中詳實紀錄各層面的品質、品管流程與辦法、以及(或者)與資料相關的品質測量數據。若標示無誤,則進一步進行資料認證及權利鑑定。大部分的博物館與植物標本館早有標示的機制,會標示鑑定專家的資訊及鑑定日期(鑑定人相關訊息),但鮮少用於紀錄上的其他資訊,也少見於觀測及無證據標本之調查資料。

Users 使用者

使用者有哪些人?資料的使用者包含資料鏈(圖3)中各階段的每一個人。物種原始資料的使用者包括分類學家、管理人員、研究人員、技術人員、採集者,也包括外部或下游的使用者如政策制定者、科學家、農業學家、林業學家、園藝學家、環境管理者、非政府組織(環保團體及製造業)、醫療專業人員、藥理學家、產業專業人士、動植物園管理員、一般大眾(含家庭園藝)及社群使用者。物種出現紀錄的使用者非常廣泛,很多時候幾乎涵蓋整個社群。

在收集物種原始資料的時候,多半沒有預設使用者,而傳統上,這些資料,尤其是博物館及植物標本館內資料,最初主要的收藏目的在於提供分類學或生物地理學研究所需。這是很重要的工作,但現今這些機構的出資者多為政府單位,他們期望付出的資金有更高的報酬率,因此希望資料用途更多、價值更高。尤其政府希望藉由這些資料做出更好的環境決策、

環境管理及保育規劃(Chapman and Busby 1994)，博物館及標本館的管理者因而更不能忽視這些使用者及其需求。有了完善的意見反應機制，使用者可以回報對於資料品質的意見，這在資料品質鏈中是相當重要的連結，以下將就資料品質鏈加以詳述。



判斷使用者需求的工作既困難又辛苦，但別無他法，收獲必定也豐厚。

Principles of Data Quality 資料品質原則

經驗告訴我們，將資料視為長期的資產，並在一整合的架構中管理，將能省下可觀的資金，並不斷創造價值(NLWRA 2003)。

資料管理過程中（擷取、數位化、儲存、分析、呈現與使用），全程都應採用資料品質原則。欲提升資料品質有兩大要點：預防及修正。錯誤預防與資料收集及輸入資料庫這兩個步驟密切相關。雖然我們有能力、也應致力於錯誤的預防，但事實上，錯誤會一直存在於大型的資料集(Maletic and Marcus 2000)，因此資料驗證與修正工作不容忽視。

由於錯誤偵測(error detection)通常所費不貲，也無法保證能百分之百成功(Dalcin 2004)，因此一般認為其功能遠不如錯誤預防，但在處理典藏品如物種原始資料及物種出現資料時，錯誤偵測扮演了舉足輕重的角色(Chapman and Busby 1994, English 1999, Dalcin 2004)。



先著手設定目標，訂定政策，執行策略，勿草草進行未經計劃、協調、非系統性的資料清理。

The Vision 願景

各個組織都應該打造一個願景，以期擁有良好的資料品質，尤其是意欲對外開放資料的組織。以良好資料品質為願景，通常也會提升該組織整體的目標(Redman 2001)，更進一步改善其營運程序。設定目標時，管理人應著重於建立一個完善的管理架構，利用適當的工具、指引方針及標準，結合領導、人員、電腦硬體、軟體應用程式、品管與資料，管理資料，將其化為品質良好的資訊產品(NLWRA 2003)。

資料品質的願景將能夠：

- 促使組織考慮其長期的資料及資訊需求，思考該需求和機構未來成功的關聯；
- 往正確的方向展開行動——良好的品質；
- 為組織內部及外部之決策提供穩固的基礎；
- 讓大家正式認同資料與資訊是機構的核心資產；
- 發揮組織資料與資訊最大功能，避免資料重複建置，促進合作關係，提升資料使用的公平性；
- 最大化資訊之整合及互通性。

The Policy 政策

除設定願景外，組織也須制定政策來實現願景。訂立良好的資料品質政策能夠：

- 促使組織通盤思考品質的議題，重新檢視例行的工作業務，
- 訂定資料管理流程，
- 協助組織瞭解以下目標：
 - 減少成本，
 - 提升資料品質，
 - 增進客服與關係，
 - 改善決策過程，
- 讓使用者使用的資料更穩定，也更信賴組織提供的資料，
- 改善與組織客戶的關係與溝通情形（包含資料提供者與使用者在內），
- 提升組織在社群間的地位，以及
- 致力達成最佳實務典範，提高受資助的機會。

The Strategy 策略

由於大型機構中的資料數量龐大，需訂定策略來進行資料擷取及檢查（參見排定優先順序，在後）。一個完善的策略（包含資料輸入及品質控管）要設定短期、中期及長期目標，舉例來說(根據 Chapman and Busby 1994)：

- 短期：6 至 12 個月即可收集並檢查完成之資料(通常包括資料庫現有資料及需要較少品質檢查之新資料)。
- 中期：輸入資料庫需時約 18 個月之資料，所需資源少，品質檢查流程簡單且可由組織內部獨立完成。
- 長期：資料輸入及檢查耗時較久，需經合作完成，檢查方法複雜。可能需有系統地處理藏品，挑出：
 - 甫經重新分類、或正由機構進行分類研究之分類群
 - 重要藏品（模式標本、特殊的參考藏品等）
 - 關鍵類群（優勢科別、具國家級重要性之分類群、列名受威脅之分類群、具生態/環境重要性之分類群）
 - 關鍵地理區域之分類群（如：來自開發中國家、期能分享資料給來源國，或是對該機構具重要性之地區）
 - 與其他機構合作計畫中的分類群（如許多機構協議分享相同分類群的資料庫）
 - 從頭到尾有系統地處理藏品
 - 新進藏品（較不偏好舊藏品）

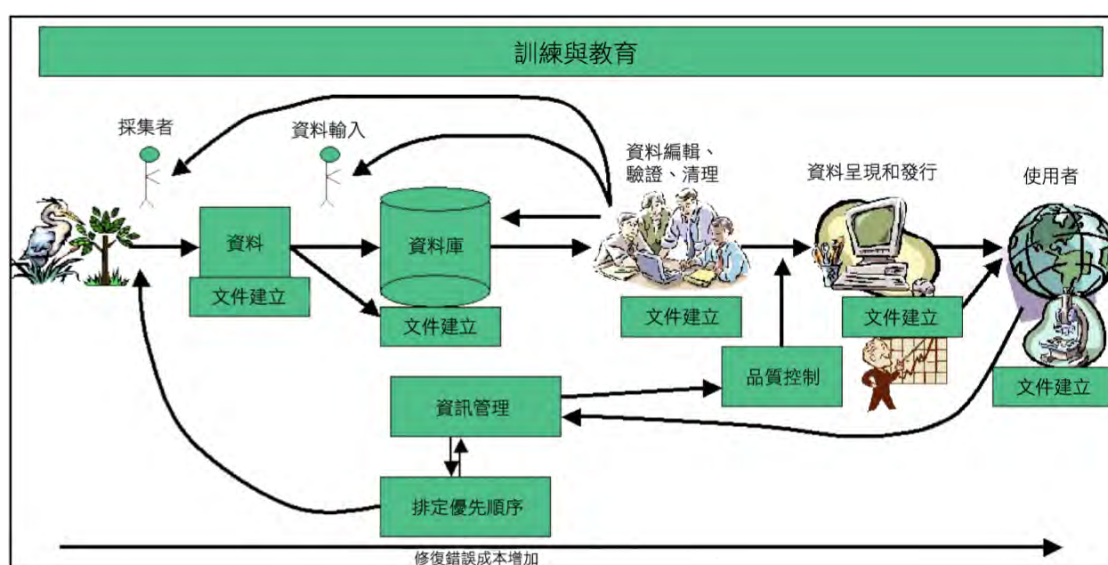
策略中應包含的資料管理原則如下(根據 NLWRA 2003)：

- 資料管理不疊床架屋
- 尋找有效率的資料收集及品管程序
- 盡可能地分享資料、資訊及工具
- 採用現行標準，或與他人一同訂定健全的新標準
- 發展網絡及合作關係
- 為資料集及管理展示週全的應用案例
- 減少重複資料收集與品質控管
- 不能只著眼於資料的立即用途，要將眼光放遠，並檢視使用者需求
- 確保實行完善的建檔及後設資料流程

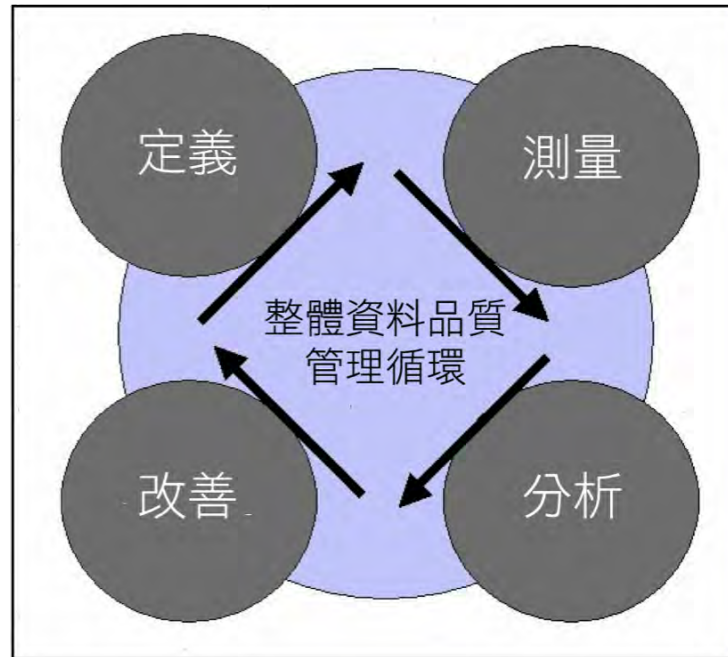
Prevention is better than cure 預防勝於治療

將藏品輸入資料庫的花費相當可觀(Armstrong 1992)，但相較於之後資料檢查及修正的支出，這只是冰山一角。預防錯誤不僅勝於事後的修正(Redman 2001)，也是目前最花費最低的選項。事後才修正錯誤，意味著錯誤的資料早在更正前已被許多分析研究採用，導致後續決策品質低落，或需重新進行分析。

然而，預防工作無法處理已存在資料庫中的錯誤，因此資料驗證與清理工作仍是資料品質管理程序中重要的一環。資料清理的重要之處在於，它能夠找出資料庫中錯誤的成因，並有後續程序確保同樣的錯誤不再發生。不過，資料清理不能單獨執行，否則問題無法根除。資料清理與錯誤預防應雙管齊下，若先進行資料清理，才考慮到預防，表示錯誤預防永遠無法有效執行，資料庫也會累積更多的錯誤。



圖三.由資訊管理鏈可知，隨著流程推移，修正錯誤所需的花費也逐漸累積。而過程中，完善的建檔、教育及訓練都是不可或缺的要素。



圖四. 整體資料品質管理循環圖(Total Data Quality Management Cycle)顯示了資料管理流程的循環特質(根據 Wang1998)

資料管理員及所有者（博物館及植物標本館等館藏機構）對其資料負有大部分的責任，儘管如此，資料提供者及使用者也同樣負有責任。



資料品質好壞的責任歸屬應為資料生產者，若無法辦到，盡可能將責任給予與資料生產過程最相關的人員。(Redman 2001)

The collector has primary responsibility 採集者擔負主要責任

資料採集者需擔負資料品質管理的主要責任。他們有責任確保：

- 標籤資訊正確，
- 標籤資訊精確地紀錄並歸檔，
- 收集地資訊盡可能地準確，同時紀錄準確度與精確度資訊，
- 詳細記載收集方法，
- 標籤或田野筆記清楚明白，
- 標籤資訊清楚易讀，使資料輸入操作員明瞭。

標籤資訊或收集者筆記裡的資訊若是不清楚或有誤，事後要再修正就變得極為困難。分類資訊這部分所受影響較小，因為在取得「證據藏品」(voucher collection)後，通常會由專家再行檢查。

此外，在收集或觀察的當下應立即紀錄採集地及附加資訊，而不要像過去一樣，等到當天晚上或回到實驗室後才補寫。



組織所獲得的大部分資料多半來自「供應者」(supplier)，因此訂定完善的資料採集程序比到流程的下游修正錯誤來得容易。

The custodian or curator has the core or long-term responsibility

管理人負有長期責任

資料管理人（博物館、植物標本館、大學、保育機構、非政府組織或私人收藏者）對資料品質的維護與改善負有長期的責任（參見 Olivieri et al. 1995, p. 623 中管理人之責任列表），只要他們仍保有對資料的責任。重要的是，資料管理機構應將資料品質管理列為首要任務，並營造資料品質管理的組織文化，讓機構裡每個人都瞭解他們分擔了資料品質的責任。管理人有責確保：

- 採集者筆記正確無誤地輸入資料庫，
- 資料取得過程中實施品質管制，
- 資料及資料品質適當地、準確地紀錄，
- 資料定時驗證，
- 資料驗證過程詳實記載，
- 資料妥善地儲存歸檔（參見下面有關資料儲存），
- 有系統地保存先前的版本，以便日後比較，並可重新檢視「未清理」之資料，
- 保持資料完整性，
- 資料具即時性及準確度，紀錄詳實，讓使用者能判定資料是否適用，
- 管理人有維護隱私權、智慧財產權、著作權及涉及傳統/原住民所有人之敏感事物之職責，
- 維護並公開資料使用條款、使用限制及已知資料不適用之處，
- 履行並遵守與資料相關之法規，
- 即時回應使用者對資料品質的意見，
- 總是保持最高的資料品質維護，
- 詳實紀錄並讓使用者知悉所有已知錯誤。



資料所有人與管理人不只有權管理並管制資料流通，更有責任進行資料管理、品管與維護。為下一代管理資料更是管理人的道德責任。

User responsibility 使用者責任

資料使用者對資料品質亦有責任。使用者需向管理人回報任何資料錯誤、缺漏、資料建檔之錯誤及以後可能需紀錄的附加資訊。通常就是當使用者在別的資料背景下檢視某筆資料時，會發現平常不會注意到的錯誤及極端值。一間博物館可能只有全部資料的一部分（例如在某州或某區域以內），惟有當資料與其他來源的資料並置時，錯誤才會顯現出來。

依機構收集資料目的不同，使用者也能貢獻於協助組織未來訂定資料收集與驗證之優先順序(Olivieri et al. 1995)。

使用者也有責任決定資料是否適用，並且避免不當地運用資料。



在管理人維護資料品質的過程中，使用者與收集者扮演了重要的角色，若資料品質高，兩者也能從中獲益。

Building of partnerships 建立合作關係

對維護資料品質而言，建立合作關係是非常有益且節省成本的方法，特別是博物館及植物標本館間常有重複的紀錄。許多圖書館間即建立了一種合作與夥伴關係，以改善圖書館資源的編目(Library of Congress 2004)，博物館與植物館應可起而效尤。這樣的合作關係可以與下列對象建立：

- 重要的資料收集者（以改善資訊流程—如訂定標準化的資料收集與報告格式、提供全球定位系統等），
- 藏有類似資料的其他機構（如：副本藏品），
- 其他具類似資料品質需求的機構，且可能會建立資料品質管制方針、工具、標準及流程，
- 關鍵的資料中間人（如全球生物多樣性資訊機構，GBIF），扮演整理並散布來自多個資料提供者的資訊的角色，
- 資料使用者（尤其在分析前或其間可能進行資料驗證的使用者），
- 統計及稽核人員，能夠改善資料管理、流通及資料品質技術的方法。



您不是唯一處理資料品質的機構。

Prioritisation 排定優先順序

為了在最短時間內取得最高品質的資料供給最多使用者，可能有必要排定資料收集及/或資料驗證的優先順序（參見以下有關完整性之討論）。為達此一目的，可能必須：

- 優先著眼於最關鍵的資料，
- 專注在不連續的小單位（分類學上的、地理上的...等），
- 排定模式標本及重要證據標本之優先順序，
- 略過沒有用途或無法保證其品質的資料（即地理參照資訊不良的紀錄—但有些地理參照不良的重大歷史資料仍具重要性），
- 考慮價值最廣、對大部份使用者最有助益、用途最多樣的資料，
- 針對能以最低成本大批清理的資料進行處理（如：利用批次作業）。



資料的價值不一。著眼於最重要的資料，若需進行資料清理的工作，則確保這個程序**永遠**不需再重覆。

Completeness 完整性

各組織應追求其資料的完整性（或在優先順序排定後，各小單元資料的完整性，如分類學上的類群、地區等），確保在彙編資料時中使用到全部合格的紀錄。開放大量不完整的資料，不如提供小單元內的完整資料，因為利用不完整資料進行的分析將不夠全面。另外，各組織也應訂定缺漏資料(missing data)政策，定義缺漏資料的門檻及因應措施，並制定紀錄資料完整性的政策（參見建檔）。

Currency and Timeliness 即時性及時效性

資料即時性(currency)或時效性(timeliness)與三個重要因素有關：

- 資料於何時收集的？
- 何時更新過資料，以反映現況？
- 資料的即時性可以持續多久？

資料即時性是使用者常提出的議題。許多管理人傾向將即時性定義為資料最早收集或調查的時間點。從收集到發表之間會有時程上的延遲（生物資料可能會非常久），發表的資訊代表的是「過去」，而非「目前」的事實。生物多樣性資料的使用者大多都瞭解這點，這也成了此類資料的價值之一，也使得與其他資料類型有所不同。

以資料品質管理的角度而言，即時性（有時稱為時效性）較常用於描述資料的「保存期限」("use-by" period)，也可用於指稱資料前次檢查或更新的時間，這可能與資料中所用的名字特別有關。資料上次更新是什麼時候？符合最新的分類嗎？在現代分類命名法規下，若將一物種分成數個小分類群，其中一個分類群會保留概念較廣泛的名稱。對使用者來說，知道資料中所用的名稱代表的是廣義或狹義的概念可能非常重要。另外，就像食品一樣，資料的即時性可作為「保存期限」的同義詞，在保存期限過後，資料管理人便不再保證紀錄上命名資料的品質。

許多資料集的時效性與即時性可能不那麼重要，或難以納入或維護，舉例來說，大型博物館或植物標本館藏品便可能如此。但另一方面，對沒有證據的觀測或調查紀錄，或未依據最近的分類修訂更新的資料而言，時效性與即時性仍十分重要。這對二手藏品(secondary collections)亦是如此，包含經外部機構整合數個資料來源而來的收藏。例如，數個開發中國的機構開放其資料予母機構後提供給 GBIF 入口網，其呈現並非動態取自來源資料庫。

Update frequency 更新頻率

資料集中資料的更新頻率與即時性及時效性有關，需正式規範並紀錄下來。這包含新資料的加入及發行修正資料的頻率，因為兩者對資料品質都有影響，故對使用者很重要。人們可不想白費力氣下載或取得一個即將更新或改善的資料庫。

Consistency 一致性

Redman (1996)指出一致性有兩個面向：「語意一致性」(semantic consistency)，即從資料可得的觀點清楚、不模糊且一致；「結構一致性」(structural consistency)指的則是資料單元的類型與屬性具有相同的基本架構與格式。舉一簡單的例子，語意一致性表示資料總是在同個欄位裡，容易尋找，如種下階層(infraspecific rank)與種下名稱(infraspecies name)分屬不同的欄位，因此我們清楚知道種下名稱的欄位只包含一個名稱或小名(epithet)（見表 1），這不會造成混淆，因為有些資料只是填一個名稱而其他則在名稱前加上「var.」或「subsp.」等前綴字（見表 2）。

表 1. 種下欄位中的語意不一致性

屬	種	種下
Eucalyptus	globulus	subsp. bicostata
Eucalyptus	globulus	bicostata

表 2. 為達到種下欄位中的語意一致性，加上種下階層的欄位。

屬	種	種下階層	種下
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

一個設計良好的關聯式資料庫不會有允許太多這類語意一致的問題，但現有的收藏機構使用的資料庫並非設計如此周全。

結構一致性則是欄位內資料一致，如種下階層欄位（表 2）中，亞種的紀錄方式永遠一樣，而非有時記成「subsp.」，有時候寫成「ssp.」、「subspecies」、「subspec.」、「sspecies」等。只要資料庫設計優良，確保屬性結構完好，就可以避免結構不一致的問題。

兩種方法和紀錄上的一致性都有其重要之處，使用者可藉此知道檢驗的種類與方法、哪裡找資訊及如何解釋重要的資訊。不過，一致性與彈性仍需相輔相成(Redman 2001)。

Flexibility 彈性

管理人員需保有資料品質管制方法的彈性，雖然大部分的生物資料本質上相當類似，不同的品質管理方法仍適用於不同地區的資料（如哪些相關的資料集可用來比對此資料）、不同的分類群（水生者之於陸生生物等），或不同的資料收集方法（觀測或調查資料之於博物館證據藏品等）。

分類學觀點事實上是假說，不同的（有效的）分類觀點（假說）可能會導致同樣的生物被不同分類學家分進不同的類別，而擁有不同名稱，但每個名稱同樣有效(Pullan et al. 2000, Knapp et al. 2004)。例如，兩位分類學家可能對某些分類群應置於那個屬各持己見，有些將其視為桉屬(*Eucalyptus*)，有些則認為其應置於傘房屬(*Corymbia*)。在實際應用上，尤其在動物學裡，通常都會接受最新的修訂，除非有充分的理由能駁斥其觀點。

彈性讓我們的觀點能隨時修正，因應新的或不同的需求。分類學資料庫工作小組²和其他機構近來的研究著重於資料庫的結構，使其能呈現不同的分類觀(Berendsohn 1997)，雖然表面上這樣的彈性似會降低品質，事實上卻能賦予使用者極大的空間，決定資料是否適用，進而提升其認定的品質。

Transparency 透明度

透明度能夠提升資料使用者進行評估時的可信度。透明度指的是錯誤不被隱瞞，並能被指出、回報，資料驗證及品質管制流程詳實地紀錄並公開，開放並鼓勵意見反應機制。

² Taxonomic Databases Working Group, TDWG。網址：<http://www.tdwg.org>。譯註：現稱為「生物多樣性資訊標準組織」，Biodiversity Information Standards。

例如，建立收集方法的文件時（尤其是針對觀測及調查資料者），透明度便十分重要，它同樣讓使用者能判斷資料是否適合其預定的用途。

Performance measures and targets 績效評估與目標

績效評估是品管流程中很有價值的部分，它讓每位使用者確認在何種程度上他們能信賴資料準確度及品質標準。績效評估包含各項統計檢測，項目包括資料（如：95%的紀錄都在紀錄位置的一千公尺之內）、品管層級（如：65%的紀錄在過去五年內經合格分類學家檢測完成；90%的紀錄在過去十年內經合格分類學家檢測完成）、完整性（所有 10 分網格皆經抽樣檢測）等等。

績效評估能夠量化資料品質，帶來以下益處：

- 使組織確定其（已建檔之）資料品質，
- 輔助整體資料管理，減少重覆冗餘，
- 協調資料品質鏈的各個面向，整合後交由不同的操作人員實行。



測量資料品質水準前，先考慮使用者會如何運用這些結果，再組織這些結果，以提升使用效率。

Data Cleaning 資料清理

資料清理的原則會在《資料清理的原則與方法》(Principles and Methods of Data Cleaning) 文件中詳述。在此只修改 Maletic and Marcus (2000) 的看法，提出大致的框架：

- 定義並決定錯誤類型
- 尋找並確定出錯誤例子
- 修改錯誤
- 建立錯誤例子及類型之紀錄文件
- 修正資料輸入流程，減少未來發生類似的錯誤



別被看來容易的資料清理工具誘惑了。它們確實相當有用，但並非長遠之計，預防錯誤的工作無以取代。

Outliers 極端值

若找出空間資料中潛在的錯誤，偵測（地理、統計及環境方面的）極端值可能是最有用的檢測方法。但檢驗過程中，不應因某筆資料是統計極端值，便毫不考慮地刪除它。或許是歷史推演、環境型態改變、人類活動遺跡等的緣故，環境資料素來就有包含統計極端值的問題，但這些資料都是非常完好的紀錄，全盤刪除極端值可能會造成珍貴的資料遺失，導致未來分析研究的偏頗。

另一方面，若使用者不確定資料的效力，可能會決定要刪除極端值。因此，極端值的確定不但能協助管理人找出潛在的錯誤，更能輔助使用者決定某筆資料是否適用於其想要進行的分析。



偵測極端值是資料驗證中很重要的方法，但並非所有的極端值都是錯誤。

Setting targets for improvement 設定改善目標

訂定簡單易量化的目標可快速改善資料品質。若計劃每六個月將地理編碼不佳的新紀錄數量減半，兩年內整體的錯誤率便能降低 94% (Redman 2001)。改善目標應著重於：

- 設定清楚積極的時程，
- 改進的速度，而非實際的品質數據，
- 明確的定義（如「地理編碼不佳」(poorly geocoded)的定義），
- 簡單實際的目標。

除逐年減少資料清理所耗費的時間外，若能改善資料輸入及驗證的技術，便可同時訂定出長遠的改善目標。



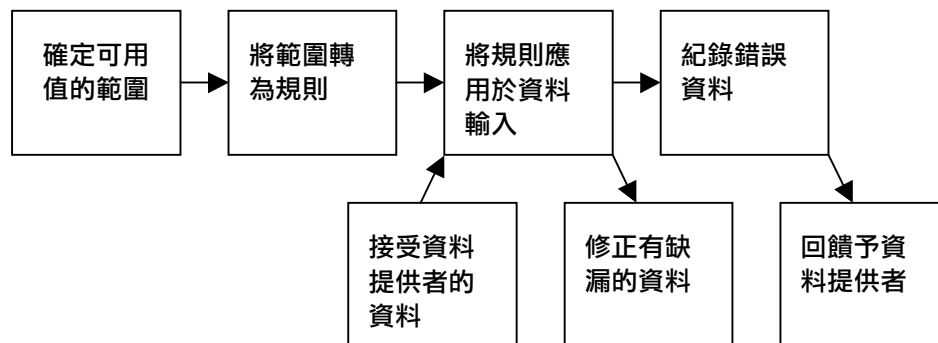
績效目標可讓組織維持固定的品質檢查與驗證作業。舉例來說，可設定資料在收到後六個月內，應歸檔並驗證完成95%的資料。

Auditability 可稽核性

管理人應清楚知道檢查過哪些資料及其檢查時間。如此可避免重複檢查，也不會忽略掉資料紀錄。最好的方法就是留下紀錄驗證的稽核存底(audit trail)。

Edit controls 編輯控制

編輯控制指的是應用規則(business rules)³中，特定欄位中可接受的數值。例如，月份的欄位中一定要是 1 到 12 間的數字，日期欄位中則是 1 至 31，最大數字為何則依月份有所不同。單一欄位(如月份)適用單變量(univariate)的規則，兩個欄位結合起來(如月份與日期)，則適用雙變量(bivariate)規則。



圖五. 編輯控制的用途(修改自 Redman 2001)

另一個例子是座標資料，經由簡易的範圍測試法（若資料以經緯度標示）可知，緯度應介於 0 至 90 度間，分秒介於 0 至 60 度。但使用二度分帶(UTM)座標的資料則較為複雜，一個資料庫可能包含來自二度分帶中一個小區域的資料，卻沒有該分帶的資料。這樣的資料如不與其他地區的資料結合，似乎還可以接受，但若試圖結合其它資料，資料幾乎不能使用。因此編輯控制需確保適當的分帶資料已包含在內。

Minimise duplication and reworking of data

將資料重複及重複作業減到最少

依企業界的經驗，使用資訊管理鏈（見圖三）可減少資料重複及重複作業的情形，且最多可降低五成的錯誤率，減少三分之二因使用不良資料造成的成本損失(Redman 2001)。這是由於資料管理及品質管制的責任劃分清楚、消弭障礙及作業等待時間、減少不同人員重複品質檢查、找出更好更精進的作業方法，效率因而提高。

Maintenance of original (or verbatim) data 維護原始資料

³ 譯注：指資料庫中，設定欄位資訊時所訂定的資料型態及預設的驗證邏輯。

收集者紀錄的原始資料或管理者事後補上的資料，都不應在編輯及資料清理過程中遺失。資料清理時，資料庫的變更應加註為附加資訊，並且保留原始資料，因為資訊一旦刪除，便難以回復。採集者及位置資訊的維護尤其重要。接手的管理人所發現的錯誤，可能不是真正的錯誤。位置名稱的改變（如捷克斯洛伐克改成捷克共和國）不只是名稱的改變，而是界線上的區別。有時瞭解原始的紀錄，而不只是修正後的資料，也是相當重要的。參見「歸檔」一節。

Categorisation can lead to loss of data and quality

資料的類型化可能造成遺失及品質損害

資料的類型化常會造成資料遺失，損害整體的資料品質。舉例來說，某資料可能紀錄了詳盡的位置資訊（甚至包括地理參照），卻儲存為網格資料。資料應盡可能以最高的解析度儲存，再依輸出需求進行歸類。例如，若使用者需要在 10x10 分的網格上畫出物種出現的分布圖，以點的形式儲存的資料便能夠輕易達成，但若是以網格形式儲存的資料，遇到更精細的需求便無法使用。此外，以不同網格尺度或不同來源歸類的資料難以（或不可能）結合，敘述性資料亦然，若區分資料類型時需要採用條件（如高於 6 公尺為喬木；小於 6 公尺為灌木），但在另一份來源的新資料中，卻以 4 公尺來界定喬木，那麼該如何處置界於 4 公尺及 6 公尺的植物？因此，資料儲存時最好標示確實的公尺數，之後再決定是喬木還是灌木。

地理編碼準確度的儲存也常發生相同的狀況。筆者總是建議地理編碼準確度建議以公尺標示，但很多資料庫都以範圍來區分（<10m, 10-100m, 100-1000m, 1000-10,000m）。若一紀錄的準確度經判定為 2km，卻只能將其放在準確度為 10km 的範圍內，當下就等於是失去了這項資訊。

Documentation 文件建立

完善的文件是資料管理中非常重要的原則。沒有完好的文件，使用者無法判定資料是否適用於預定用途，也就無法判斷資料品質。以下針對建檔將有更詳盡的討論。

Feedback 意見反應

資料管理人應鼓勵資料的使用者反應意見，並認真看待這些意見。如同前述之使用者責任中提及的，使用者在結合不同來源的資料時，發現某些錯誤類型的機會，比個別進行檢查的資料管理人高出許多。

要建立完善的意見反應機制並不容易。可以在查詢頁面增加一「意見反應鍵」，或在使用者下載資料時，寄送一附加檔案，解釋如何反應資料錯誤及意見的方法。有些相關內容在《資料清理的原則與方法》一文中有更多的說明。



開放給資料使用者及提供者有效的回饋管道，是改善資料品質既簡單又有效的機制。

Education and training 教育訓練

針對資訊鏈各階段的教育訓練能大幅提升資料品質(Huang et al. 1999)。首先訓練並指導收集者採用完善的收集程序、實作出使用者需求，其次訓練資料輸入操作員及負責資料庫例行政務的技術人員，最後教育最終使用者，讓他們了解資料的本質、限制及可能的用途。資料品質中教育訓練這一環與完善的文件息息相關。

在 MaPSTeDI 地理參照計畫(University of Colorado 2003)中，可以看到資料品質檢測與教育訓練之間如何整合。在程序中，除檢驗一定數量的地理編碼人員紀錄，對於新進的操作人員，還會安排監督人檢視其製作的前兩百筆資料之準確度。這樣一來不但能維持資料品質，更能讓操作員從錯誤中學習與改進。其後依操作員個別情況，可以進一步再確認一百筆紀錄，待操作員較熟練後，檢查方式改為隨機抽樣 10% 的紀錄筆數，最後減低至約 5%。若錯誤率持續偏高，則進行額外的檢查。

如此設計完善的流程能協助教育資料新手使用者。相反地，如果沒有制定流程，便難以確保操作員間及作業間的一致性。

Accountability 責任歸屬

整體資料品質的責任歸屬能幫助組織達到穩定的資料品質，提供錯誤回報的參考依據，亦提供建檔與查詢的聯絡窗口。



許多資料品質問題的根本原因在於訓練不足。

Taxonomic and Nomenclatural Data 分類及命名資料

品質不佳的分類學資料會「污染」相關的研究領域(Dalcin 2004)。

分類學是生物分類的理論與實務(Mayr and Ashlock 1991)。本文中提及的大部分物種資料都包含分類(或命名)資料(亦即物種名稱及其分類)，Dalcin (2004)將之稱為「分類資料域」(Classification data domain)。分類學的資料的品質及其判定方法與空間資料相當不同，前者較抽象，也更難以量化。

分類的資料包含(不一定能包含全部)：

- 命名(學名、俗名、階層、階級)
- 命名狀態(同物異名(synonym)、已接受(accepted)、模式化(typification))
- 參考依據(命名者、發表時的地點及年代)
- 判定(紀錄鑑定時的人員及時間)
- 品質欄位(判定準確度、品質鑑定者)

分類學名稱的一大錯誤來源是拼字錯誤。在分類資料庫中，若是科名或屬名等表示分類階層的學名出現拼字錯誤，比較容易發現(Dalcin 2004)。在這些情況，通常都能藉由大部分分類群的標準權威檔來處理，如「全球物種名錄建置計畫」(Species 2000) (<http://www.species2000.org>)及全球生物多樣性資訊機構的生物物種名錄計畫(ECat, <http://www.gbif.org/prog/ecat>)等提供的完整物種名錄。由於物種的名字或小名(epithets)可能隨著屬的不同而有些許變動，因此僅有種小名而未標明屬名的權威檔較不易符合需求。檢查拼字錯誤的方法之一，是利用相似度運算法，找出高度相似但不完全相同的學名，從而偵測並隔離出學名中的錯誤(Dalcin 2004, CRIA 2005)。

在資料輸入過程中利用學名(屬名、種名)及科名等清單，建立資料庫的權威檔，是迄今減少學名發生拼字錯誤機率最令人滿意的的方法。有了權威檔，在理想情況下應能幾乎根除拼字錯誤，可惜世界上大部分地區及許多主要的分類群尚未有這樣的清單。

若權威檔匯入自外部資源如「物種名錄」(Catalogue of Life)或已知生物物種名錄計畫(ECat)等，應於資料庫中紀錄其來源(Source-Id)，如權威檔來源版本變更時，資料庫也能同時合併變更而更新。希望不久，藉由使用「全球唯一識別碼」(Globally Unique Identifiers, GUIDs)⁴，這樣的工作能便利許多。

⁴ <http://www.webopedia.com/TERM/G/GUID.html>

資料的分類學品質十分仰賴分類專家的專業。分類學障礙(Taxonomic Impediment, Environment Australia 1998)以及全球各地訓練有素的分類學家日漸減少，會降低長期分類學的品質，造成原始物種資料品質下滑(Stribling *et al.* 2003)。「全球分類學倡議」(Global Taxonomic Initiative, GTI, CBD 2004)即試圖消除或改善所謂分類學障礙的情形，但這樣的問題很有可能延續至未來。另外，品質也可能隨時間衰退，尤其是沒有或未保留證據標本（如大多數的觀測及調查資料）或是分類專業知識付之闕如的領域。

機構創造高品質分類產物（包括已建立文檔的物種原始資料）的能力，受到下列因素影響（依據 Stribling *et al.* 2003）：

- 工作人員的訓練及專業程度，
- 技術文獻、參考資料、證據藏品及分類學專家是否易於取得或聯繫，
- 擁有適合的實驗室器材與設備，及
- 能夠取得網路及網路資源。

Recording of accuracy of identification etc. 鑑定準確度的紀錄

傳統上，博物館及植物館實施的鑑定管理制度(a determinavit system)⁵，讓鑽研分類群的專家不時檢查標本，判定其範圍界限（譯注：指物種觀）或進行鑑定。這經常包含在分類修訂的研究中，或由恰巧來訪機構的專家執行，檢查所有的蒐藏品。這個作法已被證明有用，卻相當耗時且雜亂無章。但自動化電腦鑑定在短期內、甚至長期看來，似乎不可能上軌道，因此目前別無他法。

另外一種選項是，在資料庫中加設欄位，填寫確定性(certainty)的指標。大部分的藏品資料庫通常都附有鑑定日期。而在此方法中，代碼欄位可以是(Chapman 2004)：

- 經世界級分類專家鑑定，高確定性
- 經世界級分類專家鑑定，具合理的確定性
- 經世界級分類專家鑑定，存疑
- 經地方分類專家鑑定，高確定性
- 經地方分類專家鑑定，具合理的確定性
- 經地方分類專家鑑定，存疑
- 經非分類專家鑑定，高確定性
- 經非分類專家鑑定，具合理的確定性
- 經非分類專家鑑定，存疑

⁵ 譯注：應為 determinative system。原文為系統，但中文應以「制度」較貼近語意。

- 經採集者鑑定，高確定性
- 經採集者鑑定，具合理的確定性
- 經採集者鑑定，存疑

以上的次序如何安排還有待商榷，同時也可討論這是否為最佳分類方法。據筆者所知有些機構有類似的欄位，但目前還沒能找到範例。植物標本館資訊交換標準第四版當中 (HISPID Standard Version 4, Conn 2000) 當中確實有一較為簡易的版本——「驗證層級標示」 (Verification Level Flag)，共有以下五種代碼：

表 3. 植物標本館資訊交換標準之鑑定層級標示(Conn 2000)

0	紀錄名稱尚未經專家核驗
1	紀錄名稱與其他已命名植物比較後決定
2	紀錄名稱經分類學家或其他合格人員以植物標本館及(或)圖書館及(或)活體而決定
3	植物名稱經參與此類群系統分類修正工作的分類學家所決定
4	紀錄為模式標本的一部份，或是由模式材料以無性生殖繁殖而來

許多機構有既定的格式紀錄確定性，如：「aff.」、「cf.」、「s. lat.」、「s. str.」、「？」等記號。雖然某些記號（如 aff.、cf.）有嚴格的定義，個人使用方式仍可能南轅北轍。狹義 (*sensu stricto*) 與廣義 (*sensu lato*) 即表示了分類學概念的變異。

此外，若非由分類專家命名，可以列出命名的依據（根據 Wiley 1981）：

- 新分類群描述
- 分類修訂
- 分類法(classifications)
- 分類檢索表
- 動、植物相研究
- 地圖集
- 目錄
- 名錄
- 手冊
- 分類學家/命名法規
- 親緣分析

比較兩篇以上的發表文獻或諮詢兩位以上的專家，可減少不確定性、改善品質。不同分類學家的鑑定各異，不表示其中一種鑑定錯誤，而只是決定分類群歸屬這種分類觀點上的歧異（比方說依據不同假說）。

Precision of identification 鑑定精確度

Stribling 等人(2003)表示，為評估鑑定精確度（Stribling 等人誤植為分類精確度），可隨機挑選不同分類學家或專家鑑定的樣本，進行比較。此外，也可比較不同機構收藏（亦鑑定過）的相同標本。但這些概念相當抽象，不確定紀錄這類資訊是否有其價值。

鑑定精確度的另一面向是標本鑑定程度。經鑑定至種或亞種的標本比鑑定至科或屬的標本資料精確度高。進行資料建檔時，應讓使用者瞭解，50%的資料只鑑定至屬的層級（動物資料常如此）。

Bias 偏差

偏差為數值產生一均勻的變動所造成的系統誤差(Chrisman 1991)，通常是因為採用一致的研究方法，而衍生出本質上為系統性的誤差。分類命名上的偏差可能是由於鑑定精確度高，準確度低，如錯誤解讀二分式檢索表(dichotomous key)或形態構造、使用無效的命名法規、依據過時的發表文獻(Stribling et al. 2003)或採用不合適的發表文獻(如：使用其他地區的植物誌來進行研究，但該區可能沒有所有與研究地區相關的分類群)。

Consistency 一致性

資料庫的分類部分，若相同的分類群有兩種以上經「接受」(accepted)之名稱，（如 *Eucalyptus eremaea* 及 *Corymbia eremaea*），便會衍生資料不一致的情形。這與分類學觀點不同、或不同拼法造成的錯誤有關（如 *Tabernaemontana hystrix*、*Tabernaemontana histrix* 及 *Tabernaemontana histrix*—CRIA 2005）。

Completeness 完整性

Motro 與 Rakov(1998 from Dalcin 2004) 將完整性指為是否具備所有資料，並將資料完整性分為檔案完整性（無紀錄遺失）及紀錄完整性（每筆紀錄皆無未知欄位）。

分類學的完整性（即名稱或分類資料庫之完整性）係指名稱包含的範圍。資料庫是否包含階層中的所有物種名稱（一直到亞種，或只有到種）？資料庫包含動物或植物界的哪些部分？資料庫中是否包含同物異名？種種資訊皆有助於使用者判定資料是否適用。例如，Dalcin (2004)將完整性分為兩類，一為命名完整性(nomenclatural completeness)，即在某種環境下，包含所有可能的名稱（如，在分類上，包含某分類群的所有名稱；或空間上，列出某地區所有名稱）；其二為分類完整性(classification completeness)，則代表與一分類群正式採用名稱相關之所有名稱（即完整的同物異名表）。

而標本或觀測資料庫之完整性可能是「是否包含所有的達爾文核心集欄位」或「是否所有的達爾文核心集欄位皆包含資料」。而特徵資料庫(character database)之完整性則為「是否包含生命史所有階段之特徵」（如植物的果實及昆蟲幼蟲之蟲齡）。

Voucher collections 證據藏品

證據藏品極為重要，但不可能所有的資料庫皆包含證據藏品。許多觀測資料庫建置過程中並未同時列入證據藏品，也不可能因政治、法律、保育等目的，每回皆能收集樣本做為證據。

能取得證據藏品進行驗證時，以物種為主的計畫在初期應訂定收集者與博物館或植物館等機構間的合作協定，以支持參考及證據藏品的存放(Brigham 1998)。協定中也應涵蓋資料歸檔及存放策略，包括資料處理或歸檔的最短時限。

Spatial Data 空間資料

空間資料引領了資料建檔標準的制定（如「空間資料轉換標準」(USGS 2004)、「歐洲空間資訊計畫」(INSPIRE, Information for Spatial Information in Europe⁶)...等），也常居於資料品質標準的前線（如地理資訊的國際標準化組織 ISO 19115 標準—後設資料⁷）。由於空間資料本質上多為數據資料，相較於分類資料，不但方便統計，也利於資料品質檢查方法的訂定（參見《資料清理的原則與方法》）。

但這不表示所有紀錄中的空間資料（Dalcin 2004 所謂「田野資料域」(field data domain)）都容易數位化或具高準確度。博物館或植物標本館中的許多歷史藏品，都只有描述收集地的基本文字，需耗費一番力氣才能將其轉化為地理編碼或座標，且許多藏品在收集時，收集者無法取得詳細的地圖，許多地名也不復見於現有的地名辭典或地圖上，使得處理更費工夫。補齊歷史紀錄的地理參照資訊，尤其是那些缺乏完善的歷史地名辭典的資料，不但需時甚久準確度也相當低。

為協助使用者進行資料的地理參照，許多工具和線上工具及指南紛紛開發出來，在《資料清理的原則與方法》一文中將詳細說明。此外，現今採集者多半利用全球定位系統，在採集的當下紀錄地理編碼。在資料擷取一章中將說明使用全球定位系統的準確度相關議題。

地理參照資訊的錯誤檢測包含：

- 檢查紀錄內部的其他訊息，或資料庫中其他紀錄—如州名、地區名等；
- 利用資料庫，檢查外部參考資料—紀錄與採集者的採集地是否一致？
- 利用地理資訊系統，檢查外部參考資料—該紀錄地點位於陸上還是海上？
- 找出地理空間中的極端值；或
- 找出環境空間中的極端值。

以上方法將於《資料清理的原則與方法》中詳述。

Spatial Accuracy 空間準確度

空間資料的位置準確度(positional accuracy)該如何評估？

相對而言，地理資訊系統圖層資料（地形圖等）的正確性較容易判定，因為在資料庫中通常會有其他幾項準確度更高的特徵，如三角量測點、道路交叉口等地點等等(Chrisman 1991)。不過，這些檢測並不簡單，記載方式也十分複雜，如美國國家地圖準確度標準(US

⁶ <http://www.ec-gis.org/inspire/>

⁷ <http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35>

National Map Accuracy Standard)。傳統上，空間準確度係利用數個完善定義的點與可接受的誤差範圍來決定，以其均方根差(root-mean-square deviation, RMSE)由零開始計算資料正確性(Chrisman 1991)。均方根差不容易應用於個別收集點，而較適於整個資料集或數位地圖。至於單一的收集點，利用簡單的點—半徑法(point-radius method)或相似的計算方法，即可估算出該點與真實地點間的距離(Wieczorek et al. 2004)。這其中有兩個因素須納入考慮，其一，這些完善定義點的準確度會影響到受測點的準確度，其二，測量方法本身的準確度與精確度亦會增加誤差。例如，若道路交叉口的準確度只有一百公尺，在將精確度納入計算前，採集點的中心即是一百公尺的圓形（參見 Wieczorek 2001）。

1998 年美國聯邦地理資料委員會(FGDC)公佈了地理座標準確度標準(GPAS, Geospatial Positioning Accuracy Standards)。標準中包含大地網絡(Geodetic Networks)及空間資料準確度兩小節(FGDC 1998)。

- 「全國空間資料準確度標準(NSSDA)利用均方根差來估計位置準確度。將資料集的座標值，與不同來源資料中相同地點但準確度較高的點座標值一一相比，計算出其差值平方的平均值，開根號後即為資料的均方根差。」
- 「地面距離準確度為 95%的信賴區間。準確度 95%的信賴區間表示，在資料集中 95%的位置與地面上實際位置的誤差等於或小於所報告的準確度。報告中的準確度反映所有的不確定性，包含因大地控制座標、編輯及地面座標值的計算。」

澳洲亦利用均方根差計算地圖準確度，準確度聲明舉例如下：

- 「本地圖完善定義細部平均水平準確度為 ± 100 公尺，垂直準確度 ± 20 公尺。」
(Division of National Mapping, Sheet SD52-14, Edition 1, 1:250,000)

任何以紙本地圖或電子地圖做地理參照的藏品，皆需附上準確度資訊。由於空間資料準確度具不確定性，準確度不可能達百分之百，因此資料準確度的記載相當重要。資訊鏈中處處可能衍生誤差，將造成最終結果的不確定性，無論是以地理資訊系統製作的地圖，或利用分布模型軟體製作的物種模型皆然(Heuvelink 1998)。

BioGeomancer project 生物地理風水師計畫

摩爾基金會(Gordon and Betty Moore Foundation)近來資助了一項計畫⁸，旨在協助改善原始物種資料之地理參照，並對準確度資料進行評估、改善及建檔。該計畫訂於 2006 年發表並發行其發展之工具。

False Precision and Accuracy 精確度及準確度錯誤

⁸ <http://www.biogeomancer.org/>

其外，亦需特別注意準確度及精確度的錯誤。許多地理資訊系統使用者未察覺空間資料準確度、誤差及不確定性的問題，常誤以為資料準確無誤。他們紀錄的準確度，常是其資料來源無法達到的標準。而許多機構也利用地理資訊系統輔助地理參照工作，將圖像放大至資料無法支援的尺度（且使用十進位度單位），亦會造成不實際的精確度。此外，利用地理定位系統紀錄採集地時，地點紀錄常準確至 1 或 2 公尺，然而事實上，攜帶式地理定位系統的準確度最多可能只能達到 10 公尺左右，這經常出現在以地理定位系統判定緯度的資料中（參見下面資料收集一節中的說明）。

Collector and Collection Data 採集者及採集的資料

採集者及採集的資料 (Dalcin 2004 的「藏品資料域」(Collection Data Domain)) 包含藏品本身的資訊——採集者、採集日期以及其他資訊如棲息地、土壤、氣候狀況、觀察紀錄等，可將分成下面幾類 (修改自 Conn 1996, 2000)：

- 採集者名、採集者編號
- 觀察紀錄等
- 採集日期/時期
- 採集方法 (觀測及調查資料)
- 相關資料

依資料類型，是博物館藏品、觀測資料或是詳細調查的結果，產生不一樣的問題。對靜態藏品如博物館藏品而言，採集者名、採集者編號與採集日期，以及相關資料中的習性、棲息地、(動物的)收集方法，皆是重要的屬性資料。觀測資料包括觀測時間長度、觀測範圍、時間 (除日期外，包括開始觀測時間與結束時間)、氣候狀況、觀察動物之性別、活動狀況等。調查資料則包含調查方法、(網格及全部區域) 大小、成果、氣候狀況、頻率、是否有證據藏品、證據藏品數以及與觀測資料相同的屬性。

Attribute Accuracy 屬性準確度

在採集資料中，影響資料品質的問題包括採集者名、編號、字母首等的紀錄方法(Koch 2003)、紀錄日期及時間的準確度、採集時紀錄習性、棲息地、土壤、植被類型、花色、性別、相關物種等相關資料的一致性。

採集資料常出現的問題為「採集者編號」，很多採集者不為藏品編列獨特的號碼，但由於這些標籤有時會用以判定收集地點、鑑定、不同機構的副本藏品等，可能會因此降低資料品質。

Consistency 一致性

有關採集的術語在使用上常無規律可循，尤其是同一資料集中的欄位很少能維持一致，遑論不同資料組。

Completeness 完整性

採集資訊的完整性常有變動。許多資料經常不會完整紀錄棲息地、採集者編號及開花等訊息，因此若要單靠藏品進行棲地的研究，並不容易。

Descriptive Data 描述性資料

描述性資料庫越來越常用於資料儲存，並常取代傳統發表方法，成為發表資料的新場域。形態學、生理學與物候學資料都屬於描述性資料的範疇。敘述性資料常可用於支序分析，也作為自動化產生描述與鑑定工具。

分類學資料庫工作小組長久以來致力於建立並提倡描述性資料庫的標準，初期建立了 DELTA 標準(Dallwitz and Paine 1986)，近來則協助成立「描述性資料結構」工作小組(Structure of Descriptive Data, <http://www.diversitycampus.net/Projects/TDWG-SDD/>)。

描述性資料的品質不一，雖然通常資料各元素都有測量，事實上當中有許多資料因為無法觀測（如歷史資料）、難以觀測（如成本太高）且（或）是非客觀事實（如顏色、豐度等主觀評估）而影響判定準確度。

描述性資料通常是以物種為單位來儲存，而較少以標本資料為單位，因此常以平均值或間距表示。如 Morse(1974, Dalcin 2004 發表)所指出，分類學資訊的可靠度本來就較標本觀測資料低。無論如何，近來開始傾向以標本資料為單位儲存這些資料（至少其中一些），品質因而提升。

Completeness 完整性

就標本而言，描述性資料紀錄的完整性取決於標本品質及年代等。例如，利用同一個標本恐怕無法紀錄果實或花卉的特徵，因此許多欄位必須留白。另外，屬性可能與特徵無關，因此不需紀錄所有的屬性。

Consistency 一致性

不一致的問題可能源自於兩個相關的資料項目。例如，兩個物種描述特徵可能記為 (Dalcin 2004)：

- 習性=草本(“HABIT=HERBACEUS”)及
- 用途=木材(“USES=WOOD”)

以不一致的格式來表示相同屬性亦會影響品質，尤其是屬性的定義不佳，或是未堅持依循一致的標準時，例如(Dalcin 2004)：

- 花色=洋紅(“FLOWER COLOUR= CARMINE”)，及
- 花色=深紅(“FLOWER COLOUR=CRIMSON”)。

採用標準術語能大大減少錯誤率及誤解程度。許多領域及學門皆開始訂定標準術語，近來聯合描述性資料庫的建立，亦提高了字彙使用的一致性。分類學資料庫工作小組制定的描述性資料架構標準(SDD, TDWG 2005)也能有助於這樣的發展。

Capturing Data 資料擷取

原始物種資料及物種出現資料的擷取方法不一，每種方法精確度與準確度各異，也各有其錯誤來源及不確定性，因此對資料最終適用性或品質有不同的影響。以下將簡要討論較常見的幾種物種資料擷取方法。

Opportunistic 機會性取得

物種出現資料多半是機會性採集的，許多皆為標本，收藏於博物館及植物標本館中。歷史資料多半只包含空間敘述文字，如「位於某鎮西北方 5 公里處」等，鮮少在採集當下即標示地理參照，而且常是由其他人在事後補上(Chapman and Busby 1994)。許多觀測紀錄如鳥類地圖集資料等，也是機會性地取得。

這類資料通常以數位化的方式批次擷取，並參考實際地圖進行地理參照，因此精確度及準確度非常低，這些資料準確度大多只達 2 至 10 公里。

Field Survey 田野調查

田野調查資料通常包含經緯度或二度分帶座標等空間參照。空間參照的準確度約 100 至 250 公尺。但需特別注意空間參考指涉的對象為何，可能不是實際觀測地點，而是樣線的中點、或網格的頂點（或中心），並不總是清楚明瞭。此外，很少紀錄含有證據藏品（亦即用作日後參照的實體藏品），因此分類學的準確度不能恆信之。而隨著調查時間越久遠，分類觀念更迭，分類學的準確度的可靠度就越低。

Broad-scale Observations 大範圍觀測

有些生物調查只紀錄特定範圍或網格內的資料。例如，國家公園內的物種調查或 10 分網格內的鳥類觀測資料（如：Birds Australia 2001, 2003）。這種紀錄的準確度可能只有大約 1 至 10 公里或 10 公里以上。

Global Positioning Systems (GPS) 全球定位系統

全球定位系統在物種資料的收集上扮演的角色越來越重要。除調查資料外，也包含機會性取得與觀測資料。

全球定位系統利用三角測量法判定地球表面某位置的地點，量測的距離即是 GPS 接收器及 GPS 衛星間的距離(Van Sickle 1996)。由於已知 GPS 衛星在於太空中的位置，因此可計算

出地面上的位置。地表上的地點位置最少需四顆 GPS 衛星方能測得(McElroy et al. 1998, Van Sickle 1996)，對今日的科技而言，大部分的地點都可接收到七個以上的衛星訊號，不會造成阻礙，但以往衛星數目有時會不夠。2000 年五月以前，大部分民間 GPS 系統還有「選擇性干擾」(Selective availability)的問題。干擾碼解除後，準確度大幅提高(NOAA 2002)。

選擇性干擾解除前，生物學家及觀測人員使用的攜帶式 GPS 接收器，準確度大約在 100 公尺以內或更差(McElroy et al. 1998, Van Sickle, 1996, Leick 1995)。其後 GPS 接收器的準確度日漸提升，在室外使用四個以上的衛星時，大多數的攜帶式 GPS 製造商可保證誤差在 10 公尺以內，而同一地點的數次觀測資料經平均後，可進一步提高準確度(McElroy et al. 1998)。現代的 GPS 接收器包含平均值演算法，可將準確度提高至 5 公尺以內。

使用差分全球定位系統(Differential GPS, DGPS)也可大幅提升準確度。差分全球定位系統參考已知地點的 GPS 基地（通常是調查控制點），以校準接收的 GPS。藉由基地及攜帶式 GPS 同時參考衛星位置，減少因大氣狀態造成的誤差，讓可攜式 GPS 將修正量適當地加在定位點上。依接收器的品質，準確度可預估界於 1~5 公尺間不等，而接收器離基地越遠，準確度越低。同樣地，透過平均值演算，亦能提升準確度(McElroy et al. 1998)。

廣域擴充系統(Wide Area Augmentation System, WAAS)是為飛機精確導航建立的 GPS 導航及降落系統(Federal Aviation Administration 2004)。WAAS 包含地面的天線，其精確的已知位置能提昇 GPS 的準確度。此外，也正在開發類似的「區域擴充系統」(Local Area Augmentation System, LAAS)，以提供更細緻的精確度。

即時差分全球定位系統(Real-time Differential GPS, McElroy et al. 1998)或靜態全球定位系統(Static GPS, McElroy et al. 1998, Van Sickle 1996)能接收到準確度更高的訊號。靜態全球定位系統使用高精確度的儀器與技術，一般只有調查人員使用。澳洲利用這些科技實施的調查，報告的準確度皆在公分的範圍內。但由於所費不貲，而且一般對精確度沒有這樣的要求，這兩項科技不太可能廣泛應用於生物紀錄收集。

要達到上述的準確度，GPS 接收器上方不能有障礙物或反射表面，且能清楚看見地平線（例如，在濃密的林蔭下 GPS 接收不良）。GPS 接收器要能夠接收到至少四個幾何排列位置適當的衛星訊號，而最佳幾何排列位置是「一個衛星位於正上方，其餘三個沿地平線呈等距排列」(McElroy et al. 1998)。GPS 接收器也要設定為適合該地區及紀錄用的單位基準。

GPS 高度：大多生物學家不太瞭解經由 GPS 測得的高度，事實上是與「地心基準」(Earth Centric Datum)相對（因此與地球的橢圓表面有關），而非平均海平面或是像「澳洲高度基準」

(Australian Height Datum)等標準高度基準。以澳洲為例，GPS 接收器顯示的高度與海平面差距可以從-35 到+80 公尺，且其變化常無法預測 (McElroy et al. 1998, Van Sickle 1996)。

Data Entry and Acquisition 資料輸入及取得 (取得電子資料)

資料輸入及取得本來就容易出錯。(Maletic and Marcus 2000)

Basic data capture 基本資料擷取

擷取資料的第一步通常是蒐集來自標本標籤、日誌、田野筆記、收藏登錄手冊或目錄卡中的資訊。這樣的工作可由專門或一般的資料輸入員進行，亦可直接掃描資訊。若要減少資料輸入過程產生的誤差，可藉由二次鍵入、或利用具有掃描檢查的學習及訓練軟體，並讓專家或主管隨機取樣檢驗輸入的資料。（參照下面 MaPSTeDI 守則。）

User-interfaces 使用者介面

建立使用者資料輸入介面，也可減少因輸入產生誤差的機率。許多機構請一般的員工或志工輸入資料，故建立一個簡單（非技術性的）使用者介面，讓資料輸入人員得以上手，可提高資料輸入的正確率。這樣的介面有助於資料輸入，不但可快速搜尋權威檔欄位、資料庫內現有的條目、其他相關資料庫等，甚至可利用 Google 一類的搜尋引擎，當輸入員無法辨識標籤、或難以決定什麼欄位要輸入什麼資料時，可提供正確的拼法和術語。有時也可藉由資料庫的設計，加以應用此方法；亦即，設計結合了權威表格及下拉式選單（項目選單），減少資料輸入員的麻煩，不須決定要輸入什麼名字、地點或棲息地。

Geo-referencing 地理參照

地圖是傳遞資訊最有效的方法之一，單憑此點，便能說明為何近來博物館的與植物標本館各項標本資料，在地理參照和資料庫建置、擷取地理參照的觀察資訊上，都有所增加。地圖資料處理能力之提升，讓我們得以針對錯誤和有疑義之處，進一步研究、辨認、呈現、紀錄與修正(Spear et al. 1996)。如此一來，也提供了有效的方法以呈現、傳達資料內有疑義之處，進而能讓使用者決定資料的品質和適用性。

擷取電子資料並附上地理編碼(將資料以地理參照表示)，有可能是項艱難費時的任務。MaPSTeDI 計畫(University of Colorado 2003)的結果顯示，稱職的操作人員每五分鐘便完成一筆紀錄的地理參照。其他研究(Armstrong 1992, Wieczorek 2002)則顯示，地理參照可能耗時更久，例如，美國的操作員每小時可完成 MANIS 資料庫中九筆紀錄，北美洲其他國家只完成六筆，而其他國家的人則僅能完成三筆(Wieczorek 2002)。

MaNIS/HerpNet/ORNIS

地理參照方針

<http://manisnet.org/manis/GeorefGuide.html>

MaPSTeDI

MaPSTeDI 的地理參照

<http://mapstedi.colorado.edu/geo-referencing.html>

為協助資料管理人的地理參照工作，許多極佳的方法與導引應運而生。柏克萊大學脊椎動物博物館的 John Wieczorek 編寫了「地理參照導引」(Georeferencing Guidelines)，與「山區平原時空資料庫資訊學倡議」⁹同為目前該議題最完善的研究，可供讀者參考。這些導引都提及了如何判定文字敘述性地點的位置準確度及精確度、使用不同基準產生的不確定性、利用不同地圖比例尺的影響等。這兩者均相當詳實，希望讀者將其視為與本文的附件。

此外，也有許多線上工具能協助使用者確定地理編碼，如某距離外的某地，或已知地點的某方位。這在《資料清理的原則與方法》一文中將另行說明。

BioGeoMancer

(Peabody Museum of Natural History) (皮巴蒂自然史博物館)

<http://www.biogeomancer.org/>

geoLoc

(Reference Centre for Environmental Information) (環境資訊參考中心)

<http://splink.cria.org.br/tools/>

Error 誤差

以上提及的工具對減少誤差及品質提升皆非常有效。但沒有一項地理編碼方法能全然免除誤差。如 MaPSTeDI 導引中所述：

「由於地理編碼方法不全然是一門科學，且沒有藏品的地理編碼能百分之百正確，品質檢查能大幅提升藏品地理編碼正確的比例。因此每個計畫在規劃地理編碼程序時，都應將品質檢查納入考量。」(University of Colorado2003)

常見的地理編碼錯誤，來自於全盤接受電子地名辭典的資料。有時候這些辭典是依據紙本地圖出版品而建構，因此地點的位置標示於左下方；事實上，左下方標明的應是地名而非

⁹ MaPSTeDI, Mountains and Plains Spatio-Temporal Database Informatics Initiative

位置(1988 年以前澳洲國土資訊處(Australian Land Information Group)建置的「澳洲地名辭典」即是一例)。希望大部分的地名辭典都已改正,但有些博物館及植物標本館的資料可能已根據錯誤的辭典進行地理參照。因此應隨機挑選採集地資訊,與地名辭典或正確的大範圍地圖對照,以檢驗這類紀錄的準確度。



標籤資訊數位化後,獨立進行地理參照的工作,會更迅速有效率。如此一來,資料庫可將藏品依採集地、採集者、採集日期等分類,也能更有效地使用地圖,以取得地理編碼資訊。此外,同一採集地點的不同紀錄也無須重複地編碼。

Documenting Data 資料建檔

後設資料是描述資料的資料，描述了為特定目的所收集的資料之特徵。

(ANZLIC 1996a)

在資料集及資料紀錄兩層級上，都可達到完善的資料建檔。

後設資料提供了資料內容、廣度、可存取性、即時性、完整性、符合目的及適用性的訊息。藉由後設資料，使用者可瞭解資料集的品質，在使用前確定資料是否適用。良好的後設資料能促進資料流通、搜尋與擷取。後設資料通常指涉整個資料集，但有些人將建立在紀錄層級資料的描述文檔（如準確度紀錄）視為紀錄層級的後設資料。無論名稱為何，資料集層級與紀錄層級的資料建檔是否完善，都至關重要。

所有的資料都有誤差，沒有資料得以倖免，重要的是，要知道錯誤為何，並知悉誤差是否在使用目的之可接受範圍內。在這種情況下，整體資料的後設資料扮演關鍵角色，而資料「適用性」一詞，正是因後設資料的建置而凸顯。90 年代初期，空間資訊的適用性概念尚未受到重視，直至中期，文獻中才開始以此處的概念提到(Agumya and Hunter 1996)。

僅限於資料集層級的資訊紀錄，無法隨時提供使用者所需的資訊。紀錄層級的誤差紀錄，對判定資料（尤其是物種資料）適用性極為重要。若能取得這樣的資料，使用者便能指定所需資料的條件，如準確度高於 5000 公尺的紀錄等。自動化地理參照工具也應在輸出資料中包含一欄位，標示計算出的準確度。

資料使用者也應瞭解適用性的概念。使用者從資料庫取得物種出現資料時，格式常記為「紀錄編號 x, y」，不論是否附帶有準確度的資訊。座標資料皆標示為點，但鮮少、也幾乎從不代表實際地點。有些資料在輸入資料庫時，只任意標示上一個地點（如：標示為「南美洲」的藏品標籤），並在準確度的欄位填上 5,000,000 公尺。確實有資料庫是這麼做的！若得到的是這樣的紀錄，並採用其標示的任意地點，會造成重大的誤導。因此使用者需注意資料是否包含準確度的欄位，並遵循使用建議。若身為資料提供者，且提出了資料標準報告，則應強制要求送出資料時，須同時包含準確度欄位。



資料建檔時應包含詳實的後設資料，讓第三者使用時無須再參考資料的原始來源。

圖 6. 利用 MaPSTeDI 搜尋工具找資料。

<http://www.geomuse.org/mapstedi/client/textSearch.html>. 圖中顯示，在紀錄層級的檔案中，可依準確度搜尋資料。

為讓使用者能夠判斷資料品質是否符合其目的，需紀錄空間資料的準確度、精確度與誤差。建檔時應（至少）包括下列各項：

- 資料集標題
- 資料來源
- 資料歷史（資料自收集始或傳遞之後所經過的處理程序）
- 準確度（位置、時間及屬性）
- 邏輯一致性
- 日期及資料保存期限（資料即時性、狀態、更新頻率）
- 資料欄位定義
- 收集方法
- 完整性
- 使用說明與限制（如版權、授權合約限制等）
- 管理人及聯絡資訊

這些術語值得定義，因為不是所有的資料管理人都熟悉。這些術語大部分用於資料庫的資料收集，而非個人的採集物資料。

Positional accuracy 定位準確度

定位準確度指的是座標描述地點與實際位置的相近度(Minnesota Planning 1999)。情況允許時，應註明用來判斷座標位置的大地基準。

此外亦建議資料庫欄位中顯示每筆紀錄的定位準確度。作法有很多，有些資料庫使用代碼，不過一般偏好利用簡單的度量值來表示該筆紀錄的估計準確度(Chapman and Busby 1994, Conn 1996, 2000, Wieczorek et al. 2004)。這對為特定目的而提取資料的使用者而言，特別重要，比方說使用者可能只想要準確度高於 2000 公尺的資料。有時候在紀錄層級的資料中納入地理參照資料判定方式的欄位，也可能很有價值。例如：

- 使用差分全球定位系統
- 受到選擇性干擾影響的攜帶式全球定位系統（如 2002 年以前的系統）
- 參照比例尺為 1:100000 之地圖，並利用可快速識別的地點特徵進行三角測量
- 以「方位推估法」(dead reckoning)繪製的地圖
- 從遠處（如直升機上）繪製的地圖
- 利用點一半徑算法，以地理參照軟體自動取得
- 利用地名辭典，包含地名、日期、及地名辭典版本

Attribute accuracy 屬性準確度

屬性準確度意指資料中的特徵描述與真實世界中相比，經評估後的正確可信度。理想的屬性描述應包含屬性列表及其個別的準確度。如：

紀錄由經驗豐富之觀察人員提供。將屬性資料與博物館或植物標本館藏的證據標本相驗後，準確度提高。約 40%的植物紀錄、51%兩棲類、12%哺乳類、18%爬蟲類及 1%的鳥類紀錄皆經證據標本查驗(SA Dept. Env. & Planning 2002)。

Lineage 世系

此處世系指的是資料的來源以及資料成為目前的狀態前經過的各個處理步驟。世系可能包含資料收集方法（如「資料自 10x10 公尺網格中收集」）及資料所使用過的驗證測試等資訊等。資料處理歷史可能包括：

- 資料擷取方法
- 中間的處理步驟及方法
- 產生最終產品的方法
- 任何驗證資料的步驟

例如：

資料自 20x20 公尺的固定樣區中收集。同時亦取得了物種總數、結構及棲地資料，資料經「雙向指標種分析法」(TWINSPAN)，依物種歸類的相似度來分群。

Logical consistency 邏輯一致性

邏輯一致性簡單地評估了資料內各項目間的邏輯關係。雖然大多數的博物館與植物標本館資料中，有些項目之間可能不具相關性，有些觀測資料（如國家公園或生態區域的物種清單等）或調查資料則有。而數位化儲存的空間資料，可進行邏輯一致性的自動化檢測，例如：

- 是否標示了所有的點線面，是否有所重複？
- 線是否相交於節點，或非刻意地交叉？
- 多邊形的邊線是否閉合？
- 所有的點線面是否符合拓樸學的關係？

邏輯一致性也適用於資料集中有其他邏輯關係的物件，因此任何關係的檢測結果都應描述紀錄下來。例如日期紀錄在數個欄位的情況，若欄位中說明收集日期在 a 年與 b 年間，但另一欄位卻表示屬性的紀錄日期不在此範圍內，邏輯一致性便無法成立。或者，紀錄在地理範圍外，像是一個欄位中表示收集地點在巴西，另一欄位標示的經緯度卻在巴拉圭，便是這兩欄位間的邏輯不一致。資料檢測的記載，是後設資料中很重要的一部分，可能包括地理資訊系統領域相同目的之「區中點」(point-in-polygon)。《資料清理的原則與方法》一文中將詳細說明檢測方法。

Completeness 完整性

完整性是指相對於資料或資料集及所能包含的最大容納量而言，目前所涵蓋時空範圍的比例。在判定資料品質時，必須有完整性的記載。舉例如下：

南緯 30 度以北區域記載完整，唯南緯 30 ~ 40 度間紀錄零散。

資料集只包含 1995 年間的紀錄，且大多為機會性地取得。主要來自新南威爾斯，但也包含其他州的紀錄。

從使用者的角度而言，完整性包含「所需的所有資料」(English 1999)，也就是說，使用者需要知道資料庫是否包含他們研究所需的所有欄位，各欄位是否完整。例如欲進行長時間的屬性比較時，若資料庫的資料只追溯至某年，便可能無法滿足其需求(參見上面第二例)。

Accessibility 易得性

對使用者而言，有價值的資料需容易取得。並非所有的資料都在網路上流通，要取得某些資料，可能需連絡管理人以取得使用同意，或取得所需資料的 CD 片。資料流通及使用狀況的記載也相當重要，讓使用者得以取得資料，因此也是資料品質的一環。易得性的紀錄可包括：

- 資料事務的連絡資訊
- 流通狀況
- 取得方式（若可經電子流程取得）
- 資料格式
- 注意事項
- 版權聲明
- 費用（如果有）
- 使用限制

Temporal accuracy 時間準確度

時間準確度為時間資訊的準確度。例如，「資料之準確度只到月份」。某些資料庫的「日」欄位不能留白，因此當缺少此項資料時，系統會自動在日期欄位中填入「1」。這可能造成精確度的假象。有時只知道紀錄的年份，資料庫便自動記為該年的 1 月 1 日，若使用者研究植物花期或鳥類的遷徙模式，必須要知道這樣的情況，瞭解就他們的研究而言，這筆資料的品質低，並不適用，便可排除該紀錄。

Documenting validation procedures 記載驗證程序

要瞭解資料中存有哪些誤差，誤差的記載是關鍵之一。若資料經過品質檢測與修正，卻未紀錄下來，使用處不大，尤其當施行檢查者非資料的原作者時，驗證紀錄更為重要。因為後人認為的錯誤很可能不是錯誤，修正會造成新的錯誤。此外，我們無法負擔資源浪費，因此檢驗工作也不能一再重複進行。使用者進行資料品質檢查後，可能會指出一些有疑問的紀錄，其後經檢查可能發現是非常好的紀錄或是真的極端值，倘若這次的檢查沒有記載下來，下次可能有其他人質疑同樣的紀錄，並將其排除在外，或浪費寶貴的時間重新檢查。這是基本的風險管理，所有資料管理人與使用者都應定期實施。建立完善的紀錄，其價值與需求的重要性不必多說了。它讓使用者明可白資料為何，品質如何及合用之處，也讓資料管理人能夠追蹤其資料，免於浪費資源在重新檢查疑似錯誤。

Documentation and database design 文件紀錄及資料庫設計

為確保誤差詳實記載，可在資料庫設計與建置初期將其納入考量，增設品質/準確度欄位，如位置或地理編碼準確度、地理編碼與海拔資訊來源、資訊紀錄者——座標資訊是採集者利用 GPS 紀錄的，抑或是資料輸入人員事後利用某比例尺的地圖加上的？海拔高度是否為利用數值地形模型(DEM)自動產生？若答案是肯定的，數值地形模型的資訊的來源為何？日期及尺度為何等。以上的訊息對未來判定這些資料對某用途是否有價值時十分重要，而使用者屆時也能自行做出決定。

「若分類學的資料集未記載其效能特性，資料使用者以該資料進行生物評估時，需特別謹慎小心。」(Stribling et al. 2003)

Storage of data 資料儲存

資料儲存對資料品質有諸多影響。大部分的影響不甚明顯，但在設計儲存容器（資料庫）及資料鏈中的一個單元時，仍須將其列入考量。

資料庫的選擇與建構牽涉的範圍甚廣，無法在此一一涵蓋，應另行研究討論。全球生物多樣性資訊機構授權的一項研究中，即檢視了「藏品管理軟體」(Collection Management Software, Berendsohn et al. 2003)，可供讀者參考。

因為與資料品質相關，本節將探討資料儲存的一些主要原則。

Backup of data 資料備份

資料定期備份可幫助確保維持品質，組織需進行即時災害復原及備份程序，因為每次的資料遺失或損壞，都將伴隨資料品質的流失。

Archiving 歸檔

資料歸檔（包含資料淘汰及丟棄）是資料及風險管理中需要特別注意的面向。資料歸檔對各大學、非政府組織及私人收藏者而言，是資料管理中的首要議題。大學的人員流動率高，研究資料的儲存經常是分散的——常存於研究人員的個人電腦或資料櫃中，因此資料若未詳實建檔，其可用性及易得性會迅速流失。研究人員離職後，由於沒人知道資料為何或在乎花精神維護，資料遭丟棄的情形更是屢見不鮮。因此大學特別需要一套完善的資料建檔及歸檔策略。

未在大型機構服務的研究員，也應確保在身後或不再進行這類研究後，其資料仍能繼續維護或歸檔。同樣地，非政府組織若沒有長期資金支助資料儲存，應與適合且有長期資料管理策略（包括歸檔）、且可能對其資料有興趣的機構合作。

近年來隨著分散式通用資訊擷取協定(DiGIR)/達爾文核心集及 BioCASE/ABCD¹⁰協定的制定，資料歸檔日漸容易。這些協定使得各機構、大學系所或個人可將資料庫以某種格式輸出，再以 XML 格式儲存於自己的網頁或傳送至母機構。如此一來，資料便能夠永久保存，並且（或者）經由 GBIF 入口網站的搜尋程序流通。

在全球資訊網中，資料的清理、刪除及歸檔對也十分重要。遭設立者棄置或資料老舊過時的網站，在網路空間留下許多「數位碎屑」(digital debris)（說法名稱各異）。組織的資訊

¹⁰ <http://www.tdwg.org>; <http://www.gbif.org/links/standards>

管理鏈中應包含資料歸檔策略，實體的資料歸檔主題太廣泛，無法在此詳述，但近來美國國家標準與技術中心¹¹與資訊與圖書資源委員會¹²發表了一份文件，說明利用 CD 及 DVD 存檔的方法(Byers 2003)，是相當有價值的摘要，可供讀者參考。



(因法律或其他原因)無用的資料不應銷毀，或在沒有盡其用時(包括歸檔)，讓資料冒遭棄置的危險。(NLWRA 2003)

Data Integrity 資料完整性

資料的完整性，即資料沒有在未經授權的情況下遭擅改或毀壞，亦未受到意外或惡意修改、改寫及損壞(如病毒侵襲或電壓突波)。

資料因時而異，紀錄中的分類學資訊可能因重新判定而改變，但使用者期望電腦系統能維持資料的完整性，且認為電腦不會不慎或錯誤地改變資料值。當資料不再完整，無意或錯誤的改變發生時，稱為資料損毀(data corruption)。



資料完整性的保持有賴完善的資料管理、儲存、備份及歸檔。

Patterns of error 錯誤類型

如同所有資料庫，分類學的及物種出現資料都容易發生內容錯誤。English (1999)提出以下幾種錯誤類型，稱之為資料缺陷(data defect)，其後經 Dalcin (2004)應用於分類學的資料庫。以下採用 English (1999)的釋義，並就 Chapman (1991)、澳洲虛擬植物標本館(Australian Virtual Herbarium¹³)資料庫及巴西的物種連結(speciesLink)¹⁴舉例說明：

- **值域重複**(Domain value redundancy)—當資料未標準化或存在異名時，可能會有兩個以上的值或代碼代表相同意義。若術語未標準化，或來源不同的資料整編沒有做好，敘述性資料便容易發生重複。
- **資料數值缺漏**(Missing Data Values)—資料欄位應填而未填。這包括必填欄位，以及收集當下不必填，但事後處理需填的欄位。地理參照或座標值(經緯度)都是。
- **資料值不正確**(Incorrect Data Values)—可能肇因於打字錯誤、資料輸入位置錯誤、誤解收集之資料、無法辨認標籤的字跡或打字員不知道必填欄位該填什麼。錯誤的

¹¹ United States National Institute of Standards and Technology

¹² The Council on Information and Library Resources

¹³ <http://www.cpbr.gov.au/avh/>

¹⁴ <http://specieslink.cria.org.br/>

資料值是最明顯但也是最常見的錯誤，可能會影響到每個欄位的每個值。與此相關的還有分類學的及命名資料庫的學名拼字錯誤，也相當常見（參見本文相關討論），此外也包括地理參照欄位補零等問題。

- **非單純資料值(Nonatomic Data Values)**—一個欄位中輸入一個以上的資料（如：一個欄位中填了屬名、種名及收集人名，或將層級及種下名稱填在同個欄位）。這種錯誤常起源於資料庫設計不良，這類錯誤可能會對資料整合造成實際的困難。

表 4. 非單純資料值的例子

屬	種	種下
Eucalyptus	globulus	subsp. bicosata
科	種	
Myrtaceae	Eucalyptus globulus Labill.	

- **值域錯亂(Domain schizophrenia)**—在欄位中填入的值不符合該欄位用途，導致欄位中資料性質不一。

表 5. 值域錯亂的例子

科	屬	種
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

- **重複出現(Duplicate Occurrences)**—相同實體的紀錄重複出現。最典型的例子為，相同學名有不同拼法或其他有效的命名。使用者在搜尋學名或結合不同資料庫的資料時可能會遭遇困難。舉例如下：
 - *Phaius tancarvilleae*
 - *Phaius tankervilliae*
 - *Phaius tankervilleae*
 - *Phaius tankervilleae*
 - *Phaius tankervillae*
 - Brassicaceae/Cruciferae（同義字；兩個名稱都為國際植物命名法規承認）
- **資料值不一致(Inconsistent Data Values)**—兩個相關的資料庫中資料更新不一致或不同步，如活體藏品與植物標本館資料庫，或博物館藏品資料庫與相關的影像資料庫。
- **資料品質汙染(Information Quality Contamination)**—將準確的資料與不準確的資料結合時，造成品質汙染。如將鑑定至亞種的資料與只鑑定至種的資料結合。

Spatial Data 空間資料

空間資料的儲存包含地點資訊（地點的敘述性資訊）及座標資訊（地理參照資料），後者通常以座標組的形式（東位及北位）。許多資料庫除了關於地點的自由描述之外，也正在

增加如最近的地名、距離、方向等已區分或區隔過的地點資料。現在也有許多計畫致力於改善這些地點資料的自由描述，增設區隔後資料的欄位，來輔助地理參照的流程。最近由摩爾基金會資助的「生物地理風水師計畫」¹⁵即是一例。

輸入至資料庫的地理參照（或座標）資訊，其格式通常是經緯度（球面座標系統），或是二度分帶(UTM)（或相關的）座標（平面座標系統）。像經緯度這樣的球面座標系統以球體表示地球表面，需以投影方式延展才能呈現在平面地圖上。球面座標的方格面積大小不一，經緯度間的距離依其離赤道或離兩極的遠近各異。平面座標系統較接近等積投影(equal area projections)，可用以量測或計算面積大小。

現在許多機構所輸入的資料，皆包含度、分、秒或度及十進位分(如許多GPS採用的)，再由資料庫轉換為十進位角度儲存。為方便轉換以應用於全球資訊系統，一般而言，資料最好儲存為十進位度，以達到可能的最高準確度。

若資料受限於單一的二度分帶(UTM)，機構常將資料存為二度分帶座標。這樣的好處是，如此一來資料以面積為準，每個網格都是正方形或長方形，容易呈現在平面地圖上，也方便距離及面積的計算。但若資料以二度分帶（或相關的）座標系統儲存時，應同時儲存分帶(Zone)資料，免除日後與其他來源或機構資料結合時會有困難。

Decimal Degrees 十進位度

許多資料庫儲存的十進位度，可能會造成上述所說的錯誤精確度。儲存及流通資料時，應考慮資料的精確度，資料庫不應允許超過資料庫中最高精確度的資料。大多數的生物資料，精確度最高約至小數以下第四位（約 10 公尺）。

Datums 基準

可能的大地基準有很多。地球並非正圓，而是橢圓，但要為橢圓表面找到適合的座標系統並非易事(Chapman et al. 2005)。為了解決這個問題，「基準」的概念應運而生。基準是一系列的點，用來表示圓形上的位置落在橢圓上的參照地點。過去，地球上的不同區域有不同的參照系統，在人造衛星誕生後，由於地球中心固定，得以製作出全球的參照系統或基準。若利用不同的大地參照得出的經緯度，在地表上的差距可達 400 公尺以上(Wieczorek 2001)。

由於這些差異，資料庫應紀錄其使用的基準，否則當資料結合時，同一地點的紀錄可能會出現很大的差異。

¹⁵ <http://www.biogeomancer.org/>

Manipulation of spatial data 空間資料操作

有許多種操作空間資料的方式，有些會影響到空間資料的準確度，有些則否。以下是一些影響到位置準確度的例子。

Conversion of data from one format to another 資料格式轉換

在收集、儲存、和使用物種及物種出現資料的過程中，最常見的資料轉換，或許是地理編碼的轉換，即從度/分/秒轉換到十進位度(DMS to DD)或從二度分帶座標轉到十進位度(UTM to DD)。其他還有地點描述中的英哩/公里轉換，或是在海拔及深度紀錄中將英呎轉為公尺等。

這些都是很簡單的轉換，卻有可能因為誤用精確度而造成準確度的假象。例如，採集地紀錄為海拔 250 英呎（採集者可能是指 200 到 300 英呎之間），若換算為公制會變成 76.2 公尺（到小數點後一位），或是化為整數 76 公尺。較好的作法是記為 80 公尺，更好的作法是加上±20 公尺的準確度範圍。精確度的誤用可能會造成準確度增加的假象，事實是品質降低。

Datums and Projections 基準與投影

由於缺少一標準的轉換方式，在轉換資料的大地基準時，可能會產生很大的誤差（參考 Wieczorek 2001 關於基準及其對資料品質影響的討論）。許多國家和地區現在都將該區資料轉為單一標準，如世界測量座標系統(World Geodetic System¹⁶, WGS84)，或是相近的基準（如澳洲地理基準，與 WGS84 相差約 10 公分，歐洲的 EUREF89 與 WGS84 則相差約 20 公分）。在某些情況下，若資料的準確度只有約 5 或 10 公里時，基準的轉換也許不是很必要。但如果處理的資料是 10 到 100 公尺的準確度，基準轉換便相當重要，因為有時可能會產生 400 公尺的差距(Wieczorek 2001)。

同樣地，處理多邊形的地圖資料（例如國家公園的資料）時，一定要注意投影方法轉換時可能產生的誤差（例如「亞伯等積投影」轉換為地理座標系投影）。有一些標準公式可計算轉換時產生的錯誤，而資料附帶的後設資料也應反映此資訊。

Grids 網格

¹⁶ 譯注：原文為 World Geodetic Datum，因 WGS84 之 S 為 System 之字首，故改之。

資料從向量格式轉為柵格(raster)或網格格式時，準確度和精確度都會流失。這是因為柵格檔案中網格的大小要趨近向量資料(Burrough and McDonnel 1998)。將資料重新轉回向量格式無法重建精確度和準確度。Chapman 等人 (2004)有更多關於使用及轉換柵格資料及尺度方面的問題討論。

Data Integration 資料整合

若是地理資料集彼此間有不一致的情形，便難以整合。這些不一致可能來自於資料的空間及屬性特徵，可能需運用各種不同且耗時的方法予以修正(Shepherd 1991)。不一致性可能源於：

- 不同的紀錄或量測技術（如觀測資料中的面積及時期）、調查方法（如網格大小、樣線寬度）或資料類型（如不同的資料類型定義）
- 量測或調查方法錯誤（如抄寫、資料紀錄及鑑定的錯誤）
- 解析度（時間、空間或屬性）不同
- 定義模糊不清
- 物件不清楚（如土壤或植被界線、物種鑑定層級不同，有些至種、有些至亞種、有些只至屬）
- 術語及命名法使用或解讀不同（如不同的分類觀念）
- 全球定位系統設定（基準、座標系統等）的差異

整合的問題，在下列情況中更嚴重：

- 類型不同（如：博物館物種資料混合調查及觀測資料）
- 來自不同管轄（如：調查方法可能因而不同）
- 取自多重來源
- 包含多種尺度
- 包含不同資料類型（地圖、標本、影像等）
- 來自不同時期
- 儲存於不同類型的資料庫或媒體中（如：有些資料庫軟體不允許無效值，即空白的欄位）
- 欄位劃分不一致（如：某資料集將學名寫在同個欄位中，其他可能分為屬、種等欄位。）



資料管理人若遵循並利用一致的資料儲存標準，資料整合便能促進品質提升。

Representation and Presentation 表示及呈現

無論現有資料品質如何，都應建立一套方法，讓資料發揮最大的功用。另外，為維持其可靠性，必須驗證資料，或附加資訊以說明其可靠程度。(Olivieri et al. 1995)

科學家與科學機構在瞭解、解釋、量化及評估生物多樣性的過程中，逐漸被視為資訊提供者。這樣的認可來自於他們能夠提供可靠可用的資訊予決策者、管理人、大眾與其他人。若因資料庫管理不良，造成資訊模稜兩可、語焉不詳、東缺西漏、相互矛盾、錯誤百出，將影響其作為資訊提供者及科學權威的聲譽(Dalcin 2004)。

生物科學資料數位化的一個關鍵目的，在於提供使用者以最具成本效益的方法，來查詢及分析資料。因此其成功與否，取決於是否能讓使用者準確地從資料瞭解生物的世界。但生物的世界極為複雜，必得經歸納簡化、取其梗概才易明白呈現(Goodchild et al. 1991)。呈現方式很多，可透過地理資訊系統、環境模型工具、決策支持系統等，但應用時必須抽樣測量其中的變異，描述並顯示出誤差及不確定性，這方面，在一個好的案例出現之前，仍需我們投注更多心血(Goodchild et al. 1991)。

生物學是首先利用誤差長條圖(error bar)及各式統計學量度制定誤差報告方法的學科。說明資料的誤差並非不好，反之，誤差估計提供了極重要的資訊，讓我們得以正確解讀資料(Chrisman 1991)。交付物種資料時，也需要類似的誤差說明技巧，如此使用者才能有同樣的能力，正確解讀並使用資料。



有效的資料品質計畫於公於私皆能讓機構及個人免於面臨窘境。

Determining Users' Needs 判定使用者需求

要判定使用者的需求並不容易，而要訂定詳盡的需求並架構資料以滿足之，則更加困難。也很重要，得找到主要的使用者共同釐清需求並據以開發。良好的使用者需求能促進資料收集、管理的效率及整體資料品質。

Relevancy 相關性

相關性與「品質」息息相關，指的是資料與使用需求的相關性。例如可能只是簡單地要嘗試使用一植物誌在其它非撰寫時所針對的區域，因為除此之外也沒有其他文本可供查詢；或資料所在的投影區與所需不同，得耗費一番工夫使資料有用且具「相關性」。

Believability 可信度

可信度即使用者認為資料可信的程度(Dalcin 2004)。可信度受使用者對資料適用性的看法與評估影響，可能基於使用者過去經驗，或與一般接受的標準比較(Pipino et al. 2002)。資料集的聲譽有時取決於使用者認定的資料可信度（因此亦取決於其適用性），而常可由完善的記載文件得到改善。

Wang 等人(1995)以圖表將這些議題繪製為層級關係，並展現可信度及名聲等之間的關聯。

Living with uncertainty in spatial data 與空間資料的不確定性共處

不確定性，尤其是空間資料的不確定性，是無法改變的事實，但一直以來資料中的不確定性未詳實記載，使用者也不易察覺。簡單好用的桌上型測繪系統大量流通，讓非 GIS 專業人士也能夠檢視分析資料的空間關係，卻通常並未使用適當的尺度(Chapman et al. 2005)，也忽略了資料中既有的空間誤差與不確定性(Chapman 1999)，在某些情況下，這可能會造成嚴重的資料誤用，帶來慘痛的結果(Redman 2001)。近來越來越多的簡易線上地圖服務，仿效傳統桌上型地理訊息系統，讓使用者得以檢視分析空間資料，由服務提供者控制顯示的資料圖層與比例。在不久的未來，隨著網路地圖服務(Web Mapping Services)的發展，這樣的服務會日益增加。地圖發行人控制資料圖層及比例（如：使用者放大檢視地圖時，不同的圖層自動出現或消失），可減少一些可能的錯誤發生。

資料的不確定性應藉由完善的後設資料或顯示與呈現的方式建檔。至於物種與物種出現資料，不確定性的顯示技術仍有賴更多研究，才能展現出準確度足跡(footprints of accuracy)。藏品紀錄不應只以一個經緯度的點表示，而應包含紀錄準確性，因此以足跡的形式呈現，可以是一個圓或一個橢圓，也甚至包含可能性高低的資訊(Chapman 2002)。

很重要的是，知悉資料及其空間與屬性限制的人，應藉由資料的描述文件及其流通，協助使用者決定資料是否適用於其目的。

Visualisation of error and uncertainty 誤差及不確定性的顯示

雖然現行有一些令人興奮的新方法來顯示物種資料誤差。完善誤差顯示方法的建立尚待努力(e.g. Zhang and Goodchild 2002)。也許最簡單的方法是將誤差圖層疊合於地理資訊系統上，像是以深淺不一的陰影來標示出地圖上不同區域的可信度。這樣的方法在製圖界行之有年。另一個技巧則是利用不同的記號，如以虛線與實線、大小或密度不一的點，來表示品質或準確度高低。這樣的疊合方法也能顯示出誤差來源的線索，對資料的驗證與檢查而言是相當有價值的工具。

誤分類矩陣法(misclassification matrix)以行表示期望值，以列表示觀察值，這在能進行此類統計的情況中相當有用。在此算法中，行的錯誤是「漏判」(errors of omissions)，列的錯誤是「誤判」(errors of commission)(Chrisman 1991)。此法不能直接適用於物種出現資料，但對長期的出沒調查資料等，仍相當有用。

Risk Assessment 風險評估

決策者希望獲得確定感，但自然系統千變萬化的特性，難以符合他們的期望。風險評估技術讓決策者及環境管理人更能預估確定性與風險，讓環境相關決策更具確定性。而就物種而言，物種實際出現地點的資訊常不足，可用「可能出現」區域來替代，但在所有可能出現的區域中，有些地區較其他出現機率更高(Chapman 2002)。

一般而言，風險的概念包含兩個要素——事情發生的可能性及規模，以及發生的時間與後果(Beer and Ziolkowski 1995)。物種資料的風險評估很廣，從備份執行缺乏而在大火後失去資料，到資料品質低落造成環境決策錯誤的風險都包含在內。若有資訊指出某地有受威脅物種出沒，估算禁止當地建設可能造成的成本多寡，便是物種風險評估的一例。因此在某些環境中，政府逐漸採用預警原則(precautionary principle)來推行重要的環境決策。

Legal and moral responsibilities 法律及道德責任

物種資料的品質與呈現牽涉到一些法律及道德責任，包括：

- 版權與智慧財產權；
- 隱私權；
- 真實標示；
- 敏感分類群資料品質呈現限制；
- 原住民權利；
- 責任義務；
- 警告及免責聲明(Caveats and Disclaimers)

資料的版權與智慧財產權可記載於附加資訊。每筆紀錄處理方式不同，但應在紀錄層級，或後設資料中應記載下來。

有一些國家近來頒布隱私權的法令，資料管理人應注意這些條款。當資料交換跨越國界或流通於網路上時，更需特別注意。在某些國家，除非獲得允許，否則個人資訊不得儲存於資料庫中或使之流通。這對物種出現資料的附加資訊會有什麼影響目前仍不明朗，但管理人應知曉相關議題，並預先做好準備。

良好的資料品質量測及後設資料，符合「真實標示」的概念。至少就目前的法律而言，「真實標示」只限於食物產品的規範，但在有關全球空間資料基礎建設(Global Spatial Data Infrastructure, Nebert and Lance 2001, Lance 2001)、美國全國空間資料基礎建設(National Spatial Data Infrastructure for the USA, Nebert 1999)、澳洲紐西蘭空間資料基礎建設(Australian and New Zealand Spatial Data Infrastructure, ANZLIC 1996b)等論文中，皆提及過這個概念。在全球空間資料基礎建設(Lance 2001)文中，建議空間資料流通交換機制(Spatial Data Clearinghouse)應包含「免費的廣告方法，讓世人能取得遵循真實標示原則的資料」，在澳洲紐西蘭的那份文件中提到：

『土地及地理資料品質標準可是敘述性或說明性，或兩者結合。敘述性的標準係基於「真實標示概念」，要求資料建立者提供已知的資料品質資訊。如此，資料使用者方能在資訊充足的情況下，判定「資料適用性」。』

敏感物種資料品質呈現受限可能發生在收集地資訊「模糊不清」時，例如瀕危物種或貿易敏感物種的切確位置資訊有所限制。這會造成資料品質降低，因此應清楚記載，讓使用者瞭解手中資料為何，並自行決定資料是否適用。

原住民權利(Indigenous rights)也可能影響到資料品質，可能有些資訊因原住民敏感問題受到限制。因此需註明「為維護原住民權益，某些資料受限」。

1998 年，Epstein 等人檢視了空間資料使用與法律責任的議題，列出以下要點：

- 空間資訊的誤差，「極有可能」引發訴訟，造成個人及組織的名聲與誠信受損。
- 訴訟中，以往的免責聲明可能無法構成有力的辯護。
- 為減少責任義務，組織可能需要維持高品質的文件記載，「盡其所能，竭其所知」、適當並真實地標示其產品。

警告及免責說明是資料品質記載中很重要的部分。除保護管理機構外，也應讓使用者瞭解資料品質，及得以從品質預期的事項。



為建立資料的組織與團體進行評價時，大多依據其資料的資訊流通度及資料品質。能發表、分享、流通、整合並使用資料的組織受惠最多。(NLWRA 2003).

Certification and Accreditation 認證及認可

物種出現資料能夠且應該認證嗎？隨著各機構釋出越來越多的資料，使用者想知道哪家機構可以信賴、哪間遵循資料品質管制流程。使用者是否只能仰賴有名的機構，抑或較不知名的機構也收藏有可靠的資料？而在知名機構可取得的資料中，哪些可信賴，哪些不能？使用者找尋資料來源時，名聲(reputation)常是關鍵性的決定因素，但名聲是個主觀的概念，也無法作為決策與行動穩固的立論基礎(Dalcin 2004)。這是我們學科領域所欲見之事嗎？完善的後設資料與資料品質程序記載，常能將名聲等主觀因素，化為使用者進行科學理性的評估時，能夠仰賴的依據。也許我們該訂定一套認證與認可的流程，讓使用者知悉哪些組織符合最低資料品質記載標準與流程。

一旦達成共識建立品質認證，便可促進整體資料品質的提升，使用者對資料價值也越有信心。另一方面，經認可之組織也可能獲得更多資金補助。Dalcin (2004)指出，「分類學資料的品質認證包含三個層面：原始資料來源（原料）、資訊鏈（流程）及資料庫（產品）。」

Peer Review of databases 資料庫同儕審查

資料庫同儕審查的機制可應用於物種資料庫。這樣的同儕審查程序，可以納入上述的認證流程，也許會涉及品質管制流程、文件記載與後設資料、更新及意見反應機制等。

Conclusion 結論

每位資訊專家都致力於避免不必要的錯誤。直截了當地點出錯誤，可將錯誤限制於可接受的範圍內，但錯誤不是花小錢或輕易地就能避免的。

(Chrisman 1991)。

資料品質及錯誤更正的重要性怎麼強調都不為過。如同通篇一再強調，這兩項工作舉足輕重，能夠賦予資料真正的價值，有助制定完善的環境決策及管理方式。資料品質對所有資料都是很重要的議題，無論是博物館或植物標本館藏品、觀測紀錄、調查資料或物種清單皆然。世界各地越來越多政府要求提高資料品質、完備資料建檔作業，諸如：

- 澳洲聯邦及地方政府強力指示要改善公共服務，更有效地運用包括資料及資訊等資源。
- 越來越多人瞭解到，公有資料應妥善管理，開放大眾取用，以發揮資料功能，大眾也才能瞭解資料製作及維護費用的必要性。
- 消費者施加更多的壓力，希望能以少許費用或免費，更便捷地取用正確的資料及資訊。
- 政府間開始重視合宜地處理資料及整合的需求，以提升資料的效率與價值。
- 資料相關性漸受重視，新的藏品、調查或資料管理與出版作業皆然。

資料品質的需求無庸置疑，但許多資料管理者都以為系統內的資料正確無誤——或只有無關緊要的小錯誤。其實，所有的資料都有錯誤與不確定性，所有的錯誤也都會影響到資料最終的用途及結果。為提升資料品質所進行的資料取得與管理，都是資料管理中重要的部分。負責物種出現資料的組織應檢查並改善資訊品質鏈中的每個步驟，其文件記載也非常重要，讓使用者能瞭解資料，決定資料是否合用，也就決定了資料品質。

人為因素可能是對空間資訊的準確度及可靠度最大的潛在威脅，同時也是確認可靠度、瞭解任何空間資料固有缺失的一項因素。(Bannerman 1999)

Acknowledgements 致謝

世界各地許多同行和機構皆對本文以不同的方式貢獻許多。有些提供直接的幫助，有些和筆者進行了為期三十年的討論，有些則經由論文發表或將訊息開放流通，間接協助撰寫本文。

特別感謝的是，過去和現在於巴西坎皮納斯環境資料中心(CRIA)以及澳洲坎培拉環境資源資訊網(ERIN, Environmental Resources Information Network)服務的同仁，感謝他們提供寶貴的意見、工具、理論，並與筆者的試探想法相互激發構思。針對環境資訊的誤差及準確度議題討論多年，墨西哥國家保護生物多樣性委員會、美國堪薩斯大學、澳洲國家科學及工業研究機構(CSIRO)、美國科羅拉多大學、康乃狄克州的皮巴蒂(Peabody)自然史博物館、美國加州大學柏克萊分校以及其它許多單位人士等開創性的工作，也幫助我們瞭解今日的物種資料品質管理。感謝各位貢獻的意見與建設性的指教。此外，感謝 Town Peterson 及其他美國堪薩斯大學的同行、美國康乃狄克衛斯連大學的 Barry Chernoff、美國耶魯大學的 Read Beaman、美國加州大學柏克萊分校的 John Wieczorek 及 Robert Hijmans、荷蘭阿姆斯特丹分類鑑定專家中心(ETI)的 Peter Shalk 及其他人、加州科學學會的 Stan Blum 以及丹麥哥本哈根 GBIF 的同仁，在與筆者討論過程中，提供了許多想法與挑戰，造就了本文中許多想法。但任何錯誤、缺漏或爭議的責任均由筆者承擔。

感謝各位在本文編輯過程的批評指教，尤其是 GBIF 自然歷史藏品資料數位化小組委員會的委員：柏林-達勒姆植物園及博物館的 Anton Güntsch、西班牙馬德里皇家植物園 Francisco Pando、南非普利托利亞美國農業部動植物檢疫局(USDA-Aphis, Pretoria, South Africa)的 Mervyn Mansell、美國堪薩斯大學的 A. Townsend Peterson、芬蘭土庫大學的 Tuuli Toivonen、美國華盛頓特區史密遜研究院的 Anna Wietzman 以及比利時全球生物多樣性資訊機構的 Patricia Mergen。

全球生物多樣性資訊機構的 Larry Speers 促成了本文的誕生，並從頭至尾給予指導。

最後，感謝巴西 FAPESP/Biota 計畫的機會與支持，讓筆者在 2003 及 2004 年旅居巴西期間，得以延伸對資料品質管理的想法，也感謝全球生物多樣性資訊機構的支持與鼓勵，使本文得以完成。

References 參考文獻

- Agumya, A. and Hunter, G.J. 1996. Assessing Fitness for Use of Spatial Information: Information Utilisation and Decision Uncertainty. Proceedings of the GIS/LIS '96 Conference, Denver, Colorado, pp. 359-70
- ANZLIC. 1996a. ANZLIC Guidelines: Core Metadata Elements Version 1, Metadata for high level land and geographic data directories in Australia and New Zealand. ANZLIC Working Group on Metadata, Australia and New Zealand Land Information Council.
<http://www.anzlic.org.au/metaelem.htm>. [Accessed 14 Jul 2004]
- ANZLIC 1996b Spatial Data Infrastructure for Australia and New Zealand. Discussion Paper.
www.anzlic.org.au/get/2374268456. [Accessed 1 Jul 2004].
- Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* 5(1): 80-88.
- Bannerman, B.S., 1999. Positional Accuracy, Error and Uncertainty in Spatial Information. Australia: Geoinnovations Pty Ltd.
<http://www.geoinnovations.com.au/posacc/patoc.htm> [Accessed 14 Jul 2004].
- Beer, T. & Ziolkowski, F. (1995). Environmental risk assessment: an Australian perspective. Supervising Scientist Report 102. Canberra: Commonwealth of Australia.
<http://www.deh.gov.au/ssd/publications/ssr/102.html> [Accessed 14 Jul 2004]
- Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283-309.
- Berendsohn, W., Güntsch, A. and Röpert, D. (2003). Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections. Copenhagen, Denmark: Global Biodiversity Information Facility.
http://circa.gbif.net/Members/irc/gbif/digit/library?l=/digitization_collections/contract_2003_report/ [Accessed 16 Mar. 2005].
- Birds Australia. 2001. Atlas of Australian Birds. Search Methods. Melbourne: Birds Australia.
<http://www.birdsaustralia.com.au/atlas/search.html> [Accessed 30 Jun 2004].
- Birds Australia. 2003. Integrating Biodiversity into Regional Planning – The Wimmera Catchment Management Authority Pilot Project. Canberra Environment Australia.
<http://www.deh.gov.au/biodiversity/publications/wimmera/methods.html>. [Accessed 30 Jun 2004].
- Brigham, A.R. 1998. Biodiversity Value of federal Collections in Opportunities for Federally Associated Collections. San Diego, CA, Nov 18-20, 1998.
- Burrough, P.A., McDonnell R.A. 1998. Principals of Geographical Information Systems: Oxford University Press.
- Byers, F.R. 2003. Care and Handling of CDs and DVDs. A Guide for Librarians and Archivists. Washington, DC: National Institute of Standards and Technology and Council on Library and Information Resources.
<http://www.itl.nist.gov/div895/carefordisc/CDandDVDCareandHandlingGuide.pdf> [Accessed 30 Jun 2004].
- CBD. 2004. Global Taxonomic Initiative Background. Convention on Biological Diversity.
<http://www.biodiv.org/programmes/cross-cutting/taxonomy/default.asp> [Accessed 13 Jul 2004].
- Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 in Lowell, K. and Jatón, A. eds. Spatial accuracy assessment: Land information uncertainty in natural resources. Chelsea, MI: Ann Arbor Press.
- Chapman, A.D. 2002. Risk assessment and uncertainty in mapped and modelled distributions of threatened species in Australia pp 31-40 in Hunter, G. & Lowell, K. (eds) Accuracy 2002 – Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Melbourne: Melbourne University.
- Chapman, A.D. 2004. Environmental Data Quality – b. Data Cleaning Tools. Appendix I to Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota. FAPESP/Biota process no. 2001/02175-5 March 2003 – March 2004. Campinas, Brazil: CRIA 57 pp.
http://splink.cria.org.br/docs/appendix_i.pdf [Accessed 14 Jul. 2004]
- Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 in Miller, R.I. (ed.). Mapping the Diversity of Nature. London: Chapman and Hall.
- Chapman, A.D., Muñoz, M.E. de S. and Koch, I. 2005. Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* 2: 24-41.
- Chrisman, N.R. 1983. The role of quality information in the long-term functioning of a GIS. Proceedings of AUTOCART06, 2: 303-321. Falls Church, VA: ASPRS.
- Chrisman, N.R., 1991. The Error Component in Spatial Data. pp. 165-174 in: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) Geographical Information Systems Vol. 1, Principals: Longman Scientific and Technical.

- Conn, B.J. (ed.) 1996. HISPID3. Herbarium Information Standards and Protocols for Interchange of Data. Version 3. Sydney: Royal Botanic Gardens.
- Conn, B.J. (ed.) 2000. HISPID4. Herbarium Information Standards and Protocols for Interchange of Data. Version 4 – Internet only version. Sydney: Royal Botanic Gardens. <http://plantnet.rbgsyd.nsw.gov.au/Hispid4/> [Accessed 30 Jun. 2004].
- Cullen, A.C. and Frey, H.C. 1999. Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs. New York: Plenum Press, 335 pages.
- CRIA 2005. speciesLink. Dados e ferramentas – Data Cleaning. Campinas, Brazil: Centro de Referência em Informação Ambiental. <http://splink.cria.org.br/dc/> [Accessed 4 Apr. 2005].
- Dalcin, E.C. 2004. Data Quality Concepts and Techniques Applied to Taxonomic Databases. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf [Accessed 7 Jan. 2004].
- Dallwitz, M.J. and Paine, T.A. 1986. Users guide to the DELTA system. CSIRO Division of Entomology Report No. 13, pp. 3-6. TDWG Standard. <http://biodiversity.uno.edu/delta/> [Accessed 9 Jul 2004].
- Davis R.E., Foote, F.S., Anderson, J.M., Mikhail, E.M. 1981. Surveying: Theory and Practice, Sixth Edition: McGraw-Hill.
- DeMers M.N. 1997. Fundamentals of Geographic Information Systems. John Wiley and Sons Inc.
- English, L.P. 1999. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York: John Wiley & Sons, Inc. 518pp.
- Environment Australia. 1998. The Darwin Declaration. Canberra: Australian Biological Resources Study. <http://www.biodiv.org/programmes/cross-cutting/taxonomy/darwin-declaration.asp> [Accessed 14 Jul 2004].
- Epstein, E.F., Hunter, G.J. and Agumya, A.. 1998, Liability Insurance and the Use of Geographical Information: International Journal of Geographical Information Science 12(3): 203-214.
- Federal Aviation Administration. 2004. Wide Area Augmentation System. <http://gps.faa.gov/Programs/WAAS/waas.htm> [Accessed 15 Sep. 2004].
- FGDC. 1998. Geospatial Positioning Accuracy Standards. US Federal Geographic Data Committee. http://www.fgdc.gov/standards/status/sub1_3.html [Accessed 14 Jul. 2004].
- Foote, K.E. and Huebner, D.J. 1995. The Geographer's Craft Project, Department of Geography, University of Texas. <http://www.colorado.edu/geography/gcraft/contents.html> [Accessed 14 Jul 2004].
- Gad, S.C. and Taulbee, S.M. 1996. Handbook of data recording, maintenance, and management for the biomedical sciences. Boca Raton: CRC Press.
- Goodchild, M.F., Rhind, D.W. and Maguire, D.J. 1991. Introduction pp. 3-7 In: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) Geographical Information Systems Vol. 1, Principals: Longman Scientific and Technical.
- Heuvelink, G.B.M. 1998. Error Propagation in Environmental Modeling with GIS: Taylor and Francis.
- Huang, K.-T., Yang, W.L. and Wang, R.Y. 1999. Quality Information and Knowledge. New Jersey: Prentice Hall.
- Juran, J.M. 1964. Managerial Breakthrough. New York: McGraw-Hill.
- Knapp, S., Lamas, G., Lughadha, E.N. and Novarino, G. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. Phil. Trans: Biol. Sci. 359(1444): 611-622.
- Koch, I. (2003). Coletores de plantas brasileiras. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 26 Jan. 2004].
- Lance, K. 2001. Discussion of Pertinent Issues. pp. 5-14 in Proceedings USGS/EROS Data Center Kenya SCI Workshop, November 12 2001. http://kism.icconnect.co.ke/NSDI/proceedings_kenya_NSDI.PDF [Accessed 1 Jul 2004].
- Leick, A. 1995. GPS Satellite Surveying: John Wiley and Sons, Inc: New York.
- Library of Congress. 2004. Program for Cooperative Cataloging. Washington, DC. US Library of Congress. <http://www.loc.gov/catdir/pcc/> [Accessed 26 Jun 2004].
- Lunetta, R.S. and Lyon, J.G. (eds). 2004. Remote Sensing and GIS Accuracy. Boca Raton, FL, USA: CRC Press.
- Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis pp. 200-209 in Proceedings of the Conference on Information Quality (IQ2000). Boston: Massachusetts Institute of Technology. <http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf> [Accessed 21 November 2003].
- Mayr, E. and Ashlock, P.D. 1991. Principles of systematic zoology. New York: McGraw-Hill.
- McElroy, S., Robins, I., Jones, G. and Kinlyside, D. 1998. Exploring GPS, A GPS Users Guide: The Global Positioning System Consortium.
- Minnesota Planning. 1999. Positional Accuracy Handbook. Using the National Standard for Spatial data Accuracy to

- measure and report geographic data quality. Minnesota Planning: Land Management Information Center. http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf [Accessed 14 Jul. 2004]
- Morse, L.E. 1974. Computer programs for specimen identification, key construction and description printing using taxonomic data matrices. *Publs. Mich. St. Univ. Mus., biol. ser.* 5, 1–128.
- Motro, A. and Rakov, I. 1998. Estimating the Quality of Databases. *FQAS* 1998: 298-307
- Naumann, F. 2001. From Database to Information Systems – Information Quality Makes the Difference. IBM Almaden Research Center. 17 pp.
- Nebert, D. and Lance, K. 2001. Spatial Data Infrastructure – Concepts and Components. *Proceedings JICA Workshop on Application of Geospatial Information and GIS*. 19 March 2001, Kenya. <http://kism.icconnect.co.ke/JICAWorkshop/pdf/Ottichilo.pdf> [Accessed 1 Jul 2004].
- Nebert, D. 1999. NSDI and Gazetteer Data. Presented at the Digital Gazetteer Information Exchange Workshop, Oct 13-14, 1999. Transcribed and edited from audiotape. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/session3/nebert.htm [Accessed 1 Jul 2004].
- NLWRA. 2003. Natural Resources Information Management Toolkit. Canberra: National Land and Water Resources Audit. <http://www.nlwra.gov.au/toolkit/contents.html> [Accessed 7 Jul 2004].
- NOAA. 2002. Removal of GPS Selective Availability (SA). http://www.ngs.noaa.gov/FGCS/info/sans_SA/ [Accessed 15 Sep 2004].
- Olivieri, S., Harrison, J. and Busby, J.R. 1995. Data and Information Management and Communication. pp. 607–670 in Heywood, V.H. (ed.) *Global Biodiversity Assessment*. London: Cambridge University Press. 1140pp.
- Pipino, L.L., Lee, Y.W. and Wang, R.Y. 2002. Data Quality Assessment. *Communications of ACM* 45(4): 211-218.
- Pullan, M.R., Watson, M.F., Kennedy, J.B., Raguenaud, C., Hyam, R. 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* 49: 55-75.
- Redman, T.C. 1996. *Data Quality for the Information Age*. Artech House, Inc.
- Redman, T.C. 2001. *Data Quality: The Field Guide*. Boston, MA: Digital Press.
- SA Dept Env. & Planning. 2002. Opportunistic Biological Records (OPPORTUNE). South Australian Department of Environment and Heritage. <http://www.asdd.sa.gov.au/asdd/ANZSA1022000008.html> [Accessed 14 Jul. 2004].
- SEC 2002. Final Data Quality Assurance Guidelines. United States Securities and Exchange Commission. <http://www.sec.gov/about/dataqualityguide.htm> [Accessed 26 Jun 2004].
- Shepherd, I.D.H. 1991. Information Integration and GIS. pp. 337-360 in: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems Vol. 1, Principals*: Longman Scientific and Technical.
- Spear, M., J.Hall and R.Wadsworth. 1996. Communication of Uncertainty in Spatial Data to Policy Makers in Mowrer, H.T., Czaplewski, R.L. and Hamre, R.H. (eds) *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*, May 21-23, 1996. Fort Collins, Colorado. USDA Forest Service Technical Report RM-GTR-277.
- Stribling, J.B., Moulton, S.R. II and Lester, G.T. 2003. Determining the quality of taxonomic data. *J. N. Amer. Benthol. Soc.* 22(4): 621-631.
- Strong, D.M., Lee, Y.W. and Wang, R.W. 1997. Data quality in context. *Communications of ACM* 40(5): 103-110.
- Taulbee, S.M. 1996. Implementing data quality systems in biomedical records pp. 47-75 in Gad, S.C. and Taulbee, S.M. *Handbook of data recording, maintenance, and management for the biomedical sciences*. Boca Raton: CRC Press.
- TDWG. 2005. TDWG Working Group: Structure of Descriptive Data (SDD). Taxonomic Databases Working Group (TDWG). <http://160.45.63.11/Projects/TDWG-SDD/> [Accessed 4 Apr. 2005].
- University of Colorado. 2003. MapSTeDI. Georeferencing in MapSTeDI. Denver, CO: University of Colorado. <http://mapstedi.colorado.edu/georeferencing.html> [Accessed 30 Jun. 2004].
- USGS. 2004. What is SDTS? Washington: USGS. <http://mcmcweb.er.usgs.gov/sdts/whatsdts.html> [Accessed 30 Jun. 2004].
- Van Sickle, J. 1996. *GPS for Land Surveyors*: Ann Arbor Press, Inc: New York.
- Wang, R.Y. 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM* 41(2): 58-65.
- Wang, R.Y., Storey, V.C., Firth, C.P., 1995. A frame-work for analysis of data Berkeley: University of California, Berkeley - MaNIS <http://manisnet.org/manis/GeorefGuide.html> [Accessed 26 Jan. 2004].
- Wieczorek, J. 2002. Summary of the MaNIS Meeting. American Society of Mammalogists, McNeese State University, Lake Charels, LA, June 16, 2002. Berkeley: University of California, Berkeley - MaNIS. <http://manisnet.org/manis/ASM2002.html> [Accessed 30 Jun. 2004].

- Wieczorek, J., Guo, Q. and Hijmans, R.J. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal for GIS* 18(8): 754-767.
- Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.
- Zhang, J. and Goodchild, M.F. 2002. *Uncertainty in Geographic Information*. London: Taylor and Francis.
- quality research, *IEEE Transactions on Knowledge and Data Engineering* 7: 4, 623-640.
- Wieczorek, J. 2001. *MaNIS: Georeferencing Geo-referencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS <http://manisnet.org/manis/GeorefGuide.html> [Accessed 26 Jan. 2004].
- Wieczorek, J. 2002. Summary of the MaNIS Meeting. American Society of Mammalogists, McNeese State University, Lake Charels, LA, June 16, 2002. Berkeley: University of California, Berkeley - MaNIS. <http://manisnet.org/manis/ASM2002.html> [Accessed 30 Jun. 2004].
- Wieczorek, J., Guo, Q. and Hijmans, R.J. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal for GIS* 18(8): 754-767.
- Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.
- Zhang, J. and Goodchild, M.F. 2002. *Uncertainty in Geographic Information*. London: Taylor and Francis.

Index 索引

- accessibility 易得性, 48
- accountability 責任歸屬, 24
- Accreditation 認可, 61
- accuracy 準確度
 - accuracy 準確度, 3
 - attribute 屬性的, 34, 46
 - documentation of 建立文件, 45
 - false 錯誤, 33
 - positional 位置的, 32, 46
 - recording of
 - taxonomic data 分類學的資料, 27
 - spatial 空間的, 32
 - temporal 時間的, 48
- archiving 歸檔, 50
- attribute accuracy 屬性準確度, 34, 46
- audit trail 稽核存底, 22
- bias 偏差, 29
- BioGeomancer 生物地理風水師計畫, 33
- caveats and disclaimers 警告及免責說明, 61
- Certification 認證, 61
- classification data domain 分類資料域, 26
- Collection data domain 藏品資料域, 34
- collection data 藏品資料, 34
- collector 採集者
 - collector 採集者, 13
 - responsibility of 責任, 13
- completeness 完整性, 16, 29, 35, 36, 47
- consistency 一致性, 17, 29, 34, 36
 - semantic 語意的, 17
 - structural 結構的, 17
- copyright 版權, 60
- currency 即時性, 16
- data cleaning 資料清理, 20
- data custodian 資料管理人, 14
- data management 資料管理, 22
- data quality 資料品質
 - policy 政策, 11
 - principle 原則, 1
 - strategy 策略, 11
 - vision 願景, 10
- data user 資料使用者
 - responsibility of 責任, 15
- databases 資料庫
 - peer review of 同儕審查, 61
- data 資料
 - archiving 歸檔, 50
 - backup of 備份, 50
 - believability 可信度, 58
 - capture 擷取, 38, 41
 - categorization of 類型化, 23
 - collection 藏品, 34
 - collector 採集者, 34
 - consistency 一致性, 34, 36
 - descriptive 描述性, 36
 - documentation of 文件記載, 23
 - entry 輸入, 41
 - grid 網格, 56
 - integration 整合, 56
 - integrity 完整性, 51
 - nomenclatural 命名的, 26
 - observational 觀測的, 38
 - opportunistic 機會性地, 38
 - presentation 呈現, 57
 - relevancy 相關性, 58
 - representation 表示, 57
 - spatial 空間的, 31, 53
 - storage 儲存, 50
 - survey 調查, 38
 - taxonomic 分類學的, 26
 - uncertainty 不確定性, 58
- decimal degrees 十進位度, 53
- DELTA standard 戴爾他標準, 36
- descriptive data 描述性資料, 36
- Differential DGPS 差分全球定位系統, 39
- documentation 建檔；建立文件
 - database design 資料庫設計, 49
 - documentation 建檔；建立文件, 44
 - validation procedures 驗證程序, 48
- domain schizophrenia 值域錯亂, 52
- Domain value redundancy 值域重複, 51
- duplicate data records 資料重複紀錄, 52
- duplication 重複
 - minimisation of 最小化, 22
- edit controls 編輯控制, 22
- education 教育, 24
- error prevention 預防錯誤, 12

error 誤差
 error 誤差, 7
error 錯誤
 documentation of 建立文件, 45
 patterns 類型, 51
 visualisation 顯示；視覺化, 59
Federal Geographic Data Committee (FGDC) 美國聯邦地理資料委員會, 32
feedback 反應意見, 24
Field data domain 田野資料域, 31
fitness for use 適用性, 4, 5, 44
flexibility 彈性, 19
gazetteers 地名辭典
 electronic 電子式, 43
geodetic datums 大地基準, 53, 55
Geodetic Networks, 32
Georeferencing Guidelines 地理參照導引, 42
geo-referencing 地理參照, 41
Geospatial Positioning Accuracy Standards (GPAS) 地理座標準確度標準, 32
Global Positioning System, GPS 全球定位系統, 38
Global Positioning Systems 全球定位系統, 31
identification precision 鑑定精確度, 29
inconsistency 不一致, 29
inconsistent data values 資料值不一致, 52
incorrect data values 資料值不正確, 52
Indigenous rights 原住民權利, 60
Information for Spatial Information in Europe 歐洲空間資訊計畫, 31
Information Management Chain 資訊管理鏈, 13, 22
information quality contamination 資料品質汙染, 52
Intellectual Property Rights 智慧財產權, 60
ISO 19115 for Geographic Information – Metadata 地理資訊的 ISO 19115 後設資料, 31
legal responsibilities 法律責任, 59
lineage 世系, 46
logical consistency 邏輯一致性, 47
MaPSTeDI Guidelines 山區平原時空資料庫資訊學倡議, 42
metadata 後設資料, 44
missing data values 資料數值缺漏, 51
moral responsibilities 道德責任, 59
nomenclatural data 命名資料, 26
nonatomic data values 非單純資料值, 52
outlier detection 極端值偵測, 21
partnerships 合作關係, 15
performance measures 績效評估, 20
positional accuracy 位置準確度, 32, 46
precision 精確度, 3
precision 精確度
 documentation of 建立文件, 45
 false 錯誤, 33
 numerical 數值的, 3
 statistical 統計的, 3
primary species data 原始物種資料, 3
principles of data quality 資料品質原則, 10
prioritisation 優先順序, 16
quality assurance 品質保證, 6
quality control 品質管制, 6
quality 品質, 4
Real-time Differential GPS 即使差分全球定位系統, 39
resolution 解析度, 3
risk 風險
 assessment 風險評估, 59
selective availability 選擇性干擾, 39
spatial accuracy, 32
Spatial Data Transfer Standards 空間資料交換標準, 31
spatial data 空間資料, 31, 53
species-occurrence data 物種出現資料, 3
Static GPS 靜態全球定位系統, 39
Structure of Descriptive Data 描述性資料結構, 36
targets 目標
 setting of 訂定, 21
Taxonomic Database Working Group, TDWG 分類學資料庫工作小組, 36
Taxonomic Impediment 分類學障礙, 27
taxonomy 分類學, 26
temporal accuracy 時間準確度, 48
threatened species 瀕危物種, 60
timeliness 時效性, 16
Total Data Quality Management Cycle 整體資料品質管理循環, 13
trade sensitive species 貿易敏感物種, 60
training 訓練, 24
transparency 透明度, 20
truth in labelling 真實標示, 8, 60
uncertainty 不確定性, 6
update frequency 更新頻率, 17
User Interface 使用者介面, 41
voucher collections 證據藏品, 30
Wide Area Augmentation System, WAAS 廣域擴充系統, 39