書目資料庫品質控制之探討

——中華民國期刊論文索引的經驗

宋美珍 國家圖書館閱覽組編輯

一、前 言

書目資料庫的製作是一 長期且連續性的工作，在經過資料庫規劃、系統設計、資料蒐集整理、分析、輸入、校正等連串的工作流程後，才能完成資料庫的實體呈現，此一製作過程可說是智力與勞力、軟體與硬體協力配合的結果。然而，書目資料庫就像其他的資訊產品一樣，除了內容必須符合使用者需求外，還必須注意產品的品質控制才能獲得使用者的滿意與信賴，因此書目資料庫在追求顧客滿意品質服務的目標時，品質管理就扮演關鍵角色，其目的在藉由不斷改善資料庫產品與服務的過程，以滿足或超越使用者的期望與需求，使得資料庫得以永續發展。

二、誰是品質責任者？

在資訊社會中資訊創造者（information creators，如作家及其他之知識生產者）、資訊生產 者（information producer，如資料庫製作者及出版者等）、資訊傳播者（information distributors，如網路服務業者、線上系統業者）與資訊使用者（end-users）共同組成了一條資訊鍊（information chain），在這條資訊供應鍊上的任何一方都必須依賴其他一方提供無錯誤的品質服務。（註1）亦即任一方必須承擔應有的品質控制的責任，在過去資訊生產者與資訊傳播者的角色很容易區別，資訊生產者如資料庫製作者（database producer），將資訊經選擇、整理、分析、儲存；再交由資訊傳播者如線上資料庫服務業者（online vender），做其它的加值並提供檢索服務，因此對於資料庫的品質控制責任，若以使用者角度來看，因爲不清楚角色的分工，且多以檢索介面及檢索效果評斷資料庫的優劣，故直覺認爲品質控制是資訊傳播者的責任；反之，對資訊傳播者而言，資料庫的內容是由資訊生產者提供，因此資料內容的品質控制，是生產者的責任，故認爲其沒有責任替資訊生產者進行資料品質的檢核與建議，至於使用者的檢索需求，則往往在資訊傳播者成本經濟效益的考量下被忽視。近年來由於資訊技術、設備與通訊網路日益普及且價格持續滑落，愈來愈多的資訊生產者開始兼具資訊傳播者的角色，由自已來經營自已的資料庫，直接提供顧客服務，例如：IAC、UMI、H. W.Wilson、EBSCO等，原本都是資料庫的製作者，現在則透過自設的線上資訊系統提供資料庫檢索服務，對於這類的業者，自然直接擔負資料庫的品質責任。

三、品質控制的內容與作法

書目資料庫製作者要如何進行品質的控制？對於書目資料庫品質的定義又是什麼？從有關資料庫品質控制的文獻，我們可以發現大部分對於資料庫品質的定義，多來自於使用者或資料庫提供者端對於資料庫服務的要求與評估內容，內容則多以館藏圖書目錄（尤其是國家書目）的品質控制爲研究重點，布萊恩（Phillip Bryant）在「國家書目服務的品質」一文中，認爲所謂的品質，應具備以下幾點要求：1.正確性；2.一致性；3.新穎性；4.符合功能要求。（註2）1990年南加州線上資料庫使用者聯盟（Southern California Online Users' Group, SCOUG）則列出十一項評估線上資料庫品質的項目：1.資料的一致性；2.資料範圍；3.時效性；4.價格；5.正確性；6.檢索功能；7.系統反應時間；8.整合性；9.輸出方式；10.說明文件；11.顧客服務。（註3）從各界對書目資料庫的品質要求來看，書目資料庫品質控制可從以下幾方面進行：

1. 建立標準化作業流程，縮短資料處理時間，掌握時效性。

2. 訂定並確實遵守各項規範與政策，以確保資料的一致性與完整性。

3. 提供足夠且完整的系統文件與使用手冊，作爲工作人員資料處理的依據，及使用者檢索時的輔助工具。

4. 提升組織人員的素質，建立自保品質的信念。

5. 透過內外部稽核的機制檢核資料內容的正確性與系統的穩定性。

6. 透過資料庫使用調查與研究，了解使用者的需求與滿意度，並蒐集資訊評估資料庫檢索品質。

7. 建立改善所發現問題的機制，落實品質改善工作。

四、中華民國期刊論文索引資料庫的經驗

本館中華民國期刊論文索引資料庫自民國72年從印刷型式索引轉變成爲電子資料庫已邁入第20個年頭，民國82年本資料庫以光碟資料庫型態，開始商品化對外發行後，透過各種行銷服務與推廣，在直接面對國內外廣大的顧客（使用者）對產品的批評與滿意程度的反應下，更加體認作爲一個資料庫製作者所必須承擔產品的品質控制責任，乃積極透過各種方式進行資料庫品質改善工程。以下從資料庫內容、主題索引方式、檢索的功能性以及文獻傳遞服務四方面，說明本資料庫近年來在品質控制方面所面臨的問題與實際因應作法，提供各界參考。

1. 資料庫內容（Content）

資料內容是書目資料庫的靈魂，因此資料品質（quality of data）的控制乃成爲書目資料庫品質管理的基礎。對於資料內容的品質要求，一般由以下幾方面加以評估：

（1）資料範圍

傳統的紙本式索引通常在凡例或使用說明中都有該索引收錄資料的主題學科、資料類型、選擇標準等相關之編輯政策（editorial policy），其內容多十分詳細，使用者可以了解該索引的特色與收錄資料範疇，作爲查檢資料的參考。例如「中華民國期刊論文索引」在其紙本索引的凡例中說明收錄民國59年以來的期刊論文篇目，因此使用者如果要查檢民國50年的期刊論文篇目就不能使用此索引。其實電子書目資料庫在製作時也都訂有相關的的編輯政策，但大多數資料庫反而忽略在系統檢索畫面的相關說明或簡介中完整描述資料範圍，或是提供了錯誤的資料範圍，造成使用者無法查檢到所需資料（因爲資料庫未收錄該類型或該語文或時間的文獻），因此有使用上的困惑，進而對資料庫產生負面的評價，這點是資料庫在製作時必須注意的問題。此外，一但訂出了相關的收錄政策，就不輕易變更，以免資料收錄範圍產生不一致性的現象。

本資料庫目前提供檢索的版本有五種，提供檢索範圍並不十分一致，故在各種版本資料庫的簡介中都力求詳細說明資料內容收錄範圍，避免造成使用者的困擾，除了資料內容範圍之外，在許多使用者與圖書館的建議下，也提供本資料庫所收編期刊一覽的詳細資料。

（2）完整性

指包括資料類型、區域範圍或語文的完整性。全球每日生產各類型的書目資料量十分龐大，任何書目資料庫都不可能也沒有必要將所有資料收錄完整，但若爲專題性的資料庫，在考量一次檢索所有資料（one-stop-searching）的使用需求下，或爲了建立其獨有不可取代的價值，必須考慮該主題收錄資

料的完整性。例如一個以臺灣史研究爲專題的書目資料庫，若僅收錄期刊文獻單一資料類型，就不如一個同主題資料庫但收錄圖書、期刊、博碩士論文的資料庫來的完整。當然資料收集的完整性還得視機構資源的支援是否充足。以本系統爲例，它是一個綜合性期刊文獻資料庫，因此在資料完整性的控制乃側重於避免遺漏重要期刊（尤其是新出版或未正式出版期刊的掌握）、具學術性及參考價值的文獻以及卷期收錄的完整。

（3）新穎性

更新或死亡（update or dead）是電子資料庫狀態與價值最好的寫照，一般使用者在評估電子資料庫時，最重視的也往往是資料的新穎性與更新速度。目前書目資料庫的發行媒介主要分爲光碟（cdrom）、連線系統主機（online host）或鏡錄（mirror site）等三種方式，版本的多樣化，雖然有利於產品的行銷，然而由於更新方式與頻率的不同，也會增加資料庫的維護成本，以本系統爲例，目前資料庫的使用版本分爲四種，在資料更新頻率上亦有所差異，故在縮短資料更新時間，維持資料庫的新穎性，避免不同版本資料的不一致性，是本資料庫品質控制要項之一，爲達此目標，首先找出造成處理瓶頸的作業項目與原因，並進行作業方式變更與工作流程重新規劃。近二年來，在未增加人力情況下，不但持續增加收錄資料量並維持更新的速度外，各版本資料內容的時間範圍也已趨近一致性。

（4）正確性

爲了維持資料庫資料內容的正確性，避免使用者因資料庫提供錯誤訊息的指引而徒勞，資料庫製作者必須致力將錯誤率降至最低。分析書目資料庫常見的錯誤主要有：著錄資料項目錯誤、繕打錯誤及主題分析的錯誤；針對以上類型的錯誤，目前可由人工或程式自動查核方式來輔助錯誤資料的發現與校正。以本資料庫爲例，在人工處理方面，除了由不同人員逐筆校對三次外，還由較有經驗的工作人員以逐項（不同欄位）瀏覽索引檔的方式發現資料的異狀，此外還利用原有紙本彙編出版格式的季刊與年彙編輸出功能，全面檢核某區間內資料的正確性。這種校正方式可發現連載文章是否完整、期刊刊名與出版卷期是否正確，以及作者名稱有誤（尤其是期刊原始資料印刷錯誤）、分類是否一致等問題。但單由工作人員來查核資料的正確性可能仍有所遺漏，近年來也藉由使用者反饋機制（feed back）的設計，利用線上即時回報功能的設計（如e-mail），由使用者主動告知，協助錯誤資料的更正，也獲致不錯的效果，同時也增加與使用者的互動及溝通的管道。

另外也利用程式自動查核，針對具有規則性的資料欄位進行資料檢查，檢核項目有ISSN、ISBN的欄位長度與規則、是否具備必備欄位、分欄與指標的合理性、拼字的檢查、出版日期的合理性、重複資料的比對等。程式自動查核又分爲線上存檔即時查核與批次資料查核，兩者檢查的項目則互有差異。

（5）一致性

指收錄資料政策、資料款目輸入方式與規則的一致性，其中最容易引起品質問題的就是資料的選擇政策。以期刊索引資料庫爲例，其選擇政策可分爲期刊主體與期刊文獻個體兩方面，除了期刊目次（TOC）類型的資料庫，對於期刊文獻個體方面採全部收錄的原則外，一般索引資料庫都有其文獻收錄的政策，尤其近年來書目文獻產生的速度相當快，連帶使得書目資料庫所收錄的書目記錄快速成長，以本資料庫爲例 ，近十年來資料庫所收錄的期刊種數與文獻篇目的資料量成長平均是過去的1.5至2倍，在考慮資料的內容品質與對使用者的參考價值前提下，勢必訂定選擇政策以維持資料庫的品質。然而「選擇」的過程目前仍以人工方式進行，難免會受到個人主觀意識的影響，故在資料選擇上，如何維持一致性，並避免遺漏重要文獻，除了確實遵守已制訂的收錄標準外，在資料的選擇上目前除主編外，也由其它工作同人

在資料審查過程中進行評估與建議。

2. 主題索引方式

在書目資料庫的製作過程中，主題索引一向是最重要卻也最困難的一部分，歐美各國書目資料庫大多起源於紙本的索引摘要，其特點在主題分析多採控制詞彙（如標題、分類號或主題詞語），此一發展已有近百年的歷史，其目的在經由對索引詞目的控制，達到主題詞表達的一致性，這種方式最大問題在於它是一種勞力密集的工作，人力與製作時間成本偏高。而大部分的資料庫製作者都是希望在最短時間內處理最大量的資料，因此過去多年來，資料庫製作者多致力於自動化索引技術的開發與利用，試圖取代控制詞彙的主題分析方式，以節省主題索引製作的時間與成本。例如關鍵詞擷取（keyword extraction）及自由文檢索（free-text search）技術都是該需求下的產品。以本資料庫爲例，原本僅以分類作爲主題分析方式，但觀察使用者的檢索行爲（利用檢索欄位次數統計），我們發現分類號是本資料庫所有欄位中被檢索率最低的一個，造成這種現象除了使用者缺乏可使用的參考工具（如完整的分類表）之外，不熟悉這些分類表的組織體系與主題的歸類都是主要原因。此外，我們也發現使用者最常使用的主題檢索方式是利用關鍵詞進行查詢，故於82年開始除原有人工分類索引外，同時採用自動斷詞擷取關鍵詞，85年起並增錄作者自設之關鍵詞（author keyword）。之後再觀察使用者的檢索紀錄，關鍵詞欄位的檢索果然成爲使用率最高的一項，且數量超過其他欄位甚多。然而自動索引方式（或關鍵詞的擷取）雖然縮短了資料進行主題索引的時間成本，但在檢索結果精確率的表現卻不盡理想，尤其是本資料庫提供關鍵詞擷取的欄位內容有限（如：篇名、並列篇名、內容註、附錄），故關鍵詞擷取的數量有限，無法完整表達文獻的內容主題，爲了改善這個問題，本系統從資料內容的加值（如增加摘要、電子全文及作者自訂的關鍵詞）與改善檢索功能（如增加參照查詢、相關資料再查詢、檢索結果相關排序等）雙管齊下，以提昇主題索引的品質與檢索的精確率。自去年8月新功能推出以來，獲得不錯的反應，未來將持續回溯與新增全文與摘要資料的輸入以增加關鍵詞擷取的內容外，並將進行參照詞庫整理工作，解決自然詞彙語意表達不一致性的問題，達到持續提升主題索引品質的目標。

3. 檢索功能性

資料庫系統的穩定性與檢索功能的完整性，不僅影響檢索的結果也是評估資料庫服務品質的主要指標之一，故資料庫除了必須維持硬體（包括主機與網路設施 ）、軟體的穩定性，保證資料庫系統的可檢索性，尤其對於透過連線主機（online host）或鏡錄型式（mirror site）使用的資料庫服務，更需隨時監測系統的使用狀況，若有不正常的現象，應當在最短時間之內排除問題，以免損害使用者的權益。以本資料庫爲例，自從開始提供線上文獻傳遞服務功能以來，線上的使用者（全國87個光碟版鏡錄使用單位未計算） 每月至少10萬人次上網，平均查詢次數爲25萬次，線上文獻傳遞篇次亦超過1萬篇以上，造成系統與網路負荷過重，連帶影響系統的穩定性，對資料庫服務品質的控制產生了強大的壓力，爲了解決此一問題，除了陸續增加硬體設施，擴大網路頻寬外，亦致力改善影像全文傳送速度，以提昇顧客的滿意度。

此外，在檢索功能的設計除充分考慮使用者的一般需求外，並參考各種「使用者研究」、「使用者檢索行爲」的研究結果，定期評鑑資料庫的各項檢索介面與功能設計作爲系統改善的參考，近年來本資料庫就新增多項功能，如以瀏覽（browsing）取代部分檢索功能、增加瀏覽相關檢索工具（如分類表、代碼表等）、提供線上互動式操作說明、增加不同格式的輸出方式（如電子郵遞、檔案下載）等，使之更符合使用者檢索習慣。目前本系統每年度都訂有系統更新功能與維護計畫，以配合資訊技術的發展，持續進行系統功能的改善，滿足使用者新增的需求。

4. 文獻傳遞服務

IFLA在其1998年公布的“Functional requirements for bibliographic records: final report”報告中提出，爲滿足使用者檢索書目資料時的資訊需求，書目紀錄必須具備：查詢（to find）、辨識（to identify）、選擇（to select）及取得（to quire or obtain）四種功能性需求，（註4）故系統若能提供文獻內容的傳遞服務，或提供與館藏書目（OPAC）的連結，則必能提昇使用者的滿意度，進而提昇資料庫的服務品質。

本資料庫自民國87年開始提供文獻傳遞服務，所提供的原文內容有電子全文（HTML）及影像檔（TIFF）兩種檔案格式。在影像與電子全文的製作、蒐集及文獻傳遞服務在品質控制方面的問題有以下幾端：

（1）電子全文顯示的穩定性

目前透過人工搜尋整理並登錄網址方式，連結（hyperlink）網路期刊的電子全文網址共有一百餘種近九千篇文章。透過網址連結的最大優點就是可以立即提供使用者免費完整的資料內容，但是其最大的問題就是來自該相對網址的穩定性，例如遇有網址變更、檔案已移除、目的伺服器或網路連線異常等，都會使得電子全文傳送失敗，雖然大多不可歸責於本系統，但以使用者角度來看，這都是資料庫本身要承擔的品質責任，故如何維持網址的有效性與正確性，是提供類似延伸服務時要同時考慮的要點，爲了解決此問題，將於本年7月完成新功能程式設計，定時針對相關網址的連結狀況進行自動偵測查核，並產生維護訊息檔，提示工作人員進行相關網址的更正維護，以提高網址連結的正確度與完整性。

（2）影像全文的索引連結

除了電子全文檔案的連結，本資料庫主要傳遞的文獻格式是以影像全文爲主，目前已掃瞄的期刊超過2000種，掃瞄的頁數也已經超過370萬頁，預計在民國90年以前可以突破500萬頁，約可達現有索引資料庫篇目頁數的二分之一。

經過兩年左右的影像掃瞄與文獻傳遞服務經驗，在影像全文的製作與提供方面，主要的品質問題有影像索引編碼的正確性與影像掃瞄的規格與品質檢核。

a.影像索引的編碼

由於本系統之影像掃瞄方式係以批次（輪轉式）整本回溯掃瞄爲主，單篇掃瞄爲輔（平臺式即時掃瞄），在期刊影像與書目索引之間必須建立「自動連結比對」的索引資訊，才能在檢索結果顯示時自動比對影像資料庫索引檔，並直接傳送該篇文獻內容。中央標準局在民國85年9月公布「期刊與圖書之文章書目識別號」國家標準（CNS13774），此標準目的在：利用文數字或特殊符號所組成之書目識別號（biblid），來呈現期刊及圖書中之單篇文章，並藉由此標準化且唯一識別號之標示，可利於資料之檢索、比對、傳遞、抽印本識別及文件訂購之處置。其中期刊文章書目識別號由國際標準期刊號（ISSN）、出版年、刊期及頁碼組成。（註5）全國期刊文獻若能遵循此標準之規範，建立標準單一識別碼當有利於文獻傳遞服務及期刊影像資源之共建共享。惟在實行上目前仍有一些問題待以解決，以本資料庫爲例，在收錄的期刊中具有國際標準期刊號者不超過二分之一，且其中有問題者（如刊名已變更仍沿用原號碼、自編ISSN、以ISBN或GPN作爲ISSN、同一機構不同刊物用相同ISSN等）更是時時可見，此外國內期刊對卷期的編立、出版年月、頁碼等標示更無規範化而十分混亂，故若以前述國家標準的編碼作爲本系統書目資料與影像資料的「連結識別」，在實行上有相當的困難，乃另編「國家圖書館期刊影像掃瞄編碼原則」制訂相關編碼規則，迄今配合掃瞄資

料整理的進行，此一編碼原則已經過七次的修正，據此原則影像掃瞄資料存檔時會產生相關索引，而書目檢索與顯示程式部分也依此原則解析書目資料中的資料來源項（如刊名、卷期、出版年月、頁碼），作爲是否已有影像資料顯示的判斷。但由於資料篇目的建檔與影像掃瞄是非同步進行，且本資料庫資料建檔已有30年歷史，輸入規則與格式的變動，都可能造成篇目索引與影像索引無法或錯誤連結狀況，爲改善此現象，除對影像編碼與篇目索引讀取規則持續配合修正外，也配合現有掃瞄規則，回溯修改舊有資料內容。

b.影像掃瞄規格與品質檢核

本系統之影像檔規格爲300dpi G4黑白之壓縮Tiff檔，每月的資料產出量約爲16萬頁，爲了確保資料的掃瞄品質，訂有相關的品質檢核規範，同時採人工與程式比對兩種方式進行。在影像品質方面除利用程式全面快速檢核外，同時由人工抽驗，依中國國家標準之「數值檢驗抽樣程序及抽樣表」（CNS2779 Z4006）規定之III級一般檢驗水準進行抽驗；另外在掃瞄資料索引檔的品質控制方面，則自掃瞄期刊送件前整理開始，即針對期刊原始資料狀況、索引編碼方式訂有相關處理程序與規範，並由本館及委外廠商進行重覆檢核，避免因掃瞄編碼錯誤造成影像連結錯誤，讓使用者權益受損。

c.影像資料的傳送速度

由於本國著作權的相關法令並未針對文獻的電子化重製與利用網路傳遞文獻有明確的說明或規範，故在參酌法界專家的意見與妥善維護作者及出版社的權益原則下，採用原樣影像掃瞄重製方式，將原始期刊製作爲影像檔以提供文化典藏與研究用途之文獻傳遞服務。採用影像檔並透過指定影像瀏覽器（viewer）可避免使用者的不當重製與改作，保持作者原著之全貌，但由於掃瞄檔案較一般文字檔大，故在網路傳送時，在速度上不甚理想，爲快速存取資料，改善影像傳輸速度，在資料儲存方面使用階層性儲存管理（Hierarchical Storage Management, HSM），將儲存成本低存取速度較慢的光碟櫃作爲儲存的最下層，而以儲存成本高存取速度較快的硬碟作爲儲存最上層，當下層的資料第一次被取用時，會被移往最上層，而每日因訂購而即時掃瞄的文件也被放在最上層，當上層的使用空間到達設定的標準時，最不常使用的檔案會被移除，利用檔案的移動節省資料存取時間加快傳送速度。此外，今年3月間將影像瀏覽軟體進行改版，大幅改善影像傳送的速度。

五、結論

臺灣地區自建書目資料庫的發展已有10多年的歷史，早期以政府及學術機構爲主，建置的資料庫多以機構內服務爲主，近年來在網際網路與電子化政府爲民服務運動的推動下，許多由政府機構與圖書館建置的書目資料庫，也已開放於網路提供各界免費檢索，另一方面由廠商自建的商品化書目資料庫也正蓬勃發展中。資料庫發展不但類型與內容趨於多元化，檢索介面也更便利，然而資料量豐富、檢索介面具親和性或檢索功能強是否就是一具有高價值或高品質的資料庫呢？在掌握資訊就掌握競爭優勢的現代社會，利用書目資料庫進行資訊檢索，已成爲資訊需求者尋求資訊來源的主要途徑之一，故如何提供正確有品質的資料內容與服務品質，是書目資料庫製作者不可規避的責任。書目資料庫的品質管理是一個以使用者（顧客）滿意爲服務目標，持續改善的工作，它更是一個團隊工作，必須由經營者、主管與所有組織人員協力配合，凝聚共識，並透過不斷的品質改善過程，才能提升資料庫的品質而永續發展。在國內相關議題的探討並不多見，本文的內容旨在與大家一起分享中華民國期刊論文索引資料庫進行品質改善與提升的一點經驗，祈起拋磚引玉之效，共同努力向上提升本國書目資料庫的品質。

註釋：

註1. Rittbergers, M.；Rittbergers, W.,“Measuring quality in the production of

databases,” Journal of Information Science 23：1（1997）：25.

註2. 林淑芬、許靜芬，〈NBINet系統資料庫品質管理問題之探討〉，華文書目資料庫合作發展研討會（國家圖書館、漢學研究中心，民國88年8月30日至9月1日），頁1。

註3. Basch, R.,“Measuring the quality of the data: report on the fourth annual SCOUG retreat,”Database Searcher（Oct.1990）：18-23

註4. Functional Requirements for Bibliographic Records: final report. IFLA UBCIM Publications-New Series vol.19, p.7

註5. 行政院國家科學委員會科學技術資料中心編，《書目資料庫製作：文獻分析與處理》，臺北市：編者，民國81年，頁11。

註6. 鄭恆雄、許令華合編，〈圖書館相關標準〉，《第三次中華民國圖書館年鑑》，臺北市：國家圖書館，民國88年，頁719。

參考書目

Garman, Nancy.“Online then and now,” Online 20：4（Jul. / Aug.96）

Notess, Greg R.,“Tips for evaluating Web Databases,”Database Magazine 21：1（Apr./May98）：69-72.

O'Neill, Edward Y.； Vizine-Goetz, Diane.“Quality Control in Online Databases,” in Annual Review of Information Science and Technology, vol.23（1988）：125-156.

Rittberger, M.； Rittberger, W.,“Measuring quality in the production of databases,” Journal of Information Science 23：1（1997）：25-37.

Tenopir, Carol.,“Database producers go online,” Library Journal 121：6（Apr.96）：31-32

Tenopir, Carol.,“Moving toward quality,” Library Journal 8：10（Jun.93）：86-87.

Tenopir, Carol.,“Online databases,” Library Journal 124：8（May.99）：36-37.

Tenopir, Carol.,“Human or automated, indexing is important,” Library Journal 124:18（Nov.99）：34-35.