圖書與資訊學刊
33(May '00)66-80
ISSN 1023-2125

館藏數位化的程序及其問題

The Procedure and Problems of Digital Conversion for Library Collection

陳 秀 慧
Hsiu-hui Chen

淡江大學教育資料科學學系碩士班
Institute of Educational Media and Library Science
Tamkung University

【摘要 Abstract】

圖書館為保存原件資料及提昇服務品質，逐漸將「珍貴」館藏予以數位化，在進行數位化前，應考慮著作權問題、保存優先性、儲存媒體、索引及檢索方式等項目，此外，本文將透過康乃爾大學的計劃，介紹「完整資訊擷取」的概念與所需的程序，所謂「完整資訊擷取」是指希望在數位轉換的過程中，原件資料的內容不會流失；另透過由英國伯明罕大學、牛津大學、里茲大學、曼徹斯特大學所進行的 ILEJ 計劃，對數位化程序及其可能發生的問題加以探討。

In order to preserve original data and to enhance the quality of service, library begins to converse valuable collections into electronic form. Before doing this, there are several issues to be noticed first, such as: copyright of original data; what should be conversed first; what kind of storage media, index system, and access methods to be used.. This article introduces two programs: the University of Cornell's program and the Internet Library of Early Journals （ILEJ） program.

The program of University of Cornell brings the concept of "Full Information Capture ". That means the result of digital conversion should be exactly the same data as the original ones.

The ILEJ program is the first stage of Electronic Libraries Programme （eLib）, organized by University of Birmingham, University of Leeds, University of Manchester, and University of Oxford. In this program, they defined standard procedures for data conversion and discussed the possible problems which may be encountered in those procedure.

關鍵詞 Keyword

館藏數位化 完整資訊擷取 康乃爾大學 ILEJ 計劃 光學字元辨識

Digital Conversion; Full Information Capture; University of Cornell; Internet Library of Early Journals （ILEJ）; OCR

壹、前言

在電腦技術的快速發展下，各領域不斷將電腦應用在各層面，圖書館亦不例外，從早期的自動化系統，到現在的「數位圖書館」或「虛擬圖書館」，圖書館界算是跟得上時代，並且開始著重使用者的需求，由於網路的發達與普及，各類圖書館似乎進入「地位保衛戰」，如何提供其他圖書館所沒有的服務、如何表現館藏的特色、如何滿足「讀者」（包括實際的讀者與潛在的讀者）需求、擴大服務範圍及讀者群範圍等問題，正考驗著圖書館的管理階層，在讀者逐漸習慣使用電子資料的同時，將館藏數位化，置於網際網路上，或以電子形式提供更多讀者使用，即是提昇服務品質的方法之一。

世界各國，已進行頗多館藏數位化的計劃，如由美國國會圖書館所進行的「American Memory」計劃，而我國的國家圖書館亦不例外，如中華民國善本叢刊影像先導系統，可見館藏數位化是一個未來的趨勢，因此，對於數位化的過程及可能發生的問題，我們都有必要加以了解及研究。本文主要想藉由國外數位圖書館計劃：康乃爾大學及 ILEJ 計劃，了解圖書館館藏數位化的程序及可能遇到的問題及限制等，以作爲國內圖書館進入數位化工作時的參考。

貳、館藏數位化

一、選擇「珍貴」館藏

本文欲提出一個概念，即是「珍貴」館藏，何謂「珍貴」館藏呢？由於數位化工作，不管在經費上或人力上，都需相當的投入，因此無法也沒有必要將圖書館的館藏全部數位化，所以在決定將館藏數位化後，第一個要確定的即是「選擇數位化的館藏」。圖書館可依其目的、服務宗旨、讀者需求的不同來決定，並輔以館藏的特性，選擇數位化的館藏，所以對不同圖書館而言，或許有不同的「珍貴」館藏，因此圖書館於數位化前，要先了解館內的「珍貴」館藏爲何，在資源有限的情況下，倘若花了大筆的金錢、人力進行數位化，但並未增進服務品質，未達成數位化的目的，則是一件非常遺憾的事，因此「珍貴」館藏的選擇影響數位化工作甚大，如館藏的數量、館藏的特質等。一般而言，凡是具單一性，且有保存價值的資料，皆可視爲「珍貴」館藏可優先選擇，如善本書、絕版書等。

二、進行數位化前應注意的事項

「館藏數位化」的優點有：可使人類智識以更適合的方式保存下來、提供資源分享、便利館藏的流通及推廣、保護原件資料等，如同進行一項研究或計劃，在進行數位化工作前，必須先「規劃」，如確定目標、數位化範圍、選擇館藏等。以康乃爾大學圖書館爲例，他們在進行館藏數位化時，提出六點應注意的事項，確實可做爲參考，此六點爲：

1. 著作權問題：這是圖書館在進行數位化前首先要注意到的問題，被選擇的館藏著作權問題是否可以解決，是否容許重製等問題都需解決。
2. 優先性：由於館藏數量十分龐大，因此必須針對館藏之特性與保存情況及數位化目的，選擇「珍貴」館藏，並決定其優先順序。
3. 決定採用何種媒體做為檔案儲存媒體：選擇光碟、磁帶、電腦硬體、或其他媒體做爲儲存媒體，並評估其優缺點。
4. 決定提供查檢的程度：數位化後的資訊，能提供題名、作者等書目資料，或是可容許查檢到全文影像？
5. 索引與檢索軟體的需求：採用的索引方式爲何？是否可以全文檢索？是否需限定檢索軟

體？

6.採用何種媒體分發給國內、外圖書館或相關機構：是否將數位化的資訊分散給其他單位？是採用光碟分發給需要的單位？或提供網路檢索？①

此外，薛理桂老師另補充三點數位化前需考慮的因素：

1.使用者對於使用這些文獻的需求；

2.費用；

3.可接受的科技等。②

關於使用者需求的考量上，圖書館首先必須確認圖書館進行數位化工作的目的爲何，或爲保存珍貴文獻，或以方便使用者利用爲主，若爲前者，則使用者的需求影響程度較小，如館藏存有珍貴台灣早期作家手稿，爲避免直接使用所造成的損傷，則需進行數位化工作；若爲後者，使用者的需求影響程度較大，則必須對使用者需求加以了解及調查。在技術的限制方面，可能是數位化研究者最頭痛的，如 OCR 辨識率的問題，必須測試各種目前市場上的產品，甚至需與產品的製作公司合作發展，而這些可能會阻礙計劃的進行，不可不思量。

參、數位化技術的重要成員－光學字元辨識技術

一般而言，除了圖像式的館藏外，爲提供使用者更方便的使用數位化館藏，圖書館皆會希望數位化館藏，透過光學字元辨識（OCR），提供全文檢索，所以數位化的過程中，OCR 正確率的高低，影響計劃成功與否甚鉅，但由於目前技術上的限制，尚無法取得 100%的正確率，因此 OCR 軟體的選擇就更重要了，一個好的 OCR 軟體，其必備條件爲：

1.對文字與圖像的擷取皆有高解析度：如可支援 8 bit、200dpi（dots per inch）灰階掃描及以 256 色儲存。

2.高性能（High performance）：由於圖書館館藏大小不一，若只使用市面上所提供之 OCR 軟體，可能無法以高效率的方式擷取大型書籍，或許必須與 OCR 軟體公司合作，發展能掃描大型書籍的軟體。

3.彈性（Flexibility）：一個好的 OCR 軟體必須能夠處理很多輸入型式、必須能夠辨識手寫的文字、具有學習字詞的功能、並且允許些微的歪斜等。

4.正確率（Accuracy）：這是 OCR 最重要的部份，雖然目前的 OCR 軟體無法做到百分之百的辨識率，但在不影響利用的情形下，其正確率必須能爲使用者所接受。

5.生產力（Productivity）：由於圖書館進行數位化研究，乃爲尋求有效率的轉換方式，而 OCR 的生產力即是影響因素之一，一個好的 OCR 軟體必須能擁有快速的辨識能力。③

目前 OCR 的發展，以英文字的辨識率較佳，而中文字因有下列幾個特性：字數多、結構複雜、相似字多、手寫字形變異非常大、字形型態非常多、異體字多等，而限制中文 OCR 的發展：再加上文字經過掃描後，很容易產生雜訊、變形等問題，若是手寫的中文字，如善本書，則會因書寫人的不同，而產生更多的困難。④

肆、個案研究－康乃爾大學的經驗

一、完整資訊擷取（Full information capture）

康乃爾大學提出「完整資訊擷取（Full Information Capture）」的概念，以利用最少的經費，獲取最高的影像品質，其目標不只在以高解析度掃描，而是希望在轉換的過程中，原件資料的內容不會流失，亦即「*No more, no less*」，因爲有時

提高解析度，並不代表影像品質亦會相對提高，反而會使檔案變大，因此康乃爾大學擬研究一套符合「完全資訊擷取」概念的轉換程序，產生數位影像，及設立適當的索引，如此才能提供長期的數位化服務。

James Reilly 曾提出一個掃描前的策略，即是要「了解及喜愛你的原件（Knowing and loving your document）」，因此在大範圍掃描之前，他建議選擇具代表性的原件，作初步的掃描，了解原件主要的特色，並可請教相關的專家。此外，數位轉換需要鑒定與技術的能力，以將文件屬性（attributes）與規格（如解析度、壓縮等）加以整合。

下圖列出進行數位轉換時，必須定義的文件屬性項目，這些項目皆會影響到掃描的成本，例如，若圖像需要彩色呈現，則其成本可能會比黑白圖像的成本要高出二十倍。

表一：紙本式資料與照片的屬性

紙本式資料	照片
1.文件大小（如高 x 寬英吋）	1.格式（如 35mm、4"x5"）
2.詳細資料的大小（如以 mm 描述）	2.重製的詳細資料（detail and edge reproduction）
3.文字特徵（手稿或印刷）	3.雜訊（noise）
4.所使用的媒體（如以鉛筆撰寫）	4.動態範圍（dynamic range）
5.圖片內容	5.色調（tone reproduction）
6.色調（tone），包括顏色	6.色彩（color reproduction）
7.密度、對照、動態範圍（dynamic range）	

二、實例研究

爲了描述完整資訊擷取，康乃爾大學選擇一本於 1914 年出版，由 Andrew Boss 所著之 Farm Management（內文文字約 1.7mm，表格文字約 0.9mm）作研究，此書爲專題論文，內容包含文字、線、中間色系的圖片等，經過鑑定後，認爲文字（包括表格）及圖像對這本書而言，是很重要的，研究結果是以 600dpi、bitonal 的方式掃描，將可以完全擷取本書的文字與圖片資料。以下即是研究過程：

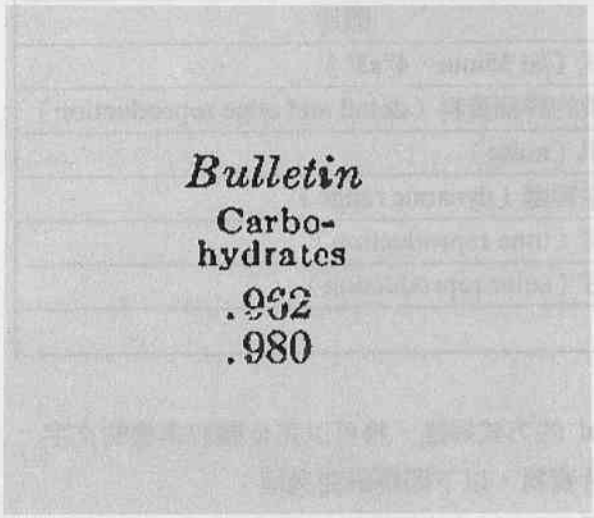
爲了比較，他們掃描了二種頁面，一爲純文字頁面；一爲包含文字與圖片的頁面，並採用不同的解析度，以檢視檔案大小、影像品質、影像於螢幕上呈現及列印後之品質，並將影像透過二種 OCR 軟體（Xerosx Textbridge 2.01 及 Calera WordScan 3.1），了解辨識結果。

表二：掃描純文字頁面之結果

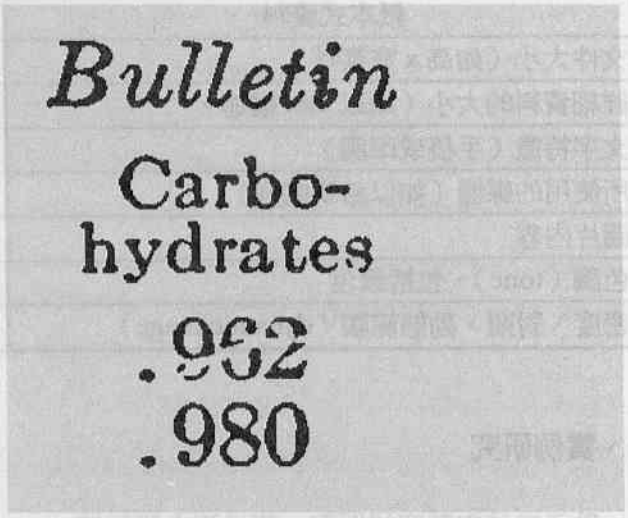
掃描選項	非壓縮檔案大小	壓縮檔案大小	品質	OCR 結果
300dpi、1-bit	380KB	31KB	易讀的（legibility）	33 個錯誤
600dpi、1-bit	1500KB	61KB	真實的（fidelity）	15 個錯誤

表二爲以 300 及 600dpi 掃描純文字資料的結果，此二解析度，爲目前最常使用的解析度，純文字頁面最小的字約 0.9mm 高，表格中的數字約 1.6mm 高，從其掃描結果看出 600dpi 檔案於壓縮前的大小是 300dpi 的四倍，而壓縮後，僅爲 300dpi 的二倍，而以 Xerox XDOS 掃描器處理時，掃描時間並沒有很大的差別。

OCR 部份，在這個例子中，對於處理 300 至 400dpi 影像的能力是可接受的，雖然以 300dpi 所產生的檔案，已是可辨別、易讀的，但它的錯誤率，卻是 600dpi 的二倍，研究發現主要的錯誤發生在 0.9mm 的文字上，並不發生在表格的數字上，因此，雖然 300dpi 所產生的影像，已可進行處理，但 OCR 軟體並不能完整地讀取它，而在原件真實性（fidelity, full capture）與易讀性（legibility, full readable）間的差異是很微小的，在 600dpi 的檔案中,可以平實地表達原件的風貌，而且 OCR 可以將斜體的「i」辨識出來，且第三行的「e」是很完整的，而 300dpi 的檔案，則不完整（如圖一、圖二）。



圖一：300dpi 掃描結果



圖二：600dpi 掃描結果

表三：包含文字與圖片的掃描結果

掃描選項	非壓縮檔案大小	壓縮檔案大小	品　質	OCR 結果
300dpi、1bit descreened/bitonal	380KB	53KB	易讀	18 個錯誤 94.1%
600dpi、1bit text setting/bitonal	1500KB	77KB	真實（文字）	2 個錯誤 99.3%
600dpi、1bit descreened/bitonal	1500KB	127KB	真實（文字／圖片）	2 個錯誤 99.3%
900dpi、1bit text setting/bitonal	3350KB	155KB	真實（文字）	不能 OCR
1200dpi、1bit text setting/bitonal	5960KB	215KB	真實（文字）	不能 OCR
300dpi、8bit grayscale	2980KB	606KB	真實（文字／圖片）	5 個錯誤 98.4%

另選取包含文字與灰階圖片的頁面進行掃描，其中 300dpi, 8bit 灰階的檔案最大，約爲 600dpi, 1bit 的五倍，且灰階掃描時間約爲 bitonal 的四倍，研究發現 600dpi, 1bit 和 300dpi, 8bit 所產生的 OCR 檔案，正確率較高，分別是 99.3%及 98.4%，而 300dpi, 1bit 的正確率只有 94.1%，比美國農業圖書館（National Agricultural Library）所訂的 95%標準還低，經過比較後 600dpi bitonal 掃描所得結果，較其他二種能符合「完整資訊擷取」的概念。⑤

由於電腦及掃描技術的發展，使得在取得高品質的數位影像與掃描媒體間的成本差異並不大。而影像處理的技術，亦漸漸在發展，只是以較緩慢的速度前進。此外，如果選擇「完整資訊擷取」概念爲中心，將原件中所有資訊數位化，必須考慮保存、檢索與經濟因素，「完整資訊擷取」可視爲保存工作的一環，如善本書等，而「完整資訊擷取」所產生的檔案，必須能滿足研究上及使用上的需求，減少使用原件的機會，因此在經費、設備與需求間必須取得平衡。

伍、他山之石－ILEJ（Internet Library of Early Journals）計劃⑥

ILEJ（Internet Library of Early Journals）是電子圖書館計劃（Electronic Libraries Programme，簡稱 eLib）的第一階段，由英國的伯明罕大學（University of Birmingham）、里茲大學（University of Leeds）、曼徹斯特大學（University of Manchester）、牛津大學（University of Oxford）等單位所組成，Joint Information System Committee（簡稱 JISC）則提供部份的資金來源。本節將由 ILEJ 計劃的 Final Report 了解其梗概。

一、研究目的

ILEJ 計劃主要是將連續出版超過二十年的期刊數位化，包括三種十八世紀及三種十九世紀的期刊，並於網路上提供使用，除此之外，這個計劃仍希望對於數位化時可能遇到的問題加以研究，包括：

1.同時使用裝訂及微縮捲片的期刊，當作數位化來源；
2.解析度、雙色調（bitonal）、或灰階的影響，及壓縮影像的品質；
3.可選擇的索引策略，包括利用 OCR 辨識全文資料、紙本式索引的重新鍵入、電子式索引的使用、結合已 OCR 文件與模糊比對軟體（fuzzy matching software）的使用；
4.透過在牛津與里茲大學的 WWW、X-Windows 伺服器展示；
5.在各學校網站（site）中資料的傳輸；
6.使用者對最後產品的接受程度；
7.輿論（critical mass）；
8.將掃描程序按比例加入大範圍數位化程序的可行性。

二、期刊的選擇標準

在考慮十九世紀前的期刊狀況及技術的限制，ILEJ 計劃訂定資料選擇的標準，最後選擇十八、十九世紀出版且連續出刊二十年以上的期刊。選擇標準分述如下：

1.涵蓋的學科範圍廣，包括科學、技術、藝術等；
2.符合英國高等教育學者的需求；
3.符合優先保存的條件；
4.字型、印刷、及紙張的品質；
5.長篇與短篇的文章結構（article formats）；
6.一欄、二欄、或三欄的排版及各種頁面的大

小；

7.文字與圖片的平衡程度（balance）；

8.在各聯盟圖書館中，複本的可得性（availability）。

表四：ILEJ 計劃所選擇的六種期刊

刊名	出版日期	格式	館藏地點
Notes and Queries	1849-1869	裝訂本	曼徹斯特大學
Blackwood's Edinburgh Magazine	1843-1863	裝訂本	伯明罕大學
The Builder	1843-1862	微縮捲片	曼徹斯特公共圖書館（Manchester Public Libraries）
Gentleman's Magazine	1731-1830	微縮捲片	劍橋大學圖書館（Cambridge University Library）
Philosophical Transactions of the Royal Society	1757-1777	裝訂本	曼徹斯特大學
Annual Register	1758-1778	裝訂本	伯明罕大學

此外，ILEJ 計劃選擇 *Illustrated London News* 爲備用的期刊，以於資源充足時，加入計劃之中，不過到計劃結束時，此期刊並沒有被利用。

三、影像處理

在開始掃描，產生影像之前，ILEJ 的人員將欲進行掃描的期刊加以整理，檢查缺期、遺失、或破損的卷期，並做記錄。計劃中共選擇二種掃描軟體，分別是建置於曼徹斯特大學的 Minolta Ps3000 Open Book Cradle Scanner ⑦，主要用來掃描裝訂期刊，另一爲建置在牛津大學的 Mekel Mx500XLG 系統，主要用來掃描微縮捲片，初期因爲掃描系統無法支援灰階（gray-scale）功能，而使工作延宕多時，直到計劃開始九個月後，才開始掃描的工作，而一直到計劃結束，Minolta 系統仍無法支援灰階功能，頗令人遺憾。主要的掃描程序由曼徹斯特大學發展後，交由伯明罕大學進行掃描工作。

影像處理包括掃描、裁切、矯正、壓縮、OCR 等步驟，並由里茲大學及牛津大學所負責，同時建立影像檔案的 Metadata 及定義單一 SGML 的識別詞（如 gm.1747.1.x.1.x.x.5），最後將檔案存於牛津大學的檔案伺服器（Hierarchical File Server，簡稱 HFS）中，其他三個學校可透過 FTP 存取這個檔案伺服器中的影像檔案。

掃描後的 TIF 檔案，在減少檔案空間的考量下，被轉換成的 GIF 檔案，並於網路中傳輸，而轉換後的解析度約在 120 至 200dpi 之間，這是爲了使每個影像檔案能以最佳的品質呈現在 800*600 像素的螢幕中。ILEJ 計劃採用 Image Magick 系統來完成這項工作，在二十四小時的工作時間內約可轉換一卷約包含 650 頁的期刊，但這個速度，並不能被 ILEJ 計劃所接受，因此他們又使用 Image Alchemy 來進行轉換，約四個小時轉換一個卷期。

四、影像品質的控制

ILEJ 計劃對於影像品質的標準是「符合目的（fitness for purpose）」，最基本的要求是必須做到在螢幕上觀看或列印出來是易讀的（legibility）、及有美感（aesthetic）的呈現，特別是在頁面的曲

度（page curvature）上，此外，OCR 的品質亦是控制的項目之一，但掃描結果中，只有二種期刊適合使用 OCR 辨識資料。ILEJ 計劃投注了很多的心力在發展一個良好的掃描程序，以獲得最佳的影像品質，但在掃描裝訂期刊時，遭遇到很多問題，如裝訂期刊的曲度（curvature）、彎曲的紙張、掃描器遠端光源所造成的陰影等，尤以掃描期刊的第一頁及最後一頁最爲嚴重。

1.解析度方面：其中四種裝訂期刊是以 400dpi 的解析度及雙色調（bi-tonal）掃描成影像，這是可以獲得的最高解析度，並且圖像的色彩是可接受的，不過這不符合康乃爾大學所訂出的標準，在擷取較小的文字時，並不能滿足需求，尤其是 *Notes and Queries* 廣告頁上的字。

2.時間方面：因裝訂期刊的不同，掃描所需時間亦不同，例如首頁與末頁較正文頁難掃描，平均一個小時約可掃描九十頁的期刊，這個時間遠比 ILEJ 在原先的計劃書中所預估的時間，整整多出三倍，而這個時間並不包含掃描前將簡單的 metadata 資料鍵入 Excel 的工作表中。

3.OCR 方面：在開始之前，ILEJ 計劃採用 Sequoia Scanfix 系統進行 TIF 影像檔矯正方向、裁切、清除斑點等工作。對於 OCR 的要求，因是否具備索引而不同，若需要提供全文的資料，則要求正確率要在 99.95%或更高，而若使用索引，則正確率可稍降至 85-99%，並且配合模糊比對的檢索功能。軟體方面，由於 ILEJ 計劃所採用的網站伺服器是 EFS 系統，開始時，亦考慮選擇能與伺服器相容的 OCR 軟體，但在經過評估後，乃選擇於 Windows 3.1 執行的 OmniPage Pro 第六版，隨後更新成第八版，並改於 Windows NT 中執行。辨識結果，較好的是 *Blackwood's*，平均正確率爲 98.5%，較差的是 *Notes and Queries*，平均正確率約低於 80%，不過部份品質較好的頁面，正確率約可高達 99%。時間方面，在使用 Pentium 90、64MB RAM 的電腦，辨識「矯正」後的約 650 頁一卷的期刊，必須花費 12.5 個小時。

五、微縮捲片掃描

於 1997 年 4 月，開始建置專用於微縮捲片的 Mekel MX500XL-G 掃描器，用以掃描 *Gentleman's Magazine*、*The Builder* 二種期刊之微縮捲片（正片），由於是微縮片，因此一些裝訂期刊所會造成的問題，如頁面彎曲等，都不會出現在掃描的過程中，而較大的問題是易讀性（legibility）的問題，所以在掃描的過程中，乃採用灰階方式，以增加易讀性，但可能會增加檔案的大小及增加掃描時間，因此必須找出一個平衡點。

掃描結果，66%的 *Gentleman's Magazine* 微縮捲片，利用 300dpi 雙色調（bitonal）的方式掃描，而且結果是可接受的、可讀的，而其他的 34%則以 100dpi、256 灰階的方式掃描。因爲 *The Builder* 的原始文件較大，且字較小，所以乃以 200dpi、256 灰階的方式掃描，而且產生了將近 10M 的檔案，但這些龐大的檔案，造成 Mekel 系統頗大的壓力。最後利用 Image Alchemy 軟體將已裁切的 bi-tonal 影像轉換成 GIF 檔案，而將以灰階方式產生的檔案轉換成 JPG 檔案。計劃初預估掃描 250 頁微縮捲片需一個小時，但掃描結果，*The Builder*（已自動裁切）每小時只可掃描 40 頁，而 *Gentleman Magazine*（不包括裁切）每小時可掃描 70 頁，與預估時間相差頗多。

六、轉換紙本式索引

使用者檢索是這個計劃的主要部份，包括：分散式的 Internet 檢索、檢索與瀏覽設備、易讀的介

面等，而製作有效率的索引是主要的元件（element）。在 ILEJ 計劃中，將提供二種檢索方式，一爲利用 OCR 將全文予以辨識，並提供全文檢索；另一爲將與期刊同時出版的索引或目次轉換成電子形式。下表即爲目前可獲得的索引或目次服務。

表五：各期刊索引及目次資料

刊　名	索引或目次
Notes and Queries	主題索引（由原始卷期中取得）
Blackwood's Edinburgh Magazine	作者及題名索引（從 *Periodical Contents Index* 中整合而來）
The Builder	無
Gentleman's Magazine	主題索引（從 1~20 卷的彙編本中取得）
Philosophical Transactions of the Royal Society	作者、題名、及主題索引（從原始卷期之主題索引及目次中取得）
Annual Register	主題索引（由原始的彙編本索引中取得）

表六：各期刊索引詞數量

期　刊　名	主題索引	作者索引	題名索引
Notes and Queries	171,390	-	-
Gentleman's Magazine	19,977	-	-
Philosophical Transactions of the Royal Society	8,861	817	917
Blackwood's Edinburgh Magazine	-	96	2,004
Annual Register	18,955	-	-

ILEJ 計劃在經過內部的測試之後，將索引建檔的工作外包給 Offshore Keyboarding Corporation 公司，每一千字約收費 0.73 到 0.85 英鎊，並以類似 SGML 的方式將索引予以建立。

里茲大學以 Excalibur EFS 軟體建置網站伺服器（Web Server），並提供模糊比對（Fuzzy Matching Capability）的檢索功能，而牛津大學則利用 PAT（Opentext）檢索引擎，提供服務，並且可連接至里茲大學的網站。ILEJ 計劃並於最後的三個月透過問卷調查與電話訪問，進行使用者評鑑的工作。

七、Metadata

在 ILEJ 計劃中，整合幾種類型的 Metadata，包括：

1.期刊卷期的基本書目資訊；

2.主題、作者、題名索引；

3.OCR 後的全文資料；

4.基本的都伯林核心集（Dublin Core metadata），包括將<META>的 HTML 標示加入 ILEJ 首頁；

5.其他資料 Metadata，如檔案解析度、所使

用的壓縮軟體等。

在開始掃描之前，掃描操作者會將期刊的簡單資料鍵於 Excel 工作表中，以作爲後續處理之參考，並作爲索引的來源。原先計劃是希望使用能與 PS3000P 掃描器相容的 SQL 資料庫管理系統，但調查發現，SQL 資料庫是一個封閉系統架構，不支援擷取 Metadata 的功能，因此只好將資料鍵入 Excel 工作表中，再以簡單的 BASIC 程式，產生符合需求的文字檔案。接著即利用 PERL Script 將文字檔案，產生 SGML 識別語及 TEI 檔案；而從微縮捲片掃描過程中，所取得的資料，亦被輸入 Excel 工作表中，並將轉換到 FoxPro 資料庫軟體，用 FoxPro Script 產生 TEI 檔案。

八、計劃的管理

計劃由里茲大學及牛津大學的二位專案管理者（Project Officers）負責，並成立一個專案執行小組，由每所大學派出二位人員參與，專案管理者是全職（full-time）的，但手上仍有其他專案正在進行，因此這二位專案管理者約將其工作時間的百分之五十挪出，進行此計劃。此外，掃描操作者（scanner operators）則不是全職的，而是由曼徹斯特、伯明罕、牛津大學分別雇用。每所大學的主要工作如下：

1.伯明罕大學：裝訂期刊掃描；
2.里茲大學：影像處理、OCR、建置 EFS 伺服器、及使用者評鑑；
3.曼徹斯特大學：裝訂期刊掃描、部份的影像處理；
4.牛津大學：微縮片掃描、建置 PAT 伺服器、影像處理、OCR、繕打（keyboarded）索引、儲存檔案。

九、計劃時程表

表六：計劃之時程表

時　間	工　作　項　目
1996.02	在牛津大學建置網站（Web Site）
1996.06	在曼徹斯特大學及伯明罕大學建置 Minolta PS3000 掃描器
1996.11	曼徹斯特大學開始進行掃描工作
1997.03	將 *Notes and Queries* 的第一卷期放置在網站上
1997.04	在牛津大學建置專用於微縮捲片的 Mekel 掃描器
1997.05	伯明罕大學開始進行 *Blackwood* 的掃描工作（至 1997.11 完成）
1997.08	曼徹斯特大學開始進行 *Philosophical Transactions* 的掃描工作（至 1997.12 完成）
1997.08	*Notes and Queries* 的十個卷期的影像開始提供使用者檢索
1997.10	牛津大學開始進行 *Gentleman's Magazine* 的微縮捲片掃描（至 1998.04 完成）
1997.12	伯明罕大學開始進行 *Annual Register* 的掃描工作（至 1998.03 完成）
1998.02	將 *Notes and Queries* 的其他卷期（共二十個卷期）提供使用者檢索
1998.04	牛津大學開始進行 *The Builder* 的掃描工作（其中的十卷期於 1998.08 完成）
1998.03~1998.08	將所有期刊檔案及其索引儲存於牛津大學的伺服器
1998.06	使用者評鑑

十、目前可得資源

目前只有其中四種期刊：*Notes and Queries*、*Philosophical Transactions of the Royal Society*、*Blackwood's Edinburgh Magazine*、*Gentleman's Magazine* 於網站上提供檢索，但理論上，可獲得的資源，如下表。

表七：可獲得的資源

期刊名	出版年	影像數量	OCR	索引
Notes and Queries	1849-69	26,254	☺	主題
Blackwood Edinburgh Magazine	1843-63	33,183	☺	
Gentleman Magazine	1731-50	14,181	😐	
Philosophical Transaction of the royal Society	1757-77	18,947	😐	主題、題名、作者
Annual Register	1758-78	12,465	😐	主題
The Builder	1843-52	5,518	😐	無

十一、成本

ILEJ 的經費 58,000 英鎊，其中 38,000 英鎊是 eLib 計劃的經費，其餘爲四所大學的經費，但上述經費不包括管理成本、建立掃描及處理程序時的內部環境及發展成本、購買及維護伺服器成本、檔案成本等，這些成本主要由各校吸收。所編列的經費主要用在：

1.影像產生：包括資料掃描前的準備、基本 Metadata 的定義、掃描的設備及人員成本。
2.索引產生：包括 OCR 及輸入（key in）成本。
3.影像轉換，影像與索引處理成本。

表八：Notes and Queries 及 Annual Register 成本明細表

	Notes and Queries		*Annual Register*	
影像數	26254		12465	
每小時掃描頁數	97		98	
項目	成本（英鎊／頁）	百分比	成本（英鎊／頁）	百分比
1.前處理／基本 Metadata	0.02	2	0.02	3
2.掃描（人員）	0.12	16	0.13	21
3.掃描（設備）	0.06	8	0.06	10
產生影像	0.20	27	0.21	34
4.OCR	0.04	5	*	*
5.索引鍵入	0.26	34	0.16	27
產生索引	0.29	39	0.16	27
6.檔案傳輸及相關活動	0.01	1	0.01	2
7.影像轉換	0.04	5	0.02	3
8.檔案結構、命名、處理	0.07	10	0.04	7
9.品質檢視	0.01	1	0.01	2
10.SGML 轉換	0.07	9	0.10	16
11.檔案索引（for Open Text）	0.05	6	0.05	8
12.設備與軟體	0.01	1	0.01	2
影像轉換與處理	0.25	34	0.24	39
總　計	0.74	100%	0.61	100%

十二、使用者評鑑與建議

ILEJ 網站免費提供給全球人士使用，並透過郵件清單（mailing list）將訊息傳送到全球。ILEJ 於計劃的最後三個月進行使用者評鑑的工作，於 1997 年 6 月至 1998 年 9 月間，統計使用 ILEJ 網頁的人數及使用情況，透過電子郵件、網站，進行問卷調查（回收率 26%），以了解使用者滿意程度，並根據回函中，選擇電話訪問對象（5 位學者及 1 位圖書館員），主要的目的是定義使用的層次、了解使用的理由、對使用電子期刊的期待等，並了解下列各項目，使用者接受的程度：

1.影像的易讀性；
2.檢索功能；
3.展示功能；
4.紙本期刊與電子期刊的使用比較。

其中使用者對 ILEJ 的建議如下：

1.對部份海外的使用者而言，檢索速度並不十分滿意；
2.在 ILEJ 網頁中，需要太多步驟才能進到主要的檢索畫面或取得資料；
3.需要於各網頁中，直接連結至 ILEJ 首頁；
4.部份圖像不易讀，並且部份需要更好的品質控制；
5.希望提供更高品質的 OCR 文件；

6.希望所有的文件皆能提供OCR的全文檢索；
7.檢索方式可否增加，如增加時間的限制；
8.因爲透過模糊比對所得的結果不如預期，所以多採取瀏覽或簡易檢索。

十三、紙本式與電子式期刊的使用比較

根據問卷調查的結果，發現ILEJ的使用者非常樂意同時使用紙本式與電子式期刊，主要是利用電子式期刊，以爲檢索之用。從網站列印次數看來，將近一半（49%）的使用者認爲不需要將期刊列印，而有 22%的使用者對列印的結果並不滿意，在ILEJ的例子中，亦不可能因電子式期刊的出現，而淘汰紙本式期刊。

十四、研究建議

1.了解原件、微縮過程、微縮片複製、影像處理等因素對影像品質的影響程度；
2.必須繼續研究由於不完整的OCR文件，所產生檢索錯誤的比率，並了解利用模糊比對是否能改善這個情況；
3.必須繼續研究提高OCR正確率的方法；
4.對微縮產品的建議爲：使用以state-of–art的微縮技術產生的高品質微縮捲片或將原件複製（影印）後掃描，放棄使用微縮產品；
5.對紙本產品的建議爲：使用灰階掃描；使用更高級的掃描設備（如Zeutschel），即使需要花費更多的經費；使用性能更好OCR軟體等；
6.使用SGML或XML爲基本Metadata架構；
7.使用者評鑑方面，建議可再延長取樣時間；
8.對使用者真正的需求再調查；
9.必須重視的是希望達到的目的，而不是技術的限制；亦即爲了達成計劃制定的目的，圖書館必須盡力地尋找目前可得的技術，而不是因爲技術上的限制，放棄或更改原先的目的。⑧

陸、結論與建議

綜合上述的討論，可了解數位化是一個重要且艱巨的工作，如館藏的品質，可得的科技等因素，都可能影響計劃的成功與否，因此圖書館在進行館藏數位化時，應注意事項及建議陳述如下：

一、有計劃的進行，建立最佳轉換程序

圖書館在決定將館藏數位化後，必須有計劃的進行這個工作，從確立數位化目的、選擇「珍貴」館藏、廣泛蒐集各國數位化資料，以了解可能遇到的問題等。在選擇館藏後，建議可以仿照康乃爾大學的模式，先擇取一個具代表性的館藏，做爲實驗的標的，進行初步的數位化工作，了解可能發生的問題、測量轉換時間、尋找所需的工具、設定各項參數，並了解將掃描程序按比例加入大範圍數位化程序的可行性，建立最佳轉換程序。

縱觀目前各國的數位圖書館計劃的目的，約可分爲二類，一爲尋求最佳轉換程序；另一爲純將館藏數位化，如本文所提的康乃爾大學與 ILEJ 計劃，前者較屬於前導型的計劃，主要目的在找出一個方法能快速，經濟地轉換資料，而ILEJ計劃除了包含上述的目的外，亦爲將館藏數位化。*以康乃爾大學的研究結果，建議使用 600dpi 雙色調（bitonal）的方式掃描，較可達到「完整資訊擷取」的目的。*

二、選擇儲存媒體

Demas 曾提出在保存印刷式的歷史文獻方面，可考慮採用「混合的方式」（hybrid approach），

針對不同媒體所具有的個別優點，建議：

1.微縮影片可提供長期保存；

2.數位化的電腦檔可依需求提供紙本式文件、透過網路查檢所需的文獻；

3.光碟產品可供全國性分發與流通。⑨

由於儲存媒體的發展快速，如早期用來大量儲存電子資料的媒體，即是磁帶，而目前較常使用的則是光碟，就連光碟也是變化快速，從早期的能寫入一次，到目前的可多次寫入，媒體的選擇的確困擾著圖書館，尤其在電子資源日趨增多的情形下，圖書館必須仔細評估、選擇儲存媒體，此外，必須注意新舊媒體的轉換問題，如果在經費的限制下，無法完全轉換成新媒體，則仍需進行轉換研究，確保目前所使用的媒體，隨時可轉換成新媒體。

三、字元辨識（OCR）

OCR 部份，英文資料部份，辨識率較高，而中文部份的辨識率，則仍需努力發展，而原始文獻本身的品質亦是影響辨識率的高低因素之一。建議可採用將原始文獻影像檔與文字辨識後的文字檔共同使用，但不進行文字校對工作，以節省人力與時間。以中文善本書爲例，若要將以書法所撰寫的古書加以 OCR，肯定會遭受到挫折，而只能等待軟體的進一步發展了。

由於國外數位圖書館計劃頗多，因此可了解英文方面的 OCR 辨識率，而中文部份，雖有部份圖書館已進行數位化工作，如國家圖書館、中央研究院，但似乎沒有相關的研究報告出版，以致無法了解中文在 OCR 方面遇到的問題及其影響。

四、使用者的閱讀習慣

在電腦與網路的快速、便利誘因下，使用者真的會放棄紙本式的媒體嗎？以 ILEJ 計劃爲例，發現使用者仍然習慣於使用紙本式館藏，但如果有電子形式的資訊可供利用或檢索，兩者相輔相成，仍是使用者樂見的，因此，在這個資訊化的時代，使用者的使用習慣影響圖書館界的服務甚鉅，甚至引導未來資訊發展至另一境界，我們可以大膽的假設，人們的閱讀習慣不容易改變，而紙本與電子資訊將繼續共存。因此圖書館在進行數位轉換之前，仍需了解使用者紙本館藏與電子館藏的需求程度。

五、保存與檢索

在開始數位化前必須先定義目標，未來的數位資料是以保存爲主，亦或是以檢索爲主。此二者的差異，在於影像品質的控制，若以保存爲主，爲避免使用原件資料，則所轉換的影像，必須能支援大部份使用者研究的需求，且 OCR 的正確率則要求較高；以檢索爲目的則不然，對影像的品質及 OCR 的正確率的要求不若以保存爲目的高（當然還是得在一定標準之上），可輔以索引或模糊比對的方式取得資料。

六、使用者評鑑

從 ILEJ 計劃的報告書，筆者發現該計劃對於使用者評鑑做得很完整，以目前圖書館界而言，數位化是一個「先進」的工作，而且還在研究階段，因此對於使用者的需求必須充分掌握，凡是一個計劃，最終還是要透過評鑑後，取得建議，以爲下次研究的參考或爲他人類似研究時的參考，因此使用者評鑑是很重要的，縱觀國內各數位化計劃，似乎沒有於階段性完成後，做使用者評鑑的工作，亦或是有做評鑑工作，但未將評鑑結果出版，是有點可惜的。

七、合作數位化

如同 ILEJ 計劃般由四所大學組成，是圖書館

進行數位化時，可考慮的合作情況，若僅單由某一單位承受掃描設備所需的經費、掃描技術、人員的獲取問題，似乎是一件困難的工作，因此可透過合作的方式，分散經費來源，或可爭取其他學術單位的支持，但在合作的過程中，則必須注意溝通管道的建立，可透過設立定期開會的委員會來解決。

綜合上面的討論可發現數位化前的研究計劃是相當重要的，透過研究計劃的進行，可找出最佳的轉換程序，同時詳細記載數位化的細節，如所需的時間、人力、經費等，以利正式轉換工作的進行，因此建議國內的圖書館在進行館藏數位化前，應先進行研究計劃，才得以節省時間、人力，不致於徒勞無功。

（收稿日期：1999 年 10 月 19 日）

註釋：

註①：薛理桂，「珍藏文獻數位化之發展現況與展望」，國立中央圖書館臺灣分館館刊 4 卷 1 期（民國 86 年 9 月），頁 18-19。

註②：同前註，頁 13。

註③：Don Keller, "High-Speed OCR Systems Have Flexibility and Accuracy," Document Image Automation 13:1(1993), p.29.

註④：潘朝陽，「OCR／中文 OCR 技術」，光學工程 47 期（民國 83 年），頁 48-49。

註⑤：Stephen Chapman and Anne R. Kenney, "Digital Conversion of Research Library Materials," D-Lib Magazine (Oct. 1996), <http://www.dlib.org/dlib/october96/cornell/10chapman.html>.

註⑥：ILEJ 的首頁：<http://www.bodley.ox.ac.uk/ilej/start.htm>.

註⑦：依其 Final Report 看來，此系統應爲類似 "翻拍" 的掃描系統，而不是目前常見的掃描器。

註⑧："Internet Library of Early Journals Final Report (January 1996~August 1998): a Project in the eLib Programme," Mar. 1999, <http://www.bodley.ox.ac.uk/ilej/papers/fr1999/fr1999.htm>.

註⑨：同註①，頁 13，引用——S. Demas, "Setting Preservation Priorities at Mann Library: a Disciplinary Approach," Library Hi Tech 12:3(1994), p.88.